https://doi.org/10.51408/issi2025_078

Leveraging Large Language Models for Post-Publication Peer Review: Potential and Limitations

Mengjia Wu[1], Yi Zhang[2], Robin Haunschild[3], Lutz Bornmann[4]

[1]*Mengjia.Wu@student.uts.edu.au*, [2]*Yi.Zhang@uts.edu.au*
Australian Artificial Intelligence Institute, Faculty of Engineering and Information Technology, University of Technology Sydney (Australia)

[3]*R.Haunschild@fkf.mpg.de*,
Max Planck Institute for Solid State Research, Heisenbergstr. 1, 70569 Stuttgart (Germany)

[4]*L.Bornmann@fkf.mpg.de*, *bornmann@gv.mpg.de*
Max Planck Institute for Solid State Research, Heisenbergstr. 1, 70569 Stuttgart (Germany)
Science Policy and Strategy Department, Administrative Headquarters of the Max Planck Society, Hofgartenstr. 8, 80539 Munich (Germany)

Abstract

Peer review is the cornerstone of scientific evaluation, ensuring the quality, accuracy, and integrity of published research. However, challenges such as reviewer bias, time constraints, and the increasing volume of submissions have strained traditional peer review systems, resulting in delays, lower-quality reviews, and reviewer fatigue. These limitations highlight the need for innovative solutions. Large language models (LLMs) have emerged as promising tools to support or potentially replace certain aspects of peer review. This study investigates the potential of LLMs to enhance post-publication peer review, offering quality assessments and recommendations for published articles. Specifically, we designed two tasks to evaluate the performance of LLMs in post-publication research evaluation: identifying high-quality articles (Task 1) and providing ratings on recommended articles (Task 2). Six versions of three generative LLMs, including open-source models such as Qwen and Llama, the closed-source GPT-4o-mini model, and four BERT-based models, were assessed using in-context learning and fine-tuning approaches. The data for training and evaluation were sourced from H1 Connect (formerly Faculty Opinions), a platform for expert recommendations in the biomedical domain. Results indicate that fine-tuning LLMs with labelled data can significantly enhance their alignment with human expert evaluations. For Task 1, fine-tuned models performed well in identifying high-quality articles with an accuracy of 84%. However, for Task 2 - rating on recommended articles - LLMs struggled to match human judgement consistently with an accuracy below 0.6, highlighting their current limitations in nuanced, context-dependent tasks.

Introduction

In the realm of academic publishing, peer review serves as the cornerstone of scientific evaluation and dissemination (Bornmann, 2008). The process ensures that manuscripts meet certain standards of quality, accuracy, and integrity (defined by a certain field, community, journal etc.). Peer review, while essential, is not without challenges. Issues such as time constraints, reviewer biases (Bornmann, 2011), and the increasing volume of submissions necessitate solutions to enhance the efficiency and effectiveness of peer review. In this context, large language models (LLMs) have emerged as a promising tool for augmenting or replacing peer review. LLMs, exemplified by OpenAI's GPT series and Google's BERT, have

demonstrated remarkable capabilities in natural language understanding and generation (ChatGPT was introduced to the public in 2022, see Farhat et al., 2023). LLMs leverage vast amounts of textual data to learn linguistic patterns and generate human-like text. Their applications span various domains, including automated content generation, research classification (Wu et al., 2024), scholarly recommendation (Jia et al., 2025), knowledge association prediction (Wu et al., 2021), sentiment analysis, and language translation. More recently, the potential of LLMs to assist in research evaluation tasks has garnered attention from researchers and practitioners alike (Thelwall, 2024a, 2024b). LLMs have been used to undertake evidence synthesis and systematic assessment tasks (Joe et al., 2024), to propose references for anonymized in-text citations (Algaba et al., 2024), to predict citation counts, Mendeley reader counts, and social media engagement (de Winter, 2024; Vital Jr et al., 2024), and to identify prominent scholars (Sandnes, 2024).

The academic publishing landscape is witnessing significant growth (Bornmann et al., 2021), with an increasing number of manuscripts submitted for review and publication. The increasing number, while reflecting the importance of scientific inquiry for society, also places immense pressure on the peer review system. Reviewers and editors, as rule volunteers, face the task of evaluating numerous manuscripts and grant proposals within limited timeframes. Furthermore, the traditional peer review process has been often criticized for its subjectivity, potential biases, and the increasing difficulty in obtaining high-quality reviews. Consequently, delays in the review process, difficulties in finding reviewers, useless reports, and reviewer fatigue have become prevalent issues. These challenges highlight the need for innovative approaches to relieve the participants (reviewers) in the peer review process.

Several studies have explored the feasibility and effectiveness of using LLMs in peer review processes (Liang et al., 2024; Liu & Shah, 2023; López-Pineda et al., 2025; Thelwall & Yaghi, 2024). These studies suggest that LLMs can assist in specific peer review tasks such as identifying errors, verifying checklists, and providing feedback, but they are not yet reliable for complete evaluations of papers or proposals. One of these studies focused on the use of LLMs, specifically GPT-4, for specific reviewing tasks such as identifying errors, verifying checklists, and choosing the better paper among pairs of abstracts (Liu & Shah, 2023). The findings suggest that while LLMs can effectively identify errors and verify checklist questions with high accuracy, they struggle with more subjective tasks like discerning the quality of papers. This indicates that LLMs can serve as valuable assistants for specific, well-defined reviewing tasks but are not yet ready to replace human reviewers entirely.

Another empirical analysis evaluated the quality of feedback generated by GPT-4 on papers (Liang et al., 2024). The study compared LLM-generated feedback with human peer reviewer feedback across thousands of papers from prestigious journals and conferences. The results show a significant overlap between the points raised by GPT-4 and human reviewers, particularly for weaker papers. An additional user study revealed that researchers found the LLM-generated feedback helpful, suggesting that LLMs can provide valuable assistance in the peer review process,

especially for researchers in under-resourced settings. The most recent study (Thelwall & Yaghi, 2024) evaluated whether ChatGPT 4o-mini can estimate the quality of papers by comparing its scores to departmental averages across 34 Units of Assessment in the United Kingdom's Research Excellence Framework (REF) 2021. The results show a generally positive correlation, with some variations, suggesting that LLMs can provide reasonable quality estimates, especially in the physical and health sciences. These assessments are based only on titles and abstracts, not comprehensive evaluations.

The previous studies on the use of LLMs in the peer review process reveal that their use holds significant promise for addressing some of the challenges associated with traditional peer review. Although LLMs may provide valuable feedback, it is essential to recognize their limitations. For example, LLMs seem to include "hallucinating" information into otherwise plausible responses (Thelwall, 2024b). Ongoing research should try to refine these models to ensure their effective and ethical use in the academic community. Building on the insights from previous studies, the current empirical investigation aims to evaluate the use of LLMs for post-publication peer review. Post-publication review, unlike traditional pre-publication review, occurs after the paper has been published, providing a platform with recommendations and quality assessments of papers. This study seeks to assess the opportunities of LLMs in enhancing post-publication peer review processes. By leveraging advanced LLMs, the study aims to explore how these models may complement human expertise and streamline the review workflow.

In this study, we designed two tasks to assess the LLMs' capabilities in post-publication research evaluation: identifying high-quality articles (Task 1) and recommended article rating (Task 2). Six versions of generative LLMs, including open-source Qwen, Llama models, and closed-source GPT-4o-mini model, in addition with four BERT-based language models, were tested under two different learning settings: in-context learning and fine-tuning, to complete the two tasks. Using data from H1 Connect (a post-publication peer review service in medicine and life sciences, formerly known as Faculty Opinions) as training and test data, we performed model comparisons on both tasks. The results revealed that, with an appropriate fine-tuning strategy, current LLMs have strong potential to serve as preliminary reviewers to identify high-quality papers (Task 1), with the fine-tuned GPT-4o-mini model achieving the accuracy of 84% and BERT models above 75%. However, the models still lack the capabilities to achieve expert-level judgment when facing more complicated tasks like article rating (Task 2), in which rating differences are more nuanced to learn.

Data and Tasks

Data source

H1 Connect is a specialized platform designed to provide expert recommendations and support research evaluation in the biomedical domain. It delivers scholarly output metadata along with expert-generated recommendations, which are enriched with detailed ratings, commentaries, and classification codes. The additional

information explains the basis for the inclusion of the papers on the platform and their relevance for the community. We selected the H1 Connect data for its extensive data coverage and rich evaluation metadata across biomedical fields, which ensures a representative and diverse dataset for comparing assessments from experts and other instruments such as bibliometrics or LLMs.

Task formulation

To examine the research evaluating capabilities of LLMs, we designed two tasks of high-quality article identification (Task 1) and recommended article rating (Task 2). To achieve the tasks, we collected two datasets from H1 Connect, with their details given in descriptions below and Table 1. Given that testing LLMs on the global dataset comes with an unneglectable burden of computational costs, we randomly sampled partial articles for each task from the entire dataset. We used the article abstracts as our input to the models due to the incomplete availability of full texts.

Task 1 - High-quality article identification: This task aims to evaluate how effectively LLMs can identify high-quality articles from a mixed pool of high- and low-quality articles, compared to the judgment of human experts. Low-quality articles are defined as those with no expert recommendations, and high-quality articles are those with three or more expert recommendations. To construct a mixed pool for testing, we compiled 4,538 articles from OpenAlex (Priem et al., 2022) – a bibliographic catalogue of scientific papers – with no expert recommendations and 4,994 articles with three or more expert recommendations. The not-recommended articles were published between 2010 and 2020 in the same journal, with the same volume and issue as the recommended papers. We excluded the journals *Science*, *Nature*, *Proceedings of the National Academy of Sciences of the United States of America*, *Science Advances*, *Nature Communications*, *Scientific Reports*, and *PLOS ONE* due to their multidisciplinary nature for the selection of not-recommended papers. The selected LLMs are required to retrieve the 4,538 high-quality articles from this pool as accurately as possible.

Task 2 - Recommended article rating: This task delves into a more detailed objective of rating research articles based on their quality and content. To avoid complications in synthesizing expert ratings, we focused on articles with only one recommendation at this stage. The data collection also follows procedures as in Task 1, resulting in 86,805 articles with a rating of 1, 54,154 articles with a rating of 2, and 11,089 articles with a rating of 3 (roughly 8:5:1). Considering the computational costs for model testing, we sampled a balanced dataset that consists of 5,000 articles from each rating of 1 (good), 2 (very good), and 3 (excellent), ending with 15,000 articles.

Table 1. Descriptions of datasets used in Task 1 and Task 2.

Task 1	# Article	**Three recommendations**		**No recommendation**
		4,994		4,538
Task 2	# Article	**1 (Good)**	**2 (Very good)**	**3 (Exceptional)**
		5,000	5,000	5,000

Methodology framework

The overall research framework is presented in Figure 1. To perform Task 1 and Task 2, we selected four BERT variant models and six generative LLMs, with details provided in the model selection section. Two representative model adaptation techniques, in-context learning (ICL) and fine-tuning, were employed to adapt the models to output the desired results for the tasks. These techniques are described in detail in the following subsections.

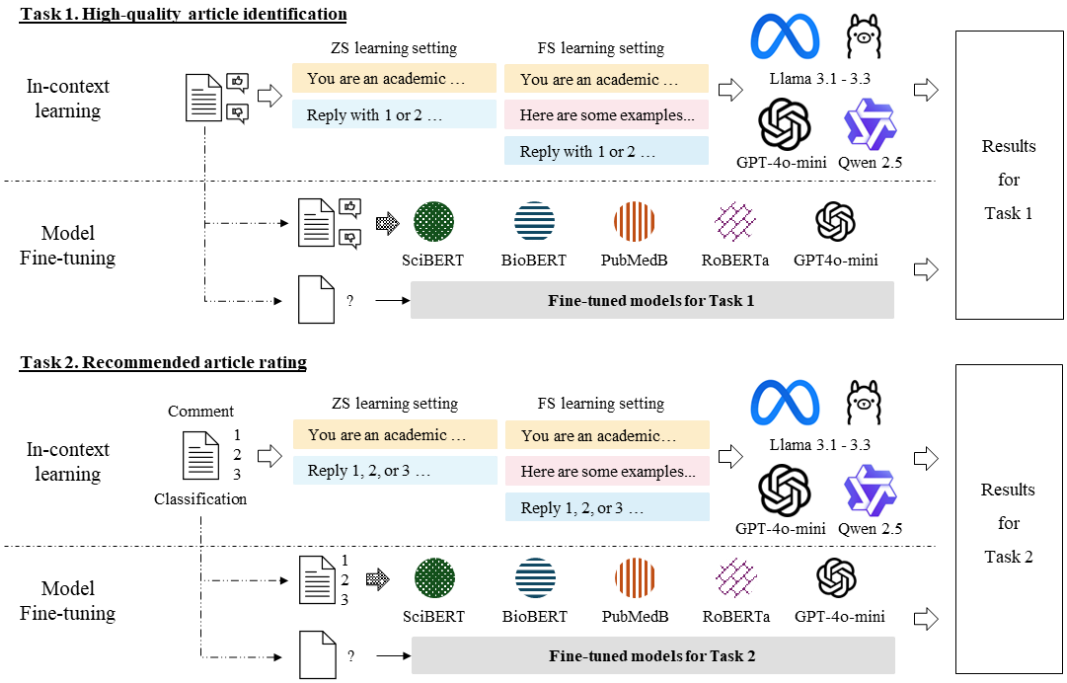

Figure 1. The overall research framework.

Model selection

Four BERT variant models: SciBERT (Beltagy et al., 2019), BioBERT (Lee et al., 2020), RoBERTa (Liu et al., 2019), and PubMedBERT (Gu et al., 2021) are encoder-only language models built on the transformer architecture, which converts the input language as embeddings for downstream analysis. The key distinction among these models lies in their training corpora and methods. SciBERT is tailored for scientific NLP tasks, pre-trained on 1.14 million scientific articles from

Semantic Scholar[1]. BioBERT extends the original BERT pretraining corpus by incorporating 29 million PubMed abstracts and full-text articles from PubMed Central[2], enhancing its performance in the biomedical domain. PubMedBERT also targets biomedical domain, but it exclusively uses PubMed abstracts and PubMed Central full-text articles for pretraining, omitting the general BERT corpus, which makes it more specialized for biomedical tasks. RoBERTa, a refined version of BERT, optimizes the pretraining procedures with modified training parameters and task settings, improving model efficiency and performance while retaining general-purpose applicability.

Current generative LLMs generally employ decoder-only architecture, enabling them to generate text sequences directly based on the given natural language input. The widespread adoption of ChatGPT has shown the remarkable capabilities of such models in language comprehension, text generation, and question-answering. Apart from GPT models, multiple big tech companies have developed and released open-source models for public access and use, represented by Llama models from Meta (formerly Facebook) and Qwen models from Alibaba. Given that, we selected multiple representative open- and closed-source models considering computing budget and time costs. For open-source models, we intentionally chose both the smallest (3B or 7B, in which B indicates billion parameters) and largest versions (70B or 72B) to test how the model size can affect evaluation results. The tested models in the final pool include: GPT-4o-mini (Achiam et al., 2023) from OpenAI, Llama 3.1-8B, Llama 3.2-3B, and Llama 3.3-70B from Meta (Dubey et al., 2024), as well as Qwen 2.5-7B and Qwen 2.5-72B from Alibaba (Yang et al., 2024).

ICL for generative LLMs

ICL is a prompt-engineering technique designed for generative LLMs (GPT-4o-mini, Llama, and Qwen models in this paper). ICL works by providing contextual information, sometimes along with task-specific input-output pair demonstrations directly in the prompts, enabling models to generate responses for given questions. Unlike fine-tuning, ICL does not alter the model's parameters; instead, it modifies the prompts to achieve more accurate outputs. This makes ICL a low-cost and user-friendly approach to leveraging LLMs. In this study, we employed two of the most prevalent ICL prompting schemes:

- Zero-shot (ZS) learning setting: In the ZS setting, the prompt only includes descriptions of the task as contextual information. The LLMs generate recommendations (Task 1) or ratings (Task 2) for each article without any additional contextual information.
- Few-shot (FS) learning setting: In the FS setting, the prompt includes both the task description and five demonstrations of input-output pairs (see the Supplementary Material) for each class. For Task 1, five recommended articles and five non-recommended articles, along with their abstracts and expert recommendations, are provided. For Task 2, five articles from each rating

[1] https://www.semanticscholar.org
[2] https://pubmed.ncbi.nlm.nih.gov

category (1, 2, and 3) are presented with their ratings. The demonstrations are selected randomly, and each inference is conducted using a different set of demonstrations.

ICL is an idealized learning setting that anticipates LLMs to complete the tasks accurately with the given contextual information (task description) or a few samples. We designed three sets of prompt templates (p1-p3) for Task 1 and Task 2 to instruct generative LLMs. The prompts and their corresponding usage for each task are provided in the Supplementary Material.

Language model fine-tuning

Fine-tuning is a model retraining method that adapts LLMs to specific tasks by updating their parameters using labelled data (in our case, the labels are article recommendations and ratings). Unlike training from scratch, fine-tuning can retain knowledge learnt during the pre-training stage in the retraining process. However, compared to ICL, fine-tuning, especially for generative LLMs, is much more computationally intensive. Additionally, fine-tuned models tend to be more task-specific, which may reduce their generalizability. This strategy can be applied to both BERT models and generative LLMs. Due to the high computational costs of fine-tuning the selected generative LLMs on local machines, we only applied this learning setting for BERT models and the GPT-4o-mini model (through the OpenAI API).

Validation metrics

Four validation metrics were employed to measure the models' performance in Task 1 and Task 2. The definitions and calculations are given as follows:

- Accuracy (A): Accuracy measures the ratio of correctly classified articles to all articles.
- Precision (P): For a specific category, P is the ratio of correctly classified articles to all articles predicted as positive for that class.
- Recall (R): For a specific category, R is the ratio of correctly classified articles to all articles that belong to that class.
- Cohen's kappa coefficient ($\boldsymbol{\kappa}$): κ measures the level of agreement between a LLM and a human expert on the classification task. It ranges from -1 to 1, with the larger value indicating higher agreement.

$$\kappa = \frac{A - p_e}{1 - p_e}$$

$$p_e = \frac{1}{N}\sum_k n_L^k n_H^k$$

N is the total number of articles and k is the number of categories to be classified (recommended or not recommended in Task 1, rating of 1, 2, or 3 in Task 2), n_L^k and n_H^k, respectively, denote the number of articles classified to category k by LLMs (L) and human experts (H). Landis and Koch (1977) characterize values < 0

as indicating no agreement and 0-0.20 as slight agreement, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1 as almost perfect agreement.

Results

Results for high-quality article identification (Task 1)

For the ICL strategies, we tested all generative LLMs on all articles in Task 1 (a total of 9,532 articles). The predictive results are shown in Figures 2 and 3, where the green and red areas represent the outputs recommended and not recommended respectively, and deep and light colors refer to articles correctly or wrongly classified (the sum of each bar may be slightly smaller than 9,532 due to a few invalid answers from LLMs).

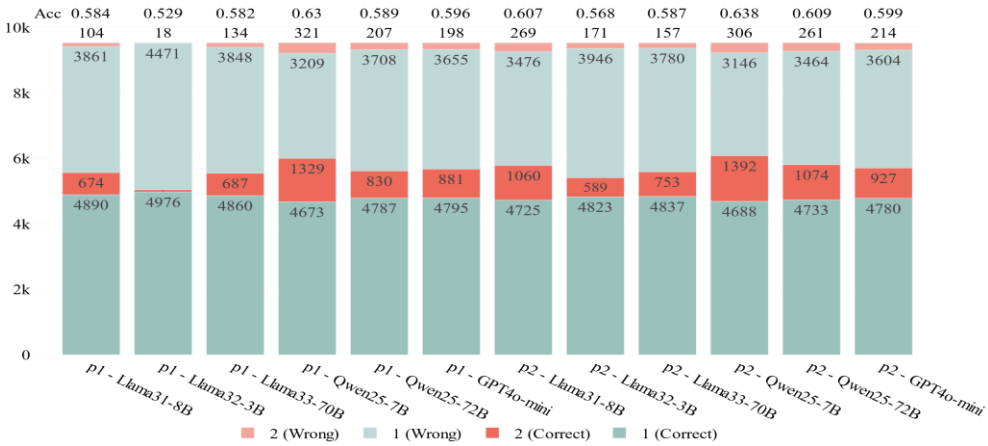

Figure 2. Model results using the ZS learning setting in Task 1.

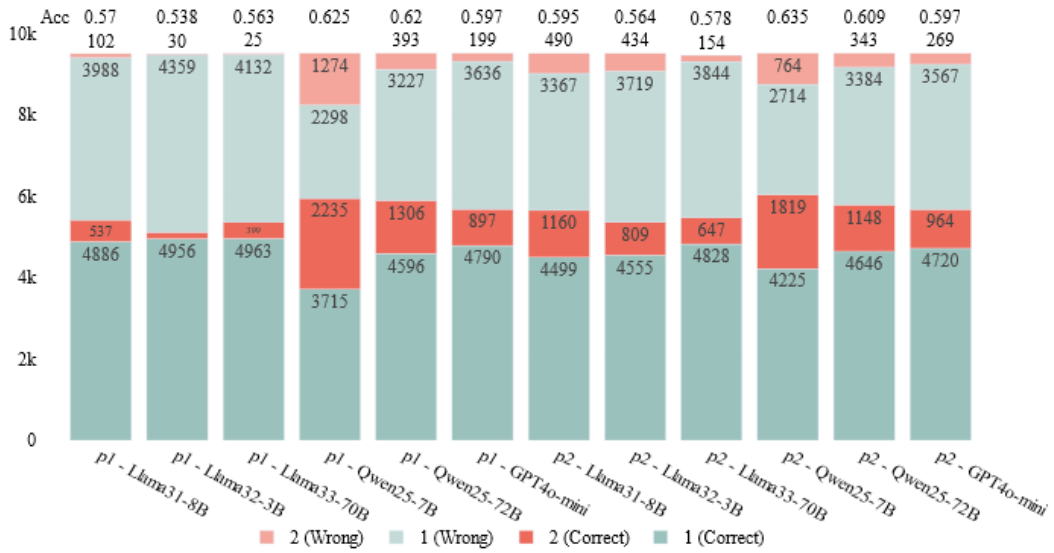

Figure 3. Model results using the FS learning setting in Task 1.

In both Figures 1 and 2, p1 refers to the prompt without evaluation criteria details given, and p2 refers to the prompt with evaluation criteria details (see the Supplementary Material). The accuracy, precision, and recall metrics for the generated answers are provided in Table 2.

Under ICL settings, the overall accuracy of tested LLMs is around 0.6, which is barely satisfying for a binary classification task. It can be observed from Figures 2 and 3 that most LLMs are inclined to generate biased positive answers (recommended) for articles – even though half of them did not receive any recommendations from human experts. This tendency is also reflected in the generally low recall rate for the "not recommended" class in Table 2. Besides, the performance of the closed-source model, GPT-4o-mini, does not show significant advancements compared to other open-source language models.

Despite that, the model outputs are also subject to which prompt and what learning setting were used. In Figures 2 and 3, the accuracies of most models increase when changing the prompt from p1 to p2, i.e., using more detailed evaluation criteria in the prompt. Details of the evaluation criteria are essentially critical for LLMs to give more accurate justification for article recommendations.

However, switching from ZS to FS setting, i.e., providing some examples to LLMs, does not let LLMs make more accurate recommendations. It increased the ratio of articles predicted as "not recommended", but the accuracy did not improve accordingly. In other words, showing both positive and negative samples, i.e., articles recommended and not recommended by human experts to LLMs, can help them to produce more critical opinions, but the alignment with human experts still struggles. This indicates that article evaluation can be a complex and long content-dependent task – realizing human-level judgment may still require a deeper understanding of articles than a few examples can provide.

When comparing results from smaller versions of models to larger versions under ICL settings, the accuracies did not show significant improvements – in most cases, the accuracy dropped slightly. Although it has been proven that larger models can perform significantly better in most generalized tasks (Touvron et al., 2023; Yang et al., 2024), our results indicate that model size is not a decisive factor in this pure binary classification task of differentiating recommended and not recommended articles under ICL settings.

The results of the fine-tuned models are presented in Table 2. Under the fine-tuning learning setting, both generative LLMs and BERT models are retrained to learn patterns for recommending articles from labelled data and then used to predict unseen records. We split the dataset into an 80% training set and a 20% test set. The optimal learning rate and number of training epochs were empirically determined by monitoring the training and validation loss.

Table 2. LLM results for Task 1 under ZS, FS, and fine-tuning learning settings*.

Setting	Prompt	Model	A	κ	P (Y)	R (Y)	P (N)	R (N)
ICL	ZS	p1	Llama 3.1-8B	0.584	0.133	0.559	0.979	**0.866**	0.149
			Llama 3.2-3B	0.529	0.012	0.527	**0.996**	0.788	0.015
			Llama 3.3-70B	0.582	0.13	0.558	0.973	0.837	0.151
			Qwen 2.5-7B	0.630	0.235	0.593	0.936	0.805	0.293
			Qwen 2.5-72B	0.589	0.147	0.564	0.959	0.8	0.183
			GPT-4o-mini	0.596	0.160	0.567	0.960	0.816	0.194
		p2	Llama 3.1-8B	0.607	0.186	0.576	0.946	0.798	0.234
			Llama 3.2-3B	0.568	0.099	0.55	0.966	0.775	0.130
			Llama 3.3-70B	0.587	0.14	0.561	0.969	0.827	0.166
			Qwen 2.5-7B	**0.638**	**0.253**	**0.598**	0.939	0.82	**0.307**
			Qwen 2.5-72B	0.609	0.191	0.577	0.948	0.804	0.237
			GPT-4o-mini	0.599	0.168	0.57	0.957	0.812	0.205
	FS	p1	Llama 3.1-8B	0.57	0.102	0.551	0.980	0.84	0.119
			Llama 3.2-3B	0.538	0.031	0.532	0.994	0.845	0.036
			Llama 3.3-70B	0.563	0.087	0.546	**0.995**	**0.941**	0.088
			Qwen 2.5-7B	0.625	0.24	**0.618**	0.745	0.637	**0.493**
			Qwen 2.5-72B	0.620	0.215	0.587	0.921	0.769	0.288
			GPT-4o-mini	0.597	0.164	0.568	0.96	0.818	0.198
		p2	Llama 3.1-8B	0.595	0.163	0.572	0.902	0.703	0.256
			Llama 3.2-3B	0.564	0.095	0.551	0.913	0.651	0.179
			Llama 3.3-70B	0.578	0.118	0.557	0.969	0.808	0.144
			Qwen 2.5-7B	**0.635**	**0.253**	0.609	0.847	0.704	0.401
			Qwen 2.5-72B	0.609	0.19	0.579	0.931	0.77	0.253
			GPT4o-mini	0.597	0.164	0.57	0.946	0.782	0.213
Fine-tuned on the training set (80% data)			SciBERT	0.785	0.564	0.764	0.863	0.817	0.696
			BioBERT	0.789	0.574	0.778	0.845	0.804	0.725
			RoBERTa	0.761	0.512	0.726	**0.885**	0.825	0.62
			PubMedBERT	0.802	0.599	0.784	0.866	**0.827**	0.728
			GPT-4o-mini	**0.84**	**0.679**	**0.878**	0.811	0.802	**0.872**

* Note: Results in bold font indicate the best accuracy, underlined results are the second best. We separated the comparison by experimental settings (ZS, FS and fine-tuning).

The fine-tuned GPT-4o-mini achieved the highest accuracy among all models, including the fine-tuned BERT models, which utilize encoder-only architectures optimized for tasks like text understanding and classification rather than generation. This result highlights the superiority of larger-scale LLMs in handling versatile tasks and supports the scaling law in language models (Kaplan et al., 2020), which suggests that model performance improves to some extent with increasing size.

BERT models typically have around 110 million parameters, while GPT models often utilize models with billions of parameters.

To compare the inter-model agreement on Task 1, we depicted the heatmap based on the pairwise κ of model outputs in Figure 4 – the darker the red, the higher the agreement is between the models. The overall agreement with human experts is the same as reflected by accuracy: Fine-tuned models are generally above 0.6 but models under the ICL settings are all lower than 0.3. Regarding the inter-model agreement, fine-tuned models show satisfying moderate agreements above 0.6, following the interpretation of Landis and Koch (1977). Notably, some models under the ICL settings also exhibit good inter-model agreement (above 0.6), including Qwen 2.5-72B, GPT-4o-mini, and Llama 3.3-70B, which are all LLMs in their larger versions. These results indicate that larger models may have more consistent behaviors when dealing with the less complicated Task 1. The results should be interpreted against the backdrop of results on the agreement of reviewers from the (pre-publication) peer review process. The results of a meta-analysis of Bornmann et al. (2010) including several primary journal peer review studies show that the agreement between reviewers assessing the same manuscript is low (in general): The pooled κ across 48 studies is 0.17. The results for the agreement of human experts and models are relatively high in this study compared to the results from the meta-analysis.

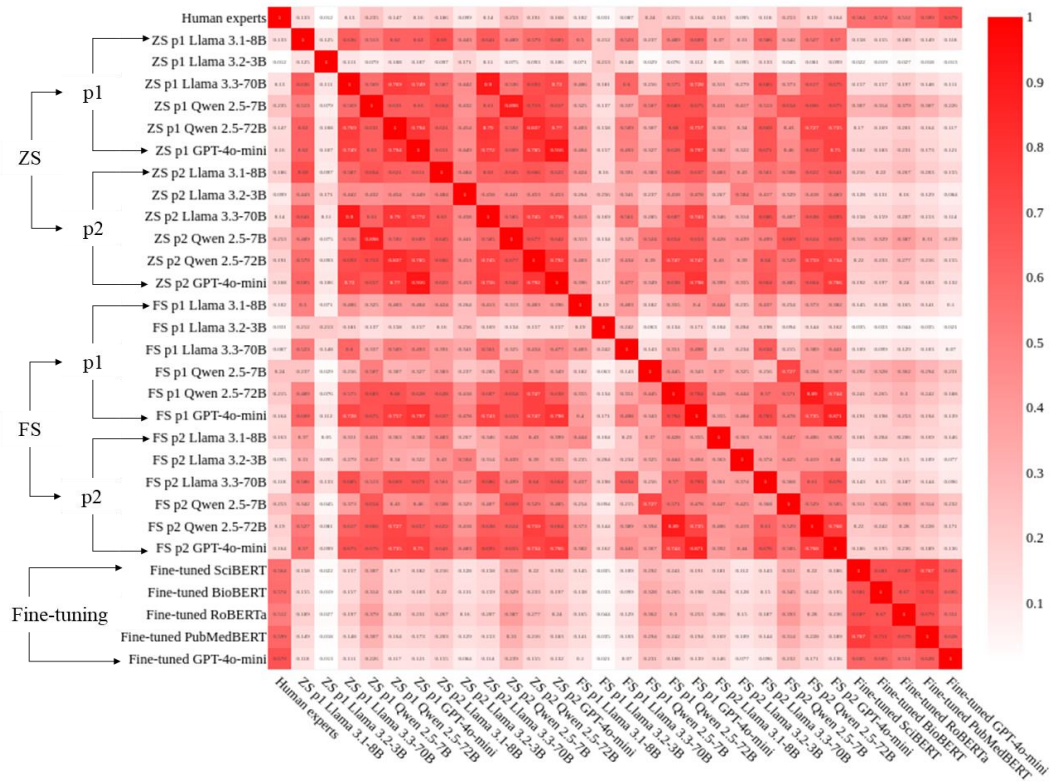

Figure 4. The heatmap of κ between LLM outputs in Task 1.

Results for recommended article rating (Task 2)

In Task 2, prompt p3 (see the Supplementary Material) was used to instruct LLMs to give ratings of 1, 2, and 3 for each article provided. The results are presented in Figure 5 and Table 3. In Figure 5, the colors represent three different ratings: Green – 1, Red – 2, and Yellow – 3 (the sum of each bar may be slightly smaller than 15,000 due to a few invalid answers from LLMs). The overall low accuracy below 0.4 highlights the challenge of differentiating article ratings under the ICL settings. Among the models, Qwen 2.5-72B achieved the highest accuracy but still presented a relatively biased preference for ratings of 2 and 3. Llama 3.1-8B within the FS setting yielded rather balanced predictions but suffered from lower accuracy. The other models, excluding Llama 3.1-8B and Llama 3.2-3B models under the FS setting, tend to show the inclination to ratings of 2 and 3.

Unlike Task 1, switching from the ZS to the FS setting significantly altered the outputs of most models, but the direction of this change depends on which specific model is used: Llama 3.1-8B produced much more balanced results with the few samples provided, Llama 3.2-3B changed its main preference from ratings of 2 to 1, results from Llama 3.3-70B did not change much, FS increases the number of ratings of 2 and 3 for Qwen 2.5-72B and GPT-4o-mini. However, the accuracy of all model outputs still did not improve much. Despite those changes, the results endorse our previous claim in Task 1: The regular FS learning setting is not an effective learning strategy for research evaluation tasks.

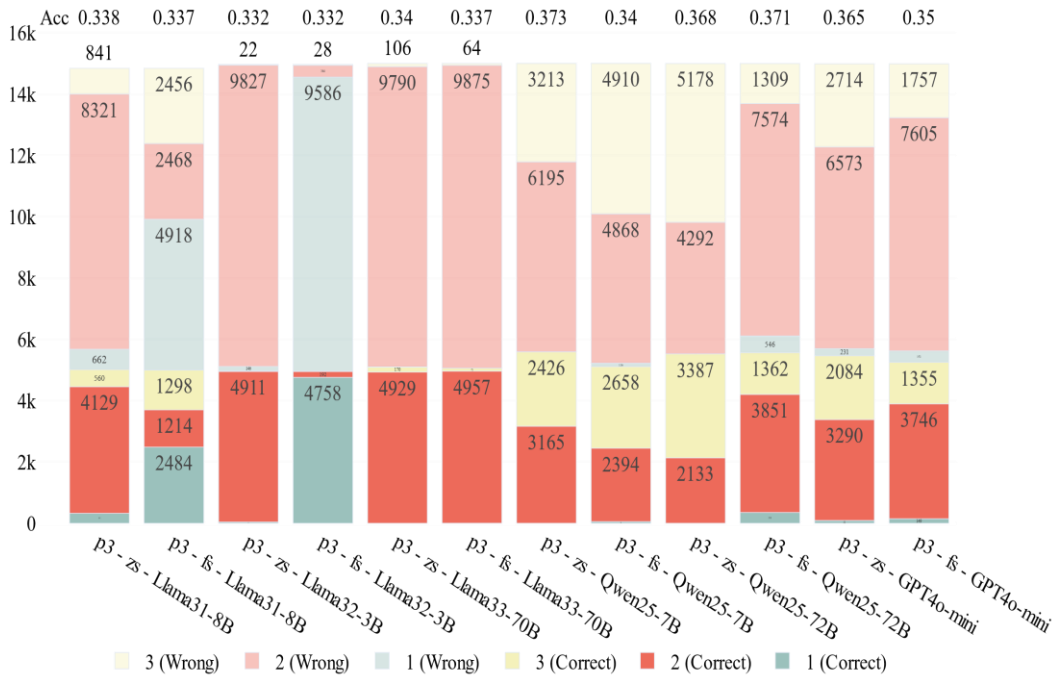

Figure 5. LLM results for Task 2 under ZS and FS learning settings.

Table 3. LLM results for Task 2 under ZS, FS, and fine-tuning settings*

	Setting	Model	A	κ	P1*	R1*	P2	R2	P3	R3
ICL	ZS	Llama 3.1-8B	0.334	0.007	0.329	**0.065**	0.332	0.826	0.4	0.112
		Llama 3.2-3B	0.332	0.001	0.22	0.008	0.334	0.985	0.421	0.003
		Llama 3.3-70B	0.34	0.01	<1e-3	<1e-3	0.335	**0.986**	**0.616**	0.034
		Qwen 2.5-7B	**0.373**	**0.059**	**1**	<1e-3	**0.338**	0.633	0.43	0.485
		Qwen 2.5-72B	0.368	0.052	0.2	<1e-3	0.332	0.427	0.396	**0.678**
		GPT-4o-mini	0.364	0.047	0.286	0.019	0.333	0.658	0.434	0.417
	FS	Llama 3.1-8B	0.336	0.005	0.335	0.502	0.329	0.245	0.347	0.262
		Llama 3.2-3B	0.331	0.002	0.332	**0.952**	0.333	0.038	0.349	0.003
		Llama 3.3-70B	0.337	0.006	**0.750**	0.001	0.334	**0.991**	**0.602**	0.019
		Qwen 2.5-7B	0.34	0.011	0.318	0.011	0.33	0.479	0.351	**0.532**
		Qwen 2.5-72B	**0.371**	**0.057**	0.392	0.07	**0.337**	0.77	0.51	0.272
		GPT-4o-mini	0.35	0.025	0.282	0.03	0.33	0.749	0.436	0.271
Fine-tuning	Test set	SciBERT	0.453	0.176	0.466	0.621	0.344	0.263	0.525	0.463
		BioBERT	0.458	0.182	0.459	0.68	**0.361**	0.208	0.515	0.47
		RoBERTa	0.452	0.172	0.442	**0.719**	0.357	0.162	0.518	0.455
		PubMedBERT	0.461	0.187	0.464	0.68	**0.361**	0.231	0.527	0.456
		GPT-4o-mini	**0.463**	**0.195**	**0.533**	0.466	0.348	0.395	0.527	0.526
	Extra test set	SciBERT	0.493	0.111	0.629	0.712	0.394	0.186	0.146	0.349
		BioBERT	0.499	0.122	0.627	0.716	0.418	0.183	0.162	**0.402**
		RoBERTa	0.492	0.098	0.616	**0.728**	0.383	0.156	0.153	0.357
		PubMedBERT	0.512	0.14	0.635	0.715	**0.453**	0.23	0.158	0.361
		GPT-4o-mini	**0.561**	**0.165**	**0.653**	0.682	0.431	**0.493**	**0.444**	0.036

* Note: P1 and R1 respectively refers to the precision and recall of category 1. Results in bold font indicate the best accuracy, underlined results are the second best. We separated the comparison by experimental settings (ZS, FS and fine-tuning).

In addition to the standard test set, we created an extra test set for Task 2 to validate the performance of the fine-tuned models in real-world settings. The new test set is a dataset that simulates the imbalanced distribution of ratings in real-world scenarios – containing 1,675 records of rating 1, 1,076 records of rating 2, and 249 records of rating 3. This corresponds roughly to a 8:5:1 ratio as introduced in the full dataset we collected.

The results indicate that the fine-tuned GPT-4o-mini achieved the overall best performance on both test sets, especially on the extra real-world simulated test set. The second best-performing fine-tuned model is PubMedBERT, the BERT variant

trained specifically on PubMed articles corpora. Generally, all the language models tested on this task presented an accuracy lower than 0.6 and a κ agreement with human experts below 0.2. The low measures indicate that in Task 2, the differences between the three ratings are much more nuanced than in Task 1. It seems that Task 2 is more challenging for language models to learn different articles' quality based on their abstracts.

The inter-model κ agreement of Task 2 is visualized in Figure 6. Compared to Task 1, the agreement among fine-tuned models and models under ICL settings both dropped to lower than 0.6 and 0.2. Despite generally low agreement of LLMs under ICL settings, Qwen 2.5-72B and GPT-4o-mini still showed relatively high agreement with each other.

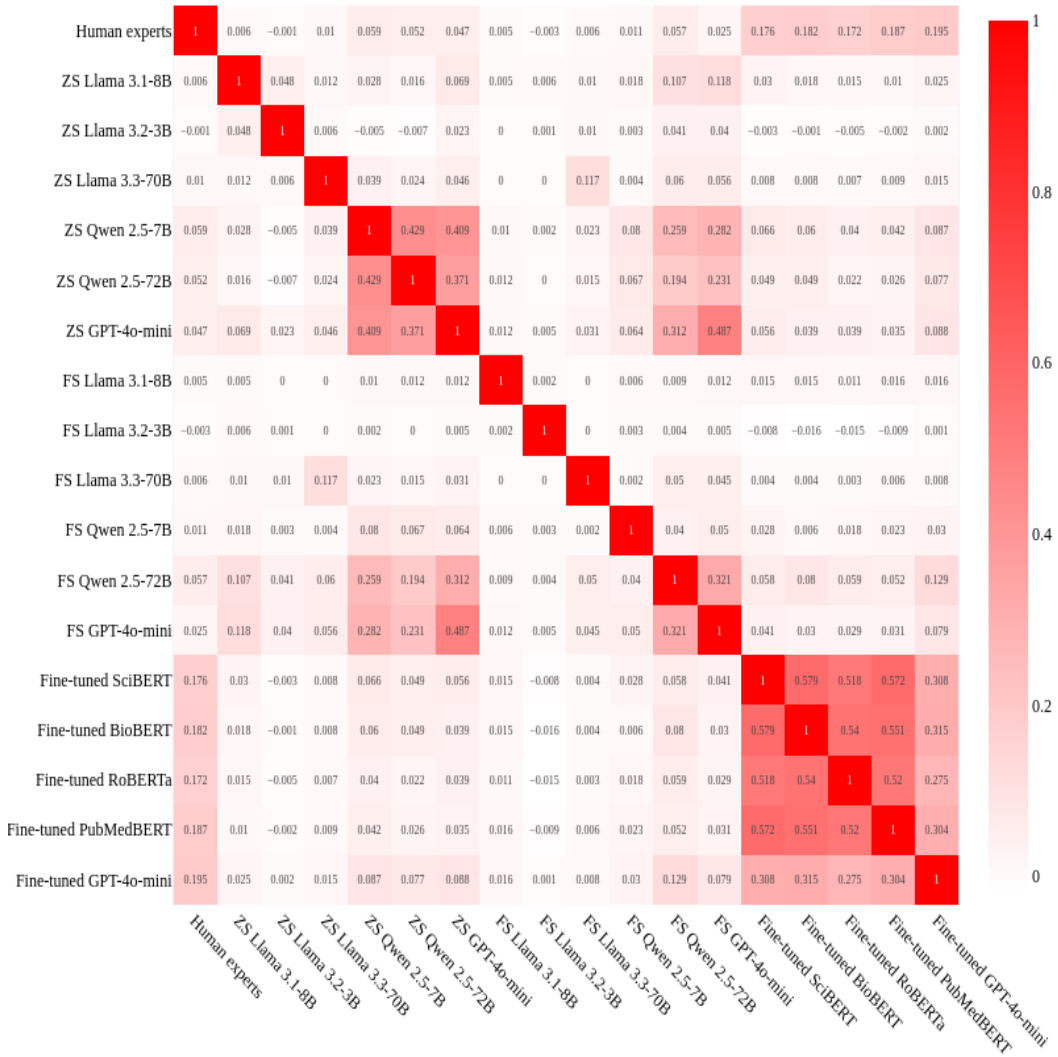

Figure 6. The heatmap of κ between LLM outputs in Task 2.

Conclusions

In this study, we performed a thorough comparison of current LLMs' performance on research evaluation tasks under ICL (ZS and FS) and fine-tuning learning settings, providing insights into leveraging LLMs for post-publication review and rating. Overall, our results demonstrate that LLMs fine-tuned with partial human expert annotations can serve as a preliminary tool for initial research evaluation. However, more complicated tasks, like rating articles on a specific scale, are more challenging and may require more resources and sophisticated methodologies. More specifically, the key findings of this study are as follows:

- Among the three model learning settings, fine-tuning works significantly better and aligns with expert opinions the most, but this comes with a trade-off of requiring a certain amount of existing training data. The idealized settings of utilizing LLMs, like ZS and FS, which anticipate LLMs to perform evaluation independently or with very limited contextual information, are still compromised in their alignment with human experts in real-world practice.
- Among the fine-tuned models, GPT-4o-mini is the best among the tested LLMs, including BERT-based models and open-sourced generative LLMs.
- Under the fine-tuning setting, LLMs can offer relatively satisfying performance on identifying high-quality articles (Task 1) with very little training data but may struggle to accurately rate recommended articles (Task 2). The selected LLMs, even after fine-tuning, are still prone to giving biased answers that are different from those of human experts.

Limitations and future directions

Certain limitations come with this work. First, we did not apply fine-tuning strategies on open-source LLMs like Qwen and Llama due to the restraints from high computational resource requirements, leading to the lack of comparison of those options in our study. Second, in this paper, we only fed article abstracts to LLMs for evaluation, which contain very concise and limited information and may be insufficient for evaluating the overall quality of research articles. Third, LLMs are the only knowledge sources for performing research evaluation tasks. No external data sources, which can be academic knowledge graphs containing more enriched information, have been leveraged. Aiming to equip LLMs with better capabilities and accuracy of research evaluation, the future directions of this study will spread to three perspectives: (1) employ more computational resources to realize fine-tuning on open-source LLMs, (2) develop a work pipeline for multi-modal LLMs to systematically process article full texts with figures and tables affiliated, and (3) incorporate external data resources with LLMs to realize enriched context-aware evaluation.

Acknowledgments

Mengjia Wu and Yi Zhang are supported by the Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia, in conjunction with the National Science Foundation (NSF) of the United States, and CSIRO-NSF

#2303037. We would like to thank H1 staff for providing data access. Access to OpenAlex bibliometric data has been supported via the German Competence Network for Bibliometrics (Schmidt et al., 2024), funded by the Federal Ministry of Education and Research (grant number: 16WIK2101A).

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., & Anadkat, S. (2023). *GPT-4 technical report* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2303.08774

Algaba, A., Mazijn, C., Holst, V., Tori, F., Wenmackers, S., & Ginis, V. (2024). *Large language models reflect human citation patterns with a heightened citation bias* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2405.15739

Beltagy, I., Lo, K., & Cohan, A. (2019). *SciBERT: A pretrained language model for scientific text* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1903.10676

Bornmann, L. (2008). Scientific peer review. An analysis of the peer review process from the perspective of sociology of science theories. *Human Architecture - Journal of the Sociology of Self-Knowledge*, *6*(2), 23-38. http://www.okcir.com/HAVI2SPRING2008.html

Bornmann, L. (2011). Scientific peer review. *Annual Review of Information Science and Technology*, *45*, 199-245. https://doi.org/10.1002/aris.2011.1440450112

Bornmann, L., Haunschild, R., & Mutz, R. (2021). Growth rates of modern science: A latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, *8*(1), 224. https://doi.org/10.1057/s41599-021-00903-w

Bornmann, L., Mutz, R., & Daniel, H.-D. (2010). A reliability-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants. *PloS one*, *5*(12), e14331.

de Winter, J. (2024). Can ChatGPT be used to predict citation counts, readership, and social media interaction? An exploration among 2222 scientific abstracts. *Scientometrics*, *129*(4), 2469-2487. https://doi.org/10.1007/s11192-024-04939-y

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., & Fan, A. (2024). *The llama 3 herd of models* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2407.21783

Farhat, F., Sohail, S. S., & Madsen, D. Ø. (2023). *How trustworthy is chatGPT? The case of bibliometric analyses*. Retrieved August, 19 from https://doi.org/10.20944/preprints202303.0479.v1

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, *3*(1), 1-23.

Jia, R., Wu, M., Ding, Y., Lu, J., & Zhang, Y. (2025). *HetGCoT-Rec: Heterogeneous Graph-Enhanced Chain-of-Thought LLM Reasoning for Journal Recommendation* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2501.01203

Joe, E. T., Koneru, S. D., & Kirchhoff, C. J. (2024). *Assessing the effectiveness of GPT-4o in climate change evidence synthesis and systematic assessments: Preliminary insights*. Retrieved July, 23 from https://arxiv.org/abs/2407.12826

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling laws for neural language models* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2001.08361

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234-1240.

Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D. Y., Yang, X., Vodrahalli, K., He, S., Smith, D. S., & Yin, Y. (2024). Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI*, *1*(8), AIoa2400196.

Liu, R., & Shah, N. B. (2023). *Reviewergpt? an exploratory study on using large language models for paper reviewing* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2001.08361

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *Roberta: A robustly optimized bert pretraining approach* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1907.11692

López-Pineda, A., Nouni-García, R., Carbonell-Soliva, Á., Gil-Guillén, V. F., Carratalá-Munuera, C., & Borrás, F. (2025). Validation of large language models (Llama 3 and ChatGPT-4o mini) for title and abstract screening in biomedical systematic reviews. *Research Synthesis Methods*, 1-11.

Priem, J., Piwowar, H., & Orr, R. (2022). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2205.01833

Sandnes, F. E. (2024). Can we identify prominent scholars using ChatGPT? *Scientometrics*, *129*(1), 713-718. https://doi.org/10.1007/s11192-023-04882-4

Schmidt, M., Rimmert, C., Stephen, D., Lenke, C., Donner, P., Gärtner, S., Taubert, N., Bausenwein, T., & Stahlschmidt, S. (2024). *The data infrastructure of the German Kompetenznetzwerk Bibliometrie: An enabling intermediary between raw data and analysis*. Zenodo. Retrieved October 28, 2024 from https://doi.org/10.5281/zenodo.13935407

Thelwall, M. (2024a). Can ChatGPT evaluate research quality? *Journal of Data and Information Science*, *9*(2), 1-21.

Thelwall, M. (2024b). *Quantitative Methods in Research Evaluation Citation Indicators, Altmetrics, and Artificial Intelligence* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2407.00135

Thelwall, M., & Yaghi, A. (2024). *In which fields can ChatGPT detect journal article quality? An evaluation of REF2021 results* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2409.16695

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., & Azhar, F. (2023). *Llama: Open and efficient foundation language models* arXiv. https://doi.org/10.48550/arXiv.2302.13971

Vital Jr, A., Silva, F. N., Oliveira Jr, O. N., & Amancio, D. R. (2024). *Predicting citation impact of research papers using GPT and other text embeddings* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2407.19942

Wu, M., Sivertsen, G., Zhang, L., Qi, F., & Zhang, Y. (2024). *Scientific progress or societal progress? A language model-based classification of the aims of the research in scientific publications.* 28th International Conference on Science, Technology and Innovation Indicators (STI2024), Berlin, Germany.

Wu, M., Zhang, Y., Zhang, G., & Lu, J. (2021). Exploring the genetic basis of diseases through a heterogeneous bibliometric network: A methodology and case study. *Technological Forecasting and Social Change*, *164*, 120513.

Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., & Huang, F. (2024). *Qwen2 technical report* [Preprint]. arXiv. https://doi.org/https://doi.org/10.48550/arXiv.2407.10671

Supplementary Material

Prompt for Task 1

p1 (used for ZS and fine-tuning):

You are an academic expert in the biomedical field, evaluating research articles based on scientific rigor, replicability, data analysis, and study limitations. You will summarize each article as "recommend" or "not recommend" by reading the abstracts.
Reply with 1 for recommending this article and 2 for not recommending it.
Reply with 1 or 2 and nothing else.

p2 (used for ZS and fine-tuning):

You are an academic expert in the biomedical field, evaluating research articles based on scientific rigor, replicability, data analysis, and study limitations. You will summarize each article as "recommend" or "not recommend" by reading the abstracts.

- *Scientific rigor is the strict application of the scientific method to ensure robust and unbiased experimental design, methodology, analysis, interpretation and reporting of results.*
- *Replicability is obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data.*
- *Data analysis is the practice of working with data to glean useful information, which can then be used to make informed decisions.*
- *Study limitations are the constraints placed on the ability to generalize from the results, to further describe applications to practice, and/or related to the utility of findings that are the result of the ways in which you initially chose to design the study, or the method used to establish internal and external validity or the result of unanticipated challenges that emerged during the study.*

Reply with 1 for recommending this article and 2 for not recommending it.
Reply with 1 or 2 and nothing else.

Additional FS demonstrations:

Here are some examples from human expert recommendations:

Traditional epidural techniques have been ... [Abstract with three or more recommendations]

1 *+ 4 more abstracts in this category*

The nature of "climate change" will differ ... [Abstract with no recommendations]

2 *+ 4 more abstracts in this category*

Prompt for Task 2

p3 (used for rating classification)

You are an academic expert in the biomedical field, evaluating research articles based on scientific rigor, replicability, data analysis, and study limitations. The definitions of the evaluation dimensions are as follows:

- *Scientific rigor is the strict application of the scientific method to ensure robust and unbiased experimental design, methodology, analysis, interpretation and reporting of results.*
- *Replicability is obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data.*
- *Data analysis is the practice of working with data to glean useful information, which can then be used to make informed decisions.*
- *Study limitations are the constraints placed on the ability to generalize from the results, to further describe applications to practice, and/or related to the utility of findings that are the result of the ways in which you initially chose to design the study, or the method used to establish internal and external validity or the result of unanticipated challenges that emerged during the study.*

You will summarize your rating using 1, 2, or 3, representing "Good," "Very Good," and "Exceptional" quality. Just reply with 1, 2, or 3 and nothing else.

Additional FS demonstrations:

Here are some examples from human expert recommendations:

Significance Fluorescent auxin ... (Abstract with rating of 1)

1 *+ 4 more abstracts in this category*

Several surveys have observed different degrees ... (Abstract with rating of 2)

2 *+ 4 more abstracts in this category*

The long-standing belief that nitrogen-containing ... (Abstract with rating of 3)

3 *+ 4 more abstracts in this category*