Journal of the American Medical Informatics Association, 2023, 1–13
doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

A Novel Question-Answering Framework for Automated Citation Screening Using Large Language Models

Opeoluwa Akinseloyin,[1] Xiaorui Jiang[1] and Vasile Palade[1]

[1]Centre for Computational Science and Mathematical Modelling, Coventry University, Puma Way, CV1 2TT, Coventry, United Kingdom

*Xiaorui Jiang. xiaorui.jiang@coventry.ac.uk

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Objective: This paper aims to address the challenges in citation screening (a.k.a. abstract screening) within Systematic Reviews (SR) by leveraging the zero-shot capabilities of large language models, particularly ChatGPT.
Methods: We employ ChatGPT as a zero-shot ranker to prioritize candidate studies by aligning abstracts with the selection criteria outlined in an SR protocol. Citation screening was transformed into a novel question-answering (QA) framework, treating each selection criterion as a question addressed by ChatGPT. The framework involves breaking down the selection criteria into multiple questions, properly prompting ChatGPT to answer each question, scoring and re-ranking each answer, and combining the responses to make nuanced inclusion or exclusion decisions.
Results: Large-scale validation was performed on the benchmark of CLEF eHealth 2019 Task 2: Technology Assisted Reviews in Empirical Medicine. Across 31 datasets of four categories of SRs, the proposed QA framework consistently outperformed other zero-shot ranking models. Compared with complex ranking approaches with iterative relevance feedback and fine-tuned deep learning-based ranking models, our ChatGPT-based zero-shot citation screening approaches still demonstrated competitive and sometimes better results, underscoring their high potential in facilitating automated systematic reviews.
Conclusion: Investigation justified the indispensable value of leveraging selection criteria to improve the performance of automated citation screening. ChatGPT demonstrated proficiency in prioritizing candidate studies for citation screening using the proposed QA framework. Significant performance improvements were obtained by re-ranking answers using the semantic alignment between abstracts and selection criteria. This further highlighted the pertinence of utilizing selection criteria to enhance citation screening.

Key words: Automated Systematic Review, Citation Screening, ChatGPT, Zero-Shot Ranking, Question Answering

Introduction

A Systematic Review (SR) in medical research is the highest form of knowledge synthesis of all available medical evidence from relevant publications on a specific topic. SR follows a principled pipeline, including candidate study retrieval, primary study selection, quality assessment, data extraction, data synthesis, meta-analysis, and reporting [1]. Because of its thoroughness and reliability, SR underpins evidence-based medicine [2]. It shapes medical research and practice by informing researchers of the state-of-the-art knowledge and knowledge gaps as well as health practitioners and policymakers of the best clinical practice [3].

SR also faces tremendous challenges at each step. For instance, it is time-consuming, expensive and resource-intensive to select primary studies, a.k.a. *citation screening*, due to the massive volume of retrieved candidate studies, often at tens of thousands [4, 5]. It is further worsened by involving multiple human annotators, which is required to reduce bias and disparities [6]. This compound complexity calls for innovative solutions to automate or semi-automate citation screening [1] to minimize the time delays and costs of this manual screening tasks [7], which is the focus of the current paper. Figure 1 shows an example of citation screening, where the abstract of an included study is matched against the selection criteria defined in the SR protocol.

© The Author 2023. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Confidential: For Review Only

Machine learning has been the focus of research in automating citation screening [1, 7, 8]. Firstly, a small set of studies are selected for human annotation, and then a classifier is trained. Typically, active learning is adopted to improve the classifier iteratively. Obviously, the quality of the initial annotations plays an important role. However, choosing initial annotations is a problem of zero-shot setting and has not been explored at all. Another disadvantage is that this approach is not generalisable, and each SR topic requires training a bespoke classifier from scratch.

An alternative perspective was to treat citation screening as a ranking problem a.k.a. *reference prioritisation* [7], incorporating approaches from the information retrieval (IR) community [9, 10, 11, 12, 13, 14, 15, 16]. One advantage of this approach is that it can utilise additional information about an SR, which is converted into queries to enhance screening performance. Such information could be review title [9, 10], original Boolean queries (for candidate study retrieval) [17], research objectives [18, 16], or a set of seed studies [15, 19]. Another advantage is the possibility of training a cross-topic ranker to generalise to diverse SR topics.

The above analysis motivated us to explore the emerging capabilities of Large Language Models (LLMs), particularly ChatGPT in the current paper, to facilitate citation screening. Indeed, the recent successes in text ranking [20, 21] suggest LLMs potentially could be used as an alternative AI-based reviewer due to their strong zero-shot capabilities [22]. This could either save at least one human reviewer's time or, less radically, suggest a good initial training set for human verification.

In addition, we witness a severe lack of study about using selection criteria in automated citation screening (except [23]). Indeed, it is the selection criteria that set up the grounds for human reviewers' decision-making. Unfortunately, only a few studies initiated similar attempts [23, 24, 25], and neither the effectiveness of their methods nor the comprehensiveness of their experiments could provide convincing conclusions about the feasibility of LLMs in this task. The current paper presents a pioneering LLM-based framework for facilitating automated citation screening to fill this gap.

Our contributions can be summarised in three folds. (1) We proposed the first comprehensive LLM-assisted question-answering framework for automated citation screening in a zero-shot setting. (2) We developed the first generalisable approach to utilising selection criteria to enhance citation screening performance. (3) We performed the first comprehensive empirical study on well-known benchmark datasets and demonstrated the high potential of the proposed approach for citation screening.

Background Study

Automating in Citation Screening

Efforts to automate systematic reviews using machine learning have surged recently. Kitchenham and Charters' presented a good survey of such attempts in software engineering [26]. In evidence-based medicine, Cohen et al. was the seminal work of citation screening using machine learning [27], while Marshall and Wallace advocated active learning techniques for citation screening [28]. Examples like RobotReviewer [29, 30] and TrialStreamer [31] showcased the power of integrating AI into the review process, with RobotReviewer claiming to reach accuracy comparable to human reviewers. Despite the progress, challenges persist, including labour-intensive labelling and the risk of overlooking relevant studies [32]. Acknowledging the limitation of full automation, tools like Rayyan and Abstracker leverage natural language processing (NLP) algorithms to partially automate article screening [33].

Machine Learning for Citation Screening

The biggest challenge is handling large document volumes, particularly in non-randomized controlled trials lacking database filters [34]. For instance, EPPI-Centre reviews often screen over 20,000 documents, necessitating more efficient approaches [35]. Efforts include refining search queries, balancing precision and recall, and leveraging resource-efficient recall-maximizing models with NLP [36].

The initial approach involves training a classifier to make explicit include/exclude decisions [27, 36, 37, 38, 39, 40, 41]. Many classifiers using this approach inherently generate a confidence score indicating the likelihood of inclusion or exclusion (similar to the ranking in the second approach). Generally, this approach requires a labelled dataset for training, hindering the assessment of work reduction until after manual screening. Research within this paradigm primarily focuses on enhancing feature extraction methods [27, 39] and refining classifiers [40]. Van Dinter et al. [8] analyzed 41 studies in medicine and software engineering, revealing Support Vector Machines and Bayesian Networks as standard models and Bag of Words and TF-IDF as prevalent natural language processing techniques. Despite advancements, a dearth of deep neural network models explicitly designed for the systematic review screening phase is noted. The most prominent challenges include handling extreme data imbalance favouring (at least close to) total recall of relevant studies.

Ranking Approaches to Citation Screening

The second approach entails utilizing a ranking or prioritization system [9, 10, 11, 12, 13, 14, 15, 16, 35, 42]. This approach might necessitate manual screening by a reviewer until a specified criterion is met. This approach can also reduce the number of items needed to be screened when a cut-off criterion is properly established [35, 42, 43]. In addition to reducing the number needed to screen, other benefits of this approach include enhancing reviewers' understanding of inclusion criteria early in the process, starting full-text retrieval sooner, and potentially speeding up the screening process as confidence in relevance grows [7]. This prioritization approach also aids review updates, enabling quicker assimilation of current developments. Various studies reported benefits from prioritization for workflow improvement, emphasizing efficiency beyond reducing title and abstract screening workload [44, 45].

Active learning in Citation Screening

It's crucial to note that the last approach, active learning, aligns with both strategies above [36, 35, 46]. This involves an iterative process to enhance machine predictions by interacting with reviewers. The machine learns from an initial set of include/exclude decisions human reviewers provide. Reviewers then judge on a few new samples, and the machine adapts its decision rule based on this feedback. This iterative process continues until meeting a specified stopping criterion. While the classifier makes final decisions for unscreened items, human screeners retain control over the training process and the point at which manual screening concludes. Wallace et al. implement

Confidential: For Review Only

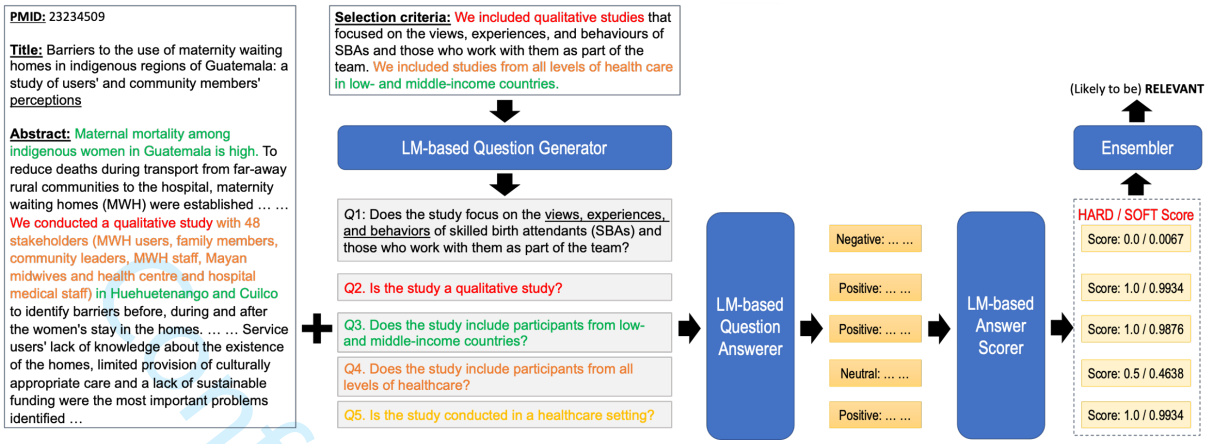


Fig. 1: Illustration of LLM-assisted automated citation screening.

active learning-based article screening using Support Vector Machines [36]. Notable tools include Abstrackr [38] and ASReview [47]. Various active learning strategies existed [7]. For instance, Marshall and Wallace [28] proposed a variant based on certainty, continuously training the classifier on manually screened articles and reordering unseen articles based on predicted relevance.

Large Language Models for Citation Screening

Recent advancements in LLMs, notably demonstrated by ChatGPT, have brought about a revolutionary paradigm shift across disciplines [48, 49]. LLMs have shown impressive generalisability across diverse domains and strong zero-/few-shot reasoning capabilities [48, 50]. Leveraging LLMs, like ChatGPT, holds promise for SRs, which however remains underexplored [7, 8]. This gap underscores the need for a comprehensive investigation into LLMs' potential in automating SRs, e.g., citation screening in the current paper.

There are some initial attempts to evaluate ChatGPT in automated SR, such as automating search queries [51]. Alshami et al. [52] utilized ChatGPT for automating the SR pipeline; however, their approach did not follow the norm of citation screening, making it incomparable to existing methods. Notably, the effectiveness of ChatGPT in automating citation screening has received limited attention, with only two studies [53, 54], which, unfortunately, lack consideration for achieving high recall, making them less suitable for real-world scenarios.

Materials and Methods

Overview

Our framework utilizes ChatGPT's zero-shot learning to assess if a candidate study's abstract aligns with the SR protocol's selection criteria. These criteria outline aspects of the selected studies. The provided sentence explains that in Figure 1, the red-highlighted text, "We included qualitative studies," serves as an example illustrating an inclusion criterion. This criterion specifies that only studies with a qualitative nature will be selected. Theoretically, all inclusion criteria should be met for the study to be included in the SR.

Our novel method frames automated citation screening as a question-answering (QA) task. Each selection criterion is treated as a question to be addressed using LLMs like ChatGPT. These models have showcased impressive question-answering abilities across diverse domains and tasks, including encoding clinical knowledge and achieving success in medical licensing exams [55, 56, 57, 58].

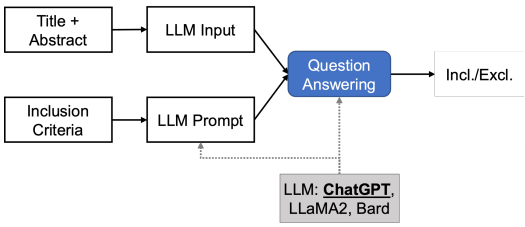
An initial experiment using the whole selection criteria as one comprehensive question (Figure 2a) proved ineffective. LLMs excel at answering focused and clearly described questions. Hence, our improved approach involves breaking down the selection criteria into several K questions (the LM-based Query Generator component in Figure 1), prompting LLMs to answer each question (the LM-based Question Answerer in Figure 1), and combining the answers for each question (the Ensembler in Figure 1).

Figure 2b details our proposed QA framework for citation screening. We begin with a Question Generator to convert the selection criteria into a set of questions. Optionally, a Question Analyser may be employed to correctly combine the question answers, considering, for instance, that answers A_1 and A_2 for questions Q_1 and Q_2 represent an inclusion and exclusion criterion, respectively, and the correct combination is $A_1 \cap \neg A_2$. Subsequently, each question is addressed by a trained Question Answerer to determine if the corresponding selection criterion is met, with each answer converted into a numeric score. Optionally, an Answer Re-ranking component can either be pre-trained on a large corpus of SRs or task-specifically trained with human reviewers' feedback. Finally, the Ensemble component combines the answers to all questions, and a final decision is made using predefined rules. Each component will be detailed in subsequent sections.

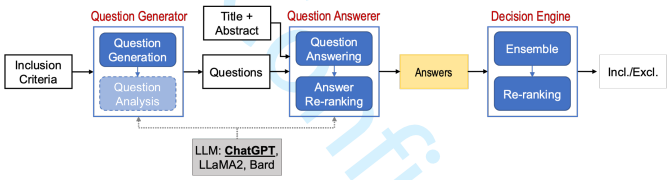
Question Generation

A substantial body of research exists on automated question generation from natural language text [59]. These methods often rely on manually crafted rules or a trained model, typically a fine-tuned pre-trained language model. While these question generation models have demonstrated utility in domain-specific tasks, such as generating questions about product descriptions for matching purchase inquiries [60] or creating questions about academic materials to assess learning outcomes [61], generalizing them to the vast diversity of SR topics presents challenges. Therefore, we entrust the question generation task to ChatGPT.

Confidential: For Review Only



(a) Naive approach for LLM-based automated citation screening using selection criteria



(b) An aspect-based approach for LLM-assisted automated citation screening using selection criteria

Fig. 2: Methodological framework for LLM-assisted automation screening.

A naive approach to question generation involves prompting ChatGPT to generate questions from the given paragraph about the selection criteria of a systematic review. However, this uncontrolled method often generates numerous questions, many subsumed by others or deemed too trivial to be meaningful.

To enhance the quality of generated questions, we constrained ChatGPT to produce no more than K questions, aiming to minimize redundancy. Based on an analysis of the lengths of selection criteria in our dataset's SRs, $K = 5$ proved sufficient for most SRs. Each sentence in the selection criteria often aligns well with a distinct criterion. In rare cases with more than 5 sentences, ChatGPT intelligently combined two sentences into one question. Figure 3a depicts the utilized prompt, and an example is shown in Figure 1.

Role: You are a researcher screening titles and abstracts of scientific papers.

Task: Using the inclusion criteria in the brackets generate 5 unique yes or no questions which encompasses the entire inclusion criteria without any duplicate or unneccessary questions to ascertain if papers meets the inclusion criteria!

Criteria: `{criteria}`

(a) Prompt for question generation

Role: You are a researcher screening titles and abstracts of scientific papers for the systematic review '{review_title}'

Task: Analyse the abstract below within the brackets and answer the question below. Taking a step-by-step approach towards reasoning and answering the question. The answer should be in either a positive, neutral, or negative sentiment format.

Abstract: `{abstract}`

Question: `{question}`

(b) Prompt for question answering

Fig. 3: Prompt design for LLM-assisted automated citation screening

Question Answering

The Question Answerer evaluates the relevance of each abstract to every selection criterion, formulated as Yes/No questions. We prompt ChatGPT to return three types of responses Figure 3b:

- **Positive:** The abstract explicitly addresses the question, offering information that aligns with the criteria posed by the question.
- **Neutral:** The information in the abstract is inqdequate or too ambiguous for ChatGPT to derive a confirmative answer.
- **Negative:** A clear NO answer to the question, indicating irrelavance to the specified criteria.

Answer Representation Approaches

Two distinct techniques represent answers, namely the Hard Answer and Soft Answer methods. These methods conceptualize answer representation as a generative sentiment classification problem, leveraging the capabilities of the BART model inspired by its recent successes in various sentiment classification tasks [62].

- **Hard Answer:** This method involves BART determining the sentiment category of each answer (Positive, Neutral, Negative). Traditionally, rejecting an abstract occurred if one question had a Negative answer, leading to low recall due to small errors in ChatGPT responses. Instead, we convert the discrete sentiment categories to semantic scores (e.g., 1, 0.5, and 0), enabling the Ensemble to combine answers into a final decision.
- **Soft Answer:** ChatGPT often justifies its answer, contributing to quantifying its confidence level in the provided answer. In the Soft Answer method, the sentiment score for each answer is the probability of BART classifying the ChatGPT-generated answer sentence as positive.

Decision Engine

Ensemble

The answer scores for each selection criterion are "averaged." This mean score provides a quantitative representation of the relevance of the abstract. Candidate studies are then ranked in descending order based on these mean scores. To enhance screening further, a significant contribution involves re-ranking candidate studies based on how well abstracts are semantically aligned with the selection criteria.

Re-ranking

Several methods are available for embedding selection criteria and abstracts. Given the emphasis on the capabilities of LLMs in this paper, GPT Embeddings [63] are chosen. Two approaches to re-ranking are defined: Abstract-level re-ranking and answer-level re-ranking.

- **Abstract-Level Re-Ranking:** This method first aggregates the mean answer score across all questions and then averages the mean score with the similarity score.
- **Answer-Level Re-Ranking:** In this more advanced method, the cosine similarity evaluates how well an abstract aligns with each generated question, enhancing the answers to each selection criterion. Each Yes/No question is matched against the abstract, and the cosine similarity is averaged with the corresponding answer score. This results in K re-ranked answer scores. Then the K scores are averaged, and the mean score is

Confidential: For Review Only

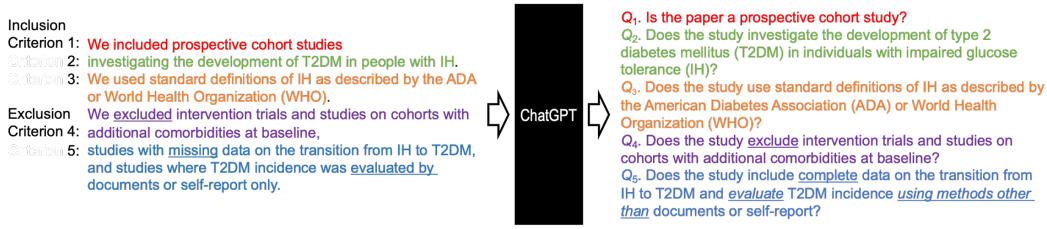


Fig. 4: Example of handing exclusion criteria.

further averaged with the overall alignment between selection criteria and abstract. This hierarchical ensemble effectively enhances the overall precision of document re-ranking by considering confidence in answering each question, aligning with each selection criterion, and adhering holistically to selection criteria.

Experimental Setup

Dataset and Evaluation

This study utilized datasets of CLEF eHealth 2019 Task 2: Technology Assisted Reviews in Empirical Medicine (TAR2019). This dataset provides valuable insights into the prevailing scientific consensus on various topics, making it a suitable resource for evaluating reranking methodologies in systematic reviews [64].

We employed the TAR2019 test set comprising 31 SRs categorized into Intervention (20), DTA (8), Qualitative (2), and Prognosis (1). We refrained from using the training set, aiming to highlight the effectiveness of our zero-shot methodology that eliminates the need for prior training [65, 25].

We used the review titles from the TAR2019 datasets to identify selection criteria. Seven evaluation metrics were employed, including the rank position of the last relevant document (L_Rel), Mean Average Precision (AP), Recall at $k\%$ ($k = 5$, 10, 20, 30), and Work Saved Over Sampling (WSS) at $k\%$ ($k = 95\%$, 100%). Notably, $WSS@k$ measures the screening workload saved by halting the examination process once $k\%$ of relevant documents are identified, compared to screening the entire document set [27].

Baseline Model

The baseline models, serving as a comparative benchmark, were based on submissions to the TAR2019 workshop [66], which encompass UvA [67], UNIPD [68], and Sheffield [17]. Additionally, we considered the nine models evaluated by Wang et al [25]. Unlike our fully automated model, many workshop submissions employ an iterative ranking system, making them semi-automated. To comprehensively assess performance, we implemented two IR baselines of our own. One is cosine similarity between selection criteria and abstract based on GPT embeddings [63], named GPT_Cosine_Similarity. The other is BM25 [69], using selection criteria as a query. The variants of our own approach are summarised below:

- **GPT_QA_Soft/Hard**: Soft/Hard answer representation, without re-ranking.
- **GPT_QA_Soft/Hard_Abstract_Level**: Soft/Hard answer representation, with abstract-level re-ranking.
- **GPT_QA_Soft/Hard_Answer_Level**: Soft/Hard answer representation, with answer-level re-ranking.

am

Results

Prognosis

Our proposed methods demonstrated promising results on the Prognosis dataset (Table 1). Notably, in terms of L_{Rel}, for which a lower value signifies superior performance, answer-level re-ranking methods showcased the most impressive results among our proposed methods: 2333 for GPT_QA_Soft_Answer_Level and 2373 for GPT_QA_Hard_Answer_Level. Our methods also achieved MAP scores from 0.350 to 0.430, underscoring the models' efficiency in prioritizing candidate studies. This outshined numerous IR methods (UNIPD and Sheffield variants).

The proposed methods sustained their superiority in $R@k\%$. When $k \in \{5, 10\}$, our re-ranking methods (the last four rows in Table 1) consistently outperformed UNIPD and Sheffield submissions. A promising finding was that our best re-ranking method successfully suggested 65% of total positive samples (included studies) for classifier training when only 10% of total samples needed to be verified by human reviewers. This is a testament to the capacity of selecting positive samples from highly imbalanced data. The best $WSS@95$ of our zero-shot approaches reached 55.5% on Prognosis, and notably $WSS@100$ was significantly better than most baselines except 2018_stem_original_p50_t1500 which used relevance feedback.

From the holistic view of the evaluation metrics, our answer-level re-ranked models stood out as cutting-edge solutions, achieving either competitive or new state-of-the-art results.

Qualitative

The results are presented in Table 2. Similarly, the L_{Rel} metric highlighted that both our abstract-level and answer-level re-ranking methods (the last four rows in Table 2) are particularly effective. The MAP results of our proposed models consistently outperformed the IR baselines of UNIPD and Sheffield. UvA showed the best performance. Note that the UvA approaches used relevance feedback to improve the ranking performance, so it was not purely a zero-shot problem in our sense. Further discussions can be found in the Discussions section.

Regarding recall, our models showed promising results when $k = 5$, meaning that our methods identified more than half of the positive samples in the top 5% ranked list. This allows us to significantly reduce the effort in annotating the initial dataset for training a citation screener. Although some IR methods showed higher recall when $k \geq 10$, our methods outperformed all baselines in $R@30\%$ and showed significant performance gains in terms of both $WSS@95$ and $WSS@100$ over all baselines,

Confidential: For Review Only

Table 1. Results obtained on the Prognosis data. The zero-shot ranking models are emboldened.

Paper	Models	L_Rel	MAP	R@5%	R@10%	R@20%	R@30%	WSS@95	WSS@100
UvA	abs-hh-ratio-ilps	2885	0.673	0.562	0.714	0.875	0.911	0.591	0.143
	abs-th-ratio-ilps	2537	0.628	0.521	0.682	0.818	0.927	0.566	0.247
UNIPD	2018_stem_original_p10_t400	2967	0.235	0.214	0.484	0.812	0.901	0.567	0.119
	distributed_effort_p10_t1500	2594	0.235	0.214	0.484	0.812	0.896	0.554	0.230
	2018_stem_original_p10_t1000	2644	0.235	0.214	0.484	0.812	0.896	0.554	0.215
	2018_stem_original_p10_t200	2911	0.242	0.214	0.536	0.812	0.901	0.530	0.135
	2018_stem_original_p10_t500	2920	0.235	0.214	0.484	0.812	0.891	0.560	0.133
	2018_stem_original_p10_t300	2955	0.239	0.214	0.547	0.818	0.891	0.556	0.122
	2018_stem_original_p10_t1500	2578	0.235	0.214	0.484	0.812	0.896	0.554	0.234
	distributed_effort_p10_t1000	2563	0.235	0.214	0.484	0.812	0.896	0.554	0.239
	2018_stem_original_p10_t100	2802	0.259	0.286	0.562	0.797	0.891	0.600	0.168
	baseline_bm25_t500	3343	0.071	0.057	0.130	0.281	0.422	0.084	0.007
	distributed_effort_p10_t300	2964	0.235	0.214	0.484	0.812	0.906	0.567	0.120
	2018_stem_original_p50_t1000	2556	0.221	0.214	0.484	0.740	0.870	0.571	0.241
	distributed_effort_p10_t100	2789	0.252	0.250	0.568	0.786	0.875	0.594	0.172
	2018_stem_original_p50_t200	2911	0.242	0.214	0.536	0.812	0.901	0.530	0.135
	baseline_bm25_t1000	3346	0.070	0.057	0.130	0.276	0.396	0.057	0.006
	distributed_effort_p10_t500	2708	0.235	0.214	0.484	0.812	0.891	0.566	0.196
	baseline_bm25_t300	3350	0.071	0.057	0.135	0.276	0.385	0.104	0.005
	baseline_bm25_t100	3350	0.066	0.047	0.130	0.255	0.365	0.059	0.005
	2018_stem_original_p50_t400	2955	0.231	0.214	0.484	0.807	0.896	0.556	0.122
	2018_stem_original_p50_t300	2955	0.239	0.214	0.547	0.818	0.891	0.556	0.122
	2018_stem_original_p50_t100	2802	0.259	0.286	0.562	0.797	0.891	0.600	0.168
	distributed_effort_p10_t200	2968	0.240	0.214	0.542	0.807	0.906	0.548	0.119
	baseline_bm25_t400	3347	0.071	0.057	0.130	0.281	0.417	0.109	0.006
	2018_stem_original_p50_t1500	1975	0.219	0.214	0.484	0.740	0.828	0.500	0.413
	2018_stem_original_p50_t500	2660	0.228	0.214	0.484	0.807	0.891	0.576	0.210
	baseline_bm25_t1500	3346	0.070	0.057	0.130	0.276	0.396	0.050	0.006
	baseline_bm25_t200	3350	0.069	0.057	0.125	0.266	0.385	0.111	0.005
	distributed_effort_p10_t400	2920	0.235	0.214	0.484	0.812	0.891	0.560	0.133
Sheffield	sheffield-baseline	2990	0.126	0.146	0.255	0.448	0.594	0.247	0.112
	sheffield-relevence_feedback	2775	0.141	0.151	0.307	0.484	0.646	0.305	0.176
Proposed Method	**GPT_Cosine_Similarity**	3160	0.178	0.200	0.305	0.495	0.647	0.239	0.053
	BM25	3337	0.074	0.089	0.132	0.279	0.416	0.020	0.000
	GPT_QA_Soft	3211	0.350	0.395	0.563	0.784	0.832	0.434	0.037
	GPT_QA_Hard	3338	0.315	0.395	0.395	0.753	0.753	0.417	0.060
	GPT_QA_Soft_Abstract_Level	2467	0.417	0.395	0.647	0.795	0.879	0.523	0.261
	GPT_QA_Hard_Abstract_Level	2398	0.417	0.395	0.637	0.789	0.884	0.543	0.282
	GPT_QA_Soft_Answer_Level	2373	0.430	0.400	0.653	0.800	0.884	0.543	0.289
	GPT_QA_Hard_Answer_Level	2333	0.429	0.400	0.642	0.789	0.884	0.555	0.301

including the relevance feedback approaches by UvA. The results are overall encouraging, showing that the proposed QA-based prioritization framework potentially applies well to qualitative SRs, too. However, a conclusive statement requires more empirical studies, which will be one direction of our future work.

Diagnostic Test Accuracy (DTA)

Table 3 shows satisfactory results on DTA. The top-5% ranked list of our best models covered as many as 45% positive samples. They outperformed all IR methods except abs-hh-ratio-ilps by UvA, leading to better MAP over the latter, and approached the fine-tuned models in $R@5\%$. This implies the feasibility of our approaches for reducing the human effort in annotating an initial training set with a reasonable number of included studies, compared to random sampling, which requires annotating 45% of total samples to reach the same size of included studies. Although our models underperformed the best UvA variant in $R@5\%$, they started to excel the latter when $k \geq 20$, resulting in a better MAP and significantly higher WSS.

On the other hand, we notice that although some UNIPD and Sheffield submissions performed better than our best models in recall (when $k > 10$) and WSS (when $R = 95$ or 100), our methods were consistently stronger than the baselines without relevance feedback (the rows in bold) by large margins. This justifies the superiority of LLMs as a zero-shot citation screening method. We anticipate that the screening performance can be further improved through an iterative question-answering conversion by providing proper feedback to LLMs. Similar ideas have been proven effective

For Peer Review Only

Table 2. Results obtained on the Qualitative data. The zero-shot ranking models are emboldened.

Paper	Models	L_Rel	MAP	R@5%	R@10%	R@20%	R@30%	WSS@95	WSS@100
UvA	abs-hh-ratio-ilps	1796	0.204	0.478	0.655	0.876	0.929	0.417	0.397
	abs-th-ratio-ilps	2564	0.187	0.487	0.628	0.805	0.920	0.398	0.215
UNIPD	2018_stem_original_p10_t400.out	2547	0.109	0.496	0.717	0.779	0.894	0.302	0.183
	distributed_effort_p10_t1500.out	2544	0.109	0.496	0.743	0.770	0.885	0.268	0.168
	2018_stem_original_p10_t1000.out	2662	0.109	0.496	0.743	0.770	0.885	0.273	0.141
	2018_stem_original_p10_t200.out	2934	0.089	0.478	0.522	0.699	0.805	0.216	0.101
	2018_stem_original_p10_t500.out	2535	0.109	0.496	0.743	0.770	0.894	0.301	0.185
	2018_stem_original_p10_t300.out	2660	0.103	0.496	0.655	0.752	0.858	0.303	0.159
	2018_stem_original_p10_t1500.out	2534	0.109	0.496	0.743	0.770	0.885	0.268	0.170
	distributed_effort_p10_t1000.out	2469	0.109	0.496	0.743	0.770	0.885	0.295	0.199
	2018_stem_original_p10_t100.out	2996	0.071	0.327	0.416	0.637	0.796	0.186	0.090
	baseline_bm25_t500.out	2700	0.051	0.274	0.425	0.469	0.611	0.412	0.256
	distributed_effort_p10_t300.out	2518	0.109	0.496	0.743	0.770	0.894	0.309	0.193
	2018_stem_original_p50_t1000.out	2438	0.116	0.496	0.743	0.920	0.947	0.357	0.194
	distributed_effort_p10_t100.out	2920	0.083	0.416	0.469	0.681	0.814	0.258	0.106
	2018_stem_original_p50_t200.out	2934	0.089	0.478	0.522	0.699	0.805	0.216	0.101
	baseline_bm25_t1000.out	3040	0.055	0.274	0.425	0.496	0.788	0.239	0.101
	distributed_effort_p10_t500.out	2641	0.109	0.496	0.743	0.770	0.894	0.295	0.162
	baseline_bm25_t300.out	2697	0.049	0.274	0.372	0.451	0.628	0.294	0.257
	baseline_bm25_t100.out	2700	0.056	0.301	0.389	0.637	0.743	0.399	0.256
	2018_stem_original_p50_t400.out	2566	0.109	0.496	0.717	0.779	0.894	0.293	0.174
	2018_stem_original_p50_t300.out	2687	0.103	0.496	0.655	0.752	0.858	0.290	0.147
	2018_stem_original_p50_t100.out	2996	0.071	0.327	0.416	0.637	0.796	0.186	0.090
	distributed_effort_p10_t200.out	2762	0.104	0.496	0.673	0.761	0.867	0.303	0.135
	baseline_bm25_t400.out	2700	0.052	0.274	0.434	0.469	0.619	0.417	0.256
	2018_stem_original_p50_t1500.out	1970	0.116	0.496	0.743	0.920	0.965	0.356	0.301
	2018_stem_original_p50_t500.out	2576	0.110	0.496	0.743	0.788	0.894	0.283	0.168
	baseline_bm25_t1500.out	3039	0.055	0.274	0.425	0.496	0.779	0.240	0.101
	baseline_bm25_t200.out	2698	0.053	0.274	0.381	0.619	0.726	0.395	0.256
	distributed_effort_p10_t400.out	2636	0.109	0.496	0.743	0.770	0.894	0.301	0.165
Sheffield	sheffield-relevance_feedback.out	2940	0.060	0.274	0.549	0.717	0.832	0.185	0.103
	sheffield-baseline	3031	0.051	0.265	0.451	0.619	0.743	0.135	0.082
Proposed Method	**GPT_Cosine_Similarity**	2256	0.082	0.173	0.478	0.559	0.618	0.303	0.289
	BM25	2704	0.037	0.078	0.146	0.191	0.259	0.135	0.135
	GPT_QA_Soft	1786	0.157	0.537	0.614	0.673	0.959	0.599	0.484
	GPT_QA_Hard	1784	0.110	0.478	0.582	0.900	0.959	0.650	0.485
	GPT_QA_Soft_Abstract_Level	1683	0.159	0.509	0.605	0.673	0.959	0.595	0.511
	GPT_QA_Hard_Abstract_Level	1675	0.200	0.509	0.609	0.678	0.959	0.608	0.514
	GPT_QA_Soft_Answer_Level	1682	0.159	0.505	0.600	0.673	0.959	0.576	0.507
	GPT_QA_Hard_Answer_Level	1684	0.157	0.514	0.600	0.678	0.959	0.601	0.507

on different NLP tasks [70, 71, 72]. Meanwhile, it is worth noting the DTA dataset has been generally considered a very difficult dataset [73].

Intervention

Table 4 shows the results on Intervention. Our methods exhibited exceptional performance across all metrics. The high recall values at different thresholds underscored the effectiveness of our proposed model, consistently outperforming all models except BioBERT-Tune. Our best models, namely the answer-level re-ranking methods GPT_QA_HARD_Answer_Level and GPT_QA_Soft_Answer_Level, also achieved better MAP results than most baselines, and the abstract-level re-ranking method GPT_QA_HARD_Answer_Level rivalled the robust UvA systems. Notably, our best models excelled in L_{Rel}, and the results of $WSS@95$ and $WSS@100$ outperformed most baseline models. In summary, the comprehensive assessment across diverse metrics and datasets reinforced the standing of our proposed methods as state-of-the-art solutions.

Confidential: For Review Only

Table 3. Results obtained on the DTA data. The zero-shot ranking models are emboldened.

Paper	Models	L_Rel	MAP	R@5%	R@10%	R@20%	R@30%	WSS@95	WSS@100
UvA	abs-hh-ratio-ilps	2420	0.493	0.589	0.682	0.789	0.834	0.406	0.304
	abs-th-ratio-ilps	2676	0.399	0.418	0.536	0.661	0.734	0.312	0.253
UNIPD	2018_stem_original_p10_t400.out	1190	0.229	0.448	0.634	0.818	0.895	0.662	0.512
	distributed_effort_p10_t1500.out	1111	0.229	0.445	0.630	0.814	0.895	0.652	0.513
	2018_stem_original_p10_t1000.out	1141	0.229	0.445	0.630	0.814	0.893	0.658	0.509
	2018_stem_original_p10_t200.out	1282	0.229	0.445	0.634	0.823	0.891	0.660	0.507
	2018_stem_original_p10_t500.out	1200	0.229	0.445	0.634	0.818	0.893	0.662	0.509
	2018_stem_original_p10_t300.out	1280	0.229	0.452	0.627	0.816	0.893	0.660	0.500
	2018_stem_original_p10_t1500.out	1126	0.229	0.445	0.630	0.814	0.895	0.657	0.514
	distributed_effort_p10_t1000.out	1109	0.229	0.445	0.630	0.814	0.895	0.649	0.514
	2018_stem_original_p10_t100.out	2024	0.221	0.418	0.609	0.791	0.868	0.525	0.399
	baseline_bm25_t500.out	2470	0.119	0.236	0.402	0.548	0.650	0.390	0.252
	distributed_effort_p10_t300.out	1111	0.232	0.445	0.630	0.814	0.886	0.649	0.528
	2018_stem_original_p50_t1000.out	1127	0.229	0.445	0.630	0.811	0.893	0.652	0.528
	distributed_effort_p10_t100.out	1271	0.204	0.439	0.614	0.770	0.839	0.610	0.468
	2018_stem_original_p50_t200.out	1291	0.229	0.445	0.634	0.820	0.898	0.660	0.499
	baseline_bm25_t1000.out	2395	0.119	0.236	0.389	0.543	0.659	0.396	0.260
	distributed_effort_p10_t500.out	1116	0.229	0.445	0.630	0.814	0.891	0.634	0.521
	baseline_bm25_t300.out	2493	0.119	0.239	0.405	0.541	0.652	0.391	0.244
	baseline_bm25_t100.out	2130	0.120	0.239	0.414	0.564	0.659	0.394	0.295
	2018_stem_original_p50_t400.out	1189	0.229	0.448	0.634	0.816	0.891	0.654	0.527
	2018_stem_original_p50_t300.out	1272	0.229	0.452	0.627	0.814	0.893	0.656	0.518
	2018_stem_original_p50_t100.out	2027	0.222	0.418	0.609	0.786	0.868	0.549	0.394
	distributed_effort_p10_t200.out	1194	0.225	0.445	0.632	0.811	0.877	0.663	0.509
	baseline_bm25_t400.out	2492	0.119	0.239	0.405	0.539	0.650	0.386	0.246
	2018_stem_original_p50_t1500.out	1056	0.229	0.445	0.630	0.814	0.898	0.651	0.537
	2018_stem_original_p50_t500.out	1200	0.229	0.445	0.634	0.809	0.889	0.649	0.524
	baseline_bm25_t1500.out	2476	0.119	0.236	0.389	0.541	0.652	0.364	0.254
	baseline_bm25_t200.out	2253	0.120	0.234	0.405	0.550	0.652	0.409	0.278
	distributed_effort_p10_t400.out	1116	0.231	0.445	0.630	0.814	0.886	0.634	0.528
Sheffield	sheffield-Log_likelihood.out	1964	0.222	0.305	0.450	0.641	0.730	0.475	0.375
	sheffield-Odds_Ratio.out	2250	0.175	0.220	0.336	0.525	0.675	0.451	0.338
	sheffield-baseline.out	2184	0.248	0.382	0.561	0.707	0.805	0.490	0.347
	sheffield-Chi_Squared.out	1972	0.234	0.350	0.527	0.668	0.759	0.487	0.381
Wang et al.	**BM25**	2723	0.119	0.213	0.329	0.528		0.314	0.208
	BERT	2514	0.092	0.132	0.238	0.391		0.258	0.210
	BERT-M	3234	0.096	0.079	0.198	0.379		0.263	0.123
	BioBERT	3264	0.081	0.129	0.229	0.337		0.137	0.095
	BlueBERT	3771	0.069	0.026	0.053	0.105		0.023	0.016
	PubMedBERT	3331	0.104	0.123	0.214	0.312		0.202	0.098
	BERT-Tuned	1400	0.223	0.439	0.601	0.762		0.587	0.460
	BERT-M-Tuned	1178	0.254	0.447	0.590	0.754		0.615	0.500
	BioBERT-Tuned	853	0.318	0.500	0.671	0.817		0.686	0.585
Proposed Method	**GPT_Cosine_Similarity**	1154	0.271	0.477	0.628	0.782	0.851	0.600	0.513
	BM25	2461	0.164	0.334	0.472	0.654	0.723	0.351	0.233
	GPT_QA_Soft	1979	0.255	0.319	0.495	0.674	0.765	0.408	0.334
	GPT_QA_Hard	1983	0.228	0.367	0.468	0.673	0.776	0.364	0.303
	GPT_QA_Soft_Abstract_Level	1491	0.301	0.384	0.574	0.705	0.810	0.473	0.422
	GPT_QA_Hard_Abstract_Level	1583	0.310	0.387	0.573	0.727	0.820	0.454	0.396
	GPT_QA_Soft_Answer_Level	1136	0.315	0.438	0.593	0.766	0.858	0.556	0.506
	GPT_QA_Hard_Answer_Level	1176	0.322	0.450	0.595	0.791	0.873	0.536	0.491

For Review Only

Table 4. Results obtained on the Intervention data. The zero-shot ranking models are emboldened.

Paper	Models	L_Rel	MAP	R@5%	R@10%	R@20%	R@30%	WSS@95	WSS@100
UvA	abs-hh-ratio-ilps	958	0.567	0.518	0.628	0.736	0.813	0.526	0.480
	abs-th-ratio-ilps	986	0.556	0.478	0.576	0.692	0.774	0.535	0.450
UNIPD	2018_stem_original_p10_t400.out	985	0.280	0.307	0.502	0.663	0.744	0.632	0.511
	distributed_effort_p10_t1500.out	981	0.280	0.306	0.499	0.664	0.745	0.633	0.517
	2018_stem_original_p10_t1000.out	977	0.280	0.306	0.499	0.664	0.745	0.630	0.510
	2018_stem_original_p10_t200.out	1180	0.280	0.312	0.501	0.671	0.775	0.617	0.488
	2018_stem_original_p10_t500.out	975	0.280	0.306	0.502	0.662	0.742	0.630	0.514
	2018_stem_original_p10_t300.out	1141	0.280	0.313	0.496	0.665	0.771	0.617	0.494
	2018_stem_original_p10_t1500.out	952	0.280	0.306	0.499	0.664	0.745	0.630	0.522
	distributed_effort_p10_t1000.out	992	0.279	0.306	0.499	0.664	0.745	0.620	0.492
	2018_stem_original_p10_t100.out	1153	0.274	0.306	0.483	0.639	0.737	0.540	0.474
	baseline_bm25_t500.out	1233	0.222	0.191	0.282	0.410	0.515	0.435	0.394
	distributed_effort_p10_t300.out	974	0.276	0.306	0.499	0.664	0.733	0.592	0.481
	2018_stem_original_p50_t1000.out	836	0.290	0.306	0.498	0.688	0.795	0.643	0.542
	distributed_effort_p10_t100.out	1114	0.248	0.315	0.444	0.604	0.704	0.458	0.372
	2018_stem_original_p50_t200.out	1185	0.290	0.312	0.499	0.693	0.792	0.630	0.481
	baseline_bm25_t1000.out	1241	0.222	0.191	0.282	0.408	0.524	0.446	0.392
	distributed_effort_p10_t500.out	991	0.278	0.306	0.499	0.664	0.743	0.606	0.483
	baseline_bm25_t300.out	1262	0.222	0.187	0.286	0.410	0.523	0.440	0.398
	baseline_bm25_t100.out	1397	0.223	0.186	0.291	0.429	0.557	0.414	0.368
	2018_stem_original_p50_t400.out	985	0.290	0.307	0.501	0.685	0.767	0.646	0.514
	2018_stem_original_p50_t300.out	1144	0.290	0.313	0.495	0.682	0.788	0.639	0.497
	2018_stem_original_p50_t100.out	1150	0.284	0.306	0.483	0.653	0.752	0.556	0.481
	distributed_effort_p10_t200.out	965	0.271	0.306	0.482	0.651	0.752	0.560	0.445
	baseline_bm25_t400.out	1242	0.222	0.191	0.286	0.412	0.523	0.434	0.393
	2018_stem_original_p50_t1500.out	796	0.290	0.306	0.498	0.688	0.785	0.642	0.553
	2018_stem_original_p50_t500.out	1001	0.290	0.306	0.501	0.691	0.779	0.650	0.505
	baseline_bm25_t1500.out	1203	0.222	0.191	0.282	0.411	0.533	0.453	0.399
	baseline_bm25_t200.out	1263	0.222	0.189	0.284	0.417	0.535	0.438	0.396
	distributed_effort_p10_t400.out	981	0.277	0.306	0.499	0.663	0.734	0.595	0.483
Sheffield	sheffield-Log_likelihood.out	1132	0.293	0.258	0.378	0.583	0.695	0.458	0.381
	Sheffield-Odds_Ratio.out	1070	0.261	0.267	0.404	0.569	0.700	0.462	0.384
	Sheffield-baseline.out	1276	0.245	0.220	0.334	0.507	0.653	0.470	0.386
	sheffield-Chi_Squared.out	1149	0.262	0.238	0.360	0.537	0.687	0.469	0.415
Wang et al.	**BM25**	1716	0.211	0.305	0.399	0.554		0.351	0.296
	BERT	1399	0.160	0.211	0.328	0.504		0.362	0.333
	BERT-M	1837	0.177	0.195	0.355	0.527		0.323	0.266
	BioBERT	1833	0.146	0.135	0.198	0.307		0.159	0.163
	BlueBERT	2057	0.046	0.028	0.051	0.107		0.008	0.036
	PubMedBERT	1975	0.078	0.050	0.091	0.275		0.121	0.094
	BERT-Tuned	1375	0.281	0.374	0.527	0.659		0.363	0.301
	BERT-M-Tuned	1572	0.334	0.402	0.565	0.706		0.446	0.362
	BioBERT-Tuned	707	0.456	0.581	0.737	0.842		0.646	0.579
Proposed Method	**GPT_Cosine_Similarity**	920	0.315	0.401	0.544	0.722	0.797	0.552	0.499
	BM25	1545	0.146	0.191	0.300	0.497	0.667	0.270	0.238
	GPT_QA_Soft	1055	0.411	0.469	0.610	0.738	0.856	0.486	0.416
	GPT_QA_Hard	1159	0.356	0.444	0.578	0.759	0.847	0.466	0.397
	GPT_QA_Soft_Abstract_Level	934	0.440	0.494	0.663	0.800	0.873	0.534	0.459
	GPT_QA_Hard_Abstract_Level	976	0.443	0.505	0.687	0.777	0.856	0.532	0.458
	GPT_QA_Soft_Answer_Level	801	0.450	0.526	0.697	0.816	0.881	0.600	0.526
	GPT_QA_Hard_Answer_Level	806	0.447	0.527	0.697	0.808	0.869	0.592	0.519

Confidential: For Review Only

Discussion

Selection Criteria

To further evaluate the usefulness of selection criteria, we implemented our own BM25 baseline using selection criteria as a query, and we observed competitive performances. Particularly it significantly outperformed the BM25 baselines of UNIPD and Wang et al. on the DTA dataset. On Intervention, it was at least on par with or slightly better than most other BM25 baselines except Wang et al. The result underscored the validity of using selection criteria to guide citation screening.

Question Generation and Answering

We manually checked the question qualities and found notable strengths and occasional challenges of ChatGPT in question generation. Figure 4 illustrates how ChatGPT was smart enough to translate a lengthy exclusion criterion into two relevant questions, Q_4 and Q_5, demonstrating its nuanced understanding of the complex semantics of the sentence. The questions were effectively formatted so that a POSITIVE response consistently signifies compliance with a selection criterion, be it inclusion or exclusion.

Occasionally, ChatGPT failed to generate completely independent questions. This led to redundant or overlapped questions, introducing biases in combining answer scores. Occasionally, ChatGPT struggled to address the "OR" clause in a long selection criterion sentence. It was split into separate questions, which was problematic. In such cases, matching one question should give a POSITIVE score, but the NEGATIVE answers to other questions generated from the OR clause might underestimate the final score. These issues imply areas of improvement in ensuring robust question generation and precise answer interpretation for citation screening. We postulate that a viable solution is to train a good question generator and analyzer to tackle these issues. Alternatively, it is sensible for human reviewers to scrutinize and correct the generated questions before sending them to LLMs to answer.

While the current paper deliberately limited answers to a simple form, it is worthwhile to consider incorporating explanations of LLM-generated answers in future iterations. Providing insight into ChatGPT's reasoning process can enhance transparency and facilitate a better understanding of the model's decision-making which is essential for instilling user confidence in the outputs of automated citation screening to encourage technology acceptance [74, 75]. In addition, it contributes to refining model performance through iterative conversation with LLMs by giving user feedback on model answers and their explanations [70].

Answer Re-Ranking

Taking a holistic view, i.e., averaging model performances over all four categories of datasets, the answer-level re-ranking methods consistently outperformed our other models across all metrics. This superiority is attributed to the additional granularity gained by considering the alignment of each question with the abstracts of candidate studies. Compared with other zero-shot models, our methods achieved substantial improvements, showcasing both the effectiveness of the proposed questions-answering framework and the utility of ChatGPT as a zero-shot ranker for automated citation screening.

When pitted against trained models employing relevant feedback or fine-tuning, our methods still held a solid ground competitively. Our best models achieved very promising performances in L_{Rel}, $R@5\%$, and WSS. This is a useful merit as our zero-shot method fits well into the real-world citation screening task, starting with no annotation of included/excluded studies. Although the UvA variants and BioBERT-Tuned models often resulted in better performances in MAP and occasionally in recall at different thresholds, our models were still demonstrated to be competitive, highlighting their brilliance requiring no prior training. Therefore, our models are better generalizable to all SR categories, especially when lacking comprehensive datasets for relevance feedback or fine-tuning. Nevertheless, it is always valuable to fine-tune LLMs for each SR topic to benefit our proposed answer re-ranking methods.

Conclusion

This paper proposed an effective LLM-assisted question-answering framework to facilitate citation screening and advance automated systematic review. Extensive experiments emphasized the particular pertinence of selection criteria of included studies to automated citation screening and ChatGPT's proficiency in understanding and utilizing selection criteria to prioritize candidate studies. Specifically, ChatGPT was able to correctly capture and handle complex semantics like several juxtaposed criteria with a logical OR relationship, significantly outperforming other zero-shot baselines. The positive results of L_{Rel} (position of the last relevant study), $R@5\%$ (recall at top 5%), $R@10\%$, $WSS@95$ (Workload Saved over Sampling at 95% recall level), and $WSS@100$ not only showed the competency of the proposed framework as a zero-shot citation screening methodology but also indicated its potential use in reducing human effort in building a high-quality dataset for training a citation screener.

References

1. Guy Tsafnat, Paul Glasziou, Miew Keen Choong, Adam Dunn, Filippo Galgani, and Enrico Coiera. Systematic review automation technologies. *Systematic reviews*, 3:1–15, 2014.
2. S Gopalakrishnan and P Ganeshkumar. Systematic reviews and meta-analysis: understanding the best evidence in primary healthcare. *Journal of family medicine and primary care*, 2(1):9, 2013.
3. Hamideh Moosapour, Farzane Saeidifard, Maryam Aalaa, Akbar Soltani, and Bagher Larijani. The rationale behind systematic reviews in clinical medicine: a conceptual framework. *Journal of Diabetes & Metabolic Disorders*, 20:919–929, 2021.
4. Ian Shemilt, Nada Khan, Sophie Park, and James Thomas. Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews. *Systematic reviews*, 5:1–13, 2016.
5. Matthew Michelson and Katja Reuter. The significant cost of systematic reviews and meta-analyses: a call for greater involvement of machine learning to assess the promise of clinical trials. *Contemporary clinical trials communications*, 16:100443, 2019.
6. Julian PT Higgins, Sally Green, et al. Cochrane handbook for systematic reviews of interventions. 2008.

Confidential: For Review Only

7. Alison O'Mara-Eves, James Thomas, John McNaught, Makoto Miwa, and Sophia Ananiadou. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic reviews*, 4(1):1–22, 2015.
8. Raymon van Dinter, Bedir Tekinerdogan, and Cagatay Catal. Automation of systematic literature reviews: A systematic literature review. *Information and Software Technology*, 136:106589, 2021.
9. Amal Alharbi, William Briggs, and Mark Stevenson. Retrieving and ranking studies for systematic reviews: University of sheffield's approach to clef ehealth 2018 task 2. In *CEUR workshop proceedings*, volume 2125. CEUR Workshop Proceedings, 2018.
10. Amal Alharbi and Mark Stevenson. Ranking abstracts to identify relevant evidence for systematic reviews: The university of sheffield's approach to clef ehealth 2017 task 2. In *Clef (working notes)*, 2017.
11. Gordon V Cormack and Maura R Grossman. Technology-assisted review in empirical medicine: Waterloo participation in clef ehealth 2017. *CLEF (working notes)*, 11, 2017.
12. Gordon V Cormack and Maura R Grossman. Systems and methods for conducting a highly autonomous technology-assisted review classification, March 12 2019. US Patent 10,229,117.
13. Maura R Grossman and Gordon V Cormack. Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. *Richmond Journal of Law & Technology*, 17(3):11, 2011.
14. Maura R Grossman, Gordon V Cormack, and Adam Roegiest. Automatic and semi-automatic document selection for technology-assisted review. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 905–908, 2017.
15. Grace E Lee and Aixin Sun. Seed-driven document ranking for systematic reviews in evidence-based medicine. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 455–464, 2018.
16. Harrisen Scells, Guido Zuccon, Anthony Deacon, and Bevan Koopman. Qut ielab at clef ehealth 2017 technology assisted reviews track: initial experiments with learning to rank. In *Working Notes of CLEF 2017-Conference and Labs of the Evaluation Forum [CEUR Workshop Proceedings, Volume 1866]*, pages 1–6. Sun SITE Central Europe, 2017.
17. Amal Alharbi and Mark Stevenson. Ranking studies for systematic reviews using query adaptation: University of sheffield's approach to clef ehealth 2019 task 2 working notes for clef 2019. In *Working Notes of CLEF 2019-Conference and Labs of the Evaluation Forum*, volume 2380. CEUR Workshop Proceedings, 2019.
18. Harrisen Scells, Guido Zuccon, Bevan Koopman, Anthony Deacon, Leif Azzopardi, and Shlomo Geva. Integrating the framing of clinical questions via pico into the retrieval of medical literature for systematic reviews. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2291–2294, 2017.
19. Shuai Wang, Harrisen Scells, Ahmed Mourad, and Guido Zuccon. Seed-driven document ranking for systematic reviews: A reproducibility study. In *European Conference on Information Retrieval*, pages 686–700. Springer, 2022.
20. Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, et al. Large language models are effective text rankers with pairwise ranking prompting. *arXiv preprint arXiv:2306.17563*, 2023.
21. Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. Large language models are zero-shot rankers for recommender systems. *arXiv preprint arXiv:2305.08845*, 2023.
22. Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
23. Oana Frunza, Diana Inkpen, Stan Matwin, William Klement, and Peter O'blenis. Exploiting the systematic review protocol for classification of medical abstracts. *Artificial intelligence in medicine*, 51(1):17–25, 2011.
24. Kentaro Matsui, Tomohiro Utsumi, Yumi Aoki, Taku Maruki, Masahiro Takeshima, and Takaesu Yoshikazu. Large language model demonstrates human-comparable sensitivity in initial screening of systematic reviews: A semi-automated strategy using gpt-3.5. *Available at SSRN 4520426*.
25. Shuai Wang, Harrisen Scells, Bevan Koopman, and Guido Zuccon. Neural rankers for effective screening prioritisation in medical systematic review literature search. In *Proceedings of the 26th Australasian Document Computing Symposium*, pages 1–10, 2022.
26. Barbara Kitchenham and Pearl Brereton. A systematic review of systematic review process research in software engineering. *Information and software technology*, 55(12):2049–2075, 2013.
27. Aaron M Cohen, William R Hersh, Kim Peterson, and Po-Yin Yen. Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association*, 13(2):206–219, 2006.
28. Iain J Marshall and Byron C Wallace. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic reviews*, 8:1–10, 2019.
29. Iain J Marshall, Joël Kuiper, Edward Banner, and Byron C Wallace. Automating biomedical evidence synthesis: Robotreviewer. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2017, page 7. NIH Public Access, 2017.
30. Iain J Marshall, Joël Kuiper, and Byron C Wallace. Robotreviewer: evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association*, 23(1):193–201, 2016.
31. Iain J Marshall, Benjamin Nye, Joël Kuiper, Anna Noel-Storr, Rachel Marshall, Rory Maclean, Frank Soboczenski, Ani Nenkova, James Thomas, and Byron C Wallace. Trialstreamer: A living, automatically updated database of clinical trial reports. *Journal of the American Medical Informatics Association*, 27(12):1903–1912, 2020.
32. Carlos Francisco Moreno-Garcia, Chrisina Jayne, Eyad Elyan, and Magaly Aceves-Martins. A novel application of machine learning and zero-shot classification methods for automated abstract screening in systematic reviews. *Decision Analytics Journal*, page 100162, 2023.
33. Mourad Ouzzani, Hossam Hammady, Zbys Fedorowicz, and Ahmed Elmagarmid. Rayyan—a web and mobile app for systematic reviews. *Systematic reviews*, 5:1–10, 2016.

Confidential: For Review Only

34. Tanja Bekhuis and Dina Demner-Fushman. Screening nonrandomized studies for medical systematic reviews: a comparative study of classifiers. *Artificial intelligence in medicine*, 55(3):197–207, 2012.
35. Ian Shemilt, Antonia Simon, Gareth J Hollands, Theresa M Marteau, David Ogilvie, Alison O'Mara-Eves, Michael P Kelly, and James Thomas. Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. *Research Synthesis Methods*, 5(1):31–49, 2014.
36. Byron C Wallace, Thomas A Trikalinos, Joseph Lau, Carla Brodley, and Christopher H Schmid. Semi-automated screening of biomedical citations for systematic reviews. *BMC bioinformatics*, 11(1):1–11, 2010.
37. Stan Matwin, Alexandre Kouznetsov, Diana Inkpen, Oana Frunza, and Peter O'Blenis. A new algorithm for reducing the workload of experts in performing systematic reviews. *Journal of the American Medical Informatics Association*, 17(4):446–453, 2010.
38. Byron C Wallace, Kevin Small, Carla E Brodley, Joseph Lau, and Thomas A Trikalinos. Deploying an interactive machine learning system in an evidence-based practice center: abstrackr. In *Proceedings of the 2nd ACM SIGHIT international health informatics symposium*, pages 819–824, 2012.
39. Georgios Kontonatsios, Sally Spencer, Peter Matthew, and Ioannis Korkontzelos. Using a neural network-based feature extraction method to facilitate citation screening for systematic reviews. *Expert Systems with Applications: X*, 6:100030, 2020.
40. Raymon van Dinter, Cagatay Catal, and Bedir Tekinerdogan. A decision support system for automating document retrieval and citation screening. *Expert Systems with Applications*, 182:115261, 2021.
41. Xiaonan Ji, Alan Ritter, and Po-Yin Yen. Using ontology-based semantic similarity to facilitate the article screening process for systematic reviews. *Journal of biomedical informatics*, 69:33–42, 2017.
42. David Martinez, Sarvnaz Karimi, Lawrence Cavedon, and Timothy Baldwin. Facilitating biomedical systematic reviews using ranked text retrieval and classification. In *Australasian document computing symposium (adcs)*, pages 53–60, 2008.
43. James Thomas and Alison O'Mara-Eves. How can we find relevant research more quickly? *NCRM Newsletter: MethodsNews*, 2011.
44. Aaron M Cohen, Kyle Ambert, and Marian McDonagh. Cross-topic learning for work prioritization in systematic review creation and update. *Journal of the American Medical Informatics Association*, 16(5):690–704, 2009.
45. Aaron M Cohen, Kyle Ambert, and Marian McDonagh. Studying the potential impact of automated document classification on scheduling a systematic review update. *BMC medical informatics and decision making*, 12:1–11, 2012.
46. Byron C Wallace, Kevin Small, Carla E Brodley, Joseph Lau, Christopher H Schmid, Lars Bertram, Christina M Lill, Joshua T Cohen, and Thomas A Trikalinos. Toward modernizing the systematic review pipeline in genetics: efficient updating via data mining. *Genetics in medicine*, 14(7):663–669, 2012.
47. Rens Van De Schoot, Jonathan De Bruin, Raoul Schram, Parisa Zahedi, Jan De Boer, Felix Weijdema, Bianca Kramer, Martijn Huijts, Maarten Hoogerwerf, Gerbrich Ferdinands, et al. An open source machine learning framework for efficient and transparent systematic reviews. *Nature machine intelligence*, 3(2):125–133, 2021.
48. Murray Shanahan. Talking about large language models. *arXiv preprint arXiv:2212.03551*, 2022.
49. Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv:2302.09419*, 2023.
50. Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.
51. Shuai Wang, Harrisen Scells, Bevan Koopman, and Guido Zuccon. Can chatgpt write a good boolean query for systematic review literature search? *arXiv preprint arXiv:2302.03495*, 2023.
52. Ahmad Alshami, Moustafa Elsayed, Eslam Ali, Abdelrahman EE Eltoukhy, and Tarek Zayed. Harnessing the power of chatgpt for automating systematic review process: Methodology, case study, limitations, and future directions. *Systems*, 11(7):351, 2023.
53. Eugene Syriani, Istvan David, and Gauransh Kumar. Assessing the ability of chatgpt to screen articles for systematic reviews. *arXiv preprint arXiv:2307.06464*, 2023.
54. Eddie Guo, Mehul Gupta, Jiawen Deng, Ye-Jean Park, Mike Paget, and Christopher Naugler. Automated paper screening for clinical reviews using large language models. *arXiv preprint arXiv:2305.00844*, 2023.
55. Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.
56. Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.
57. Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022.
58. Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, David Chartash, et al. How does chatgpt perform on the united states medical licensing examination? the implications of large language models for medical education and knowledge assessment. *JMIR Medical Education*, 9(1):e45312, 2023.
59. Ruqing Zhang, Jiafeng Guo, Lu Chen, Yixing Fan, and Xueqi Cheng. A review on question generation from natural language text. *ACM Transactions on Information Systems (TOIS)*, 40(1):1–43, 2021.
60. Yang Deng, Wenxuan Zhang, Qian Yu, and Wai Lam. Product question answering in e-commerce: A survey. *arXiv preprint arXiv:2302.08092*, 2023.
61. Xiangjue Dong, Jiaying Lu, Jianling Wang, and James Caverlee. Closed-book question generation via contrastive

Confidential: For Review Only

learning. *arXiv preprint arXiv:2210.06781*, 2022.

62. Nehal Muthukumar. Few-shot learning text classification in federated environments. In *2021 Smart Technologies, Communication and Robotics (STCR)*, pages 1–3. IEEE, 2021.
63. Ryan Greene, Ted Sanders, Lilian Weng, and Arvind Neelakantan, Dec 2022.
64. Giorgio Maria Di Nunzio and Evangelos Kanoulas. Special issue on technology assisted review systems, 2023.
65. Alessio Molinari and Evangelos Kanoulas. Transferring knowledge between topics in systematic reviews. *Intelligent Systems with Applications*, 16:200150, 2022.
66. Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. Clef 2019 technology assisted reviews in empirical medicine overview. In *CEUR workshop proceedings*, volume 2380, page 250, 2019.
67. Dan Li and Evangelos Kanoulas. Automatic thresholding by sampling documents and estimating recall. In *CLEF (Working Notes)*, 2019.
68. Giorgio Maria Di Nunzio. A distributed effort approach for systematic reviews. ims unipd at clef 2019 ehealth task 2. In *Clef (working notes)*, 2019.
69. Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.
70. Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*, 2023.
71. Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, et al. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *arXiv preprint arXiv:2310.01783*, 2023.
72. Haodi Zhang, Min Cai, Xinhe Zhang, Chen Jason Zhang, Rui Mao, and Kaishun Wu. Self-convinced prompting: Few-shot question answering with repeated introspection. *arXiv preprint arXiv:2310.05035*, 2023.
73. Mariska MG Leeflang, Jonathan J Deeks, Yemisi Takwoingi, and Petra Macaskill. Cochrane diagnostic test accuracy reviews. *Systematic reviews*, 2(1):1–6, 2013.
74. Annette M O'Connor, Guy Tsafnat, James Thomas, Paul Glasziou, Stephen B Gilbert, and Brian Hutton. A question of trust: can we build an evidence base to gain trust in systematic review automation technologies? *Systematic reviews*, 8(1):1–8, 2019.
75. Xiaorui Jiang. Trustworthiness of systematic review automation: An interview at coventry university. 2022.

Confidential: For Review Only