LGAR: Zero-Shot LLM-Guided Neural Ranking
for Abstract Screening in Systematic Literature Reviews

Christian Jaumann[1,2]   **Andreas Wiedholz**[1]   **Annemarie Friedrich**[2]

[1] XITASO GmbH, Germany   [2] University of Augsburg, Germany

`christian.jaumann|annemarie.friedrich@uni-a.de andreas.wiedholz@xitaso.com`

Abstract

The scientific literature is growing rapidly, making it hard to keep track of the state-of-the-art. Systematic literature reviews (SLRs) aim to identify and evaluate all relevant papers on a topic. After retrieving a set of candidate papers, the abstract screening phase determines initial relevance. To date, abstract screening methods using large language models (LLMs) focus on binary classification settings; existing question-answering (QA) based ranking approaches suffer from error propagation. LLMs offer a unique opportunity to evaluate the SLR's inclusion and exclusion criteria, yet, existing benchmarks do not provide them exhaustively. We manually extract these criteria as well as research questions for 57 SLRs, mostly in the medical domain, enabling principled comparisons between approaches. Moreover, we propose LGAR, a zero-shot LLM-Guided Abstract Ranker composed of an LLM-based graded relevance scorer and a dense re-ranker. Our extensive experiments show that LGAR outperforms existing QA-based methods by 5-10 pp. in mean average precision. Our code and data is publicly available.[1]

1 Introduction

The scientific literature grows rapidly, by an estimated average of roughly 7% per year over the past 22 years (Oliveira et al., 2022), with 3.3 million articles having appeared worldwide in 2022 (Schneider, 2023). Keeping track of the state-of-the-art of a particular research topic is hence inherently difficult. Systematic Literature Reviews (SLRs) are an academic standard of investigating the current state-of-the-art in a specific field. They follow a strict protocol (e.g., Kitchenham and Brereton (2013)). Since conducting SLRs is typically very expensive and time-consuming (Michelson and Reuter, 2019), efforts have been made to automate various parts of the process. Most existing automation
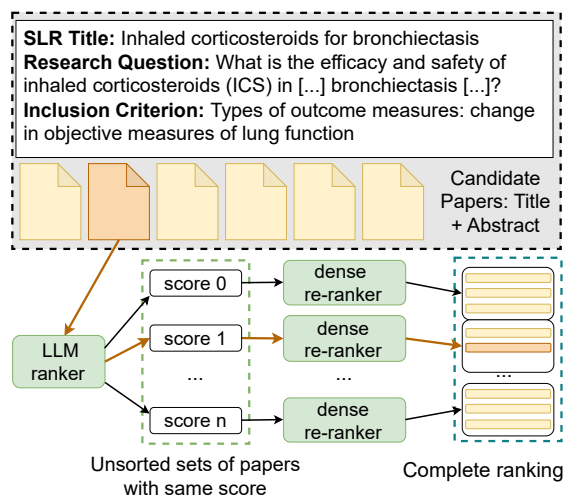

Figure 1: **LGAR method for abstract screening** consisting of a **2-stage ranking approach**: (1) the LLM ranker makes use of a realistic specification of the SLR to assign high-level graded relevance scores, (2) dense re-rankers produce a fully ordered list in scalable time.

approaches focus on facilitating the selection of relevant articles, one of the most time-consuming steps (van Dinter et al., 2021). Large language models (LLMs) offer unprecedented means and zero-shot capabilities for interpreting text in complex retrieval and ranking scenarios such as abstract screening, where a set of inclusion and exclusion criteria and research questions have to be compared to an abstract. Recent studies suggest that methods based on LLMs can reduce up to 60-70% of the workload necessary for paper selection in the abstract screening phase of SLRs (thi Tran et al., 2023; Sandner et al., 2024).

Due to the difficulty of leveraging LLMs for ranking in a scalable manner (Zhu et al., 2023), almost all prior work on abstract screening with LLMs works in a binary classification setting, refraining from inducing a ranking of candidate papers. In practice, a ranked list is much more useful, because the decision of the cut-off point can be left

---

[1] https://github.com/XITASO/lgar/

7910

to the human user flexibly. The only existing LLM-based ranker (Akinseloyin et al., 2024)[2] generates questions for SLR criteria and generates answers for each abstract. The answers are converted to numeric scores using a sentiment classifier, extended with embedding-based similarities, and averaged. While promising, this pipeline-based approach suffers from error propagation in the question generation and sentiment analysis steps.

Existing abstract screening datasets (De Bruin et al., 2023; Kanoulas et al., 2019) provide only part of the SLR criteria. None of them provides the research questions underlying the SLRs. Prior ranking studies approximate the information need by using the SLR's title and Boolean document retrieval queries (e.g., Wang et al., 2023a,c), or use the incomplete set of criteria specified in the SLR abstracts (Akinseloyin et al., 2024). To evaluate LLM-based rankers using the same SLR descriptions as humans do, we manually collect criteria and research questions exhaustively from the full texts of 57 SLRs and validate our collection process. This important dataset extension will be made publicly available to foster future work.

We propose LGAR, a novel zero-shot LLM-Guided Abstract Ranker, which directly leverages LLMs to provide graded relevance judgments. A dense ranker (Nogueira et al., 2020) then re-ranks the sets of abstracts with the same score. Our experiments using our new complete set of criteria and recent open-weights models for all baselines show that LGAR outperforms the existing state-of-the-art system (Akinseloyin et al., 2024) by a large margin (5-10pp. in mean average precision).

Our contributions are as follows: (1) We provide the exhaustive and validated set of inclusion and exclusion criteria, as well as research questions for two SLR datasets, enabling realistic experiments. (2) We propose LGAR, a zero-shot system for LLM-based abstract screening that performs robustly across datasets and LLM families. (3) Our extensive experimental evaluation highlights the superior performance of LGAR compared to existing zero-shot and fine-tuned methods. We perform a detailed ablation study for the task of abstract ranking with LLMs. We find that using large scales with up to 20 options is most effective. This differs from general ranking tasks, where scales with up to five choices have been reported to be optimal.

---

[2]The approach was also recently tested on four SLRs on the impact of climate change on diseases Sujau et al. (2025).

2 Background and Related Work

This section provides a brief overview of SLR automation with a focus on abstract screening.

SLR automation. In the first stage of an SLR a set of candidate papers is retrieved from scientific databases, typically using Boolean queries. The automatic formulation of these queries based on review protocol text has been approached using semantic parsing (Alharbi et al., 2018; Scells et al., 2020b,c), rank fusion (Scells et al., 2020a), and LLM-based generation (Wang et al., 2023b,a). SLRs have also been supported by applying summarization methods (Bui et al., 2016) and by extracting fine-grained entity types and relations (Panayi et al., 2023). Most active learning (AL) tools for SLR automation use traditional machine learning (ML), e.g., AS Review (van de Schoot et al., 2021), Abstrackr (Wallace et al., 2012), Colandr (Cheng et al., 2018), FASTREAD (Yu et al., 2018), Rayyan (Ouzzani et al., 2016), or RobotAnalyst (Przybyła et al., 2018). Carvallo and Parra (2019) compare neural classifiers in an AL setup. AL assumes substantial training data. In contrast, we focus on zero-shot settings.

Abstract screening. Early approaches for binary relevance classification for abstract screening apply traditional ML techniques (Kim and Choi, 2014; Khabsa et al., 2016; Matwin et al., 2010; Matwin and Sazonova, 2012). Qin et al. (2021) fine-tune several BERT models (Devlin et al., 2019) as binary classifiers and combine their predictions. First learning-to-rank setups for abstract screening use non-neural techniques (Anagnostou et al., 2017; Lagopoulos et al., 2018). Lee (2017) combines sentence-level relevance scores computed by a convolutional neural network. Wang et al. (2023c) use a cross-encoder approach following monoBERT (Nogueira et al., 2019). The first abstract screening methods using LLMs all operate in classification settings, applying ChatGPT in a zero-shot setting (Syriani et al., 2023; Guo et al., 2023). Robinson et al. (2023) also fine-tune Llama (Touvron et al., 2023) and Guanaco (Dettmers et al., 2024) to generate binary inclusion labels. Wang et al. (2024b) compare a larger set of open-weights LLMs on a calibrated binary classification task.

Ranking with LLMs. Integrating LLMs with ranking approaches is a highly active research area that has just recently emerged (Zhu et al., 2023). LLMs can be integrated as query rewriters, as dense retrievers, as "readers" to generate natural language

Dataset	# SLRs	# Papers	Domain	Avg. inclusion rate (%)
SYNERGY	26	169288	medicine (21), psychology (6), computer science (2), chemistry (2), biology (3), mathematics (1)	4.31
CLEF TAR2019				
– DTA	8	30521	medicine (diagnostic technology assessment)	7.05
– Intervention	20	41996	medicine (clinical intervention trials)	5.49
– Qualitative	2	6536	medicine (qualitative studies)	1.32
– Prognosis	1	3367	medicine (prognosis)	5.70

Table 1: Overview of labeled benchmark datasets for automating systematic literature reviews. Counts for the domains in SYNERGY do not add up to 26 because several SLRs fall under multiple scientific disciplines.

answers, or for data augmentation. We here focus on their application in the ranking stage, where they can be applied using a *pointwise* (Sun et al., 2023), *pairwise* (Qin et al., 2024), or *listwise* approach (Tang et al., 2024). The latter two are inefficient for long candidate lists, which is usually the case in abstract screening. In pointwise ranking, the generation likelihood of the query given the document can be used as relevance score (Sachan et al., 2022), or the LLM can be prompted to generate a relevance score (Liang et al., 2023). Zhuang et al. (2024) show that adding intermediate relevance options can improve ranking performance over binary setups. As abstract screening queries are long and complex, we turn to relevance generation.

To the best of our knowledge, the only existing *ranking* approach for abstract screening using LLMs (Akinseloyin et al., 2024) uses a QA setup. For each inclusion/exclusion criterion, a question is generated using ChatGPT, and the LLM is prompted to answer them for each paper. Answers are converted to numerical scores based on whether a BART (Lewis et al., 2020) model trained on sentiment analysis assigns *Positive* (1), *Neutral* (0.5), or *Negative* (0). The final score for each paper is computed by averaging these scores with embedding similarities of the questions, the original criteria, and the abstract. Akinseloyin et al. (2024) report that while promising, their pipeline-based approach suffers from error propagation. Our own system, LGAR, uses a more effective combination of zero-shot LLM-based pointwise relevance generation and monoT5-based re-ranking.

3 Dataset Extensions

In this section, we describe our extension of two abstract screening benchmarks with inclusion/exclusion criteria and research questions to facilitate more realistic evaluation setups. We extend the SYNERGY dataset (De Bruin et al., 2023) and the test set of the CLEF Technological Assisted Re-

view 2019 (TAR2019) dataset (Task 2) (Kanoulas et al., 2019).[3] Both datasets are in English and provide the title of each SLR and a set of papers labeled with binary relevance judgments.[4] Table 1 gives an overview of the corpus statistics. The inclusion rate specifies the percentage of abstracts included in the SLR out of the original candidate set retrieved using Boolean queries. The majority of SLRs address topics in medicine.

3.1 Extension with Inclusion/Exclusion Criteria and Research Questions

For SLRs, inclusion and exclusion criteria as well as research questions are specified explicitly as a basis for decision making. Akinseloyin et al. (2024) use the criteria section from the structured abstracts (but no research questions) for TAR2019. However, the full set of criteria and research questions is not always spelled out comprehensively in the SLR's abstract (e.g., it is not the case for SYNERGY).

To provide the models with a realistic evaluation setup, which can be reproduced and re-used by future work, we manually retrieve the inclusion and exclusion criteria and research questions from the full texts of all 57 SLRs of both datasets. If the review itself does not explicitly mention research questions, we infer them from its objectives or extract them from the review protocols which are provided for the TAR2019 dataset. We focus on primary research questions and objectives. Some SLRs also provide sub-questions and secondary objectives. We do not extract these because they are not typically defined at the start of the abstract screening phase, but more often during the refinement phase of the SLR when several articles have already been checked in greater detail.

---

[3]License SYNERGY: CC0 1.0 Universal; License TAR2019: MIT

[4]The SYNERGY dataset has very recently been extended with inclusion and exclusion criteria also collected from full texts, this is concurrent to our contribution. They do not provide research questions.

In the full text, the criteria are often combined with explanations, definitions, and reasons for specifying a particular criterion. We only extract the criteria, separating them from surrounding descriptive text, and reformulate them as key points. In the medical domain, there are sometimes multiple criteria describing the type of studies or type of participants. We summarize these into one criterion describing all possible types of the respective category. We ensure that the extracted text does not refer to other material in the review, such as figures or tables. In very rare cases, SLRs also formulate inclusion and exclusion criteria based on metadata such as publication dates which are taken care of during document retrieval. Our set of extracted criteria focuses on semantic information needs.

3.2 Validation

The extraction of criteria and research questions for all SLRs was performed by the first author of this paper. We validate the extracted information on 15 randomly selected SLRs (5 SLRs from each the SYNERGY, DTA and Intervention dataset). The second author of this paper extracted the criteria and research questions with the same protocol without prior inspection of the data extracted by the first author. We align the extracted criteria and research questions using the cosine similarities of their BERT embeddings and then verify manually whether they indeed correspond to each other.

For the 15 SLRs in our validation study, the first and second author each extracted 23 research questions, which matched exactly. This high inter-rater agreement stems from the structured nature of medical and computer science SLRs, which often present a section titled "Objectives" or explicitly state research questions, which allows straightforward extraction. Incidentally, each author extracted 99 inclusion/exclusion criteria, of which 97 match. Disagreements can be narrowed down to one author including more details to an criterion than the other one or including/neglecting a criterion that was not mentioned explicitly as criterion but described as important factor in the review process. Overall, we include that inclusion/exclusion criteria and research questions have been extracted reliably for our extension of the SYNERGY and CLEF TAR2019 datasets.

4 LLM-Guided Abstract Ranking

We propose LGAR, a ranking model comprised of a two-stage process to determine the relevance of a screened paper, as illustrated in Figure 1. In the first stage, the LLM assigns each paper a relevance score on a scale from 0 to k, similar to the LLM-based ranking approach *DIRECT(0,k)* of Guo et al. (2024).

Ranking for abstract screening requires producing an ordered list of potentially thousands of papers. The differentiation that LLMs are able to produce is limited to the range of a dozen relevance scores (Zhuang et al., 2024), i.e., the first ranking step will produce many ties. We hence produce the global ranking by additionally ranking the papers within each set of papers that have obtained the same relevance score by the LLM using a zero-shot dense ranking model. Our proposed two-stage ranker is modular, providing for modular combinations of LLMs and dense rankers.

4.1 Graded Relevance Judgments by LLM

In the first ranking stage, the LLM is prompted with the SLR's criteria, as well as research questions (shown in Figure 2). The general instruction encoded in the system message suggests that a paper should only be considered relevant if all inclusion criteria but none of the exclusion criteria are met. The method is configurable to use different prompting techniques, but the key elements remain the same. The system message provides the title and research questions of the SLR and the associated ranking task, in addition to a definition of relevance in the context of the SLR. The user message contains a more specific task description, instructing the LLM to determine a degree of relevance in the form of assigning a score from a Likert scale using the provided selection criteria, and to answer in a predefined format. It also provides the title and the abstract of the paper.

Prompting Techniques. LLM responses are highly dependent on the type of prompt used (Brown et al., 2020; Wu et al., 2023; Wang et al., 2023d). Therefore, we test several different prompting techniques, i.e., zero-shot, Chain of Thought (CoT), and CoT with self-consistency (n=3). We also tried 2-shot variants without observing any improvements (see Appendix A.4). In self-consistency setups, the relevance scores of each paper are averaged over three runs. For CoT prompting, we instruct the model to "think step by step" (Kojima et al., 2022).

Figure 2: Zero-shot prompts used in LGAR.

Output Parsing. To make the responses as deterministic and reproducible as possible, we use a temperature of 0. The scores are extracted using regular expression and we verify that they are in the desired relevance scale. If score extraction fails, and when using self-consistency, the temperature is set to 0.5 to encourage more diverse results. In the rare case that the LLM does not respond in the expected format, the request is repeated up to three times, without providing the model's previous faulty response. If the response is still not in the expected format, the paper is assigned the average of the sum of all LLM generated relevance scores for the respective SLR.

4.2 Neural Dense Re-Ranking

In the second stage, we re-rank the papers within the groups generated by the LLM with neural dense retrievers backboned by pretrained language models such as monoBERT (Nogueira and Cho, 2019) or ColBERT (Khattab and Zaharia, 2020). These cross- and bi-encoders based on BERT still rank among the state-of-the-art for re-ranking tasks. Each query-document pair is the concatenation of the query followed by the title and abstract of the paper in question, truncated to 512 wordpiece tokens. As query, we use (1) only the SLR's title following Wang et al. (2023c) or (2) the title and the research questions of the SLR. We use neural re-rankers that were fine-tuned on the general retrieval dataset MS MARCO (Nguyen et al., 2016), but not specifically tuned for abstract screening.

5 Experiments

To compare LGAR to existing approaches, we conduct an extensive experimental evaluation using our newly constructed benchmark. In this section, we describe our experimental design, including our selection of models, our choice of evaluation metrics, and the baselines. We then present our experimental results, followed by several ablation studies.

5.1 Model Selection and Experimental Settings

To enhance the reproducibility of our results, our model selection of instruction-following LLMs is constrained to open-weights alternatives. We focus on Llama3.3-70B-Instruct (Dubey et al., 2024), which is one of the best performing open-weights models for its size at the time of writing. To assess the extent of performance disparities among diverse model families and varying model sizes, we conduct comparative analyses with the following models: Llama3.1-8B-Instruct, Qwen2.5-32B-Instruct,[5] Qwen2.5-72B-Instruct,[6] and Mistral-Large-Instruct-2411 (123B).[7] We run all models using 16-bit quantization.

We report test results on the entire SYNERGY dataset (which does not provide any training data), and on the test portion of the TAR2019 dataset.[8] To develop LGAR, we performed prompt-engineering using Llama3.3-70B-Instruct in interactive experiments using SLRs taken from an additional dataset in the medical domain (Guo et al., 2023) (see Appendix A.1 in particular for the exploration of different scale sizes) and one unpublished SLR. We also run the zero-shot dense rankers on our tuning data, identifying monoT5 (Nogueira et al., 2020) in the 3B variant as most effective (also see Appendix A.1). In our ablation tests, we additionally evaluate monoBERT (large) (Nogueira and Cho, 2019) and ColBERTv2 (Santhanam et al., 2022).

5.2 Evaluation Metrics

We adopt the evaluation metrics of the CLEF TAR 2019 challenge (Kanoulas et al., 2019): mean average precision (MAP), recall at $k\%$ of top-ranked abstracts ($R@k\%, k = 1, 5, 10, 20, 50$), and work

---

[5]https://huggingface.co/Qwen/Qwen2.5-32B-Instruct
[6]https://huggingface.co/Qwen/Qwen2.5-72B-Instruct
[7]https://huggingface.co/mistralai/Mistral-Large-Instruct-2411

[8]In alignment with the general evaluation protocol used in the CLEF TAR 2019 Shared Task, we only use the test split of the CLEF TAR dataset comprising 31 SLRs for evaluating our zero-shot approach. We use the train split of 100 additional SLRs to replicate the fine-tuning of BERT-based models for comparing different model types.

7914

Dataset	Model	MAP	TNR@95%	R@1%	R@5%	R@10%	R@20%	R@50%	WSS@95%	WSS@100%
SYNERGY	random reranking	4.9	6.2	1.2	5.1	10.4	20.0	50.3	3.0	3.0
	BM25 (T+R)	12.5	36.3	9.0	26.4	40.0	54.4	80.3	33.6	21.8
	monoT5 (T)	18.8	52.3	14.1	37.5	54.5	69.1	91.1	50.0	35.4
	Akinseloyin et al. (ours) [†]	34.0	63.6	25.8	55.0	67.8	80.5	93.1	59.5	**50.6**
	LGAR (T, monoT5) [†]	36.8	63.8	32.7	57.7	71.0	81.2	94.3	63.6	46.5
	LGAR (T+R, monoT5)	**40.7**	**67.0**	34.1	**59.3**	**72.0**	81.9	**94.4**	**65.2**	49.3
	+ CoT	38.8	62.9	33.4	**59.3**	70.3	81.4	92.8	62.5	45.8
	+ CoT (n=3)	**40.7**	66.4	**34.7**	58.6	70.7	**82.6**	93.5	64.5	50.4
	LGAR (T+R, random rerank)	38.9	59.8	33.3	58.4	70.1	79.7	91.4	58.6	40.9
TAR2019	random reranking	6.7	9.9	1.1	5.5	10.6	20.2	52.0	6.6	7.0
	BM25 (T+R)	21.1	44.8	12.6	34.2	46.2	60.5	82.5	40.9	30.4
	monoT5_3B (T)	30.1	63.2	17.7	44.6	59.3	73.4	94.0	58.2	50.7
	Akinseloyin et al. (theirs)	42.8	68.6	23.9	53.8	70.8	83.8	96.0	63.7	55.2
	Akinseloyin et al. (ours) [†]	45.1	70.8	26.8	57.5	72.9	83.4	**96.6**	66.1	53.0
	LGAR (T, monoT5) [†]	48.4	76.4	26.0	60.9	75.7	86.8	96.2	69.9	**61.1**
	LGAR (T+R, monoT5)	*50.6*	**76.5**	29.3	**63.6**	**76.7**	**88.3**	95.8	**71.2**	60.9
	+ CoT	43.5	72.0	25.7	57.5	69.8	86.0	96.5	67.0	60.5
	+ CoT (n=3)	44.8	74.9	26.1	57.9	74.2	85.7	95.8	68.8	59.4
	LGAR (T+R, random rerank)	**50.7**	64.8	**29.7**	62.1	75.1	86.0	93.7	61.4	42.9

Table 2: Zero-shot models; recall values are macro-averages over SLRs. All LLM-based models use Llama3.3-70B. T = title of the SLR, R = research questions of the SLR. Akinseloyin et al. (2024): ours = uses our annotated criteria, theirs = uses their criteria. In all combinations, monoT5 uses the title as query and the LLM the scale 0-19. [†] indicates that the approaches get the same information on SLR and paper.

saved over sampling (WSS) at different recall levels ($WSS@r\%$, $r = 95, 100$), defined as follows by Cohen et al. (2006):

$$\text{WSS@r\%} = \frac{\text{TN} + \text{FN}}{\text{N}} - (1 - r) \qquad (1)$$

WSS aims to estimate the reduction in human screening workload for a given recall level. However, it has some drawbacks, such as that its value range depends on the inclusion rate of the respective SLR, making comparisons across multiple SLRs unfair (Kusa et al., 2023). Therefore, we additionally provide the min-max normalized version of WSS proposed by Kusa et al. (2023), which mitigates these issues and reports the True Negative Rate at recall level r:

$$\text{nWSS@r\%} = \text{TNR@r\%} = \frac{\text{TN}}{\text{TN} + \text{FP}} \qquad (2)$$

To compute results, we use the evaluation scripts provided in CLEF TAR 2018 with one adaptation: we found that they compute the average recall of multiple SLRs using micro-averaging:

$$R@k\% = \frac{\sum_{\{\text{SLRs}\}} \text{TP}}{\sum_{\{\text{SLRs}\}} (\text{TP} + \text{FN})} \qquad (3)$$

This gives more weight to larger SLRs or SLRs with a higher inclusion rate. To obtain a more realistic estimate of performance for an unseen SLR, we instead follow Akinseloyin et al. (2024)

by first computing recall scores per SLR, and then macro-averaging over the obtained scores.

5.3 Baselines

As non-neural information retrieval baseline, we provide the results of Okapi **BM25** (Robertson and Zaragoza, 2009). We also report the results of **monoT5**, which is similar to the zero-shot model of Wang et al. (2023c) that uses BERT-base.

The closest existing work to ours is that of Akinseloyin et al. (2024), who use a QA setup and (except for a small case study) ChatGPT. For a fair comparison, we replicate the *GPT_QA_Soft_Both_ReRank* variant of their model (which also uses selection criteria for each SLR) using Llama3.3-70B-Instruct and the embeddings proposed by Wang et al. (2024a) (instead of the closed-source embeddings they use). To increase reproducibility, we use a temperature of 0 instead of 0.2 as in the original work. Our replication outperforms their scores provided for the TAR2019 dataset on all metrics except WSS (which is a somewhat unreliable metric and only reported for comparison with prior work). Detailed scores are provided in Appendix A.2.

5.4 Main Results and Findings

Table 2 compares the effectiveness of LGAR using Llama3.3-70B, relevance scale 0-19, and monoT5

(title-only) as re-ranker with our baselines. On both the SYNERGY and the TAR2019 datasets, zero-shot monoT5 using only the title as query outperforms BM25 that compares the titles and abstracts of each paper to the title and research questions of the SLR. For reasons explained in section 5.2, we focus on TNR instead of WSS when interpreting our results. Overall, LGAR achieves the strongest performance over all models, outperforming the QA-based model of Akinseloyin et al. (2024) by a large margin.

Impact of dense re-ranking. Our ablation removing monoT5, randomly ranking the papers within each set, shows that in terms of MAP, the difference is small for SYNERGY and negliglible for TAR2019, which is somewhat expected as the LLM-based relevance judgment defines the overall order of the ranking. Nevertheless, the dense re-ranker contributes to workload savings by moving up articles that were incorrectly assigned a low relevance score by the LLM, which is particularly highlighted by the the strong increase of around 7-12% in TNR. Similarly, LGAR profits from having access to the SLR's research questions (this difference is more pronounced on SYNERGY).

Comparison to QA-based system. LGAR (T, monoT5) and the QA model of Akinseloyin et al. (2024) (ours) differ with respect to model architecture and prompts, yet both receive exactly the same information about the SLR and the papers. Even in this setting (without using research questions), LGAR outperforms the QA model, showing that the performance increase is not merely due to additional information in the prompt but due to model architecture. However, given that the research questions should be available at the stage of abstract screening when conducting an SLR and that adding them appears to enhance the performance of our approach, it seems reasonable to make use of them.

CoT prompting. Instructing the model to think step by step does not improve the performance of LGAR. Using self-consistency to make up for erroneous runs improves the situation in particular for SYNERGY, but does not lead to any performance improvements over standard LGAR. Detailed information can be found in Appendix A.4.

5.5 Ablation Tests and Analysis

We now perform an extensive exploration into various settings of LGAR. Detailed scores are provided in Appendix A.3.

Dataset	Model	MAP	TNR@95%	R@20%
SYNERGY	Llama3.1-8B	31.9	61.3	80.9
	Llama3.3-70B	40.7	67.0	81.9
	Mistral-123B	38.8	65.8	82.0
	Qwen2.5-32B	38.7	63.2	78.6
	Qwen2.5-72B	**41.3**	**68.5**	**83.0**
TAR2019	Llama3.1-8B	42.3	69.6	86.1
	Llama3.3-70B	**50.6**	**76.5**	88.3
	Mistral-123B	48.4	70.8	86.4
	Qwen2.5-32B	47.6	73.4	84.6
	Qwen2.5-72B	50.2	73.0	**89.0**

Table 3: Comparison of LLMs. Scale: 0-19. Re-ranker: monoT5 with title-only as query. 0-shot prompting.

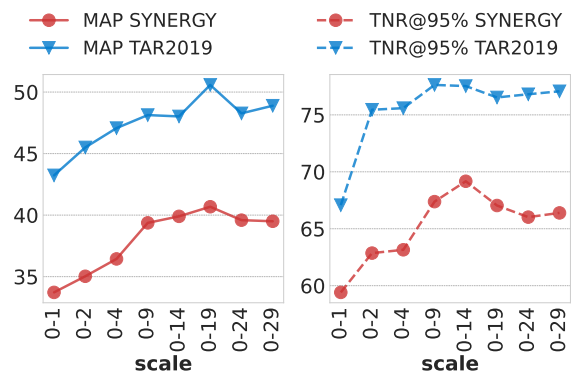### 5.5.1 Effectiveness of different LLMs

Figure 3: Comparison of scales in terms of MAP and TNR@95%. The optimum is around 0-14 and 0-19.

Table 3 reports the effectiveness of using different LLMs with a relevance scale of 0-19 and monoT5 (title-only) as re-ranker. Performance degrades for smaller models, but using the larger Mistral-123B did not provide any benefits over Llama3.3-70B. Qwen2.5-72B excels on SYNERGY and also performs strongly on TAR2019. Overall, LGAR (which was developed for Llama-3.3-70B) is relatively robust with regard to the underlying LLM (within the same model sizes).

5.5.2 Impact of scale size

We here conduct a principled analysis of various choices for scales. Figure 3 reports the performance of LGAR using several different scale sizes. Zhuang et al. (2024) found that for general ranking tasks, 0-4 worked best. By contrast, for our complex ranking task, it is beneficial to increase the scale size even further. In terms of workload savings, 0-14 performs best, while the best ranking is produced by 0-19. Increasing the scale size beyond 0-19 does not further enhance the performance. Interestingly, this replicates our initial findings on
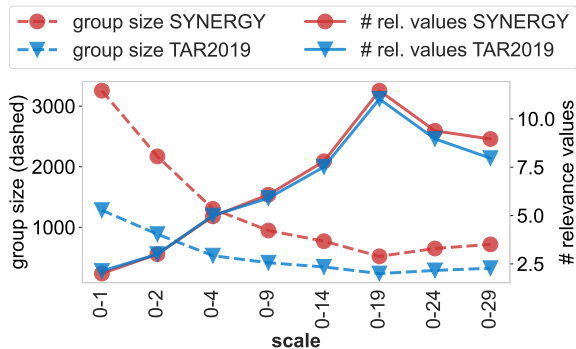
Figure 4: Average number of different relevance scores actually used by LLM and average group size by scale, all SLRs of SNYERGY and TAR2019, Llama3.3-70B as LLM ranker and monoT5 (title-only) as re-ranker.
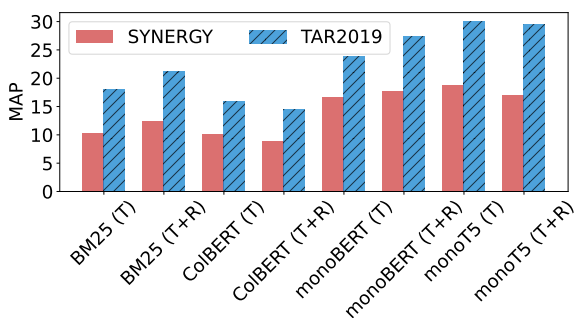

Figure 5: MAP of zero-shot dense rankers (without LGAR). T = title / R = research questions used as query.

Dataset	Model	MAP	TNR@95%	R@20%
Interv.	BioBERT_ft (T)	45.9	70.8	84.2
	LGAR (T)	55.0	**75.0**	**88.9**
	LGAR (T+R)	**56.8**	75.0	88.4
DTA	BioBERT_ft (T)	35.4	76.0	82.5
	LGAR (T)	34.5	**78.1**	84.2
	LGAR (T+R)	**38.1**	76.8	**85.0**

Table 4: Comparison to fine-tuned ranker. BioBERT_ft is our replication of Wang et al. (2023c). Re-ranker for LGAR: monoT5.
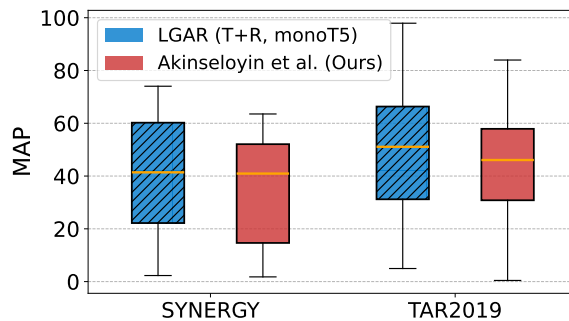
Figure 6: Boxplots of MAP scores of LGAR (T+R, monoT5) and Akinseloyin et al. (2024) (ours) on both datasets.

our tuning data, validating our choice of 0-19. As can be seen in Figure 4, the number of different relevance scores actually used by the LLM drops for larger scales. The average number of papers ending up in one group is smallest for 0-19, which means in the 0-19 setting, the LLM takes over most ranking workload compared to other settings.

5.5.3 Effectiveness of dense rankers

In Table 4, we compare the performance of BM25 with that of three zero-shot dense rankers (without using the LLM step) using two different queries: title-only (T) and title and research questions (T+R). Details are reported in the Appendix A.1. On both the SYNERGY and the TAR2019 dataset, the best configuration by far uses monoT5 with only the title as query. While BM25 and monoBERT can deal with the longer input query (T+R), the performance of ColBERT and monoT5 degrades over only using the title.

5.5.4 Comparison to fine-tuned model

To put LGAR's strong zero-shot performance into perspective, we compare it to the fine-tuned model of Wang et al. (2023c). We replicate the model to report all metrics in the exact same setting. Our replicated model slightly outperforms the results reported by Wang et al. (2023c) (see Appendix A.2). We fine-tune BioBERT (large) (Lee et al., 2020) using the framework of Gao et al. (2021) and the training splits for the Intervention and DTA subset of TAR2019 (there is no training data for the rest of TAR2019 and for SYNERGY). Our zero-shot LGAR model outperforms the fine-tuned model by 7-10pp. in terms of MAP (see Table 4).

5.5.5 Performance distribution across SLRs

Figure 6 compares the distribution of MAP scores achieved by LGAR with that by our replication of the approach of Akinseloyin et al. (2024) on both datasets. MAP scores differ greatly between SLRs, which is due to their very different inclusion rates. LGAR appears to have an even greater range of scores due to its higher peak MAP scores. Overall, LGAR achieves at least the same performance as the QA-based approach, with a clear tendency to be better.

6 Discussion and Outlook

Our results suggest that LLMs are very effective in interpreting inclusion and exclusion criteria for

abstract screening. They provide a good initial categorization. Re-ranking is particularly effective for finding relevant papers accidentally categorized as irrelevant by the LLM, thus significantly contributing to workload savings.

Compared to the approach of Akinseloyin et al. (2024), it does not seem necessary to split the criteria into several parts, which significantly reduces the number of LLM requests. At the same time, LGAR is less "explainable" than models evaluating every criterion separately. We hypothesize that LGAR performs more robustly precisely because it is not forced to evaluate each criterion separately, which likely introduces many noisy scores in the QA-based model. For many criteria, the abstract may not contain sufficient information to allow a meaningful assessment. The QA-based baseline is supposed to generate *Neutral* in such cases, yet, detecting when information is insufficient to answer a question is yet a difficult task for LLMs (Ji et al., 2022; Gautam et al., 2023). In future work, it would be interesting to further support the human user by identifying the criteria that led to including or excluding a paper.

7 Conclusion

In this paper, we have proposed LGAR, a novel LLM-Guided Abstract Ranker for SLRs. LGAR is based entirely on zero-shot open-weights models, yet outperforms models specifically tuned for this task. We contribute a dataset extension of exhaustively extracted criteria and research questions and compare to baselines in a fair way, using the same backbones and inputs, finding that LGAR performs best on two abstract screening datasets.

Limitations

The SYNERGY and TAR2019 datasets have been created based on actual SLRs, i.e., the set of relevant papers has been extracted from the SLRs' bibliographies. Hence, strictly speaking, both datasets that we use in this study provide relevance labels that correspond to the full text screening stage. Yet, as redistributing the full text of scientific publications is not generally possible due to license restrictions, both datasets only provide titles and abstracts of the papers. Hence, it is theoretically possible that some papers were still included after the abstract screening stage of an SLR, but excluded in the full text screening stage. In these cases, the title and abstract of the paper, which are the basis for

abstract screening, may not contain the necessary information. Despite this minor approximation, the SYNERGY and the TAR2019 datasets nevertheless provide suitable benchmarks for automatic abstract screening methods, an assumption that we share with prior work (Akinseloyin et al., 2024; Wang et al., 2023c,a, 2024b).

Another limitation of the evaluation is that even though the evaluation was conducted with 57 different SLRs, the approach could only be tested mainly in the medical domain. This is due to the composition of the datasets that are widely used when evaluating automation methods for SLRs. Although other domains are represented in the SYNERGY dataset, we believe that the number of SLRs for these respective domains is too small to give a thorough answer to the question of whether the performance of the approach is equivalent to that in the medical domain.

In addition, since many of these SLRs were published several years ago, some of their data may have already been exposed to the LLMs in their training. Hence, we can not exclude the possibility of data contamination.

Furthermore, since LLMs can perpetuate and amplify biases present in the training data, LGAR may inadvertently reinforce these biases when ranking abstracts. This could lead to certain research being unfairly prioritized or overlooked, potentially skewing the outcomes of SLRs.

Risks. When using this approach in conducting an SLR, the main risk is to overlook a paper that is relevant to the research questions being addressed in the SLR. Missing a relevant paper is problematic when writing an SLR, which can happen when defining a cut-off in the ranked list when applying LGAR to the abstract screening stage of an SLR. However, we argue that this risk is also present when this is done manually, since manual screening of hundreds or thousands of papers is also error-prone.

Acknowledgments

References

Opeoluwa Akinseloyin, Xiaorui Jiang, and Vasile Palade. 2024. A question-answering framework for automated abstract screening using large language models. *Journal of the American Medical Informatics Association*, page ocae166.

Amal Alharbi, William Briggs, and Mark Stevenson. 2018. Retrieving and ranking studies for systematic reviews: University of sheffield's approach to CLEF ehealth 2018 task 2. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*, volume 2125 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Antonios Anagnostou, Athanasios Lagopoulos, Grigorios Tsoumakas, and Ioannis P. Vlahavas. 2017. Combining inter-review learning-to-rank and intra-review incremental training for title and abstract screening in systematic reviews. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017*, volume 1866 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Duy Duc An Bui, Guilherme Del Fiol, John F. Hurdle, and Siddhartha Jonnalagadda. 2016. Extractive text summarization system to aid data extraction from full text in systematic review development. *J. Biomed. Informatics*, 64:265–272.

Andres Carvallo and Denis Parra. 2019. Comparing word embeddings for document screening based on active learning. In *Proceedings of the 4th Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2019) co-located with the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019), Paris, France, July 25, 2019*, volume 2414 of *CEUR Workshop Proceedings*, pages 100–107. CEUR-WS.org.

SH Cheng, C Augustin, A Bethel, D Gill, S Anzaroot, J Brun, B DeWilde, RC Minnich, R Garside, YJ Masuda, et al. 2018. Using machine learning to advance synthesis and use of conservation and environmental evidence. *Conservation Biology*, 32(4):762–764.

Aaron M. Cohen, William R. Hersh, K. Peterson, and Po-Yin Yen. 2006. Research paper: Reducing workload in systematic review preparation using automated citation classification. *J. Am. Medical Informatics Assoc.*, 13(2):206–219.

Jonathan De Bruin, Yongchao Ma, Gerbrich Ferdinands, Jelle Teijema, and Rens Van de Schoot. 2023. SYNERGY - Open machine learning dataset on study selection in systematic reviews.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and et al. 2024. The Llama 3 Herd of Models. *CoRR*, abs/2407.21783.

Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021. Rethink training of BERT rerankers in multi-stage retrieval pipeline. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part II*, volume 12657 of *Lecture Notes in Computer Science*, pages 280–286. Springer.

Vagrant Gautam, Miaoran Zhang, and Dietrich Klakow. 2023. A lightweight method to generate unanswerable questions in english. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Eddie Guo, Mehul Gupta, Jiawen Deng, Ye-Jean Park, Mike Paget, and Christopher Naugler. 2023. Automated paper screening for clinical reviews using large language models. *CoRR*, abs/2305.00844.

Fang Guo, Wenyu Li, Honglei Zhuang, Yun Luo, Yafu Li, Le Yan, and Yue Zhang. 2024. Generating diverse criteria on-the-fly to improve point-wise LLM rankers. *CoRR*, abs/2404.11960.

Yunjie Ji, Liangyu Chen, Chenxiao Dou, Baochang Ma, and Xiangang Li. 2022. To answer or not to answer? improving machine reading comprehension model with span-based contrastive learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1292–1300, Seattle, United States. Association for Computational Linguistics.

Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. 2019. Clef 2019 technology assisted reviews in empirical medicine overview. *CEUR Workshop Proceedings*, 2380. 20th Working Notes of CLEF Conference and Labs of the Evaluation Forum, CLEF 2019 ; Conference date: 09-09-2019 Through 12-09-2019.

Madian Khabsa, Ahmed K. Elmagarmid, Ihab F. Ilyas, Hossam Hammady, and Mourad Ouzzani. 2016. Learning to identify relevant studies for systematic reviews using random forest and external information. *Mach. Learn.*, 102(3):465–482.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 39–48. ACM.

Seunghee Kim and Jinwook Choi. 2014. An svm-based high-quality article classifier for systematic reviews. *J. Biomed. Informatics*, 47:153–159.

Barbara Kitchenham and Pearl Brereton. 2013. A systematic review of systematic review process research in software engineering. *Information and Software Technology*, 55(12):2049–2075.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Wojciech Kusa, Aldo Lipani, Petr Knoth, and Allan Hanbury. 2023. An analysis of work saved over sampling in the evaluation of automated citation screening in systematic literature reviews. *Intell. Syst. Appl.*, 18:200193.

Athanasios Lagopoulos, Antonios Anagnostou, Adamantios Minas, and Grigorios Tsoumakas. 2018. Learning-to-rank and relevance feedback for literature appraisal in empirical medicine. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings*, volume 11018 of *Lecture Notes in Computer Science*, pages 52–63. Springer.

Grace Eunkyung Lee. 2017. A study of convolutional neural networks for clinical document classification in systematic reviews: Sysreview at CLEF ehealth 2017. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017*, volume 1866 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang.

2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.*, 36(4):1234–1240.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yüksekgönül, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. *Trans. Mach. Learn. Res.*, 2023.

Stan Matwin, Alexandre Kouznetsov, Diana Inkpen, Oana Frunza, and Peter O'Blenis. 2010. A new algorithm for reducing the workload of experts in performing systematic reviews. *J. Am. Medical Informatics Assoc.*, 17(4):446–453.

Stan Matwin and Vera Sazonova. 2012. Direct comparison between support vector machine and multinomial naive bayes algorithms for medical abstract classification. *Journal of the American Medical Informatics Association*, 19(5):917–917.

Matthew Michelson and Katja Reuter. 2019. 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.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pre-trained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718, Online. Association for Computational Linguistics.

Rodrigo Frassetto Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *CoRR*, abs/1901.04085.

Rodrigo Frassetto Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019. Multi-stage document ranking with BERT. *CoRR*, abs/1910.14424.

Eduardo A Oliveira, Maria Christina L Oliveira, Enrico A Colosimo, Daniella B Martelli, Ludmila R Silva, Ana Cristina Simões E Silva, and Hercílio Martelli-Júnior. 2022. Global scientific production in the pre-covid-19 era: An analysis of 53 countries for 22 years. *Anais da Academia Brasileira de Ciências*, 94(suppl 3):e20201428.

Mourad Ouzzani, Hossam Hammady, Zbys Fedorowicz, and Ahmed Elmagarmid. 2016. Rayyan—a web and mobile app for systematic reviews. *Systematic reviews*, 5:1–10.

Antonia Panayi, Katherine Ward, Amir Benhadji-Schaff, A Santiago Ibanez-Lopez, Andrew Xia, and Regina Barzilay. 2023. Evaluation of a prototype machine learning tool to semi-automate data extraction for systematic literature reviews. *Systematic Reviews*, 12(1):187.

Piotr Przybyła, Austin J Brockmeier, Georgios Kontonatsios, Marie-Annick Le Pogam, John McNaught, Erik von Elm, Kay Nolan, and Sophia Ananiadou. 2018. Prioritising references for systematic reviews with robotanalyst: a user study. *Research synthesis methods*, 9(3):470–488.

Xuan Qin, Jiali Liu, Yuning Wang, Yanmei Liu, Ke Deng, Yu Ma, Kang Zou, Ling Li, and Xin Sun. 2021. Natural language processing was effective in assisting rapid title and abstract screening when updating systematic reviews. *Journal of Clinical Epidemiology*, 133:121–129.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. 2024. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 1504–1518. Association for Computational Linguistics.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Ambrose Robinson, William Thorne, Ben P. Wu, Abdullah Pandor, Munira Essat, Mark Stevenson, and Xingyi Song. 2023. Bio-sieve: Exploring instruction tuning large language models for systematic review automation. *CoRR*, abs/2308.06610.

Devendra Singh Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3781–3797. Association for Computational Linguistics.

Elias Sandner, Bing Hu, Alice Simiceanu, Luca Fontana, Igor Jakovljevic, Andre Henriques, Andreas Wagner, and Christian Gütl. 2024. Screening automation for systematic reviews: A 5-tier prompting approach meeting cochrane's sensitivity requirement. *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pages 150–159.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.

Harrisen Scells, Guido Zuccon, and Bevan Koopman. 2020a. You can teach an old dog new tricks: Rank fusion applied to coordination level matching for ranking in systematic reviews. In *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part I*, volume 12035 of *Lecture Notes in Computer Science*, pages 399–414. Springer.

Harrisen Scells, Guido Zuccon, Bevan Koopman, and Justin Clark. 2020b. Automatic boolean query formulation for systematic review literature search. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 1071–1081. ACM / IW3C2.

Harrisen Scells, Guido Zuccon, Bevan Koopman, and Justin Clark. 2020c. A computational approach for objectively derived systematic review search strategies. In *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part I*, volume 12035 of *Lecture Notes in Computer Science*, pages 385–398. Springer.

Benjamin Schneider. 2023. Publications output: Us trends and international comparisons. science & engineering indicators 2024. nsb-2023-33. *National Science Foundation*.

Masood Sujau, Masako Wada, Emilie Vallee, Natalie Hillis, and Teo Susnjak. 2025. Accelerating disease model parameter extraction: An llm-based ranking approach to select initial studies for literature review automation. *Preprints*.

Weiwei Sun, Zheng Chen, Xinyu Ma, Lingyong Yan, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Instruction distillation makes large language models efficient zero-shot rankers. *CoRR*, abs/2311.01555.

Eugene Syriani, Istvan David, and Gauransh Kumar. 2023. Assessing the ability of chatgpt to screen articles for systematic reviews. *CoRR*, abs/2307.06464.

Raphael Tang, Xinyu Crystina Zhang, Xueguang Ma, Jimmy Lin, and Ferhan Ture. 2024. Found in the middle: Permutation self-consistency improves listwise ranking in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 2327–2340. Association for Computational Linguistics.

Viet thi Tran, Gerald Gartlehner, Sally Yaacoub, Isabelle Boutron, Lukas Schwingshackl, Julia Stadelmaier, Isolde Sommer, Farzaneh Aboulayeh, Sivem Afach, Joerg Meerpohl, Philippe Ravaud, Paris Est, and PhD Viet-Thi Tran MD. 2023. Sensitivity, specificity and avoidable workload of using a large language models for title and abstract screening in systematic reviews and meta-analyses. In *medRxiv*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Rens van de Schoot, Jonathan de Bruin, Raoul Schram, Parisa Zahedi, Jan de Boer, Felix Weijdema, Bianca Kramer, Martijn Huijts, Maarten Hoogerwerf, Gerbrich Ferdinands, Albert Harkema, Joukje Willemsen, Yongchao Ma, Qixiang Fang, Sybren Hindriks, Lars Tummers, and Daniel L. Oberski. 2021. An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, 3(2):125–133. Publisher: Nature Publishing Group.

Raymon van Dinter, Bedir Tekinerdogan, and Cagatay Catal. 2021. Automation of systematic literature reviews: A systematic literature review. *Information and Software Technology*, 136:106589.

Byron C. Wallace, Kevin Small, Carla E. Brodley, Joseph Lau, and Thomas A. Trikalinos. 2012. Deploying an interactive machine learning system in an evidence-based practice center: abstrackr. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, IHI '12, page 819–824, New York, NY, USA. Association for Computing Machinery.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.

Shuai Wang, Harrisen Scells, Bevan Koopman, Martin Potthast, and Guido Zuccon. 2023a. Generating natural language queries for more effective systematic review screening prioritisation. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP '23, page 73–83, New York, NY, USA. Association for Computing Machinery.

Shuai Wang, Harrisen Scells, Bevan Koopman, and Guido Zuccon. 2023b. Can chatgpt write a good boolean query for systematic review literature search? In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 1426–1436. ACM.

Shuai Wang, Harrisen Scells, Bevan Koopman, and Guido Zuccon. 2023c. Neural rankers for effective screening prioritisation in medical systematic review literature search. In *Proceedings of the 26th Australasian Document Computing Symposium*, ADCS '22, New York, NY, USA. Association for Computing Machinery.

Shuai Wang, Harrisen Scells, Shengyao Zhuang, Martin Potthast, Bevan Koopman, and Guido Zuccon. 2024b. Zero-shot generative large language models for systematic review screening automation. In *Advances in Information Retrieval*, pages 403–420, Cham. Springer Nature Switzerland.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023d. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Dingjun Wu, Jing Zhang, and Xinmei Huang. 2023. Chain of thought prompting elicits knowledge augmentation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6519–6534, Toronto, Canada. Association for Computational Linguistics.

Zhe Yu, Nicholas A. Kraft, and Tim Menzies. 2018. Finding better active learners for faster literature reviews. *Empir. Softw. Eng.*, 23(6):3161–3186.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *CoRR*, abs/2308.07107.

Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang, and Michael Bendersky. 2024. Beyond yes and no: Improving zero-shot LLM rankers via scoring fine-grained relevance labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 358–370, Mexico City, Mexico. Association for Computational Linguistics.

A Appendix

A.1 Initial Interactive Tuning of LGAR

In Table 5, the results of preliminary experiments are shown. These experiments were performed on the dataset provided by Guo et al. (2023), which consists of 11 medical SLRs with a total of about 50000 papers. We excluded the SLR aiha from our evaluation because the provided dataset did not contain any relevant papers. By not using our test sets, i.e., the SYNERGY and TAR2019 dataset, for the initial optimization of our approach, we aimed to make our results more realistic since we did not optimize our approach on these test sets. We also refrained from using the train set of the CLEF TAR 2019 dataset for the initial tuning, since optimizing on this dataset, which is quite similar to the corresponding test set, could give our approach an unfair advantage. Therefore, we decided to use the dataset of Guo et al. (2023), which focuses on different medical topics. Another reason for using this dataset for the initial tuning was that it is rarely used by other publications, so we wouldn't have any baselines to compare ourselves to on the dataset anyway.

The results of using different scales on this dataset (see Table 5) suggest that increasing the scale size significantly improves performance up to scale 0-14. The scale 0-19 performs only slightly worse, while further increasing the scale size does not improve performance. Since 0-19 still performs quite well, e.g., in terms of MAP, we decided to use this scale size for our main evaluation, thus giving the LLM as many degrees of freedom as possible without compromising performance.

Similarly, in preliminary experiments when using the dataset presented in Guo et al. (2023), we also investigate the performance of different dense rankers that are used in the second stage of LGAR (see Table 6). While monoT5 with title and research questions as query achieves the best ranking performance, using monoT5 with title-only as query significantly improves performance after screening about 20% of the papers, resulting in a better performance on the metrics regarding potential workload savings. Therefore, we decided to use this configuration as our dense ranker for our main evaluation. The results of these subsequent experiments confirmed our choice to use monoT5 with title-only as the query (see Table 10).

A.2 Detailed Evaluation of LGAR

In Table 7, the more detailed results of the main findings presented in this work are shown. Here, we report separately the results of all four splits with different medical topics of the TAR2019 dataset. In the main evaluation, the averaged results of all SLRs of TAR2019 are reported in order to keep the table clear. It is worth noting that the splits shown in this table are not equal in size, e.g., the Intervention split includes 20 SLRs, while the Prognosis split includes only one SLR.

A.3 Detailed Results of Ablations

Table 8 shows detailed results of using different LLMs as initial ranker of LGAR. Llama3.3-70B performs best on most metrics for the TAR2019 dataset, which is also the trend when evaluated on the SYNERGY dataset.

Table 9 shows the results of using different scales on the SYNERGY and TAR2019 datasets. The results confirm our choice to use the 0-19 scale, which was identified as one of the best performing scales in preliminary experiments (see Table 5). Figure 7 depicts the distribution of relevance scores on the relevance scale averaged over all SLRs. Most papers receive a score of 0, which is expected, since most of the papers retrieved as candidates from scientific databases are not relevant to the respective SLR.
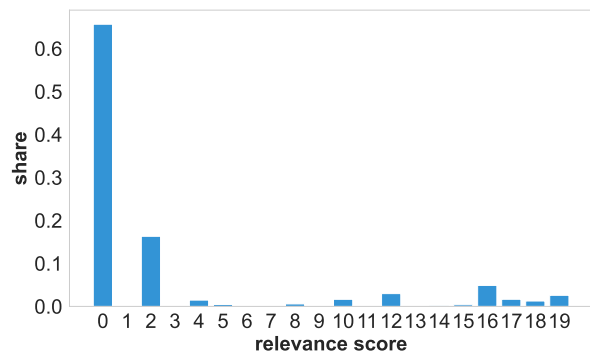
Figure 7: Distribution of relevance scores averaged for all SLRs of SYNERGY and TAR2019. LLM: Llama3.3-70B. Scale: 0-19. Re-ranker: monoT5 with title-only as query.

Table 10 shows detailed results of the comparison of dense ranking models briefly shown in Figure 5. The results confirm our choice of using monoT5 as a dense ranker with title-only as the query identified in our preliminary experiments (see Table 6). On both the SYNERGY and

Dataset	Scale	MAP	TNR@95%	R@1%	R@5%	R@10%	R@20%	R@50%	WSS @95%	@100%
Guo	0-1	44.7	51.4	26.7	59.0	69.6	79.6	93.0	46.1	24.9
	0-2	47.0	71.9	26.2	60.4	74.3	89.9	97.2	65.1	33.4
	0-4	52.2	74.8	27.7	64.2	78.5	**92.1**	97.9	68.9	34.4
	0-9	51.6	75.5	28.7	62.6	**81.4**	91.7	98.0	72.2	35.4
	0-14	**54.8**	**80.9**	29.5	**65.2**	78.1	91.9	**99.0**	**73.0**	39.1
	0-19	54.7	75.1	**30.4**	63.4	80.0	91.8	98.4	69.0	37.6
	0-24	53.8	74.8	29.0	64.0	80.4	91.8	97.9	68.0	**39.8**
	0-29	51.8	76.9	28.3	61.0	79.5	91.2	98.3	70.9	39.5

Table 5: Results of different scales on dataset of Guo et al. (2023). Re-ranker: monoT5 with title-only as query. LLM: Llama3.3-70B.

Dataset	Model	MAP	TNR@95%	R@1%	R@5%	R@10%	R@20%	R@50%	WSS @95%	@100%
Guo	BM25 (T)	9.5	32.5	2.1	19.5	30.0	47.2	76.4	27.2	16.4
	BM25 (T+R)	15.5	32.2	9.8	23.8	35.7	51.5	80.6	27.0	17.4
	ColBERT (T)	9.9	38.0	3.6	12.4	25.8	46.8	86.5	32.9	17.0
	ColBERT (T+R)	9.5	32.8	2.6	12.9	23.1	42.1	81.7	27.3	13.6
	monoBERT (T)	14.5	41.1	7.0	28.6	42.3	61.8	86.3	37.4	20.6
	monoBERT (T+R)	18.1	41.7	7.9	34.1	46.2	**63.5**	88.4	39.6	20.7
	monoT5_3B (T)	15.7	**45.3**	8.8	34.5	47.3	63.2	**89.8**	**40.8**	**21.3**
	monoT5_3B (T+R)	**19.1**	37.5	**12.3**	**37.2**	**48.6**	61.8	87.0	32.8	19.7

Table 6: Results of dense rankers on dataset of Guo et al. (2023). T = title of the SLR, R = research questions of the SLR used as query.

TAR2019 datasets, this configuration performs best on most metrics.

A.4 2-shot Experiments

Since previous research on LLMs has shown that adding examples to the prompts can be beneficial to their performance (Brown et al., 2020), we test the few-shot setting for the sake of completeness: admittedly, each paper can be considered irrelevant/relevant due to various aspects of the selection criteria, i.e., it is an extremely difficult task to generate examples that are representative for all irrelevant/relevant papers.

All abstract screening datasets provide binary relevance labels, yet we work with assigning graded relevance judgments ranging from 0 to k. We use only two examples (one positive and one negative) for few-shot prompting. In the prompt, the positive example is always followed by the negative example. In practice, a domain expert could provide these examples together with a formulation of the reasoning they applied. As an approximation, we select the few-shot examples from zero-shot runs of the respective prompting technique, i.e., our default prompt or CoT prompt, for each SLR as follows. We select the set of examples to which the zero-shot model assigns either the score 0 (for instances labeled as irrelevant in the original dataset) or k (for instances labeled as relevant in the original dataset), that exhibit the expected format, i.e., the decision of the model can be extracted using regular expressions. We then select one positive and one negative example from these sets randomly and remove them from the evaluation sets.

In **Table 11**, we compare the performance of 0-shot prompting with 2-shot, 2-shot CoT, and 2-shot CoT with self-consistency. The results suggest that adding examples does not improve the performance of LGAR, and in fact it is harmful. The reason for this could be that the randomly chosen examples are not representative of all irrelevant/relevant papers. Furthermore, since the examples are generated by the LLM, they could also contain reasoning errors. This could be improved by having domain experts generate the examples. However, the challenge of finding good representatives for an SLR remains. Therefore, we did not pursue the use of few-shot examples.

A.5 Hardware and Experiment Statistics

We ran our experiments on Nvidia A40 (40 GB) and Nvidia A100 (80 GB) GPUs, using up to 8 GPUs in parallel. The total amount of GPU hours was 7345h on A40 and about 3664h on A100.

Dataset	Model	MAP	TNR@95%	R@1%	R@5%	R@10%	R@20%	R@50%	WSS@95%	WSS@100%
SYNERGY	random reranking	4.9	6.2	1.2	5.1	10.4	20.0	50.3	3.0	3.0
	BM25 (T+R)	12.5	36.3	9.0	26.4	40.0	54.4	80.3	33.6	21.8
	monoT5 (T)	18.8	52.3	14.1	37.5	54.5	69.1	91.1	50.0	35.4
	Akinseloyin et al. (ours)†	34.0	63.6	25.8	55.0	67.8	80.5	93.1	59.5	**50.6**
	LGAR (T+R, random rerank)	38.9	59.8	33.3	58.4	70.1	79.7	91.4	58.6	40.9
	LGAR (T, monoT5)†	36.8	63.8	32.7	57.7	71.0	81.2	94.3	63.6	46.5
	LGAR (T+R, monoT5)	**40.7**	**67.0**	34.1	**59.3**	**72.0**	81.9	**94.4**	**65.2**	49.3
	+ CoT	38.8	62.9	33.4	**59.3**	70.3	81.4	92.8	62.5	45.8
	+ CoT (n=3)	**40.7**	66.4	**34.7**	58.6	70.7	**82.6**	93.5	64.5	50.4
Interv.	random reranking	6.8	10.2	1.0	5.5	10.5	20.0	52.6	6.6	7.6
	BM25 (T+R)	24.9	44.6	13.6	35.9	48.1	62.3	85.2	41.0	31.7
	Wang et al. (2023c) (0s)	16.0	-	5.4	21.0	32.8	50.4	-	36.2	33.3
	monoT5 (T)	36.0	63.7	21.7	47.5	62.6	76.8	94.0	59.0	53.8
	Akinseloyin et al. (2024)	45.0	-	-	52.6	69.7	80.8	96.2	59.2	51.9
	Akinseloyin et al. (theirs)	47.1	67.1	25.3	57.4	71.1	82.8	96.3	62.7	54.2
	Akinseloyin et al. (ours)†	50.4	68.4	**30.2**	60.6	73.2	83.8	96.9	65.2	49.8
	LGAR (T+R, random rerank)	**57.0**	61.9	29.1	**65.8**	78.0	87.6	94.2	60.4	45.4
	LGAR (T, monoT5)†	55.0	**75.0**	27.8	62.2	76.9	88.9	96.0	68.6	**63.2**
	LGAR (T+R, monoT5)	56.8	**75.0**	28.5	65.2	**78.6**	88.4	95.6	**70.7**	61.9
	+ CoT	54.3	74.2	25.6	63.0	76.6	89.0	95.4	69.1	62.2
	+ CoT (n=3)	54.5	73.8	27.0	64.8	77.6	**89.2**	94.9	68.5	**63.2**
	Wang et al. (2023c) (ft)	45.6	-	21.6	58.0	73.7	84.2	-	64.6	57.9
	BioBERT_ft (T)	45.9	70.8	24.5	55.5	73.3	84.2	**97.0**	66.8	53.6
DTA	random reranking	7.8	5.8	1.4	5.9	10.6	20.4	49.4	3.6	2.8
	BM25 (T+R)	16.7	50.5	13.9	37.6	50.1	65.0	86.0	46.4	33.1
	Wang et al. (2023c) (0s)	9.2	-	2.4	13.2	23.8	39.1	-	25.8	21.0
	monoT5 (T)	20.4	64.0	11.1	45.0	60.5	74.6	93.0	58.4	47.1
	Akinseloyin et al. (2024)	31.5	-	-	43.8	59.3	76.6	94.1	55.6	50.6
	Akinseloyin et al. (theirs)	37.3	70.6	25.0	52.5	67.0	84.4	94.1	64.5	60.3
	Akinseloyin et al. (ours)†	35.9	77.5	23.4	56.4	74.0	85.3	94.9	70.0	**62.6**
	LGAR (T+R, random rerank)	38.0	64.5	30.6	52.8	66.6	78.0	90.3	57.7	34.1
	LGAR (T, monoT5)†	34.5	78.1	25.8	57.6	73.1	84.2	95.2	71.3	57.0
	LGAR (T+R, monoT5)	**38.1**	76.8	30.8	59.9	72.9	85.0	94.6	69.5	57.3
	+ CoT	36.5	**79.4**	20.4	57.0	**75.4**	**87.0**	95.9	**71.8**	59.4
	+ CoT (n=3)	37.6	77.5	24.9	58.9	73.4	85.5	95.1	69.9	61.0
	Wang et al. (2023c) (ft)	31.8	-	26.0	50.0	67.1	81.7	-	68.6	58.4
	BioBERT_ft (T)	35.4	76.0	25.3	54.8	69.7	82.5	**96.7**	69.5	60.4
Qualitative	random reranking	1.6	24.6	0.9	3.4	11.0	21.7	58.0	20.0	21.6
	BM25 (T+R)	6.9	38.3	3.2	16.2	26.1	36.0	46.4	33.2	18.3
	monoT5 (T)	15.0	64.2	9.9	28.4	36.0	41.0	98.6	59.4	51.1
	Akinseloyin et al. (2024)	15.9	-	-	50.5	60.0	67.3	97.8	57.6	50.7
	Akinseloyin et al. (theirs)	17.8	75.8	11.3	28.4	85.1	91.9	99.6	72.0	50.3
	Akinseloyin et al. (ours)†	22.4	66.6	12.2	31.5	65.5	71.0	99.6	61.1	50.4
	LGAR (T+R, random rerank)	**29.9**	86.7	**39.2**	62.3	74.5	**98.4**	99.3	**81.6**	54.0
	LGAR (T, monoT5)†	27.3	81.0	13.5	61.5	70.5	73.6	**100.0**	75.4	71.8
	LGAR (T+R, monoT5)	29.1	86.6	38.5	62.4	69.1	98.2	**100.0**	80.6	**80.1**
	+ CoT	26.1	81.3	13.1	63.3	90.5	96.0	**100.0**	78.5	75.7
	+ CoT (n=3)	28.2	82.0	13.1	**63.7**	**91.4**	97.3	**100.0**	81.0	73.5
Prognosis	random reranking	6.3	6.3	1.4	6.0	11.0	20.0	50.8	1.8	0.7
	BM25 (T+R)	9.2	15.3	0.5	7.8	16.2	38.0	73.4	10.2	5.1
	monoT5 (T)	18.5	43.9	5.7	18.2	31.2	59.9	92.7	36.7	16.6
	Akinseloyin et al. (2024)	43.0	-	-	40.0	65.3	80.0	98.4	54.3	28.9
	Akinseloyin et al. (theirs)	49.4	68.3	13.0	43.8	66.2	82.8	97.9	60.7	43.8
	Akinseloyin et al. (ours)†	57.5	71.8	14.1	54.2	74.5	85.4	99.0	63.7	**44.6**
	LGAR (T+R, random rerank)	67.1	81.1	**15.1**	62.1	84.4	93.3	**99.8**	72.8	42.3
	LGAR (T, monoT5)†	67.7	81.2	**15.1**	62.0	83.8	93.8	99.5	74.1	29.8
	LGAR (T+R, monoT5)	**67.9**	83.6	**15.1**	61.5	84.4	93.8	99.5	74.3	32.8
	+ CoT	65.2	82.4	14.6	61.5	**84.9**	93.8	99.5	73.4	34.5
	+ CoT (n=3)	67.1	**84.9**	**15.1**	60.4	**84.9**	95.3	99.5	**75.5**	32.6

Table 7: Main results (detailed). T = title of the SLR, R = research questions of the SLR. Akinseloyin et al. (2024): ours = uses our annotated criteria, theirs = uses their criteria. In all combinations, monoT5 uses the title as query and the LLM the scale 0-19. Wang et al. (2023c): 0s=zero-shot, ft=fine-tuned. † indicates that the approaches get the same information on SLR and paper.

Dataset	LLM	MAP	TNR@95%	R@1%	R@5%	R@10%	R@20%	R@50%	WSS @95%	@100%
SYNERGY	Llama3.1-8B	31.9	61.3	22.9	53.4	67.0	80.9	93.3	60.5	45.4
	Llama3.3-70B	40.7	67.0	**34.1**	59.3	72.0	81.9	**94.4**	**65.2**	**49.3**
	Mistral-123B	38.8	65.8	33.1	**61.6**	**73.3**	82.0	93.6	63.4	47.2
	Qwen2.5-32B	38.7	63.2	33.4	57.0	68.1	78.6	93.7	61.4	45.8
	Qwen2.5-72B	**41.3**	**68.5**	32.7	60.0	72.0	**83.0**	94.1	64.8	47.7
TAR2019	Llama3.1-8B	42.3	69.6	24.7	57.5	71.5	86.1	95.2	65.3	55.1
	Llama3.3-70B	**50.6**	**76.5**	29.3	**63.6**	**76.7**	88.3	95.8	**71.2**	**60.9**
	Mistral-123B	48.4	70.8	28.5	62.7	75.3	86.4	95.9	68.1	55.1
	Qwen2.5-32B	47.6	73.4	28.6	60.4	74.0	84.6	**96.2**	68.6	58.9
	Qwen2.5-72B	50.2	73.0	**29.8**	63.1	76.1	**89.0**	95.8	68.9	56.0

Table 8: Comparison of LLMs. Scale: 0-19. Re-ranker: monoT5 with title-only as query. 0-shot prompting.

Dataset	Scale	MAP	TNR@95%	R@1%	R@5%	R@10%	R@20%	R@50%	WSS @95%	@100%
SYNERGY	0-1	33.7	59.4	29.2	55.1	67.6	79.3	93.2	58.2	41.2
	0-2	35.0	62.9	30.8	56.0	70.1	82.1	93.6	63.6	45.3
	0-4	36.4	63.2	30.6	56.6	70.0	80.6	93.7	63.8	44.9
	0-9	39.4	67.4	32.3	57.6	69.2	82.2	**94.6**	64.8	50.1
	0-14	39.9	**69.2**	33.3	57.4	69.7	**82.3**	94.5	65.5	**50.8**
	0-19	**40.7**	67.0	**34.1**	**59.3**	**72.0**	81.9	94.4	65.2	49.3
	0-24	39.6	66.0	32.7	57.8	70.9	81.5	94.4	64.5	50.0
	0-29	39.5	66.4	33.3	59.1	71.1	82.2	94.2	**66.2**	49.0
TAR2019	0-1	43.2	67.1	25.2	57.3	73.4	82.0	95.1	63.2	52.9
	0-2	45.5	75.4	25.0	60.7	75.2	86.8	95.6	69.7	59.2
	0-4	47.1	75.6	27.0	61.8	76.5	87.5	95.6	70.1	60.8
	0-9	48.1	**77.6**	27.3	61.0	76.2	88.6	**96.4**	70.7	**63.3**
	0-14	48.0	77.5	28.9	61.6	75.6	88.0	96.2	70.9	63.2
	0-19	**50.6**	76.5	29.3	**63.6**	**76.7**	88.3	95.8	**71.2**	60.9
	0-24	48.3	76.8	**29.6**	62.0	76.5	**88.8**	96.3	69.9	62.4
	0-29	48.9	77.1	27.1	62.3	76.3	88.6	**96.4**	69.6	62.8

Table 9: Comparison of different Likert scale ranges in LLM. LLM: Llama3.3-70B. Scale: 0-19. Re-ranker: monoT5 with title-only as query.

A.6 Use of AI Assistants

Copilot and ChatGPT were used only as coding assistants. This was especially useful for debugging existing code and for understanding third-party code. We have not used AI assistants for writing this paper.

Dataset	Model	MAP	TNR@95%	R@1%	R@5%	R@10%	R@20%	R@50%	WSS @95%	WSS @100%
SYNERGY	BM25 (T)	10.3	36.8	7.3	24.2	35.4	49.3	77.8	32.0	22.9
	BM25 (T+R)	12.5	36.3	9.0	26.4	40.0	54.4	80.3	33.6	21.8
	ColBERT (T)	10.2	32.4	7.4	22.9	34.2	50.7	80.3	32.9	20.0
	ColBERT (T+R)	8.9	29.7	5.1	18.2	31.7	47.5	79.4	29.6	17.0
	monoBERT (T)	16.6	44.6	13.8	33.6	47.8	64.1	87.2	42.6	30.3
	monoBERT (T+R)	17.8	42.6	**14.6**	32.6	48.4	64.3	87.1	40.6	28.6
	monoT5 (T)	**18.8**	**52.3**	14.1	**37.5**	**54.5**	69.1	**91.1**	**50.0**	**35.4**
	monoT5 (T+R)	17.0	35.0	14.4	30.3	42.3	58.6	83.0	32.5	22.6
TAR2019	BM25 (T)	18.1	38.9	9.5	30.2	40.5	57.4	81.4	36.1	28.4
	BM25 (T+R)	21.1	44.8	12.6	34.2	46.2	60.5	82.5	40.9	30.4
	ColBERT (T)	15.8	45.1	6.1	21.2	36.1	54.5	84.8	40.8	33.5
	ColBERT (T+R)	14.5	43.7	4.7	24.2	34.2	52.4	83.7	39.4	33.7
	monoBERT (T)	23.8	50.8	13.2	38.5	51.3	64.8	89.4	47.4	36.2
	monoBERT (T+R)	27.4	54.4	16.9	40.9	54.3	71.2	90.8	49.9	41.0
	monoT5 (T)	**30.1**	**63.2**	17.7	**44.6**	**59.3**	73.4	**94.0**	**58.2**	**50.7**
	monoT5 (T+R)	29.6	59.9	**19.0**	42.6	57.1	72.8	93.2	56.1	45.8

Table 10: Comparison of dense ranking models. T = title of the SLR, R = research questions of the SLR used as query. All models have been trained on MS MACRO.

Dataset	Prompt	MAP	TNR@95%	R@1%	R@5%	R@10%	R@20%	R@50%	WSS @95%	WSS @100%
SYNERGY	0-shot	**40.7**	**67.0**	**34.1**	59.3	**72.0**	81.9	**94.4**	**65.2**	49.3
	2-shot	38.7	60.5	**34.1**	60.6	71.8	80.7	92.8	59.5	44.8
	2-shot CoT	37.0	65.1	32.3	59.0	71.8	**83.4**	93.6	62.8	48.2
	2-shot CoT (n=3)	39.2	66.2	32.5	58.7	71.0	82.8	94.1	64.2	**49.9**
TAR2019	0-shot	**50.6**	**76.5**	**29.3**	**63.6**	**76.7**	**88.3**	95.8	**71.2**	**60.9**
	2-shot	45.4	72.5	27.3	57.8	76.0	85.2	95.7	68.2	57.1
	2-shot CoT	43.5	72.0	25.7	57.5	69.8	86.0	**96.5**	67.0	60.5
	2-shot CoT (n=3)	44.8	74.9	26.1	57.9	74.2	85.7	95.8	68.8	59.4

Table 11: Performance of 2-shot approaches. LLM: Llama3.3-70B. Scale: 0-19. Re-ranker: monoT5 with title-only as query.