Article

Not peer-reviewed version

Accelerating Disease Model Parameter Extraction: An LLM-based Ranking Approach to Select Initial Studies For Literature Review Automation

Masood Sujau * , Masako Wada , Emilie Vallee , Natalie Hillis , Teo Susnjak

Posted Date: 21 January 2025

doi: 10.20944/preprints202501.1513.v1

Keywords: Large Language Models in Systematic Reviews; Automated AI Literature Screening; Zero-Shot Relevancy Ranking; Climate-Sensitive Zoonotic Disease Modeling; Information Retrieval in Medical Literature; Systematic Literature Review Automation; Biomedical Text Mining for Disease Tracking; AI-Assisted Disease Surveillance

 doi:10.20944/preprints202501.1513.v1

Article

Accelerating Disease Model Parameter Extraction: An LLM-Based Ranking Approach to Select Initial Studies For Literature Review Automation

Masood Sujau [1,*], Masako Wada [1], Emilie Vallee [1], Natalie Hillis [1] and Teo Susnjak [2]

[1] School of Veterinary Science, Massey University, Palmerston North, New Zealand
[2] School of Mathematical and Computational Sciences, Massey University, Auckland, New Zealand
* Correspondence: mmsujau@massey.ac.nz

Abstract: As climate change transforms our environment and human intrusion into natural ecosystems escalates, there is a growing demand for disease spread models to forecast and plan for the next zoonotic disease outbreak. Accurate parametrization of these models requires data from diverse sources, including scientific literature. Despite the abundance of scientific publications, the manual extraction of this data via systematic literature reviews remains a significant bottleneck, requiring extensive time and resources, and is susceptible to human error. To address this challenge, a novel automated parameter extraction framework, CliZod is presented. A crucial stage of the automation process is to screen scientific articles for relevance. This paper focuses on leveraging large language models (LLMs) to assist in the initial selection and ranking of primary studies, which then serve as training examples for the screening stage. By framing the selection criteria of articles as a question-answer task and utilising zero-shot chain-of-thought prompting, the proposed method achieves a saving of at least 60% work effort compared to manual screening at a recall level of 95% (NWSS@95%). This was validated across four datasets containing four distinct zoonotic diseases and a critical climate variable (rainfall). The approach additionally produces explainable AI rationales for each ranked article. The effectiveness of the approach across multiple diseases demonstrates the potential for broad application in systematic literature reviews. The substantial reduction in screening effort, along with the provision of explainable AI rationales, marks an important step toward automated parameter extraction from scientific literature.

Keywords: large language models in systematic reviews; automated ai literature screening; zero-shot relevancy ranking; climate-sensitive zoonotic disease modeling; information retrieval in medical literature; systematic literature review automation; biomedical text mining for disease tracking; AI-assisted disease surveillance

1. Introduction

Zoonotic diseases, are becoming increasingly more prevalent due to increased interactions between, humans, livestock, wildlife, disease vectors, and pathogens, exacerbated by climate change and rapid human expansion into natural habitats [1,2]. Globally zoonotic diseases have disproportionately impacted impoverished livestock workers in low- and middle-income countries, responsible for millions of human deaths every year [3]. To address this threat, there is a mounting need for systems to forecast and model the spread of diseases, thereby aiding public health planning and supporting early warning systems [1].

Constructing such models require precise parametrisation. This necessitates data from multiple sources, including clinical records, environmental datasets, grey literature and scientific publications [4,5]. Researchers conduct a systematic literature reviews (SLR), regarded the gold standard, to reliably extract this information. The key stages of the SLR process consists of; planning and developing protocols, searching for articles, screening articles for relevance, extracting information, assessing the quality of studies and finally summarising and reporting [6,7].

 doi:10.20944/preprints202501.1513.v1

Despite the abundance of scientific publications, the manual extraction of disease parameters via SLRs remain a significant bottleneck. Conducting an SLR requires a significant time investment [8]; the process is costly, [9] and the exponential growth of publications [10] contributes to researcher fatigue and increased cognitive overload [11]. Finally, these reviews are difficult to keep current, leading to unnecessary waste and potential harm in decision-making [12]. The core issue lies in the inefficiency to robustly scale the manual processes. Automation offers a cost-efficient, faster, and more reliable solution, capable of meeting the demands for scale and quality [13–15].

Since the 2000s, systems based on natural language processing (NLP) and machine learning (ML) have emerged to automate or semi-automate key steps of SLR process[15]. Current ML based systems on the market continue to use frequency-based sparse vector representation of text or static word embeddings [16] which struggle to capture context and address challenges like polysemy. Moreover, many tools lack the capacity to generalise to other domains and have primarily been evaluated on specific research questions or target particular study designs [17].

In recent years, tools powered by transformer-based large language models (LLM), such as GPT, have gained significant popularity [16,18]. These models leverage vast amounts of text data and advanced contextual understanding, to handle complex language tasks, including screening abstracts [19–23], extracting data [23–25] and synthesise information [23,26,27]. However, integrating LLMs into SLR workflows pose major challenges [16]. They can produce plausible-sounding erroneous responses known as *"hallucinations"* [28], promote negative biases seen in training data [29], lack transparency and are considered *"black-box"* systems [30]. While these issues are significant, researchers are actively exploring various options to mitigate risks [30–32] and ensure reliable and effective integration of the technology into the workflow.

Climate change and zoonotic diseases present complex challenges that require a focused and innovative solution. While LLMs and automation tools demonstrate potential in transforming SLRs, a significant gap exists for domain-specific systems capable of managing diverse datasets. Tools that can provide scalable, accurate, and transparent actionable insights are urgently required. Therefore, this paper presents an overview of the conceptual design of *CliZod* (Climate Sensitive Zoonotic Database), a novel collaborative system for the automatic identification, extraction and storage of climate sensitive zoonotic disease parameters from scientific literature using artificial intelligence. This paper specifically focuses on the experimental validation of the relevancy ranking component located within title-abstract screening module of the system. The primary objective is to assess the effectiveness of employing an LLM-based assessor with a question-answer (QA) framework to guide the ranking of primary studies by relevance.

1.1. CliZod

The *CliZod* system is structured as a scalable and transparent modular framework to automate parameter extraction from scientific publications, specifically targeting climate-sensitive zoonotic disease modelling. The architecture comprises of a multistage workflow pipeline (Figure 1) guided by the the Vienna Principles [33]. The system is being developed as a collection of open-source modular web service components with well-defined APIs and standardised input/output formats. The intention is to promote flexibility, transparency and adaptability, allowing each component to function as a 'plug-and-play' module that could be integrated into a customizable broader workflow. The intention is to enable researchers to easily adapt, extend, or reuse components across diverse use cases, making it a scalable solution for automating parameter extraction from literature.

The process commences with the user specification module, where the system gathers information regarding the research topic, selection criteria, parameters of interest, and establishes context for subsequent stages.

 doi:10.20944/preprints202501.1513.v1

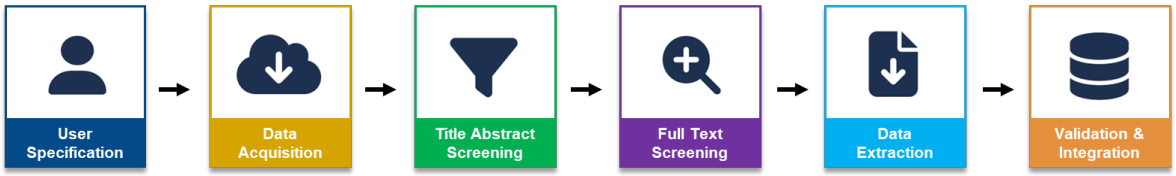

Figure 1. CliZod pipeline

In the data acquisition module, automated search queries are utilised to search and retrieve titles and abstracts from designated online journal databases, followed by de-duplication and metadata harvesting. The next module is the screening of titles and abstracts, where articles are classified according to relevance. The full text screening module attempts to automatically download the full text of relevant articles when possible and seeks to parse and refine the screening process. The parsed data that contain information from text, tables, and figures, encompassing both structured and unstructured data, is used in the data extraction module to facilitate the identification and extraction of essential parameters. In the validation and integration phase, the data undergoes consistency and quality checks, followed by a final review before integration into the database.

1.2. Relevancy Ranking

Title-abstract screening is the most time-consuming aspect of the manual SLR process [34]. Best-practice guidelines recommend that primary studies are sorted by relevance before being screened. This can improve reviewer efficiency, increase productivity, and increase motivation [35]. Presenting users with a sorted list of title-abstracts, with the most relevant studies prioritised at the top, facilitates faster and more efficient decision making, especially for highly skewed datasets, such as those in the medical domain, which contain only 3% to 6% [36] relevant documents. The output of a relevancy ranking system can be directly utilised by selecting the top-k ranked articles based on an approximation of relevant document coverage [37,38] or, alternatively, assist users in selecting *seed* or example articles for input into a further specialised binary classifier, which is the use case of interest for the CliZod project.

1.2.1. Problem Definition

The focus of this study is to investigate the utility of using an LLM to rank primary studies by relevance. Given a set of documents $\mathcal{D}$ and a set of eligibility criteria $\mathcal{C}$, the task of the relevancy ranker is to assign each document $d \in \mathcal{D}$ a relevance score $r_d \in [0, 1]$, where r_d indicates how well document d satisfies the criteria $\mathcal{C}$. A score closer to 1 indicates higher relevance, while a score closer to 0 indicates lower relevance.

The eligibility criteria constitutes a series of rules and specific requirements that a document must satisfy for inclusion in the SLR. These criteria are typically evaluated based on attributes of the SLR protocol, including review title, keywords, research question or inclusion/exclusion criteria, but could also be framed as a question-answering problem [39]. While selection criteria are a common choice for LLM based binary screening solutions [19,40–42], the QA framework - also referred to as a screening tool in manual SLR workflow and regarded best practice [35] - offers a more fine-grained, consistent and targeted approach to determining eligibility.

Using the QA framework approach, Kusa et al. describes the eligibility criteria $\mathcal{C}$ being transformed into a set of questions $\mathcal{Q} = \{q_1, \cdots, q_{|C|}\}$, where q_k corresponds to a specific criteria C. A set of predicated answers $\hat{\mathcal{A}}^d = \{\hat{a}_k^d | meets(q_k, \hat{a}_k^d)\}$ can be obtained for each document $d \in \mathcal{D}'$, where $meets(q_k, \hat{a}_k^d)$ denotes that the document d should meet the criterion expressed by q_k. The relevancy score $\hat{r}_d$ of a document can be computed on the predicate answers $\hat{\mathcal{A}}^d$ using an aggregation method such as linear weighted sum (LWS).

 doi:10.20944/preprints202501.1513.v1

1.2.2. Background

There are multiple ways to use LLMs for ranking documents, rankers [37,43–45] and assessors are [46] popular choices. Rankers determine the order of items based on their perceived value while assessors provide an evaluation or judgement of the quality or suitability of a single item.

Zero-shot rankers [47], have no need for prior training or examples and can be categorised into point-wise [45], which scores one query and document at a time, ranking the documents based on a score; pair-wise, where the model assesses a pair of documents against a query or list-wise [44], which involves presenting both query and the complete list of items in the prompt. Pair-wise and list-wise methods do not scale for long lists as is the case with SLRs where the initial search could yield a substantial list of results. Wang et al. examined the performance of zero-shot and fine-tuned point-wise neural rankers in the context of SLR document ranking and discovered that zero-shot neural rankers performed worse than traditional methods like BM25 in the absence of fine-tuning. Moreover, research utilising point-wise rankers with generative LLMs indicated that obtaining ranking scores from log-likelihood values resulted in superior ranking performance relative to employing LLM generated labels[45]. The potential for additional output from an LLM, such as reasoning text, is constrained in this approach, as each output token, the smallest text unit a model can process, must be individually parsed, ultimately increasing complexity and limiting the interpretability of the ranker. Such constraints and the reliance on fine-tuning for improved performance highlights the need for approaches that balance scalability and transparency in zero-shot settings.

Zero-shot LLM assessor [46], perform relevancy judgement on a query-document pair when provided with a set of relevancy labels. When ground truth labels are available, these assessors are frequently employed to generate annotated datasets [48]. In addition to issuing judgement, LLM assessors can offer natural language explanations to support their conclusions, they provide scalability, consistency and the potential to complement human assessors in judgement tasks [46].

Several studies investigating the application of LLMs in SLR workflow have highlighted the importance for human oversight [19,26,40,42] or the use of LLMs as assistant reviewers [19,49]. Adopting an LLM as an assessor can effectively realise these strategies. While there is no ground truth data available at the start of a review, the assessor's approach can still assign labels based on predefined eligibility criteria, establishing an initial framework to guide the ranking process and provide transparency. This study examines the effectiveness of an LLM as an assessor only, utilising its capacity to provide answer labels to questions.

Both ranker and assessor approaches can be enhanced using various strategies ranging from resource intensive options, such as pre-training and fine-tuning of models [37,50] to prompt engineering [51]. Domain-specific pre-training and fine-tuning have demonstrated substantial performance gains [27,37,52]. However, techniques such as few-shot and chain-of-thought (CoT) prompting, leveraging a *persona* and adopting fine-grained labels offer a more economical, less complex initial approach prior to pursuing advanced optimisations. While zero-shot prompting does not utilise task-specific examples, few-shot prompting incorporates both positive and negative examples, demonstrating promising results in title-abstract screening automation [21,50,53,54]. Considering that ranking occurs at the very start of the screening process, users typically lack domain-specific examples at this early stage, making few-shot prompts less practical for an initial ranking task. Chain-of-thought (CoT) prompting, instructs the model to adopt a grounded, *"step-by-step"* approach to task resolution, reducing the likelihood of *hallucinations*, improving performance and accuracy [31]. To increase transparency, this study captures the CoT reasoning so users can gain insights into the model reasoning. Additionally, integrating a *persona* can enhance a model's capacity to deliver a more tailored and consistent responses across multiple interactions, adapting to a wide range of scenarios [40,55]. Finally, recent experiments with zero-shot LLM rankers indicate that fine-grained relevancy labels help guide the model to differentiate documents more effectively [45]. This study specifically examines the impact of fine-grained labels on the ranking performance using an LLM as an assessor within a zero-shot setting for prioritising primary studies.

 doi:10.20944/preprints202501.1513.v1

1.3. Research Questions

This study aims to evaluate the effectiveness of using an LLM in the role of an assessor to assist in the ranking and initial selection of primary studies of climate-sensitive zoonotic diseases. To guide the investigation the following research questions are posed:

RQ1 How does an LLM based assessor utilising a QA framework compared to baseline models utilising review title and selection criteria?

RQ2 Does the label granularity effect the ranking performance of an LLM based assessor utilising a QA framework for climate sensitive zoonotic disease?

RQ3 Does the ranking performance of an LLM based assessor generalise across climate sensitive zoonotic disease datasets with varying relevance rate?

RQ4 Does CoT rationale provided by an LLM assist a human reviewer's ability to detect misclassifications in SLR?

1.4. Contribution

This papers contributes to enhancing information retrieval techniques using LLMs, advancing the automating of SLRs and investigates generalisable and transparent methods for ranking and reviewing literature on climate sensitive zoonotic diseases. The specific contributions can be summarised as:

- Proposing a system to extract, store and share disease model parameters for climate sensitive zoonotic diseases modelling.
- Introducing and validating the use of an LLM-based assessor for ranking primary studies by utilising a QA framework.
- Evaluate the impact of label granularity on ranking performance to improve early stage recall quality
- Demonstrate that the proposed LLM-based assessor can manage highly skewed datasets and generalise across diverse climate sensitive zoonotic disease literature.
- Validate the utility of CoT generated reasoning text for human reviewers to enhance transparency and trust.

2. Methodology

A series of experiments were conducted to investigate the research questions outlined in section 1.3. This sections provides details of the dataset, models, the evaluation metrics and the experimental design.

2.1. Dataset

The dataset used in this study is derived from a broader ongoing study examining 11 diseases (see appendix) along side 3 climatic variables, rainfall, humidity and temperature. Each combination of disease and climate variable constitutes a distinct SLR. An SLR protocol was established, outlining the selection criteria for each combination of disease and climate variable. A search was conducted in the PubMed and Scopus online journal repositories and the results were imported into a reference manager, along with titles and abstracts followed by de-duplication. A team of 5 researchers (early and mid-career researchers, postgraduate student and undergraduate student) were trained to screen abstracts using the criteria outlined in Table 1. Abstracts that met the inclusion criteria were assigned a score of 1, otherwise a score of 0. When a reviewer was unsure, abstracts were discussed with one of the mid-career researchers whom had experience in reviewing infectious diseases climate sensitivity. Abstracts that were unclear, or did not precisely meet the study selection criteria were scored 1 to be further examined in the full-text inspection stage.

An initial dataset of 2,905 title-abstracts was established, consisting of the rainfall variable and 4 zoonotic diseases, Crimean-Congo haemorrhagic fever (CCHF), Ebola virus, Leptospirosis (Lepto) and

 doi:10.20944/preprints202501.1513.v1

Rift Valley fever virus (RVF). This constituted the entirety of the screened data at that point. Ebola articles exhibited a particularly pronounced skew, with only 1.5% (13/915) deemed relevant while the remaining diseases ranged from 10.8% to 12.6% (Table 2). The titles-abstracts were exported to a CSV file along with the manually assigned scores and labels were assigned per record to indicate the target disease and climate variable (only rainfall).

Table 1. Selection criteria used in the CliZod systematic review for the abstract screening process.

Inclusion criteria	Exclusion criteria
• Primary research or meta-analysis • Assesses the relationship between the selected climate variable and either: – Disease incidence or prevalence – Pathogen survival – Transmission – Virulence – Demonstrated vector or maintenance host survival, development or distribution	• Reviews, opinions, books, editorials • Laboratory-focused studies e.g. studies to develop an appropriate culture method

Table 2. Overview of dataset composition by disease, climate variable, and relevance proportions..

Disease	Climate Variable	Relevant	Irrelevant	Total	% Relevant
CCHF	Rainfall	57	398	454	12.6%
Ebola	Rainfall	14	902	915	1.5%
Lepto	Rainfall	108	890	999	10.8%
RVF	Rainfall	63	474	537	11.7%

2.2. QA Framework

Following the abstract screening best practice guideline [35], a set of eligibility questions was developed to evaluate the relevance of title-abstracts based on the selection criteria. Distinct sets of eligibility questions were formulated for each disease and the associated climate variable (rainfall), ensuring their relevance to the SLR topic, as detailed in Table 3. It should be noted that reviewers did not use these questions during their screening, as they were developed retrospectively for the evaluation process. The guideline stipulate that the questions should be clear and concise, and must be 1) objective 2) "single-barrelled" or focused on a specific aspect of the citation 3) use a consistent sentence structure, and d) ensure responses are limited to yes, no and unsure only. Furthermore, the questions should be organised hierarchically, starting with the easiest and progressing to difficult. In this study, the question text and the number of questions assigned to each disease remained constant across experiments, while the answer labels were varied. The specific variations are detailed in section 2.4.

 doi:10.20944/preprints202501.1513.v1

Table 3. Disease-specific topics and question-based eligibility criteria for climate-sensitive zoonotic disease studies..

Disease	Topic and Eligibility Questions
CCHF	**Topic**: Impact of Climate Change on CCHF: A Focus on Rainfall **Eligibility Questions**: Q1. Does the study report on primary research or a meta-analysis rather than a review, opinion, or book? Q2. Does the study measure the incidence or prevalence or virulence or survival or transmission of Crimean-Congo haemorrhagic fever or a relevant vector (such as ticks) without specifically measuring the incidence of the pathogens? Q3. Does the research examine environmental factors such as rainfall, seasonality (e.g., wet vs. dry season) or regional comparisons impacting disease prevalence or vector distribution? Q4. Is the study focused on field-based or epidemiological research rather than laboratory method validation?
Ebola	**Topic**: Impact of Climate Change on Ebola: A Focus on Rainfall **Eligibility Questions**: Q1. Does the study report on primary research or a meta-analysis rather than a review, opinion, or book? Q2. Does the study measure the incidence or prevalence or virulence or survival or transmission of Ebola or Marburg, a relevant vector, or reservoir hosts abundance or distribution (such as bats or primates) without specifically measuring the incidence of the pathogens? Q3. Does the research examine environmental factors such as rainfall, seasonality (e.g., wet vs. dry season) or regional comparisons impacting disease prevalence or vector distribution? Q4. Is the study focused on field-based or epidemiological research rather than laboratory method validation?
Lepto	**Topic**: Impact of Climate Change on Leptospirosis: A Focus on Rainfall **Eligibility Questions**: Q1. Does the study report on primary research or a meta-analysis rather than a review, opinion, or book? Q2. Does the study measure the incidence or prevalence or virulence or survival or transmission of Leptospirosis, a relevant arthropod vector, or reservoir hosts (such as rodents) without specifically measuring the incidence of the pathogens? Q3. Does the research examine environmental factors such as rainfall, seasonality (e.g., wet vs. dry season) or regional comparisons impacting disease prevalence or vector distribution? Q4. Is the study focused on field-based or epidemiological research rather than laboratory method validation?
RVF	**Topic**: Impact of Climate Change on Rift Valley Fever Virus: A Focus on Rainfall **Eligibility Questions**: Q1. Does the study report on primary research or a meta-analysis rather than a review, opinion, or book? Q2. Does the study measure the incidence or prevalence or virulence or survival or transmission of Rift Valley fever or other vector-borne diseases (such as malaria) that share similar vectors (e.g., mosquitoes) without specifically measuring the incidence of the pathogen? Q3. Does the research examine environmental factors such as rainfall, seasonality (e.g., wet vs. dry season) or regional comparisons impacting disease prevalence or vector distribution? Q4. Is the study focused on field-based or epidemiological research rather than laboratory method validation?

2.3. Prompts

The design and choice of prompting has implications on the model response [49,56]. The prompts deployed in this study were inspired by those from previous research [19,21,31,40,49,57] and received

 doi:10.20944/preprints202501.1513.v1

iterative refinement. The prompt performance was evaluate using *Recall@k* and Mean Average Precision *MAP* metrics detailed in section 2.7.

A persona description stating, *"You are a world leading expert veterinary epidemiologist screening abstracts of scientific papers for the systematic literature review of '$topic'"* was established prior to the main instruction prompt. All experimental runs used the same persona, with the *$topic* placeholder dynamically replaced by the topics listed in Table 3.

Two main instruction prompts were designed for this study. The first prompt, TSC prompt, serves as a baseline, utilising the review title and selection criteria. This prompt is an adaptation of the *Zero-shot Framework CoT* prompt [21], directing the model to analyse the title and apply inclusion and exclusion criteria to classify a given abstract as "Definitely Include," "Probably Include," "Probably Exclude," "Definitely Exclude," or "Unsure" (see listing 1). The placeholders *$title*, *$inclusion*, *$exclusion*, and *$abstract* were populated dynamically according to the disease context of the experiment.

```
Task: You are a researcher rigorously screening titles and abstracts of scientific papers for inclusion or exclusion in a review paper titled "$topic". The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:

Inclusion criteria:
$inclusion

Exclusion criteria:
$exclusion

Abstract:
"$abstract"

We now assess whether the paper should be included from the systematic review by evaluating it against each and every predefined inclusion and exclusion criterion. First, we will reflect on how we will decide whether a paper should be included or excluded. Then, we will think step by step for each criteria, giving reasons for why they are met or not met.

We will conclude by outputting: "Definitely Include", "Probably Include", "Probably Exclude", "Definitely Exclude" or "Unsure".

Required Format:
Format the output as a JSON object with the following keys.
    "reason": Step-by-step reasoning to the question.
    "answer": "Definitely Include", "Probably Include", "Probably Exclude", "Definitely Exclude" or "Unsure".
Example Format:
{
    "reason": "<YOUR REASONING FOR THE QUESTION>",
    "answer": "<YOUR FINAL ANSWER>"
}

Strict Output Requirements:
You MUST NOT output any other text before or after the JSON.
Do NOT be chatty. Output exactly what is instructed.
```

Listing 1. TSC prompt, based on the review title and selection criteria

The second prompt, QA prompt in Listing 3 (see Appendix A), utilises the QA framework discussed in Section 2.2. This prompt instructs the model to follow a structured approach to answering the eligibility questions with respect to a given abstract. The answer labels available to the model are discussed in section 2.4. Similar to the first prompt, placeholders for *$abstract* and *$question*, are dynamically populated with title-abstract and eligibility questions based on disease context.

Both prompts employ predefined answer labels and explicit reasoning, adhering to the principles of CoT prompting [31], and aligns with the *instructive* template category, as described in [54]. Given that GPT models are *autoregressive* [50] where previous tokens influence the final answer, the prompts

 doi:10.20944/preprints202501.1513.v1

have been designed to explicitly request reasoning prior to delivering a final answer label. Finally, the prompts instructs the model to format the response as a JSON object, fields capture the reasoning text, answer and in the case of the QA framework, the question number. This structure helps ensure consistent and reliable output formatting and eliminates the need for parsing natural text while facilitating easy integration with downstream components.

2.4. Answer Labels

Table 4 lists the predefined answer schemas designed to be used with the QA and TSC prompt, restricting the model to selecting only from these options. The schema, QA-3 was adapted from the best practice guidelines [35], while QA-4 and Qa-5 were inspired by [45].

Table 4. Answer labels and scoring scales for single-level QA and TSC models..

Answer Schema	Answer Labels	Answer Score
QA-2	Yes	1.0
	No	0.0
QA-3	Yes	1.00
	Unsure	0.50
	No	0.00
QA-4	Definitely Yes	0.95
	Probably Yes	0.75
	Probably No	0.25
	Definitely No	0.05
QA-5	Definitely Yes	1.00
	Probably Yes	0.75
	Unsure	0.50
	Probably No	0.25
	Definitely No	0.00
TSC-5	Definitely Include	1.00
	Probably Include	0.75
	Unsure	0.50
	Probably Exclude	0.25
	Definitely Exclude	0.00

Additionally, a variant of QA-2 schema, termed QA-2-C, a multi level answer schema was explored by instructing the model to generate a separate confidence score (Table 5). This method aims to more accurately capture the uncertainty that the original QA-2 fails to account for, yet appears in the other schemas. Each answer label was assigned a predefined answer score to quantify its contribution to the relevancy ranking calculation (described in section 2.5), in the case of QA-2-C the confidence score was used in place of the answer score. The QA prompt was adjusted to accommodate the multi level scoring as shown in Listing 4 (see Appendix A).

Table 5. Answer labels and scoring scales for multi-level QA models.

Answer Schema	Answer Labels	Confidence Labels	Confidence Score
QA-2-C	Yes	High	1.00
		Medium	0.75
		Low	0.50
	No	High	0.00
		Medium	0.25
		Low	0.50

 doi:10.20944/preprints202501.1513.v1

2.5. Relevancy ranking

A relevancy score was computed for each title-abstract record by employing a Linear Weighted Sum (LWS) of the *answer scores* derived from all questions within the QA framework. The records are subsequently sorted in descending order based on the relevancy score. In this study, it was assumed that all questions carried equal weight. Formally, let:

- $\mathcal{Q} = \{q_1, \cdots, q_{|C|}\}$ be the eligibility questions derived from the selection criteria.
- $\hat{\mathcal{A}}^d = \{\hat{a}_k^d\}$ be the set of predicate answers for each question q_k in document d, and
- w_k be a predefined weight reflecting the importance of each question q_k.

Then, the relevancy score $\hat{r}_d$ of document d is expressed as:

$$\hat{r}_d = \sum_{k=1}^{|C|} w_q \hat{a}_k^d \tag{1}$$

where for equal weighted questions $w_k = \frac{1}{k}$ for all q. This naturally simplifies to:

$$\hat{r}_d = \frac{1}{k} \sum_{k=1}^{|C|} \hat{a}_k^d \tag{2}$$

In cases where ties occur in the relevancy score, title-abstract length was used as a tiebreaker by sorting the title-abstracts in descending order of length, assuming longer articles were more relevant.

2.6. Models

2.6.1. Baseline Models

The BM25 [58] and MiniLM v2 [59] models were utilised to establish a zero-shot baseline. BM25 is a widely used ranking algorithm used in information retrieval and evaluates a document's relevance with respect to a query. The query in this instance is a concatenation of the review title with the selection criteria. BM25 scores were computed for each title-abstract and query pair following a preprocessing step that included lower casing, punctuation elimination, stop word removal and stemming. The results were ranked according to the BM25 scores, with the text length serving as a tie breaker.

MiniLM v2 is a pre-trained sentence-transformer model optimised for embedding text and trained on a massive and diverse dataset [60]. This study used the *all-MiniLM-L6-v2* model without fine-tuning. Embedding vectors were generated for both title-abstracts and queries. Cosine similarity was computed as per equation 1, for each vector pair and subsequently normalised and sorted, with the text length serving as a tie breaker.

$$cosine\_similarity(d, q) = \frac{d \cdot q}{\|d\| \|q\|} \tag{3}$$

This methodology for establishing a baseline is aligned with the approach implemented by CSMed, a meta-dataset comprising 325 systematic literature reviews from the medical and computer scientific fields [39].

2.6.2. Large Language Models

For this experiment *GPT-4o-mini-2024-07-18*, a smaller and optimised version of *GPT-4*, is utilized to achieve a balance between cost, performance and computational efficiency [61]. The model was used in it's pretrained form without any fine-tuning and accessed via the *OpenAI API*. The sampling *temperature* value was set to 0 to improve reproducibility and ensure deterministic responses between invocations but also to ensure that it does not affect the model's ability to *self-correct*, *liuLargeLanguageModels2024*. The *max_token* parameter was set to 512 to provide sufficient context for processing the prompts and generate concise reasoning responses. The *response_format* parameter was used to ensure consistent and reliable output formatting. A JSON Schema is assigned to this parameter which

 doi:10.20944/preprints202501.1513.v1

describes the structure of the expected output, preventing the need for parsing of natural text and facilitates integration with downstream code.

2.7. Evaluation Metrics

The performance of ranking tasks in information retrieval challenges is evaluated using rank-based metrics and metrics at predefined cut-offs, such as the *top-k*% of retrieved documents [39,62]. The evaluation metrics used to measure the performance of the approach are listed below:

Recall @ k%

Recall also known as *sensitivity* provides a measure of how many relevant documents are retrieved within the *top-k*% of ranked results.

nWSS @ r%

Normalized work saved over sampling, equivalent to *True Negative Rate* (or *Specificity*) quantifies the effort or work saved by the automated system when compared to random sampling, assuming a fixed *recall* level, and can be utilized to evaluate outcomes across models and datasets [63]. Here *nWSS*% is evaluated at $r = 95\%$ and $r = 100\%$.

Average Precision (AP)

Average Precision (*AP*) represents the average of precision computed at each relevant document's position, considering all documents retrieved up to a specific rank. It combines both *precision* and *recall*, evaluating how effective documents are ranked by relevance. Unlike *precision @ k%* which is influenced by the total number the number of relevant documents, *AP* addresses this limitation by providing a more balanced assessment [62].

Mean Average Precision (MAP)

MAP represents the average *AP* across individual information needs and provides a single robust metric for evaluating the ranking quality across *recall* levels [62].

2.8. Experimental Setup

The experiments were automated using Python 3.9.21, with each experimental run organised into separate Python scripts. The baseline experiments utilised title and selection criteria, abbreviated as TSC in conjunctions with BM25, MiniLM and ChatGPT-4o-mini models. Meanwhile, the QA framework, was exclusively tested with ChatGPT-4o-mini model using all QA answer schemas list in Table 4 and Table 5.

Each execution script requires two input files: a CSV of title-abstracts and a prompt template file which gets dynamically populated with experimental context specific values for each zoonotic disease, CCHF, Ebola, Lepto, and RVF. The code then calls the OpenAI API completion endpoint using predefined model parameters and the adapted prompts. By utilizing the API, the study ensures reproducibility. All responses from the API are captured and subsequently processed to calculate a relevancy ranking.

3. Results

This section presents the results of the eight models evaluated across the four diseases: CCHF, Ebola, Lepto, and RVF evaluated against the study metric (Table 8). The following sections will examine results for each research questions.

3.1. RQ1. LLM Based QA Assessor vs Baseline

The LLM-based QA assessor models consistently outperform the TSC baseline models across all metrics, achieving higher recall, greater work savings, improved precision and reduced variability, holding true whether the baseline is generative (TSC-5) or non-generative (TSC-BM25, TSC-MiniLM).

 doi:10.20944/preprints202501.1513.v1

Figure 2, illustrates that the QA approaches exhibits superior early retrieval performance (*Recall@5%* - *Recall@20%*) compared to the baseline models, identifying at least 76% of relevant articles within the top 20% of the dataset, regardless of disease. The high *nWSS@r%* values show (Figure 3) that QA models are also better at detecting irrelevant documents early in the screening process, thereby reducing the number of items requiring manual review. The leading QA models, QA-4 and QA-5, demonstrate a minimum savings of 67% at *r=95%* across all diseases; however, performance declines when aiming for complete recall. In contrast, the TSC models display lower overall work savings, with particularly high variability in the LLM-based TSC model: it ranges from negative savings with Ebola to 46% *nWSS@95%* with Lepto.

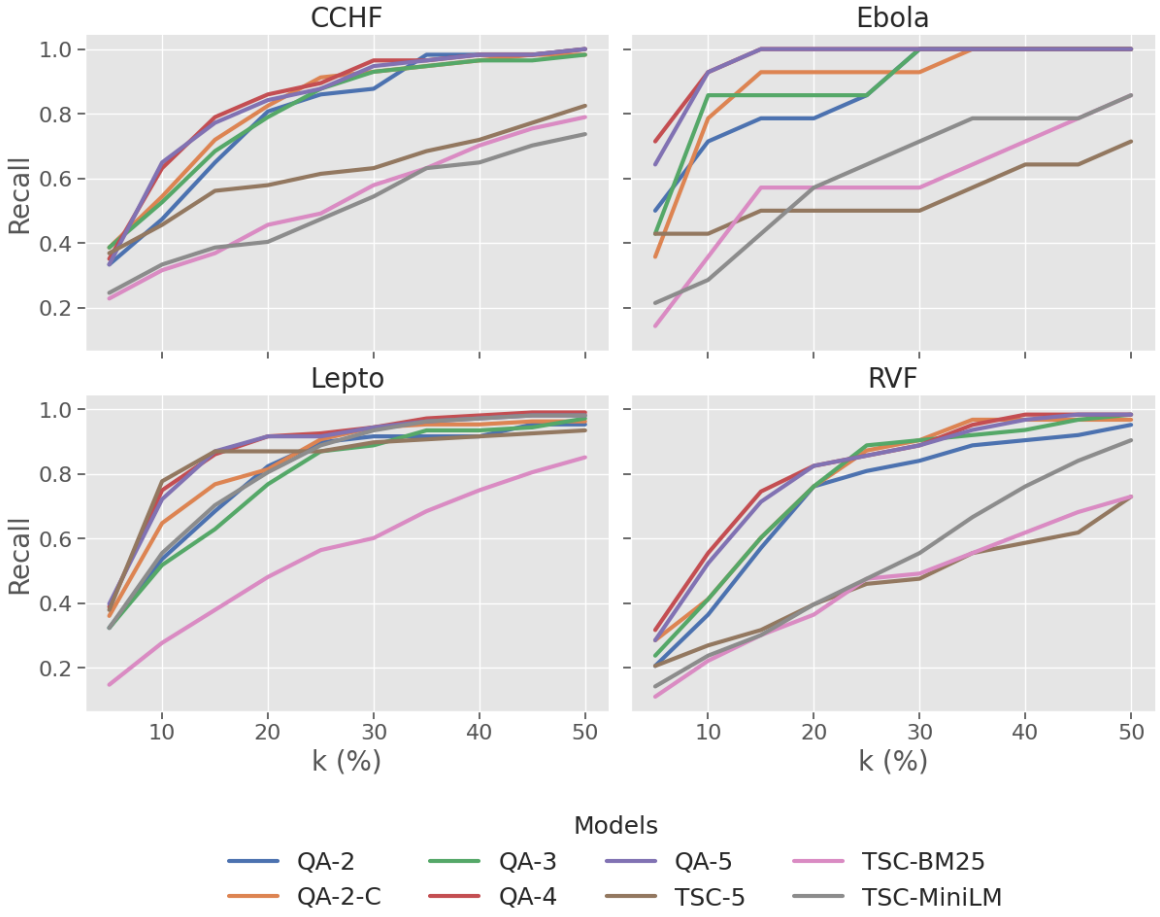

Figure 2. Plot of *Recall* at varying levels of *k*% across four zoonotic diseases

 doi:10.20944/preprints202501.1513.v1

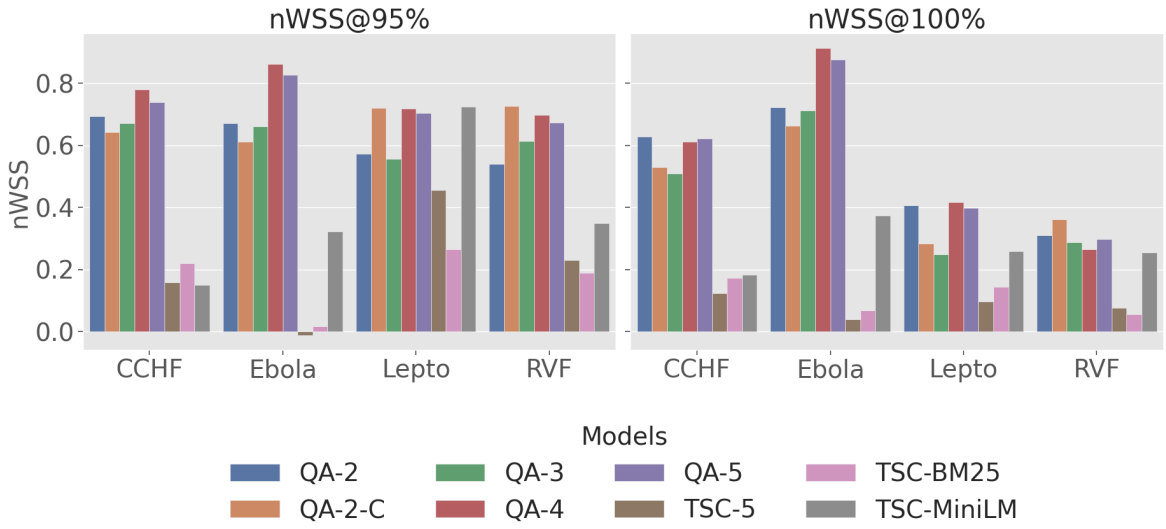

Figure 3. Plot of *nWSS* across four zoonotic diseases datasets at *recall* threshold of 95% and 100%

The *AP* metric evaluates ranking quality by measuring the number of relevant documents and how early they appear, further confirms QA models consistently high scores across all disease datasets (Figure 4). Except for Lepto, both TSC-5 and TSC-MiniLM fail to demonstrate robust ranking quality for the other diseases. In addition, the QA models exhibit less variability and more consistency across the datasets as seen in Figure 5. This box plot shows that QA models have narrower inter-quartile ranges (IQR) and shorter whiskers while the baseline models display higher medians and wider variability in rank positions.

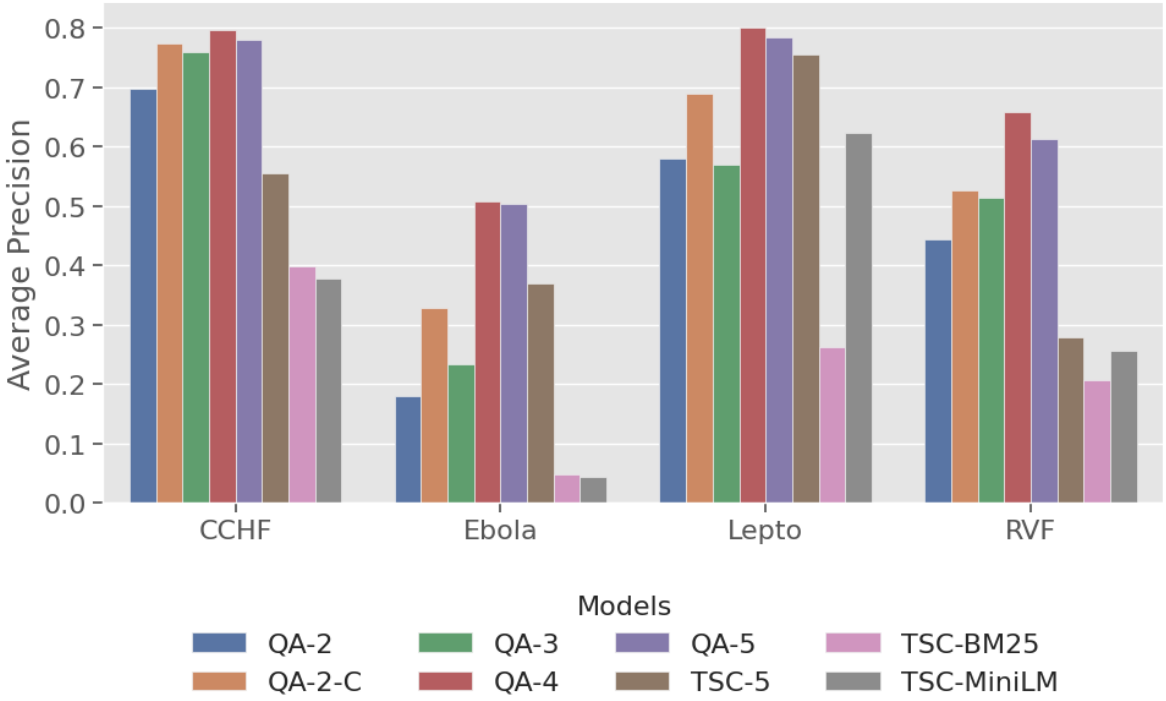

Figure 4. Plot of *Average Precision* across four zoonotic diseases

 doi:10.20944/preprints202501.1513.v1

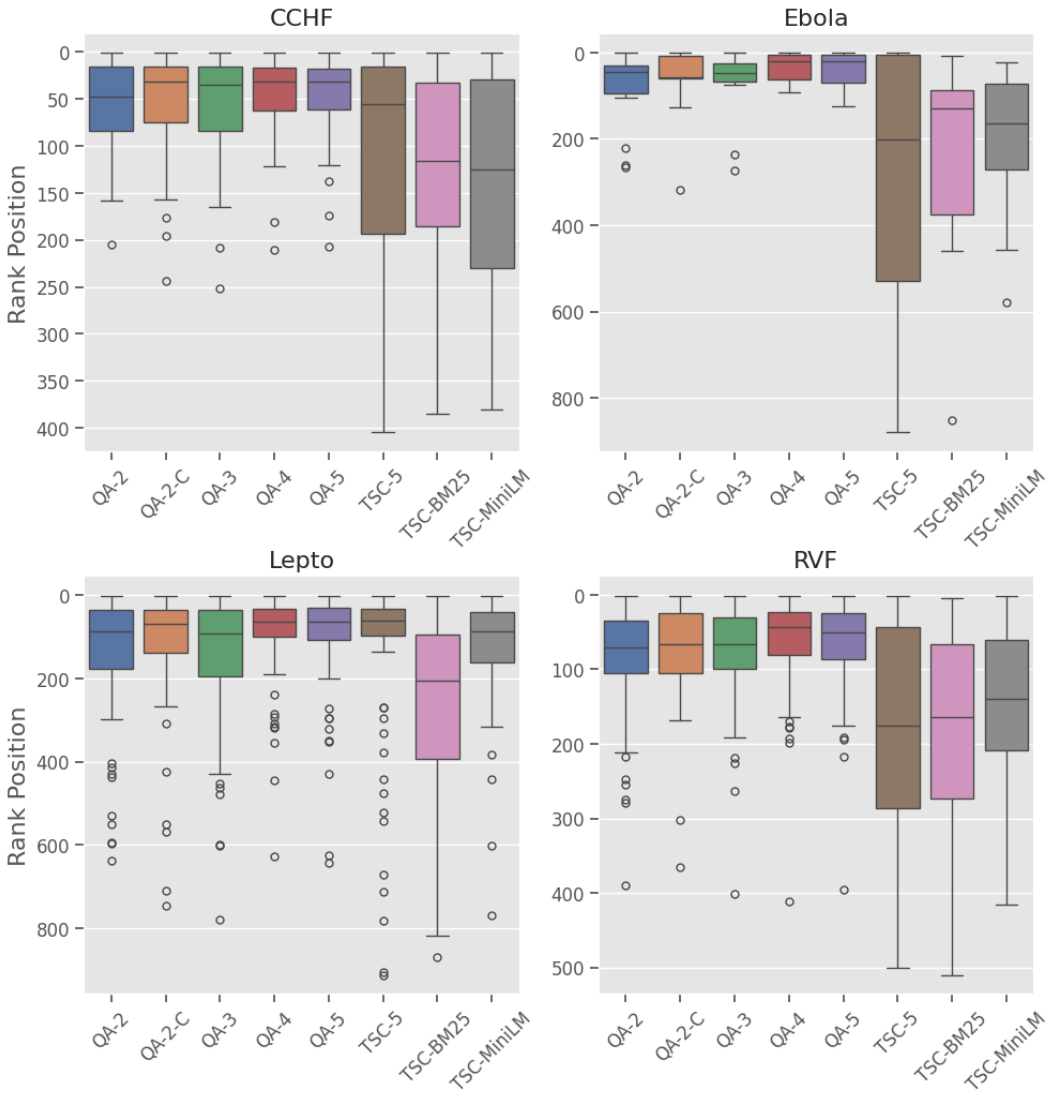

Figure 5. Plot of the distribution of rank positions of relevant title-abstract across four zoonotic diseases datasets

Finally, the *MAP* metric offers a single performance measure for each model. Table 6 indicates QA models clearly outperform the baseline models with QA-4 attaining the highest *MAP* score of 0.691. A *Precision-Recall Curve* (Figure 6), visualises overall performance by illustrating the effect of varying thresholds on the precision-recall balance. The QA models dominate the baseline models, and the LLM-based TSC-5 model notably outperforms TSC-BM25 and TSC-MiniLM, though it still lags behind QA-4 and QA-5.

 doi:10.20944/preprints202501.1513.v1

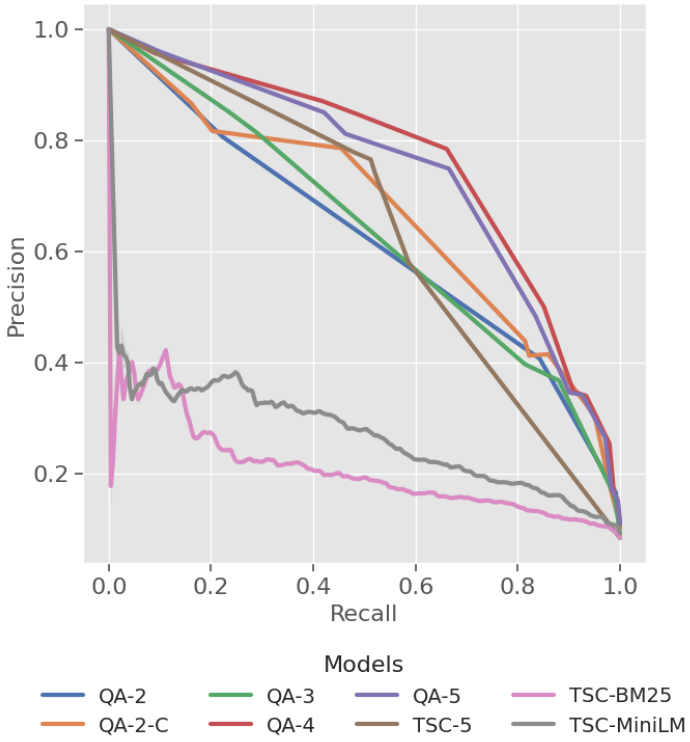

Figure 6. Plot of *precision* and *recall* across four zoonotic disease datasets

Table 6. Results for Mean Average Precision (MAP) and Area Under the Precision Recall Curve (PR-AUC) by model.

Model	MAP	PR-AUC
QA-2	0.476	0.621
QA-2-C	0.579	0.669
QA-3	0.519	0.636
QA-4	0.691	0.761
QA-5	0.670	0.740
TSC-5	0.489	0.639
TSC-BM25	0.229	0.206
TSC-MiniLM	0.325	0.271

3.2. RQ2. Effect of Label Granularity

The results indicate that label granularity strongly affects the ranking performance of an LLM-based assessor using a QA framework. Overall, performance generally improves with granularity QA-2 to QA-4, with diminishing returns observed from QA-4 to QA-5. As shown in Figure 2, single-level fine-grained labels (QA-4, QA-5) exhibit high *recall* at early ranks ($k \leq 20\%$), retrieving a substantial proportion of relevant articles. By ($k = 50\%$), they reach near complete *recall* (≥ 0.98) across all disease datasets.

QA-2-C's multi-level approach of presenting granularity by pairing binary labels with confidence scores demonstrates a middle ground between fine-grained (QA-4, QA-5) and coarse-grained (QA-2, QA-3) labels. QA-2-c lags behind QA-4, and QA-5 but all demonstrate significant work savings at 95% *recall* target over coarse-grained labels, achieving a minimum of 61%, 70% and 67% respectively. In particular, QA-4 and QA-5 perform notably well for highly skewed datasets such as Ebola, realising 83% to 88% savings at *recall* levels of 95% and 100%. Finally, the *MAP* scores in Table 6 demonstrate a clear trend indicating improved ranking performance with increased label granularity peaking at QA-4 (0.691).

 doi:10.20944/preprints202501.1513.v1

3.3. RQ3. Performance Across Zoonotic Diseases

The LLM-based QA assessor exhibit robust generalisable performance across climate sensitive zoonotice disease datasets, despite significant variation in relevancy rates. The Ebola dataset reveals a significant skew with only 1.5% of relevant records, whereas the CCHF, Lepto, and RVF datasets demonstrate moderate skews, as outlined in Table 2. This variation is apparent when evaluating model performance across the different diseases. Yet, QA based models, specifically QA-4 and QA-5, retained higher *recall@k%*, *nWSS@95%* and *AP* scores, while maintaining consistent ranking across the datasets in comparison to the baseline models.

In the Ebola dataset, QA-4 and QA-5 models demonstrate a significantly high *recall*, achieving complete *recall* at $k = 15\%$ (Figure 2), suggesting that all relevant title-abstracts were identified within the top 15% of the ranked list. The baseline models struggle to achieve a comparable level of recall, with BM25 exhibiting a consistent but significantly lower recall across all diseases. In contrast, TSC-5 and TSC-MiniLM perform well with Lepto but severely unperformed with other diseases.

For all diseases, the QA models retain a minimum of 54% work saving at *nWSS@95%*; however, performance drops when transitioning to *nWSS@100%* for all disease except Ebola, where savings slightly increase. Notably, TSC-MiniLM achieves the same work savings as QA-4 and QA-2-C *nWSS@95% = 0.72*); however, when striving for full *recall*, this value reduces dramatically proving high variability.

A consistently high *AP* across diseases indicates models that can handle disease-specific terminology and generalise well independent of relevancy rates. The leading QA models (QA-4, QA-5) achieve a minimum of 0.5 *AP* across all diseases, demonstrating their capability to manage variations in relevance rates and address disease-specific terminology. The baseline models exhibit lower average precision (AP) and greater variability in AP, suggesting limited robustness to skewed datasets and less capacity for handling disease-specific terminology.

QA models achieved the best rank distribution with the Ebola dataset (Figure 5), demonstrated by the lower median rank, tighter IQR and few outliers. Similar patterns were seen in Lepto, CCHF, and RVF, all of which suggest that QA models are superior to baseline models at prioritising relevant articles in the earlier ranks. Nevertheless, the Lepto dataset has the greatest number of outliers, which may be a sign of ambiguous abstracts, misaligned classification criteria, or model limitations.

 doi:10.20944/preprints202501.1513.v1


```
{
    "results": [
        {
            "question_number": 1,
            "reason": "The abstract describes a study that identifies mosquito larvae species and examines their habitats, indicating that it is based on original data collection rather than a review or meta-analysis.",
            "answer": "Definitely Yes"
        },
        {
            "question_number": 2,
            "reason": "The study discusses mosquito species that are vectors for diseases like Rift Valley Fever but does not measure the incidence or prevalence of the virus itself, focusing instead on the mosquito larvae.",
            "answer": "Probably No"
        },
        {
            "question_number": 3,
            "reason": "The research includes surveys conducted during dry and rainy seasons, which implies an examination of seasonal environmental factors that could impact vector distribution.",
            "answer": "Definitely Yes"
        },
        {
            "question_number": 4,
            "reason": "The study involves field surveys and ecological assessments rather than laboratory method validation, focusing on real-world data collection.",
            "answer": "Definitely Yes"
        }
    ]
}
```

Listing 2. Example JSON response for an RVF abstract

3.4. RQ4. Utility of Generated CoT Rationale

The CoT reasoning logic proved to be consistent, in line with human expectations, helpful in detecting edge cases and capable of revealing variations in misclassifications across disease datasets. Listing 2 is an example RVF response from the QA-4 model that demonstrates the logic the model provides for each title-abstract record. A senior reviewer on the team was presented with the results to manually assess the validity of the responses. For each disease, title-abstracts with relevancy scores exceeding a predetermined threshold and classified as irrelevant, as well as those below the threshold classified as relevant, were displayed in MS Excel alongside the answer scores and reasoning text for each question. The reviewer did report any significant issues in the model's reasoning and considered the explanatory text highly beneficial for comprehending the logic underlying the answer scores. Using the sort and filter facilities provided in MS Excel the reviewer was able to efficiently review and identify several misclassified items as shown in Table 7. Lepto (16), RVF (13) and CCHF(10) had the most misclassifications, highlighting the inherent variability in manual SLR screening quality but also opportunities to refine the selection criteria and eligibility questions. These findings not only highlight the potential of CoT reasoning to ensure transparency in the SLR process but also a means to iteratively enhance the ranking process.

Table 7. Misclassifications identified by the system.

Disease	Revised to include	Revised to exclude	Total
CCHF	7	3	10
Ebola	1	0	1
Lepto	13	3	16
RVF	8	5	13

 doi:10.20944/preprints202501.1513.v1

Table 8. Performance metrics for models by zoonotic diseases.

Disease	Model	Recall@k%				nWSS@r%			AP
		5	10	20	30	50	95	100	
CCHF	QA-2	0.33	0.47	0.81	0.88	1.00	0.69	0.63	0.70
	QA-2-C	0.39	0.54	0.82	0.93	0.98	0.64	0.53	0.77
	QA-3	0.39	0.53	0.79	0.93	0.98	0.67	0.51	0.76
	QA-4	0.35	0.63	0.86	0.96	1.00	0.78	0.61	0.80
	QA-5	0.33	0.65	0.84	0.95	1.00	0.74	0.62	0.78
	TSC-5	0.37	0.46	0.58	0.63	0.82	0.16	0.12	0.55
	TSC-BM25	0.23	0.32	0.46	0.58	0.79	0.22	0.17	0.40
	TSC-MiniLM	0.25	0.33	0.40	0.54	0.74	0.15	0.18	0.38
Ebola	QA-2	0.50	0.71	0.79	1.00	1.00	0.67	0.72	0.18
	QA-2-C	0.36	0.79	0.93	0.93	1.00	0.61	0.66	0.33
	QA-3	0.43	0.86	0.86	1.00	1.00	0.66	0.71	0.23
	QA-4	0.71	0.93	1.00	1.00	1.00	0.86	0.91	0.51
	QA-5	0.64	0.93	1.00	1.00	1.00	0.83	0.88	0.50
	TSC-5	0.43	0.43	0.50	0.50	0.71	-0.01	0.04	0.37
	TSC-BM25	0.14	0.36	0.57	0.57	0.86	0.02	0.07	0.05
	TSC-MiniLM	0.21	0.29	0.57	0.71	0.86	0.32	0.37	0.04
Lepto	QA-2	0.32	0.54	0.82	0.92	0.95	0.57	0.41	0.58
	QA-2-C	0.36	0.65	0.81	0.94	0.96	0.72	0.28	0.69
	QA-3	0.32	0.52	0.77	0.89	0.97	0.56	0.25	0.57
	QA-4	0.39	0.75	0.92	0.94	0.99	0.72	0.42	0.80
	QA-5	0.40	0.72	0.92	0.94	0.98	0.70	0.40	0.78
	TSC-5	0.38	0.78	0.87	0.90	0.94	0.46	0.10	0.75
	TSC-BM25	0.15	0.28	0.48	0.60	0.85	0.26	0.14	0.26
	TSC-MiniLM	0.32	0.56	0.81	0.94	0.98	0.72	0.26	0.62
RVF	QA-2	0.21	0.37	0.76	0.84	0.95	0.54	0.31	0.44
	QA-2-C	0.29	0.41	0.76	0.90	0.97	0.73	0.36	0.53
	QA-3	0.24	0.41	0.76	0.90	0.98	0.61	0.29	0.51
	QA-4	0.32	0.56	0.83	0.89	0.98	0.70	0.27	0.66
	QA-5	0.29	0.52	0.83	0.89	0.98	0.67	0.30	0.61
	TSC-5	0.21	0.27	0.40	0.48	0.73	0.23	0.08	0.28
	TSC-BM25	0.11	0.22	0.37	0.49	0.73	0.19	0.06	0.21
	TSC-MiniLM	0.14	0.24	0.40	0.56	0.90	0.35	0.26	0.26

4. Discussion

This paper presented the conceptual architecture of the CliZod system and contextualised the role of the relevancy ranking component in the broader system. The subsequent empirical findings illustrated that a zero-shot LLM-based QA assessor, leveraging fine-grained labels, can reliably rank primary studies by relevance. The model demonstrated broad generalisability across four climate-sensitive zoonotice disease datasets with varying relevancy rates. Additionally the CoT reasoning text generated by this approach provides valuable insight to human reviewers and aided in identification of misclassified records in the disease datasets.

Before further interpretation of the results, the following important limitations are note worthy. Most importantly, the study depended on a single reviewer for the manual evaluation of the model's reasoning text. No metrics were established to measure the quality or potential bias of the text generated by the LLM, which are prone to *"hallucination"*, *jiMitigatingHallucinationLarge2023*, *zackAssessingPotentialGPT42024*. Nevertheless, the results from the study demonstrated a potential use case for such data and the possibility for creating a unique dataset to enable the team to further research into reasoning and bias within LLM-based approaches.

This study used a single closed-source commercial LLM from OpenAI to conduct all experiments. Restricting the evaluation soley to ChatGPT-4-o-mini risks the generalisation to other LLMs. Ad-

 doi:10.20944/preprints202501.1513.v1

ditionally, while a fixed model version was used with the *temperature* parameter set to 0 to ensure deterministic behaviour, and prior research confirming consistent outcomes on repeated invocation [49], ongoing optimisations by OpenAI may have led to performance changes that could have influenced the results. In the future, the team intends to explore additional models, including open-source alternatives applying the same methodology.

Moreover, the eligibility questions used in the experiments were generated in retrospect following manual screening, hence there is a risk of over fitting and misalignment with broader applicability. Polanin et al. advocates for conducting a pilot screening session with a subset of abstracts to refine and validate eligibility questions prior to screening. This approach could ensure clarity and consistency of the questions while improving the effectiveness of the rankings. Such an exercise may prove beneficial, particularly for the remaining unscreened disease and climate variable articles as it could mitigate bias and facilitate the development of a more comprehensive dataset. Additionally, the collection of ground truth data at the question level would facilitate a more precise and detailed evaluation of the model and support further research.

Lastly, although the real-world setting of this study's dataset demonstrated the effectiveness of the approach, it has not been thoroughly evaluated on datasets from other domains. CSMed is an initiative to create a standardised dataset to evaluate the performance of automated SLR screening models [39], a gap highlighted in recent research [16]. It provides access to 325 SLRs in the medical and computer science domain and presents an opportunity to benchmark the approach proposed in this study and gain a broader understanding of its performance, generalisability and adaptability across domains.

Despite the limitations, this study demonstrates the effectiveness of employing a QA framework with an LLM-based assessor to robustly rank literature on four climate-sensitive zoonotic diseases by relevance. The QA models consistently outperformed all the baseline models, achieving high *Recall@k%* and *MAP* scores across all disease datasets in this study. High *recall* is crucial in SLR automation systems [64] to prevent bias and ensure all relevant articles are identified. Additionally a high *MAP* score reflects strong discrimination and resilience to imbalanced datasets [62], ensuring reviewers are presented with both manageable and highly relevant articles.

While the study results are promising, it is important to contextualise the system performance. Direct comparisons to other systems must be made with caution due to the challengingly unique climate-sensitive zoonotic disease dataset employed in this study. However, the baseline TSC-BM25 model provides a simple yet informative reference point. In the study by Wang et al. examining the performance of neural rankers on the CLEF [65] datasets, a baseline BM25 model recorded a *Recall@20%* ranging from 0.52 to 0.64, alongside a *MAP* score of 0.16 using title as the input query to the model. By comparison, TSC-BM25 in this study recorded a *Recall@20%* ranging from 0.37 to 0.57 and a slightly higher *MAP* score of 0.23 using title and selection criteria as the input query. Their most effective fine-tuned BioBERT neural ranker demonstrated a *Recall@20%* ranging from 0.82 to 0.89, and a *MAP* score of 0.381. In contrast, the QA-4 model in this study attained a *Recall@20%* ranging from 0.83 to 1.0, and a significantly higher *MAP* score of 0.670. Although the highly skewed Ebola dataset may have contributed to the elevated recall score, the findings highlight the potential of the zero-shot LLM-based QA framework for enhancing ranking performance without the need for fine-tuning. However, any conclusions should be deferred until a comprehensive evaluation with a standardised benchmarking dataset is conducted.

The adoption of a QA framework has shown to be highly effective in the current study, contrary to the findings of prior research, such as Kohandel Gargari et al. where a screening tool approach yield poor performance. Upon examining their methodology, several factors may have contributed to the discrepancy, including the complexity of the prompt, ambiguity in managing 'unclear' labels, and the assumption that the model will adhere to the embedded logic flow in the prompt without errors. Additionally, their research employed GPT 3.5, an older model, less advanced and less accurate. In contrast, the findings from Akinseloyin et al. align more closely with this study, utilising

 doi:10.20944/preprints202501.1513.v1

a similar QA framework methodology. However, their eligibility questions were derived from the inclusion/exclusion criteria using an LLM and observed that the generated questions lacked complete independence and recommended that such questions be created by humans for better reliability, a recommendation followed in this study.

In addition to the QA framework, the use of fine-grained labels appears to have a consistent positive impact on ranking tasks whether in an assessor or a ranker context. The single-level fine-grained models, QA-4 and QA-5 demonstrated superior recall, precision and work savings, with minimal variability, compared to the coarser-grained models, QA-2 and QA-3. Fine-grained labels appear to enhance the system performance by supplying the model with broader answer options, allowing to more effectively represent uncertainty. However, the benefits diminished as granularity increased beyond four levels as evident by the reduced *MAP* score for QA-5. Zhuang et al. found that fine-grained relevancy labels improved performance, with no advantage in exceeding four levels of granularity using a point-wise LLM-based ranker. While the tasks in these studies differ (assessor vs ranker), this suggests that fine-grained labels may be broadly beneficial for LLM-based ranking tasks.

Interestingly, the shift from a single-level to a multi-level labelling approach (QA-2-c) demonstrated reduced performance. Several factors could account for this outcome. Firstly, the multi-level model employs a more intricate prompt, simple prompt structures are more effective than complex ones [67]. Furthermore, a single level classification inherently captures both the relevance and uncertainty in one consolidated label (e.g. "Probably Yes") where as a two step classification introduces uncertainty post-decision, thereby splitting the context. Finally, the model's autoregressive characteristic [50], combined with single versus multiple decision-making points, may account for the observed performance differences.

The study approach also shows promising generalisability across disease datasets, despite their varied characteristics. Nicholson Thomas et al. used a selection criteria based prompt, similar to TSC-5 baseline model, to screen articles for ecosystem condition indicators. They report that to handle multidimensional topics with high precision, iterative refinement of the selection criteria was essential. While further enhancements of the selection criteria could enhance the performance of the TSC-5 model for diseases where it underperformed (other than Lepto), the QA framework offers an additional layer of flexibility. It decouples the decision making from the LLM's internal logic by decomposing the assessment into granular questions. Weak signals in the form of uncertainty labels (e.g. "Probably No" or "Unsure"), allow studies that marginally fail to still contribute to overall ranking through the LWS. This flexibility ensures relevant studies are not prematurely excluded and provides adaptability to the varied disease datasets.

Complementing the flexibility of the QA framework is the utility of reasoning text generated through chain-of-thought prompting, which provides a detailed explanation for each relevancy assessment, exposing the models' decision making process to the reviewers. For example, the reasoning text in Listing 2 illustrates the model's ability to justifying uncertain responses like "Probably No" in question 2, where it explains that although the study addresses disease vectors, it does not directly assess virus prevalence. On the spectrum of human-machine collaboration, this approach of automatically generating a first-pass judgement with rationale falls into the *human verification* category, or *human-in-the-loop* approach [46]. While chain-of-thought prompting does not ensure the accuracy of reasoning path, it does enable the model to more effectively access relevant data learnt during pre-training [31], which may explain the QA models' performance over the baseline models. Further investigation is required to assess the validity and reliability of the generated rationale.

It is recommended that tool developers test their products across a diverse range of domains, as the effectiveness of prompts is influenced by the characteristics of the dataset [49]. The dataset being curated as part of the CliZod project has the potential to provide 33 SLRs for tool developers, focusing on research in the areas intersecting climate change and veterinary epidemiology. This unique combination of domains provides additional benchmark for evaluating the efficacy of automated screening systems in handling multidisciplinary research questions.

 doi:10.20944/preprints202501.1513.v1

Finally, the approach proposed in this study can currently be used as a standalone decision aid to reduce the risk of human error and bias during the initial screening phase. The main prompt provided in the supplementary code is generic and applicable across domains. All that is needed is to formulate research-specific eligibility questions, define the review topic, and modify the persona text accordingly. Data is output as a CSV file containing ranking scores, answers, and reasoning text. It can be reviewed in a tool such as MS Excel providing an additional layer of adaptability. By prioritising and streamlining the review process, this approach allows human reviewers to focus on the edge cases, making it a practical and efficient solution for SLRs.

5. Conclusions

This study introduces CliZod, a novel automated framework for extracting modelling parameters from scientific literature to support climate-sensitive zoonotic disease modelling. It reports the empirical results of the relevancy ranking approach, leveraging an LLM as an assessor and guided by a QA framework to rank literature on four climate-sensitive zoonotic diseases and one climate variable (rainfall).

The findings of this study demonstrate that an LLM-based QA assessor using zero-shot CoT prompting can effectively and reliably rank primary studies by relevance across 4 zoonotic disease datasets with varying relevance rates. Furthermore, fine-grained QA models significantly outperform the baseline models, achieving strong *recall* at multiple thresholds, high *MAP* scores and substantial work savings of at least 70% (NWSS@95%). Additionally, the CoT reasoning text generated by this approach provides valuable insight, enhancing and aiding human reviewers in their decision making process.

Although further empirical research is necessary to validate the approach against standardised benchmark dataset, the substantial reduction in screening effort, combined with the provision of explainable AI rationales, represents an important step toward automated parameter extraction from scientific literature.

Funding: This research was funded by Wellcome Trust grant number XXX.

Data Availability Statement: All of the code used in this study is accessible in a github repository (link provided upon publishing).

Acknowledgments: We gratefully acknowledge funding from Wellcome to support the development and implementation of the CliZod system.

Conflicts of Interest: This work is funded by a Wellcome grant for the development of the CliZod system. The funding body had no role in the study design, data collection, analysis, or interpretation. We have no conflicts of interest to declare.

Abbreviations

The following abbreviations are used in this manuscript:

MDPI	Multidisciplinary Digital Publishing Institute
LLM	Large Language Model
QA	Question and answer
CCHF	Crimean-Congo haemorrhagic fever
RVF	Rift valley fever

 doi:10.20944/preprints202501.1513.v1

Appendix A. Additional Prompts and Disease Information

```
Task: Analyse the abstract below within the double quotes and answer the questions below. Take a step-by-step approach towards reasoning and then answer the questions with either [ANS-SCHEMA-LABELS] only.
Keep the reasoning short and concise and do not repeat the question.

Abstract:
"$abstract"

Question:
$questions

Required Format:
Format the output as a JSON object with a single key, "results", containing an array of objects. Each object should represent an answer to a specific question.

Answer Requirements:
Answer all questions without exceptions.
Do not add any text before or after the JSON output.

Object Structure:
Each object within "results" should contain the following fields:
    "question_number": Number of the question.
    "reason": Step-by-step reasoning to the question.
    "answer": [ANS-SCHEMA-LABELS]

Example Format:
{
    "results": [
        {
            "question_number": <THE NUMBER IN FRONT OF THE QUESTION YOU ARE ANSWERING>,
            "reason": "<YOUR REASONING FOR THE QUESTION>",
            "answer": "<YOUR FINAL ANSWER>"
        },
        # the next question number, answer and explanation
    ]
}

Strict Output Requirements:
You MUST answer all questions.
You MUST NOT output any other text before or after the JSON.
Do NOT be chatty. Output exactly what is instructed.
```

Listing 3. QA single level prompt, based on QA framework

 doi:10.20944/preprints202501.1513.v1


```
Task: Analyse the abstract below within the double quotes and answer the questions below. Take a step-by-step approach towards reasoning and then answer the questions with either "Yes" or "No" only.
Keep the reasoning short and concise and do not repeat the question. Provide a confidence score to your answer reflecting how certain you are based on the provided context and your reasoning using the confidence scale below.

Confidence scale:
Low
Medium
High

Abstract:
"$abstract"

Question:
$questions

Required Format:
Format the output as a JSON object with a single key, "results", containing an array of objects. Each object should represent an answer to a specific question.

Answer Requirements:
Answer all questions without exceptions.
Do not add any text before or after the JSON output.

Object Structure:
Each object within "results" should contain the following fields:
    "question_number": Number of the question.
    "reason": Step-by-step reasoning to the question.
    "answer": "Yes" or "No"
    "confidence_score": Confidence score for your answer

Example Format:
{
    "results": [
        {
            "question_number": <THE NUMBER IN FRONT OF THE QUESTION YOU ARE ANSWERING>,
            "reason": "<YOUR REASONING FOR THE QUESTION>",
            "answer": "<YOUR FINAL ANSWER>",
            "confidence_score": "<YOUR CONFIDENCE SCORE FOR THE ANSWER>"
        },
        # the next question number, answer and explanation
    ]
}

Strict Output Requirements:
You MUST answer all questions.
You MUST NOT output any other text before or after the JSON.
Do NOT be chatty. Output exactly what is instructed.
```

Listing 4. QA multi level prompt, based on QA framework and supports an answer and a confidence level

 doi:10.20944/preprints202501.1513.v1

Table A1. Shortlisted diseases selected for CliZod's climate-sensitive zoonotic disease modelling framework.

Disease
Crimean-Congo haemorrhagic fever
Ebolavirus
Henipah virus
Mammarenavirus
Marburg virus
MERS-CoV
Rift Valley Fever
SARS-CoV
Zika
Leptospirosis
Dengue

References

1. Ryan, S.J.; Lippi, C.A.; Caplan, T.; Diaz, A.; Dunbar, W.; Grover, S.; Johnson, S.; Knowles, R.; Lowe, R.; Mateen, B.A.; et al. The Current Landscape of Software Tools for the Climate-Sensitive Infectious Disease Modelling Community. *The Lancet Planetary Health* **2023**, *7*, e527–e536. https://doi.org/10.1016/S2542-5196(23)00056-6.
2. Allen, T.; Murray, K.A.; Zambrana-Torrelio, C.; Morse, S.S.; Rondinini, C.; Di Marco, M.; Breit, N.; Olival, K.J.; Daszak, P. Global Hotspots and Correlates of Emerging Zoonotic Diseases. *Nature Communications* **2017**, *8*, 1124. https://doi.org/10.1038/s41467-017-00923-8.
3. Grace, D.; Mutua, F.K.; Ochungo, P.; Kruska, RL.; Jones, K.; Brierley, L.; Lapar, M.L.; Said, M.Y.; Herrero, M.T.; Phuc, PM.; et al. Mapping of Poverty and Likely Zoonoses Hotspots **2012**.
4. Gubbins, S.; Carpenter, S.; Mellor, P.; Baylis, M.; Wood, J. Assessing the Risk of Bluetongue to UK Livestock: Uncertainty and Sensitivity Analyses of a Temperature-Dependent Model for the Basic Reproduction Number. *Journal of the Royal Society Interface* **2008**, *5*, 363–371. https://doi.org/10.1098/rsif.2007.1110.
5. Guis, H.; Caminade, C.; Calvete, C.; Morse, A.P.; Tran, A.; Baylis, M. Modelling the Effects of Past and Future Climate on the Risk of Bluetongue Emergence in Europe. *Journal of The Royal Society Interface* **2011**, *9*, 339–350. https://doi.org/10.1098/rsif.2011.0255.
6. Chandler, J.; Cumpston, M.; Higgins, J.P.T.; on evidence-based medicine) Li, T.W.; Page, M.J.; Thomas, J.P.o.S.R.a.P.; Welch, V.A., Eds. *Cochrane Handbook for Systematic Reviews of Interventions*, second edition ed.; Cochrane Book Series, Wiley-Blackwell, 2019.
7. Kitchenham, B.; Charters, S. Guidelines for Performing Systematic Literature Reviews in Software Engineering **2007**. *2*.
8. Tricco, A.C.; Brehaut, J.; Chen, M.H.; Moher, D. Following 411 Cochrane Protocols to Completion: A Retrospective Cohort Study. *PloS One* **2008**, *3*, e3684. https://doi.org/10.1371/journal.pone.0003684.
9. Michelson, M.; Reuter, K. 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* **2019**, *16*, 100443. https://doi.org/10.1016/j.conctc.2019.100443.
10. Bornmann, L.; Haunschild, R.; Mutz, R. Growth Rates of Modern Science: A Latent Piecewise Growth Curve Approach to Model Publication Numbers from Established and New Literature Databases. *Humanities and Social Sciences Communications* **2021**, *8*, 1–15. https://doi.org/10.1057/s41599-021-00903-w.
11. Parolo, P.D.B.; Pan, R.K.; Ghosh, R.; Huberman, B.A.; Kaski, K.; Fortunato, S. Attention Decay in Science **2015**. https://doi.org/10.48550/ARXIV.1503.01881.
12. Bashir, R.; Surian, D.; Dunn, A.G. Time-to-Update of Systematic Reviews Relative to the Availability of New Evidence. *Systematic Reviews* **2018**, *7*, 195. https://doi.org/10.1186/s13643-018-0856-9.
13. Clark, J.; McFarlane, C.; Cleo, G.; Ishikawa Ramos, C.; Marshall, S. The Impact of Systematic Review Automation Tools on Methodological Quality and Time Taken to Complete Systematic Review Tasks: Case Study. *JMIR medical education* **2021**, *7*, e24418. https://doi.org/10.2196/24418.
14. Thomas, J.; McDonald, S.; Noel-Storr, A.; Shemilt, I.; Elliott, J.; Mavergames, C.; Marshall, I.J. Machine Learning Reduced Workload with Minimal Risk of Missing Studies: Development and Evaluation of a Randomized Controlled Trial Classifier for Cochrane Reviews. *Journal of Clinical Epidemiology* **2021**, *133*, 140–151. https://doi.org/10.1016/j.jclinepi.2020.11.003.

15. Tsafnat, G.; Glasziou, P.; Choong, M.K.; Dunn, A.; Galgani, F.; Coiera, E. Systematic Review Automation Technologies. *Systematic Reviews* **2014**, *3*, 74. https://doi.org/10.1186/2046-4053-3-74.
16. Bolanos, F.; Salatino, A.; Osborne, F.; Motta, E. Artificial Intelligence for Literature Reviews: Opportunities and Challenges **2024**. https://doi.org/10.48550/ARXIV.2402.08565.
17. Khalil, H.; Ameen, D.; Zarnegar, A. Tools to Support the Automation of Systematic Reviews: A Scoping Review. *Journal of Clinical Epidemiology* **2022**, *144*, 22–42. https://doi.org/10.1016/j.jclinepi.2021.12.005.
18. Santos, Á.O.D.; Da Silva, E.S.; Couto, L.M.; Reis, G.V.L.; Belo, V.S. The Use of Artificial Intelligence for Automating or Semi-Automating Biomedical Literature Analyses: A Scoping Review. *Journal of Biomedical Informatics* **2023**, *142*, 104389. https://doi.org/10.1016/j.jbi.2023.104389.
19. Guo, E.; Gupta, M.; Deng, J.; Park, Y.J.; Paget, M.; Naugler, C. Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study. *Journal of Medical Internet Research* **2024**, *26*, e48996. https://doi.org/10.2196/48996.
20. Issaiy, M.; Ghanaati, H.; Kolahi, S.; Shakiba, M.; Jalali, A.; Zarei, D.; Kazemian, S.; Avanaki, M.; Firouznia, K. Methodological Insights into ChatGPT's Screening Performance in Systematic Reviews. *BMC Medical Research Methodology* **2024**, *24*. https://doi.org/10.1186/s12874-024-02203-8.
21. Cao, C.; Sang, J.; Arora, R.; Kloosterman, R.; Cecere, M.; Gorla, J.; Saleh, R.; Chen, D.; Drennan, I.; Teja, B.; et al. Prompting Is All You Need: LLMs for Systematic Review Screening, 2024. https://doi.org/10.1101/2024.06.01.24308323.
22. Alshami, A.; Elsayed, M.; Ali, E.; Eltoukhy, A.E.E.; Zayed, T. Harnessing the Power of ChatGPT for Automating Systematic Review Process: Methodology, Case Study, Limitations, and Future Directions. *Systems* **2023**, *11*, 351. https://doi.org/10.3390/systems11070351.
23. Fernandes Torres, J.P.; Mulligan, C.; Jorge, J.; Moreira, C. PROMPTHEUS: A Human-Centered Pipeline to Streamline Slrs with Llms. *arXiv e-prints* **2024**, pp. arXiv–2410.
24. Schmidt, L.; Finnerty Mutlu, A.N.; Elmore, R.; Olorisade, B.K.; Thomas, J.; Higgins, J.P.T. Data Extraction Methods for Systematic Review (Semi)Automation: Update of a Living Systematic Review. *F1000Research* **2023**, *10*, 401. https://doi.org/10.12688/f1000research.51117.2.
25. Polak, M.P.; Morgan, D. Extracting Accurate Materials Data from Research Papers with Conversational Language Models and Prompt Engineering. *Nature Communications* **2024**, *15*, 1569. https://doi.org/10.1038/s41467-024-45914-8.
26. Nicholson Thomas, I.; Roche, P.; Grêt-Regamey, A. Harnessing Artificial Intelligence for Efficient Systematic Reviews: A Case Study in Ecosystem Condition Indicators. *Ecological Informatics* **2024**, *83*, 102819. https://doi.org/10.1016/j.ecoinf.2024.102819.
27. Susnjak, T.; Hwang, P.; Reyes, N.H.; Barczak, A.L.C.; McIntosh, T.R.; Ranathunga, S. Automating Research Synthesis with Domain-Specific Large Language Model Fine-Tuning, 2024, [arXiv:cs/2404.08680]. https://doi.org/10.48550/arXiv.2404.08680.
28. Ji, Z.; Yu, T.; Xu, Y.; Lee, N.; Ishii, E.; Fung, P. Towards Mitigating Hallucination in Large Language Models via Self-Reflection, 2023, [arXiv:cs/2310.06271]. https://doi.org/10.48550/arXiv.2310.06271.
29. Zack, T.; Lehman, E.; Suzgun, M.; Rodriguez, J.A.; Celi, L.A.; Gichoya, J.; Jurafsky, D.; Szolovits, P.; Bates, D.W.; Abdulnour, R.E.E.; et al. Assessing the Potential of GPT-4 to Perpetuate Racial and Gender Biases in Health Care: A Model Evaluation Study. *The Lancet Digital Health* **2024**, *6*, e12–e22. https://doi.org/10.1016/S2589-7500(23)00225-X.
30. Zhao, H.; Chen, H.; Yang, F.; Liu, N.; Deng, H.; Cai, H.; Wang, S.; Yin, D.; Du, M. Explainability for Large Language Models: A Survey. *ACM Trans. Intell. Syst. Technol.* **2024**, *15*, 20:1–20:38. https://doi.org/10.1145/3639372.
31. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D.; et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Proceedings of the Advances in Neural Information Processing Systems, 2022, Vol. 35, pp. 24824–24837.
32. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.t.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Proceedings of the Advances in Neural Information Processing Systems. Curran Associates, Inc., 2020, Vol. 33, pp. 9459–9474.
33. Beller, E.; Clark, J.; Tsafnat, G.; Adams, C.; Diehl, H.; Lund, H.; Ouzzani, M.; Thayer, K.; Thomas, J.; Turner, T.; et al. Making Progress with the Automation of Systematic Reviews: Principles of the International Collaboration for the Automation of Systematic Reviews (ICASR). *Systematic Reviews* **2018**, *7*, 77. https://doi.org/10.1186/s13643-018-0740-7.

34. Scott, A.M.; Forbes, C.; Clark, J.; Carter, M.; Glasziou, P.; Munn, Z. Systematic Review Automation Tool Use by Systematic Reviewers, Health Technology Assessors and Clinical Guideline Developers: Tools Used, Abandoned, and Desired, 2021. https://doi.org/10.1101/2021.04.26.21255833.
35. Polanin, J.R.; Pigott, T.D.; Espelage, D.L.; Grotpeter, J.K. Best Practice Guidelines for Abstract Screening Large-Evidence Systematic Reviews and Meta-Analyses. *Research Synthesis Methods* **2019**, *10*, 330–342. https://doi.org/10.1002/jrsm.1354.
36. Sampson, M.; Tetzlaff, J.; Urquhart, C. Precision of Healthcare Systematic Review Searches in a Cross-sectional Sample. *Research Synthesis Methods* **2011**, *2*, 119–125. https://doi.org/10.1002/jrsm.42.
37. Wang, S.; Scells, H.; Koopman, B.; Zuccon, G. Neural Rankers for Effective Screening Prioritisation in Medical Systematic Review Literature Search, 2022, [2212.09017]. https://doi.org/10.48550/arXiv.2212.09017.
38. Mitrov, G.; Stanoev, B.; Gievska, S.; Mirceva, G.; Zdravevski, E. Combining Semantic Matching, Word Embeddings, Transformers, and LLMs for Enhanced Document Ranking: Application in Systematic Reviews. *Big Data and Cognitive Computing* **2024**, *8*. https://doi.org/10.3390/bdcc8090110.
39. Kusa, W.; E. Mendoza, O.; Samwald, M.; Knoth, P.; Hanbury, A. CSMeD: Bridging the Dataset Gap in Automated Citation Screening for Systematic Literature Reviews. *Advances in Neural Information Processing Systems* **2023**, *36*, 23468–23484.
40. Kohandel Gargari, O.; Mahmoudi, M.H.; Hajisafarali, M.; Samiee, R. Enhancing Title and Abstract Screening for Systematic Reviews with GPT-3.5 Turbo. *BMJ Evidence-Based Medicine* **2024**, *29*, 69–70. https://doi.org/10.1136/bmjebm-2023-112678.
41. Matsui, K.; Utsumi, T.; Aoki, Y.; Maruki, T.; Takeshima, M.; Takaesu, Y. Human-Comparable Sensitivity of Large Language Models in Identifying Eligible Studies Through Title and Abstract Screening: 3-Layer Strategy Using GPT-3.5 and GPT-4 for Systematic Reviews. *Journal of Medical Internet Research* **2024**, *26*, e52758. https://doi.org/10.2196/52758.
42. Sanghera, R.; Thirunavukarasu, A.J.; Khoury, M.E.; O'Logbon, J.; Chen, Y.; Watt, A.; Mahmood, M.; Butt, H.; Nishimura, G.; Soltan, A. High-Performance Automated Abstract Screening with Large Language Model Ensembles, 2024. https://doi.org/10.48550/ARXIV.2411.02451.
43. Wang, S.; Scells, H.; Koopman, B.; Potthast, M.; Zuccon, G. Generating Natural Language Queries for More Effective Systematic Review Screening Prioritisation. In Proceedings of the SIGIR-AP 2023 - Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, 2023, pp. 73–83. https://doi.org/10.1145/3624918.3625322.
44. Hou, Y.; Zhang, J.; Lin, Z.; Lu, H.; Xie, R.; McAuley, J.; Zhao, W.X. Large Language Models Are Zero-Shot Rankers for Recommender Systems. In Proceedings of the Advances in Information Retrieval; Goharian, N.; Tonellotto, N.; He, Y.; Lipani, A.; McDonald, G.; Macdonald, C.; Ounis, I., Eds., Cham, 2024; pp. 364–381. https://doi.org/10.1007/978-3-031-56060-6_24.
45. Zhuang, H.; Qin, Z.; Hui, K.; Wu, J.; Yan, L.; Wang, X.; Bendersky, M. Beyond Yes and No: Improving Zero-Shot LLM Rankers via Scoring Fine-Grained Relevance Labels. In Proceedings of the Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers); Duh, K.; Gomez, H.; Bethard, S., Eds., Mexico City, Mexico, 2024; pp. 358–370. https://doi.org/10.18653/v1/2024.naacl-short.31.
46. Faggioli, G.; Dietz, L.; Clarke, C.L.A.; Demartini, G.; Hagen, M.; Hauff, C.; Kando, N.; Kanoulas, E.; Potthast, M.; Stein, B.; et al. Perspectives on Large Language Models for Relevance Judgment. In Proceedings of the Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, New York, NY, USA, 2023; ICTIR '23, pp. 39–50. https://doi.org/10.1145/3578337.3605136.
47. Wu, L.; Zheng, Z.; Qiu, Z.; Wang, H.; Gu, H.; Shen, T.; Qin, C.; Zhu, C.; Zhu, H.; Liu, Q.; et al. A Survey on Large Language Models for Recommendation. *World Wide Web* **2024**, *27*, 60. https://doi.org/10.1007/s11280-024-01291-2.
48. Thomas, P.; Spielman, S.; Craswell, N.; Mitra, B. Large Language Models Can Accurately Predict Searcher Preferences. In Proceedings of the Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, 2024; SIGIR '24, pp. 1930–1940. https://doi.org/10.1145/3626772.3657707.
49. Syriani, E.; David, I.; Kumar, G. Screening Articles for Systematic Reviews with ChatGPT. *Journal of Computer Languages* **2024**.
50. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models Are Few-Shot Learners, 2020, [2005.14165]. https://doi.org/10.48550/arXiv.2005.14165.

51. Sahoo, P.; Singh, A.K.; Saha, S.; Jain, V.; Mondal, S.; Chadha, A. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications **2024**. https://doi.org/10.48550/ARXIV.2402.07927.
52. Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; Poon, H. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthcare* **2021**, *3*, 2:1–2:23. https://doi.org/10.1145/3458754.
53. Huotala, A.; Kuutila, M.; Ralph, P.; Mäntylä, M. The Promise and Challenges of Using LLMs to Accelerate the Screening Process of Systematic Reviews. In Proceedings of the Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering, New York, NY, USA, 2024; EASE '24, pp. 262–271. https://doi.org/10.1145/3661167.3661172.
54. Kojima, T.; Gu, S.S.; Reid, M.; Matsuo, Y.; Iwasawa, Y. Large Language Models Are Zero-Shot Reasoners. In Proceedings of the Proceedings of the 36th International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 2024; NIPS '22, pp. 22199–22213.
55. Tseng, Y.M.; Huang, Y.C.; Hsiao, T.Y.; Chen, W.L.; Huang, C.W.; Meng, Y.; Chen, Y.N. Two Tales of Persona in LLMs: A Survey of Role-Playing and Personalization. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024; Al-Onaizan, Y.; Bansal, M.; Chen, Y.N., Eds., Miami, Florida, USA, 2024; pp. 16612–16631. https://doi.org/10.18653/v1/2024.findings-emnlp.969.
56. Liu, D.; Nassereldine, A.; Yang, Z.; Xu, C.; Hu, Y.; Li, J.; Kumar, U.; Lee, C.; Qin, R.; Shi, Y.; et al. Large Language Models Have Intrinsic Self-Correction Ability, 2024, [arXiv:cs/2406.15673]. https://doi.org/10.48550/arXiv.2406.15673.
57. Spillias, S.; Tuohy, P.; Andreotta, M.; Annand-Jones, R.; Boschetti, F.; Cvitanovic, C.; Duggan, J.; Fulton, E.; Karcher, D.; Paris, C.; et al. Human-AI Collaboration to Identify Literature for Evidence Synthesis, 2023. https://doi.org/10.21203/rs.3.rs-3099291/v1.
58. Robertson, S.; Zaragoza, H. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends® in Information Retrieval* **2009**, *3*, 333–389. https://doi.org/10.1561/1500000019.
59. Wang, W.; Wei, F.; Dong, L.; Bao, H.; Yang, N.; Zhou, M. MINILM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In Proceedings of the Advances in Neural Information Processing Systems, 2020, Vol. 2020-December.
60. SentenceTransformers Documentation — Sentence Transformers Documentation. https://www.sbert.net/.
61. OpenAI Platform. https://platform.openai.com.
62. Manning, C.D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge university press: Cambridge, England, 2008.
63. Kusa, W.; Lipani, A.; Knoth, P.; Hanbury, A. An Analysis of Work Saved over Sampling in the Evaluation of Automated Citation Screening in Systematic Literature Reviews. *Intelligent Systems with Applications* **2023/05/01/May 2023///**, *18*. https://doi.org/10.1016/j.iswa.2023.200193.
64. Feng, Y.; Liang, S.; Zhang, Y.; Chen, S.; Wang, Q.; Huang, T.; Sun, F.; Liu, X.; Zhu, H.; Pan, H. Automated Medical Literature Screening Using Artificial Intelligence: A Systematic Review and Meta-Analysis. *Journal of the American Medical Informatics Association* **2022**, *29*, 1425–1432. https://doi.org/10.1093/jamia/ocac066.
65. Kanoulas, E.; Li, D.; Azzopardi, L.; Spijker, R. CLEF 2017 Technologically Assisted Reviews in Empirical Medicine Overview. In Proceedings of the Conference and Labs of the Evaluation Forum, 2018.
66. Akinseloyin, O.; Jiang, X.; Palade, V. A Question-Answering Framework for Automated Abstract Screening Using Large Language Models. *Journal of the American Medical Informatics Association* **2024**, p. ocae166. https://doi.org/10.1093/jamia/ocae166.
67. Linzbach, S.; Tressel, T.; Kallmeyer, L.; Dietze, S.; Jabeen, H. Decoding Prompt Syntax: Analysing Its Impact on Knowledge Retrieval in Large Language Models. In Proceedings of the Companion Proceedings of the ACM Web Conference 2023, New York, NY, USA, 2023; WWW '23 Companion, pp. 1145–1149. https://doi.org/10.1145/3543873.3587655.