Li *et al. Systematic Reviews* (2024) 13:219
https://doi.org/10.1186/s13643-024-02609-x

Systematic Reviews

RESEARCH

Open Access

Evaluating the effectiveness of large language models in abstract screening: a comparative analysis

Michael Li[1], Jianping Sun[2] and Xianming Tan[1,3*]

Abstract

Objective This study aimed to evaluate the performance of large language models (LLMs) in the task of abstract screening in systematic review and meta-analysis studies, exploring their effectiveness, efficiency, and potential integration into existing human expert-based workflows.

Methods We developed automation scripts in Python to interact with the APIs of several LLM tools, including ChatGPT v4.0, ChatGPT v3.5, Google PaLM 2, and Meta Llama 2, and latest tools including ChatGPT v4.0 turbo, ChatGPT v3.5 turbo, Google Gemini 1.0 pro, Meta Llama 3, and Claude 3. This study focused on three databases of abstracts and used them as benchmarks to evaluate the performance of these LLM tools in terms of sensitivity, specificity, and overall accuracy. The results of the LLM tools were compared to human-curated inclusion decisions, gold standard for systematic review and meta-analysis studies.

Results Different LLM tools had varying abilities in abstract screening. Chat GPT v4.0 demonstrated remarkable performance, with balanced sensitivity and specificity, and overall accuracy consistently reaching or exceeding 90%, indicating a high potential for LLMs in abstract screening tasks. The study found that LLMs could provide reliable results with minimal human effort and thus serve as a cost-effective and efficient alternative to traditional abstract screening methods.

Conclusion While LLM tools are not yet ready to completely replace human experts in abstract screening, they show great promise in revolutionizing the process. They can serve as autonomous AI reviewers, contribute to collaborative workflows with human experts, and integrate with hybrid approaches to develop custom tools for increased efficiency. As technology continues to advance, LLMs are poised to play an increasingly important role in abstract screening, reshaping the workflow of systematic review and meta-analysis studies.

Keywords Large language models, Abstract screening, Automated text analysis, ChatGPT, Google PaLM, Meta Llama

---

*Correspondence:
Xianming Tan
xianming@email.unc.edu
Full list of author information is available at the end of the article

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Introduction

A systematic review is a rigorous form of research that collates and synthesizes all existing evidence on a specific research question [1]. It stands as a cornerstone not just in medical research but across diverse academic disciplines. Unlike traditional literature reviews, systematic reviews follow a comprehensive and standardized process, such as the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guideline [2], designed to minimize bias and ensure reproducibility. Therefore, these reviews are recognized as one of the zenith levels of evidence in evidence-based research [3] and play a pivotal role in shaping clinical guidelines, healthcare policies, and informing medical decisions [4].

Commencing with a well-articulated research question, a typical systematic review launches an exhaustive search strategy that subsequently sweeps through databases, such as PubMed and Embase, supplemented by additional sources such as clinical trial registries and pertinent article reference lists, with aims to capture all relevant studies and mitigate bias. Predetermined inclusion and exclusion criteria guide the subsequent screening and selection of studies, encompassing facts like study design, patient demographics, and intervention types. Reviewers, working independently, appraise each study's eligibility, reconciling disagreements through discussions or third-party reviews. Data extraction and synthesis follow, either through meta-analysis or narrative synthesis, depending on the heterogeneity of the selected studies.

The practice of conducting systematic reviews has gained substantial popularity with considerable demand within the academic community. A notable reference [5] to this trend found that approximately 75 trials and potentially 11 systematic reviews are disseminated daily, based on data from around 2010. Moreover, a query for "systematic review" in the Google Scholar database yields approximately 17,000 entries that have been published within the year 2023 alone, which translates into about 51 systematic reviews per day, as observed on November 26, 2023. This expanding volume of literature underscores the critical role that systematic reviews play in the consolidation of research findings across various fields of study.

Despite their pivotal role, executing systematic reviews remains a formidable task due to the abstract screening, a key phase that can be overwhelmingly time-consuming due to its volume. For example, in Polanin et al. [6], the authors reported that the research staff screened 29,846 abstracts independently (14,923 unique citations were double screened) over the course of 189 days. In addition, the variability in reporting standards, use of jargon, and varied study designs can further complicate the abstract screening process [7]. Nevertheless, the repetitive nature of the task, combined with cognitive fatigue, can lead to human errors [8, 9]. Recent advancements in machine learning (ML) and deep learning propose possible solutions to these challenges. However, traditional ML models, while promising, require domain-specific training, a time-consuming process that often demands manual labeling of datasets [10].

Advancements in natural language processing (NLP) and artificial intelligence (AI) are opening doors to address challenges in systematic reviews. Large language models (LLMs) like ChatGPT [11], PaLM [12], Llama [13], and Claude [14] are capturing the research community's attention. Their collective potential, especially their capability to operate without exhaustive domain-specific training, makes them prime candidates for revolutionizing the systematic review process.

While each of the aforementioned AI tools brings its unique capabilities to the table, the fundamental question remains: How do they stack up, both individually and collectively, in the realm of abstract screening, against the human expert-based process? In pursuit of answers, this research seeks to thoroughly investigate the potential of ChatGPT, Google PaLM, Llama, and Claude in automating the crucial abstract screening phase integral to systematic reviews. Our goal is to rigorously compare the performance of these advanced AI-driven methods with existing machine learning (ML)-based approaches. In doing so, we aim to develop AI strategies that masterfully balance efficiency and accuracy with minimal human intervention, ultimately transforming systematic review practice across disciplines.

The use of NLP for abstract screening is established [10, 15]. However, the application of LLMs specifically for this task is a nascent field [16, 17]. This emerging area offers significant potential to improve efficiency and accuracy. Our study aims to fill this gap by providing a comprehensive analysis of LLM capabilities in abstract screening, laying the groundwork for future research and application. This is particularly relevant considering the rapid evolution of this technology, highlighting its potential to streamline systematic reviews now and in the future.

The remainder of this paper is structured to provide a comprehensive exploration of our topic. We begin with an in-depth examination of existing methods for abstract screening, including both manual and NLP-based approaches, laying the groundwork for understanding the current state of the field. We then introduce the use of large language model (LLM) tools for abstract screening, detailing our experimental design to meticulously evaluate their performance in this context. Subsequent sections present our empirical findings and results, shedding light on the capabilities and limitations of the AI tools in question. Finally, we engage in a thoughtful discussion, reflecting on the implications of our findings and considering the future trajectory of abstract screening in systematic reviews.

Existing approaches to abstract screening in systematic reviews

In the vast realm of systematic reviews, the critical task of abstract screening serves as a foundational step in curating the highest quality of evidence [2]. However, this process often presents significant challenges due to the involvement of sifting through large volumes of literature to identify those that align with predefined criteria. Over time, various methodologies, ranging from manual evaluations to sophisticated AI-driven techniques, have been proposed to address the complexities of this task. In this section, we first describe the existing approaches on their operational mechanisms and associated advantages and disadvantages.

Manual approach

Historically, the process of abstract screening was firmly rooted in manual evaluations. In this conventional approach, individual reviewers would scrutinize each abstract against predefined criteria [1]. The meticulous nature of this method required that multiple experts independently evaluate the literature to ensure both reliability and reduced biases [8]. While the depth of human expertise brought about nuanced understanding, the manual nature of this method made it both time-consuming and, at times, prone to human error [6, 9].

NLP-based approach

As technology evolved, the field witnessed the incorporation of natural language processing (NLP) to automate abstract screening [10]. In this framework, abstract text undergoes preprocessing and vectorization. Supervised machine learning models, notably the support vector machine (SVM) and the random forest (RF), are then trained on this vectorized data to classify literature based on specific criteria [15]. The strength of this approach lies in its potential for efficiency. However, its efficacy and accuracy hinges heavily on the availability of a well-curated, labeled training set.

Zero-shot classification

A more recent and innovative approach is zero-shot classification, which was notably highlighted by Xian et al. [18]. Eschewing the need for an extensive labeled training dataset, zero-shot classification offers the allure of categorizing abstracts without prior domain-specific training. By calculating a probability score for each abstract, researchers obtain a dynamic measure of its alignment with predefined categories. While it does not require training for the machine learning model, the tradeoff is a loss of sensitivity and potential omissions of pertinent studies [19]. In this study, for a given set of abstracts, we first obtain embeddings (using OpenAI's text embedding API, the babbage model) for abstracts and for a pre-specified description of an ideal study to include. We indeed use the inclusion/exclusion criteria (see "Examples of abstract screening by using LLMs," "Automated workflow for streamlining abstract screening via ChatGPT and other tools," and "Study design" sections) as the description of an ideal study to include. We then compute the cosine similarity score between each abstract (its embedding) and the pre-specified description (its embedding). We claim the top 10% (in terms of similarity) of abstracts as positive (i.e., a study should be included).

Hybrid approach

To reconcile the strengths and limitations of these models, a hybrid methodology has emerged [18]. It cleverly marries the rapid categorization capabilities of zero-shot classification with the precision of traditional machine learning models such as SVM and RF. Here, zero-shot classification provides an initial curation of all abstracts in the training set. Then, based on zero-shot classification's curation, the abstracts with classification score higher than a pre-specified threshold are selected to undergo manual review to rectify zero-shot classification. The rectified classification then serves as a foundation for training traditional models, which are then employed for screening the broader dataset (e.g., in the testing set). The hybrid approach balances the speed of zero-shot and the precision of traditional ML, and potentially offers enhanced accuracy at reduced human efforts. However, this approach involves the use of multiple methodologies and still relies on well-curated, labeled training data (in this case, a subset of the whole training set). In this study, for a given set of abstracts, we rectify the "positive" abstracts (i.e., the top 10% of abstracts) identified by zero-shot, and then fit a classification model for rectified labels to the text embeddings of the corresponding abstracts using SVM. This learned classification model is then used to predict the class of other abstracts. We added the R code for the zero-shot and hybrid approaches to the GitHub repository https://github.com/mikeli380/LLMAbstractScreening.

Active learning approach

Active learning [20–22] is an innovative approach to machine learning that optimizes the training process by allowing the model to selectively query a human annotator for labels on the most informative data points. This method is particularly advantageous in scenarios where labeled data is scarce or expensive to obtain. Active learning models can start with minimal datasets, often requiring only one relevant and one irrelevant abstract, making them particularly suitable for tasks such as abstract screening, where the acquisition of large labeled datasets can be prohibitive.

The main advantage of active learning is its efficiency. By focusing on the most informative samples, it reduces the amount of data that needs to be labeled while still effectively training the model. This can significantly reduce the time and resources required for the annotation process. However, the effectiveness of active learning depends heavily on the initial selection of samples and the criteria used to determine the informativeness of subsequent data points. If not well calibrated, the model may request labels for data that do not significantly improve its performance, resulting in an inefficient use of resources. In addition, the iterative nature of querying and updating the model based on new labels can introduce complexity into the training process, potentially requiring more sophisticated infrastructure and oversight than traditional supervised learning methods.

While this study did not test active learning approaches for abstract screening, readers interested in exploring this methodology further are referred to [23] for detailed information on the application of active learning in abstract screening contexts.

Large language models

Amidst the evolving methodologies described in the "Existing approaches to abstract screening in systematic reviews" section, modern AI tools based on large language models (LLMs), such as ChatGPT, PaLM, Llama, and Claude, are emerging as potential game-changers. Grounded in advanced language processing capabilities, these tools can be tailored to evaluate abstracts against nuanced criteria and offer detailed assessments and classifications. Their prowess signals transformative potential for abstract screening. In this section, we first present two examples to illustrate the potential of LLMs in the context of abstract screening, and then precede with our in-depth discussion on the study's workflow and design that aims to critically investigate the performance of LLMs in this domain.

Examples of abstract screening by using LLMs

In this subsection, we demonstrate the process of LLM-based abstract screening using ChatGPT with two specified abstracts: (1) Millard et al. [24] and (2) Zawia et al. [25]. Using other LLM tools such as Llama, Google PaLM, or Claude for abstract screening is similar and will be skipped for brevity.

In constructing the prompts, we adopted a standardized approach to mimic a typical interaction between a senior researcher and a research assistant. Each prompt was designed to include three critical elements: a brief statement of the topic under review, the exact eligibility criteria as specified in the study guidelines, and the text of the abstract to be reviewed. This approach was designed to simulate a realistic and practical scenario that reflects common academic practices. In addition, to ensure consistency and objectivity in the responses, we required that the results of the LLM be in a binary "yes" or "no" format, directly indicating whether each abstract met the inclusion criteria. This format was chosen to eliminate any ambiguity in the model's decisions and to facilitate straightforward analysis of its accuracy.

Abstract (1)

ChatGPT prompt: "I would like you to help me with conducting a systematic review on animal depression. I will provide the title and abstract for one journal article and would like you to screen the paper for inclusion. Here are the inclusion criteria: Does the study provide primary data of an animal model of depression or depressive-like phenotype with an appropriate control group? Here is the abstract:

> *Wistar-Kyoto (WKY) rats show endogenous depressive behavior that can be reversed by antidepressants. Given that WKYs exhibit decreased sensitivity to some antidepressants and treatment-resistant depressed patients often show hypothalamic-pituitary-thyroid (HPT) dysregulation, we examined the behavioral and HPT hormonal responses of WKYs to altered thyroid status. "Euthyroid" WKYs had elevated basal plasma TSH and T(3) levels as compared to Wistars. Hypothyroidism increased TSH levels more in WKYs than in Wistars and increased response latency in the open field test (OFT) of WKYs only. Administration of T(4) and T(3) suppressed plasma TSH equally in both strains. Wistars responded to increased T(3) levels with decreased response latency and increased activity in the OFT, but increased immobility in the forced swim test. In contrast, WKYs responded only to the high T(3) levels with decreased response latency in the OFT. These results suggest the existence of a decreased central nervous system sensitivity to thyroid hormones in WKYs that could be related to their depressive behavior."*

ChatGPT v4 response (accessed on 10/15/2023):

Given the abstract and the inclusion criteria, this study appears to meet the inclusion criteria for your systematic review on animal models of depression.

Consideration for inclusion:

1. Animal model and depressive-like phenotype:
    - The study utilizes Wistar-Kyoto (WKY) rats which are mentioned to show endogenous depressive behavior.
    - This study investigates behaviors and hormonal responses associated with depression in this particular animal model, with a focus on the role of thyroid hormones.
2. Primary data and appropriate control group:
    - The study provides primary data on behavioral and hormonal responses of WKY rats, comparing them to Wistars (likely serving as the control group), under altered thyroid status, allowing for comparisons between the depressive-like phenotype and controls.
    - The utilization of both "Euthyroid" WKYs and Wistars as controls helps in effectively investigating the behavioral changes and responses in WKY rats due to alterations in thyroid status.
3. Behavioral test indicating depression:
    - Depressive behavior and responses are assessed using standardized tests like the open field test (OFT) and the forced swim test, which are conventional methods to investigate depressive-like behaviors in animal models.
    - The mention of antidepressant's impact on WKY rats indicates a focus on depressive-like states.
4. Relevance to human condition:
    - Insights gained from this study on decreased central nervous system sensitivity to thyroid hormones and its possible relation to depressive behavior in WKYs can have implications in understanding treatment-resistant depression in humans, making it valuable for the review.

Conclusion:
Given the provided abstract, this study does seem to fit the inclusion criteria for a systematic review on animal depression, as it provides primary data on an animal model demonstrating depressive-like behavior, includes an appropriate control group, and employs standardized tests to study the depressive-like states and their relation to thyroid hormones. The detailed exploration of depressive behaviors and hormonal imbalances in relation to altered thyroid status can offer valuable insights for understanding and treating depression.

Abstract (2)

ChatGPT prompt: "I would like you to help me with conducting a systematic review on animal depression. I will provide the title and abstract for one journal article and would like you to screen the paper for inclusion. Here are the inclusion criteria: Does the study provide primary data of an animal model of depression or depressive-like phenotype with an appropriate control group? Here is the abstract:

> *The induction of the activity of cerebral ornithine decarboxylase (EC 4.1.1.17) and mRNA by electrical stimulation exhibits regional differences. The effects of the enzyme inhibitor difluoromethylornithine on these regional variations was examined. Administration of this inhibitor resulted in pronounced depression of both basal and induced activity of ornithine decarboxylase in the hippocampus. Basal activity of the enzyme in the neocortex and the cerebellum appeared to be resistant to difluoromethylornithine but the induced enzyme activity was sensitive to the effects of this inhibitor. Susceptibility to difluoromethylornithine may be directly correlated with a slower turnover rate for ornithine decarboxylase. These results suggest that ornithine decarboxylase in the hippocampus may possess a longer half-life than its counterparts in other regions of the brain. Pretreatment with difluoromethylornithine had no effect on the induced ornithine decarboxylase mRNA in the neocortex. Thus, elevated activity of ornithine decarboxylase enzyme, due to electrical stimulation, appears to not have any effect on either the transcription or the decay rate of the induced ornithine decarboxylase mRNA. These findings support the concept of region-specific regulation of cerebral ornithine decarboxylase."*

ChatGPT v4 response (accessed on 10/15/2023):

Given the provided abstract and the inclusion criteria, this study does not appear to meet the criteria for inclusion in a systematic review on animal models of depression or depressive-like phenotypes with an appropriate control group.

Consideration for inclusion:

1. Depression or depressive-like phenotype:
    - The abstract does not provide any information or data related to depression or depressive-like behavior or phenotypes in animals. It is mainly focused on the activity of cerebral ornithine decarboxylase and its regional differences in response to electrical stimulation and inhibitor administration.
2. Animal model:
    - There is no mention of the use of any specific animal model or any behavioral assays to assess depressive-like symptoms or states in the provided abstract.
3. Behavioral assessment and primary data:
    - The abstract does not contain any details about behavioral assessments related to depression or depressive-like states. It primarily provides data on the effects of enzyme inhibitor difluoromethylornithine on cerebral ornithine decarboxylase activity and mRNA, but does not relate these findings to depression or depressive-like behaviors.
4. Relevance to depression:
    - While the study explores biochemical aspects in brain regions like the hippocampus, which is implicated in mood disorders like depression, the connection between the findings and depressive symptoms or states is not established in the provided abstract.

Conclusion:
This article does not seem to fit the inclusion criteria for a systematic review on animal models of depression or depressive-like phenotypes as it lacks primary data on depressive-like behaviors or phenotypes in animals, and there is no mention of the use of animal models representative of depression or an appropriate control group for such a comparison.

From the above two examples presented, it is evident that ChatGPT adeptly comprehends the posed queries and efficiently extracted the inclusion and exclusion criteria. It not only makes clear decisions regarding study inclusion but also articulates its rationale by cross-referencing each criterion.

It is worth noting that the same ChatGPT prompt might result in slightly different responses between different users due to the variability in its response generation, but the conclusion for the eligibility of an abstract is typically consistent. In addition, enhancing the specificity of the criteria within the prompts can further refine the screening process. To illustrate, consider the following enhanced prompts:

Enhanced ChatGPT prompt for abstract (1): "I seek assistance with a systematic review focused on animal depression. I'll provide the abstract of a particular journal article and would appreciate an assessment for its inclusion based on the following criteria:

1. The study employs an animal model reflecting depression or a depressive-like phenotype.
2. It incorporates an appropriate control group.
3. Primary data is presented within the study.

and here are the exclusion criteria:

1. The study bypasses the use of an animal model.
2. The utilized animal model doesn't characterize depression or a depressive-like phenotype.
3. Absence of a control group within the study.
4. The study is primarily a review, devoid of primary data.

Here is the abstract: ...".

ChatGPT v4 response (accessed on 10/15/2023):

Let's assess the provided abstract based on the inclusion and exclusion criteria:

Inclusion criteria:

1. Animal model reflecting depression: The study uses Wistar-Kyoto (WKY) rats, which are described as showing "endogenous depressive behavior." Thus, this criterion is met.

2. Appropriate control group: The study compares WKY rats to Wistars (another strain), which serves as a comparison or control group. This criterion is met.

3. Primary data: The abstract mentions specific results such as the elevated basal plasma TSH and T(3) levels in "Euthyroid" WKYs compared to Wistars and the different responses in the open field test (OFT). This suggests the presence of primary data. Thus, this criterion is met.

Exclusion criteria:

1. Bypasses the use of an animal model: The study uses WKY rats as an animal model. Thus, this criterion is not met, which is good for inclusion.

2. Animal model doesn't characterize depression: The WKY rats show "endogenous depressive behavior." Thus, this criterion is not met, which again supports inclusion.

3. Absence of a control group: The study does use a control group (Wistars). Thus, this criterion is not met, supporting inclusion.

4. Study is primarily a review: The abstract presents experimental data and does not indicate that the study is a review. Thus, this criterion is not met, which is favorable for inclusion.

Conclusion:

Given that the abstract meets all the inclusion criteria and none of the exclusion criteria, it should be considered in your systematic review on animal depression.

With these revised prompts applied to the previously discussed abstracts, the conclusions remain consistent. Nonetheless, we contend that this enhanced structure not only provides a more rigorous framework but also aligns more closely with widely accepted practices in systematic review criteria definition.

Automated workflow for streamlining abstract screening via ChatGPT and other tools

For systematic reviews, particularly when dealing with voluminous data, efficient workflows are paramount. The ChatGPT API (application programming interface) offers a dynamic solution, enabling the automation of abstract screening on a large scale, which circumvents the labor-intensive process of manually inputting abstracts into a chatbot interface. In this subsection, we present an automated workflow for streamlining abstract screening via ChatGPT. Note, though this automated workflow uses ChatGPT as the platform, analogous workflows work for other AI platforms like PaLM, Llama, and Claude.

Automated workflow:

1. Data collection: The preliminary step entails accumulating a list of titles and abstracts. By utilizing carefully crafted keywords, we retrieve these from PubMed and other pertinent databases. This comprehensive approach ensures the potential inclusion of all relevant studies for a subsequent detailed screening. It is worth noting that while this list is expansive, most of these studies may not find their way into the final meta-analysis post-screening.
2. Automation through Python: We have devised a python script aimed at harnessing the capabilities of ChatGPT for evaluating the amassed abstracts.
    a. This script interacts with the ChatGPT API (specifically, the GPT-4 version) and, when furnished with tailored prompts, extracts structured responses from ChatGPT.
    b. Typically, the AI's response commences with a succinct summary, delves into explanations aligned with each criterion, and concludes with a decisive judgment, as exemplified by the examples in the "Examples of abstract screening by using LLMs" section.
    c. This automated process efficiently saves ChatGPT's verdicts on each abstract for ensuing analyses. For instance, it extracts the final decisions regarding the inclusion or exclusion of studies and determines the stance on each pre-specified criterion for every abstract, as exemplified by the last example in the "Examples of abstract screening by using LLMs" section.

  d. Additionally, to ascertain the efficiency and cost-effectiveness of this methodology, the script also monitors the time, token usage, and the financial implications of querying the OpenAI API.

     In essence, we envision this procedure as delegating the meticulous task of poring over scientific summaries to an AI assistant. This virtual entity meticulously sifts through each summary, determining its alignment with stipulated criteria.

3. Tuning parameters in the ChatGPT API: The effectiveness of the ChatGPT API is not only dependent on the input data; it is also significantly influenced by adjustable parameters that can refine the responses. Parameters such as temperature, top k, and top p critically affect model performance by modulating the randomness and focus of the output. While fine-tuning these parameters can improve results, it requires significant technical expertise and resources. Defaults, which are rigorously tested by developers, strike a balance between output quality and ease of use, making LLMs more accessible to a wider range of users without the need for complex parameter optimization. While customization holds promise for custom applications, the default settings provide an efficient and practical solution that facilitates wider adoption of LLM technologies.

   Given the complexity of fine-tuning these parameters to optimize performance specifically for abstract screening, our study primarily used the recommended default settings for these parameters as provided by the respective platforms (detailed in Table 1). This approach was chosen to maintain the feasibility of our experiments and to ensure that our findings are applicable to typical deployment scenarios.

The automated workflow described applies to other LLM tools or different versions of the same tools. Throughout our study, we have tested several popular LLM tools available to us. The specific LLM models and their versions used at the time of our first submission are detailed in Table 1, along with the tuning parameters.

The field of large language models (LLMs) has evolved rapidly since we started this study in 2023. New models are frequently released, and existing versions are constantly updated. To account for these advances, we have expanded our analysis to include results from the latest versions of previously studied models, as well as a few entirely new ones. As a best practice, we set the temperature parameter to 0 for all latest models. Table 1 now includes these latest models and their versions.

A brief explanation of these parameters is as follows:

- Temperature: The temperature controls the randomness of the outputs, with a range from 0 to 2. A temperature value greater than 1 is random and 0 is deterministic. The maximum temperature of 2 gives the most creative and variable outputs.
- Max length: The max length is the maximum number of tokens for the model to generate as a response. A single word is generally 2–3 tokens.
- Stop sequences: This parameter controls which tokens or phrases will stop the LLM from generating more text.
- Top p: When generating text, the model samples from the top p percentage of most likely tokens to generate. The top p is the cumulative probability cutoff for the model's selection of tokens to generate.

Table 1 LLM models/versions and tuning parameters used in our study

LLM/parameters	Temperature	Maximum length	Stop sequences	Top p	Top k	Frequency penalty	Presence penalty
Originally used LLM models (versions)							
ChatGPT4.0 (gpt-4.0–0613)	0	1000	None	1	–	0	0
ChatGPT3.5 (gpt-3.5-turbo-1106)	0	1000	None	1	–	0	0
Google PaLM 2 (text-bison-001)	0.25	2048	None	–	–	–	–
Meta Llama 2 (llama-2-70b-chat)	0.75	500	None	1	–	–	–
Latest LLM models (versions)							
ChatGPT-4-Turbo (gpt-4-turbo-2024–04-09)	0	4096	None	1	–	0	0
ChatGPT-3.5-Turbo (gpt-3.5-turbo-0125)	0	4096	None	1	–	0	0
Gemini-1.0-pro (gemini-1.0-pro-latest)	0	8092	None	1	1	–	–
Llama 3 (llama-3-70b-instruct)	0	4096	None	0.9	50	0.2	1.15
Claude 3 Opus (Claude-3-opus-20240229)	0	3000	None	–	–	–	–

Lower top p values mean sampling from a smaller, more top-weighted nucleus.

- Top k: When generating text, the model samples from the top k most likely tokens. When deciding the next word to output, the model will select the most likely word when top k is lower.
- Frequency penalty: This frequency penalty parameter controls how the LLM penalizes tokens that have already appeared in the input and output text. A frequency penalty of 0 implies that the frequency of the tokens does not impact the LLM and will generate them based on their probability.
- Presence penalty: This parameter controls how the LLM penalizes tokens that have not been used recently. A presence penalty of 0 means that the LLM does not care about the recency of the tokens and will generate them based on their probability.

Study design

In our pursuit to assess ChatGPT's proficiency in abstract screening, we selected certain benchmark databases that have existing performance data from other methodologies. This selection aids in a comparative analysis of performance.

In selecting the systematic reviews for our study, we used a systematic approach guided by specific criteria to ensure relevance and reliability. These studies were selected from the publicly available SYNERGY [23] dataset, which contains 26 systematic reviews from different disciplines. Key selection criteria included:

> Clarity and conciseness of eligibility criteria: The selected studies had well-defined and explicit eligibility criteria. This clarity is essential for accurate replication of the study selection process, which is critical for assessing the performance of LLM tools in an analogous real-world application.
>
> Completeness and cleanliness of data: We ensured that the selected reviews had complete datasets, with all necessary information on included and excluded studies clearly documented, minimizing the risk of ambiguities affecting our analysis.

In addition, to comply with the AMSTAR-2 [26] guidelines, in particular point 5, we reviewed the methodologies of these reviews to confirm the selection of studies was performed in duplicate and disagreements were resolved by consensus. While our analysis assumes that these systematic reviews adhere to high standards, we recognize the inherent limitations of using pre-existing datasets as a proxy for gold standards in the discussion section.

Databases

We picked the following 3 databases from the publicly available SYNERGY dataset [23]:

1. Bannach-Brown 2016 [27]—topic: use of animal models to study depressive behavior
    - Human-curated (gold standard) results: 1258 excluded abstracts and 230 included abstracts.
    - We randomly selected 100 excluded abstracts and 100 included abstracts for screening by LLM tools.
2. Meijboom 2021 [28]—topic: retransitioning of etanercept in patients with a rheumatic disease
    - Human-curated (gold standard) results: 599 excluded abstracts and all 32 included abstracts.
    - We randomly selected 100 excluded abstracts and 32 included abstracts for screening by LLM tools.
3. Menon 2022 [29]—topic: methodological rigor of systematic reviews in environmental health
    - Human-curated (gold standard) results: 896 excluded abstracts and 73 included abstracts.
    - We randomly selected 100 excluded abstracts and all 73 included abstracts for screening by LLM tools.

For each chosen database, abstracts were categorized as either "cases" (those included based on a gold standard) or "controls" (those excluded per the gold standard). From each category, we randomly selected 100 abstracts (we use all abstracts if there are less than 100 abstracts in that category in a database). These abstracts underwent evaluation by ChatGPT (v4.0) as per our established workflow. Subsequently, ChatGPT's decisions were juxtaposed against the gold standard to determine sensitivity, specificity, and overall accuracy. The same abstracts were also processed using other LLM tools as listed in Table 1 to record their respective verdicts.

Statistical analysis

To quantify the efficacy of ChatGPT and other AI tools for each database, we calculated the following metrics: (1) sensitivity, (2) specificity, and (3) overall accuracy, where sensitivity is defined as the number of true positives divided by the sum of true positives and false negatives, specificity as the number of true negatives divided by the sum of true negatives and false positives, and accuracy as sum of true positives and true negatives divided by the total number of abstracts. For each metric, associated 95% confidence intervals were also determined. Although it is very common in the field to report F1 score, recall rate, and precision, we believe it is more appropriate to report sensitivity and specificity given this study design. In addition, F1 score, recall rate, and precision can be derived from sensitivity and specificity.

Furthermore, to explore the potential of a unified decision-making process, we combined the decisions from all AI tools using a voting mechanism. The majority decision across the tools was taken as the final verdict for each abstract. For this consolidated approach, we again computed sensitivity, specificity, overall accuracy, and the associated 95% CIs for each database. We also explore the use of latent class analysis (LCA), a model-based clustering approach, to derive consolidated decisions. More details on this LCA approach are provided in the "Beyond majority voting" section.

For a given database, 100 cases and 100 controls yield a two-sided 95% confidence interval with a half-width equal to 0.048 when the underline sensitivity (specificity) is approximately 95%.

All statistical analyses were conducted using the R statistical software (version 4.3.1). All tests were two-sided with an alpha level set at 0.05 unless otherwise mentioned.

To improve the transparency and reproducibility of studies using AI tools, we have included the TRIPOD + AI checklist [30] in our report. This checklist has been adapted to reflect the specifics of our research, which focuses on the evaluation of large language models for abstract screening rather than diagnostic or prognostic modeling. The completed checklist is presented in Table S1, to provide readers with a comprehensive overview of our study's adherence to established reporting standards.

Results

We present the results for each of the 3 databases. For each, we first present the prompts we used when we called LLM tools to screen an abstract, then present the performance data (accuracy, sensitivity, and specificity for each method or LLM tool), followed by a summary of the performance, and a comparison of the performance of the different methods to that based on ChatGPT v4.0.

Results on the Bannach-Brown 2016 database (see Table 2)

The prompts we used for screening abstracts in this database are as follows:

> *Conduct a systematic review on animal depression. I provide the title and abstract for one journal article. Provide an overall assessment based on eligibility criteria with only one word answer yes or no with no*

Table 2 Results on the Bannach-Brown 2016 database

	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Zero-shot method[a,b]	0.839 [0.820, 0.858]	0.304 [0.246, 0.368]	0.937 [0.922, 0.950]
Hybrid method[b]	**0.961 [0.950, 0.970]**	0.843 [0.790, 0.888]	**0.982 [0.974, 0.989]**
ChatGPT (v4.0)	**0.945 [0.904, 0.972]**	0.930 [0.861, 0.971]	0.960 [0.901, 0.989]
ChatGPT (v3.5)	0.905 [0.856, 0.942]	0.940 [0.874, 0.978]	0.870 [0.788, 0.929]
Google PaLM 2[c]	0.900 [0.850, 0.938]	0.850 [0.765, 0.914]	0.950 [0.887, 0.984]
Meta Llama 2	0.780 [0.716, 0.835]	0.950 [0.887, 0.984]	0.610 [0.507, 0.706]
Majority voting	0.915 [0.867, 0.950]	**0.960 [0.901, 0.989]**	0.870 [0.788, 0.929]
LCA model[d]	0.945 [0.904, 0.972]	0.930 [0.861, 0.971]	0.960 [0.901, 0.989]
Results from the latest LLM models			
ChatGPT-3.5-Turbo	0.870 [0.815, 0.913]	0.830 [0.742, 0.898]	0.910 [0.836, 0.958]
ChatGPT-4-Turbo	0.830 [0.771, 0.879]	0.670 [0.569, 0.761]	**0.990 [0.946, 1.000]**
Gemini-1.0-pro	0.870 [0.815, 0.913]	0.750 [0.653, 0.831]	**0.990 [0.946, 1.000]**
Llama 3	0.910 [0.861, 0.946]	**0.930 [0.861, 0.971]**	0.890 [0.812, 0.944]
Claude 3 Opus	**0.920 [0.873, 0.954]**	0.900 [0.824, 0.951]	0.940 [0.874, 0.978]

[a] Zero-shot is based on Open AI's babbage embedding, and hybrid then trained the model based on curated label for the top 10% cases identified by zero-shot

[b] Performance summaries for zero-shot and hybrid method are based on all 230 positive abstracts and all 1258 negative abstracts

[c] Google PaLM 2 might generate empty responses to some abstracts, and we assumed all null outputs as missing. "Majority voting" refers to majority voting without using decisions from Google PaLM2

[d] The LCA model part, which can be seen as a more sophisticated version of "majority voting," is explained in the "Beyond majority voting" section. These notes also apply to Tables 3 and 4

> *explanation. Then, for each inclusion or exclusion criterion, answer with only one word, yes if it is included by the inclusion criterion or excluded by the exclusion criterion, and answer no if it does not meet the inclusion criterion or not excluded by the exclusion criterion. After answering all the criteria with yes or no, then provide an overall explanation.*
>
> *Here is the eligibility criteria: Inclusion Criteria: 1. Any article providing primary data of an animal model of depression or depressive-like phenotype with an appropriate control group (specified above). 2. Animals of all ages, sexes and species, where depression-like phenotype intended to mimic the human condition have been induced. Including animal models where depressive-like phenotypes are induced in the presence of a comorbidity (e.g. obesity or cancer). 3. All studies that claim to model depression or depressive-like phenotypes in animals. Studies that induce depressive behavior or model depression and that also test a treatment or intervention (prior or subsequent to model induction), with no exclusion criteria based on dosage, timing or frequency. 4. Studies measuring behavioral, anatomical and structural, electrophysiological, histological and/or neurochemical outcomes and where genomic, proteomic or metabolomic outcomes are measured in addition to behavioral, anatomical, electrophysiological, histological or neurochemical outcomes. Exclusion Criteria: 1. Review article, editorials, case reports, letters or comments, conference or seminar abstracts, studies providing primary data but not appropriate control group. 2. Human studies and ex vivo, in vitro or in silico studies. Studies will be excluded if authors state an intention to induce or investigate only anxiety or anxious behavior. Studies will be excluded if there is no experimental intervention on the animals (e.g. purely observational studies). 3. Studies that investigate treatments or interventions, but no depressive behavior or model of depression is induced (e.g. toxicity and side-effect studies). 4. Where metabolic outcome measures are the primary outcome measure of a study. Where genomic, proteomic, metabolic or metabolomic outcomes are the sole outcome measures in a study, they will be excluded.*

Here is the abstract:

Abstract X

Among all the LLM tools we tested, ChatGPT v4.0 stood out with the highest accuracy (0.945) and specificity (0.960), and satisfactory sensitivity (0.930). Combined decision using major voting seemed to improve sensitivity (0.970) considerably but did not improve specificity (0.870) much. Comparatively, the zero-shot method was less effective across these metrics, while the hybrid method attained superior accuracy (0.961) and specificity (0.982), albeit with reduced sensitivity (0.843). We acknowledge that our findings regarding the zero-shot and hybrid methods differ from those reported in [18] for the same set of abstracts. This discrepancy could be attributed to the use of distinct sets of embeddings in our analysis compared to the previous study.

For this database, the newer versions of ChatGPT (3.5-Turbo and 4-Turbo) did not improve performance over ChatGPT (v4.0). Gemini-1.0-pro (vs. PaLM 2) and Llama 3 (vs. Llama 2) improved over their older versions, but did not surpass the performance of ChatGPT (v4.0). Claude 3 performed well, but still did not surpass the performance of ChatGPT (v4.0).

Comparison between LLM tools. We compared the performance (sensitivity and specificity) between ChatGPTv4.0 and other LLM tools using the McNemar test and found that ChatGPTv4.0 performed significantly better (*p* value = 0.002) than Google PaLM 2 in terms of sensitivity; ChatGPTv4.0 performed significantly better than ChatGPTv3.5 (*p* value = 0.008) and better than Llama-2 (*p* value < 0.001) in terms of specificity. Combining the decisions of different LLM tools using majority voting did not improve the overall accuracy compared to ChatGPTv4.0. Specifically, there was no statistically significant difference (*p* value = 0.134) in sensitivity between the combined decision (majority voting) and ChatGPTv4.0, and the combined decision was significantly worse (*p* value = 0.008) than ChatGPTv4.0 in terms of specificity.

Comparison between ChatGPT v4.0 and zero-shot and hybrid methods: We assessed the performance (sensitivity and specificity) of ChatGPT v4.0 against both the zero-shot and hybrid approaches using the McNemar test. Specifically, we aligned the screening results from 100 cases and 100 controls as per the ChatGPT v4.0 method and similarly for the zero-shot and hybrid methods, testing for inconsistencies between these approaches as previously done. Our analysis revealed that ChatGPT v4.0 significantly outperformed the zero-shot method in sensitivity (*p* value < 0.001) but showed comparable effectiveness in specificity (*p* value = 0.37). Additionally, ChatGPT v4.0 demonstrated marginally superior sensitivity compared to the hybrid method (*p* value = 0.07), while its performance in specificity was similar (*p* value = 1.00).

Results on the Meijboom 2021 database (see Table 3)

The prompts we used for screening abstracts in this database are as follows:

> *Conduct a systematic review on transitioning patients from an originator to a corresponding biosimilar.*

Table 3 Results on the Meijboom 2021 database

	Accuracy (95% CI)	Sensitivity (recall) (95% CI)	Specificity (95% CI)
Zero-shot method	0.862 [0.833, 0.888]	0.125 [0.035, 0.290]	0.902 [0.875, 0.924]
Hybrid method	**0.954 [0.935, 0.969]**	0.065 [0.008, 0.214]	**1.000 [0.994, 1.000]**
ChatGPT (v4.0)	**0.848 [0.776, 0.905]**	0.812 [0.636, 0.928]	0.860 [0.776, 0.921]
ChatGPT (v3.5)	0.591 [0.502, 0.676]	0.969 [0.838, 0.999]	0.470 [0.369, 0.572]
Google PaLM 2	0.890 [0.802, 0.949]	0.647 [0.383, 0.858]	**0.954 [0.871, 0.990]**
Meta Llama 2	0.636 [0.548, 0.718]	**1.000 [0.891, 1.000]**	0.520 [0.418, 0.621]
Majority voting	0.720 [0.635, 0.794]	**1.000 [0.891, 1.000]**	0.630 [0.528, 0.724]
LCA model	0.841 [0.778, 0.904]	**1.000 [0.891, 1.000]**	0.790 [0.710, 0.870]
Results from the latest LLM models			
ChatGPT-3.5-Turbo	0.667 [0.580, 0.747]	0.970 [0.840, 0.999]	0.570 [0.467, 0.669]
ChatGPT-4-Turbo	0.840 [0.766, 0.898]	0.560 [0.374, 0.734]	0.930 [0.861, 0.971]
Gemini-1.0-pro	0.819 [0.743, 0.881]	0.380 [0.215, 0.568]	0.960 [0.901, 0.989]
Llama 3	**0.894 [0.829, 0.941]**	0.970 [0.840, 0.999]	0.870 [0.788, 0.929]
Claude 3 Opus	0.857 [0.785, 0.912]	0.910 [0.755, 0.982]	0.840 [0.753, 0.906]

I provide the title and abstract for one journal article. Provide an overall assessment based on eligibility criteria with only one word answer yes or no, with no explanation. Then, for each inclusion or exclusion criterion, answer with only one word, yes if it is included by the inclusion criterion or excluded by the exclusion criterion, and answer no if it does not meet the inclusion criterion or not excluded by the exclusion criterion.

After answering all the criteria with yes or no, then provide an overall explanation.

Here is the eligibility criteria:

Articles were included if they met the following criteria:

1. *Study involved transitioning from a TNFα inhibitor (including etanercept, infliximab, and adalimumab) originator to a biosimilar*
2. *The number of patients who retransitioned was reported or could be calculated*
3. *The article was an original research article published in a peer-reviewed journal*
4. *The article included baseline characteristics of the patients who transitioned*
5. *The article was written in English*
6. *The full-text version of the article could be obtained.*

Transitioning was defined as patients in whom the biosimilar was introduced after the originator, without treatment with other drugs in between. Retransitioning was defined as restarting the originator directly after discontinuing a biosimilar, without treatment with other drugs in between. In summary, transitioning was defined as switching from the originator to a biosimilar; retransitioning was defined as switching from the originator to a biosimilar and back to the originator. Both transitioning and retransitioning involved changes with the same active biological substance.

Here is the abstract:

Abstract X

Among all the LLM tools we tested, ChatGPT v4.0 stood out with the highest accuracy (0.840), but not with specificity (0.860) or satisfactory sensitivity (0.812). Compared to ChatGPTv4.0, combined decision using major voting did not improve overall accuracy (0.720), but improved sensitivity (1.000) at the sacrifice of specificity (0.630).

Comparison between LLM tools. We compared the performance (sensitivity and specificity) between ChatGPTv4.0 and other LLM tools using the McNemar test and found that ChatGPTv4.0 performed significantly better (p value < 0.001) than Google PaLM 2, but significantly worse than ChatGPT3.5 (p value = 0.001) and Llama 2 in terms of sensitivity; ChatGPTv4.0 performed significantly better than ChatGPTv3.5 (p value < 0.001) and better than Llama 2 (p value < 0.001), but worse than Google PaLM 2 (p value = 0.002), in terms of specificity.

Combining the decisions of different LLM tools using majority voting did not improve the overall accuracy compared to ChatGPTv4.0. Specifically, there was statistically significant difference (p value = 0.008) in sensitivity between the combined decision (majority voting) and ChatGPTv4.0, and the combined decision was not significantly worse (p value > 0.50) than ChatGPTv4.0 in terms of specificity.

For this database, the newer versions of ChatGPT (3.5-Turbo and 4-Turbo) did not improve performance over ChatGPT (v4.0), and Gemini-1.0-pro (vs. PaLM 2) did not improve performance either. However, Llama 3 (vs. Llama 2) improved over its older version and surpassed the performance of ChatGPT (v4.0). Claude 3 also slightly surpassed the performance of ChatGPT (v4.0).

Comparison between ChatGPT v4.0 and zero-shot and hybrid methods: We evaluated the performance of ChatGPT v4.0, focusing on sensitivity and specificity, in comparison with the zero-shot and hybrid approaches, employing the McNemar test as described above. In this analysis, we aligned the screening results from 32 cases and 100 controls for the tests. Our findings indicated that ChatGPT v4.0 significantly surpassed the zero-shot method in sensitivity (p value = 0.0002) and exhibited marginally improved specificity (p value = 0.099). Furthermore, ChatGPT v4.0 showed notably higher sensitivity than the hybrid method (p value < 0.001), although its specificity was comparatively lower.

Results on the Menon 2022 database (see Table 4)

The prompts we used for screening abstracts in this database are as follows:

> *"Conduct a systematic review on the methodological rigour of systematic reviews in environmental health.*
> *I provide the title and abstract for one journal article. Provide an overall assessment based on eligibility criteria with only one word answer yes or no, with no explanation.*
> *Then, for each inclusion or exclusion criterion, answer with only one word, yes if it is included by the inclusion criterion or excluded by the exclusion criterion and answer no if it does not meet the inclusion criterion or not excluded by the exclusion criterion.*

After answering all the criteria with yes or no, then provide an overall explanation.

Here are the eligibility criteria:

> *To be eligible for inclusion in the SR sample, documents had to fulfill the following criteria:*
>
> 1. *Identify explicitly as a "systematic review" in their title*
> 2. *Assess the effect of a non-acute, non-communicable, environmental exposure on a health outcome. Environmental exposures can include air and water pollutants, radiation, noise, occupational hazards, lifestyle factors (like diet or physical activity) and lifestyle choices influenced by family and peers (like substance use), social and economic factors (like stress from work or living conditions).*
> 3. *Include studies in people or mammalian models*
> 4. *Be available in HTML format*

Table 4 Results on the Menon 2022 database

	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Zero-shot method	0.882 [0.860, 0.902]	0.384 [0.272, 0.505]	0.923 [0.904, 0.940]
Hybrid method	**0.951 [0.936, 0.964]**	0.708 [0.589, 0.810]	0.971 [0.958, 0.981]
ChatGPT (v4.0)	**0.913 [0.861, 0.951]**	**0.932 [0.847, 0.977]**	0.900 [0.824, 0.951]
ChatGPT (v3.5)	0.711 [0.637, 0.777]	0.315 [0.211, 0.434]	**1.000 [0.964, 1.000]**
Google PaLM 2	0.569 [0.486, 0.648]	0.000 [0.000, 0.054]	**1.000 [0.958, 1.000]**
Meta Llama 2	0.827 [0.762, 0.880]	0.808 [0.699, 0.891]	0.840 [0.753, 0.906]
Majority voting	0.884 [0.827, 0.928]	0.808 [0.699, 0.891]	0.940 [0.874, 0.978]
LCA model	0.879 [0.830, 0.928]	0.945 [0.893, 0.997]	0.830 [0.756, 0.904]
Results from the latest LLM models			
ChatGPT-3.5-Turbo	0.753 [0.682, 0.816]	0.470 [0.352, 0.590]	0.960 [0.901, 0.989]
ChatGPT-4-Turbo	0.876 [0.817, 0.921]	0.760 [0.646, 0.852]	0.960 [0.901, 0.989]
Gemini-1.0-pro	0.927 [0.877, 0.961]	0.950 [0.872, 0.987]	0.910 [0.836, 0.958]
Llama 3	0.920 [0.869, 0.955]	0.960 [0.886, 0.992]	0.890 [0.812, 0.944]
Claude 3 Opus	0.804 [0.737, 0.861]	0.550 [0.429, 0.667]	0.990 [0.946, 1.000]

Here is the abstract:

Abstract X"

Among all the LLM tools we tested, ChatGPT v4.0 stood out with the highest accuracy (0.913) and specificity (0.932), but not with specificity (0.900). Compared to ChatGPTv4.0, combined decision using major voting did not improve overall accuracy (0.884) or sensitivity (0.808), but improved specificity (0.940).

Comparison between LLM tools. We compared the performance (sensitivity and specificity) between ChatGPTv4.0 and other LLM tools using the McNemar test and found that ChatGPTv4.0 performed significantly better than ChatGPT3.5 (p value < 0.001), Google PaLM 2, and Llama 2 (p value = 0.02) in terms of sensitivity; ChatGPTv4.0 performed worse than ChatGPTv3.5 and Google PaLM 2, in terms of specificity. Combining the decisions of different LLM tools using majority voting did not improve the overall accuracy compared to ChatGPTv4.0. Specifically, there was statistically significant difference (p value = 0.008) in sensitivity between the combined decision (majority voting) and ChatGPTv4.0, and the combined decision was not significantly different (p value = 0.134) than ChatGPTv4.0 in terms of specificity.

For this database, the newer versions of ChatGPT (3.5-Turbo and 4-Turbo) did not improve performance over ChatGPT (v4.0). However, both Gemini-1.0-pro (vs. PaLM 2) and Llama 3 (vs. Llama 2) improved over their older versions and surpassed the performance of ChatGPT (v4.0). Claude 3 also performed well but did not surpass the performance of ChatGPT (v4.0).

Comparison between ChatGPT v4.0 and zero-shot and hybrid methods: We aligned the screening results from 73 cases and 100 controls based on the ChatGPT v4.0 method, and similarly for the zero-shot and hybrid methods, to test for inconsistencies between these approaches, using the McNemar test as done in previous assessments. Our analysis showed that ChatGPT v4.0 significantly outperformed the zero-shot method in both sensitivity (p value < 0.001) and specificity (p value = 0.016). In comparison with the hybrid method, ChatGPT v4.0 also demonstrated superior sensitivity (p value < 0.001) and better specificity (p value = 0.04).

Monetary cost and time cost

To use the ChatGPT API or other LLM tools, the owners of these platforms charge a predetermined rate for access to the corresponding APIs. These fees are calculated in USD per thousand tokens, where tokens are the basic units used by these LLM platforms to quantify text length. In this context, a token can represent a word, a punctuation mark, or a character. The financial cost of screening 200 abstracts was approximately \$6 for ChatGPT v4.0, \$0.2 for ChatGPT v3.5, \$10 for Llama 2 (using Replicate), while Google PaLM 2 offered its services for free to invited developers. Thus, the cumulative cost of evaluating 200 abstracts across all platforms was approximately \$16.2. The cumulative cost of evaluating 200 abstracts across all latest models (\$3 for GPT-4-Turbo, \$0.05 for GPT-3.5-Turbo, free for Gemini-1.0-pro, \$0.05 for Llama-3, \$4 for Claude) was less, approximately \$7.1. In terms of time efficiency, processing 200 abstracts with each of these LLM tools took approximately 10–20 min using a single thread. However, it is imperative to recognize that abstract screening lends itself well to parallelization. Consequently, one could significantly speed up the process by setting up multiple threads to simultaneously screen different subsets of abstracts, thereby reducing the overall time required for completion. This parallel approach not only increases efficiency, but also ensures that large amounts of data can be processed in a timely manner, making LLM tools even more attractive for large-scale abstract screening tasks. In summary, the monetary and time costs of using LLM tools for abstract screening are negligible compared to manual labeling.

Beyond majority voting

We have expanded our analysis to include a variety of approaches for synthesizing decisions across different LLM tools. Our methodology is inspired by the concept of combining multiple diagnostic tests in the absence of a gold standard, akin to situations where human expert consensus is not available. There are several publications discussing such scenarios [31, 32], among which we proposed to use the latent class analysis (LCA) models.

Latent class analysis (LCA) is a statistical method used to identify subgroups within a population, which are not directly observed (hence "latent") [33–35]. It is particularly useful when the research interest lies in categorizing individuals into mutually exclusive groups based on their responses to multiple observed variables. In the context of abstract screening, LCA can offer a sophisticated means of integrating decisions from different LLM tools without relying on a gold standard, typically provided by human expert consensus. This approach assumes that the unobserved subgroups (or "latent classes") explain the dependence between the observed decisions made by each of the LLM tools.

Utilizing the LCA model, we treat the decisions from all LLM tools as dichotomous variables, corresponding to the adherence to each inclusion or exclusion criterion, as well as the overall decision. For instance, within the Bannach-Brown 2016 database (BB2016), there are eight criteria in total with four criteria each for inclusion and exclusion and one overall decision for inclusion, resulting in a total of nine binary items per LLM tool. In our analysis, we incorporated decisions from GPT v4.0, v3.5, and Llama 2. Decisions from Google PaLM 2 were excluded due to a high frequency (10% or more) of incomplete responses. Consequently, for the Bannach-Brown 2016 database, we worked with 27 binary items. For other databases such as Meijboom 2021 (Mj2021) and Menon 2022 (Me2022), the binary items totaled 21 and 15, respectively. It is important to note that LCA models were fitted to the binary data of each database independently.

The LCA model fitting process enables us to calculate the posterior probabilities of each abstract belonging to specific latent classes or subgroups. Abstracts are then categorized based on these probabilities, with assignment to the class for which an abstract has the highest posterior membership probability. The determination of the number of latent classes is a critical step in the LCA model fitting, which requires a priori specification. In our evaluation, we explored models with class numbers ranging from 2 to 6 and utilized the Bayesian information criterion (BIC) to identify the most "optimal" LCA model for our datasets.

Table 5 shows that after applying the Bayesian information criterion (BIC) to determine the most appropriate model for our data, we identified a 3-class model as the best fit for the binary decisions derived from the BB2016 database. Similarly, a 4-class model was optimal for the Mj2021 database, while a 3-class model was again best for the Me2022 database. The confusion matrices generated by the selected LCA models for each database provided a clear juxtaposition between the LLM-assigned classes and the actual labels of the abstracts (see Table 6).

Table 5 Negative BIC values for LCA models with different number of classes

#Class	BB2016	Mj2021	Me2022
2	3287.7	1441.0	2017.7
3	**3036.2**	1433.6	**2012.6**
4	3091.9	**1396.5**	2060.2
5	3128.7	1440.0	2112.5
6	3194.9	1531.6	2129.0

The performance metrics derived from these models are noteworthy. For the BB2016 database, should we categorize abstracts within class with label 2 as "included" and assign the others in classes with label 1 or 3 to the "excluded" category, the LCA model achieved a sensitivity rate of 93% and a specificity rate of 96%, indicating a high degree of accuracy in classifying relevant and irrelevant abstracts. In the case of the Mj2021 database, if we view class 1 as "included" group and classes 2 and 3 as "excluded" group, the model achieved a perfect sensitivity rate of 100%, meaning that it correctly identified all relevant abstracts, although the specificity was lower at 79%, indicating a higher rate of false positives. Similarly, for the Me2022 database, considering class 1 to be the "excluded" category and classes 2 and 3 to be the "included" group, the model showed a sensitivity of 94.5% and a specificity of 83%, a good balance between identifying relevant abstracts and avoiding false inclusions.

These results highlight the robustness of the latent class analysis approach to the abstract screening task, providing an effective method for classifying abstracts when a gold standard is not available. The varying levels of sensitivity and specificity across databases underscore the need to tailor the LCA model to the specific characteristics of each dataset as well as further future research.

We have incorporated results from the latest LLM models to increase the robustness and relevance of our findings. However, to maintain consistency and comparability with our original analyses, we have not incorporated these new results into the previously established majority voting or latent class analysis (LCA) approaches. Instead, we have chosen to make all raw data, including results from these newer models, freely available in our GitHub repository. This approach allows interested researchers and practitioners to conduct further investigations or apply alternative methods of analysis. By providing access to this additional data, we aim to promote transparency and enable the broader community to engage with and potentially build upon our work.

Table 6 A crosstab of class assignments based on selected LCA model

Class label	BB2016		Mj2021		Me2022	
	Excluded	Included	Excluded	Included	Excluded	Included
1	53	3	21	**32**	83	4
2	4	**93**	42	0	4	**54**
3	43	4	8	0	13	**15**
4	–	–	29	0	–	–

Discussion

This study began with a rigorous exploration of the capabilities of large language models (LLMs) in abstract screening. We used automation scripts developed in Python to interact with the APIs of several LLM tools, including Chat GPT v4.0, Google PaLM 2, and Meta Llama 2, as well as latest versions of these tools. Our central goal was to evaluate the efficiency and accuracy of these tools across three different databases of abstracts, leading us to a complex understanding of their potential in this context.

Large language models (LLMs), particularly ChatGPT, have garnered global attention since their inception. Employing LLMs for abstract screening in systematic reviews is an innovative concept [32, 33] and remains underexplored. This study presents the first comprehensive evaluation of LLMs applied to systematic review processes. The findings are encouraging, suggesting that LLMs could revolutionize abstract screening. Specifically, ChatGPT v4.0 exhibited stellar performance across three test scenarios, achieving an accuracy of at least 85%. Furthermore, it attained sensitivity and specificity rates ranging from 80% to an impressive 95%. These exceptional outcomes highlight the substantial promise of LLMs in abstract screening, offering an efficient and capable alternative to the conventional, laborious approaches that typically necessitate extensive human annotation.

However, it is important to acknowledge that we are still in the early stages of integrating LLM tools into the abstract screening field, and they are not without their imperfections—for example, occasionally excluding many (20%) relevant studies, even with the best-performed LLM tools, ChatGPT v4.0. These tools are not a universal solution to all the challenges associated with abstract screening, and they are not ready to completely replace human expertise in this area. Instead, they should be embraced as invaluable assistants in the abstract screening process.

In discussing the limitations of our study, it is important to recognize the constraints associated with our dataset selection and model configuration. We used only three databases from the SYNERGY set, limiting the generalizability of our findings across disciplines and datasets. In addition, the reliance on human-curated labels as a gold standard, while necessary, introduces potential biases due to the retrospective nature of our analysis. These labels may contain errors, and the methodology used in the original reviews to resolve discrepancies may affect the validity of our conclusions.

The parameters chosen for our LLMs—temperature, top k, top p, and prompts—were set to defaults or based on natural conversations to balance output quality with ease of use. While this approach minimizes the need for technical expertise, it may not be optimal for all screening scenarios. In addition, the reliance of LLMs on abstracts alone, as opposed to full-text evaluations, presents a fundamental challenge; critical data influencing inclusion or exclusion may not be present in the abstracts, potentially compromising screening accuracy.

In addition, the rapid development of LLMs and their “black box” nature pose challenges to the transparency that is essential in scientific settings. The environmental impact of using these computationally intensive models is also significant [36], requiring sustainability considerations. Future research should focus on refining these tools to increase transparency and efficiency, and ensure their responsible development and use in systematic reviews.

Our research suggests that LLM tools are ready to take on a role in abstract screening and are poised to have an immediate and positive impact on the process. Their integration into abstract screening can manifest itself in a few ways. They can serve as autonomous AI reviewers, adding an extra layer of scrutiny and ensuring thoroughness. Our findings suggest that a collective decision, such as one derived from majority voting, can sometimes improve sensitivity, underscoring the potential of LLM tools as a reliable safeguard against oversight, ensuring both comprehensiveness and accuracy.

In addition, LLM tools can facilitate a synergistic partnership with human experts. They are adept at identifying “high-risk” abstracts where different LLM tools have reached different judgments, flagging them for further human evaluation, and promoting a careful and error-free screening process while minimizing human effort.

Another exciting prospect is the integration of LLM tools with hybrid approaches and active learning approach. In this scenario, LLM tools could autonomously annotate abstracts in the training set, minimizing the need for human labeling. Subsequently, these labeled abstracts could be used to train custom NLP models, paving the way for a streamlined and efficient abstract screening process, resulting in significant time and resource savings. Further research is needed to understand how the “uncertainty” in LLM-based decisions, when human-curated labels are not used, affects the performance of the hybrid approach. We also note from the test examples that the performance of hybrid approaches varies widely from example to example and depending on the text embedding tools used. Extensive research is needed to perfect the hybrid approaches.

Another future research direction is to explore how to fine-tune different versions of LLM tools and how to derive collective decisions. One idea is that by using different tuning parameters, one could propose different versions of LLM tools, and if they perform similarly in terms of accuracy but give different decisions (e.g., not highly correlated), we expect that better collective decisions would be obtained, as observed in many ensemble learning experiences [37]. However, this could be costly and require a lot of exploration.

More importantly, it would be particularly valuable to prospectively test these LLM tools, or integration of these tools with other approaches like zero-shot, active learning, in real-world meta-analysis and systematic review projects. This would provide deeper insights into their practical utility and effectiveness, and a more comprehensive understanding of their impact in live research environments.

Conclusion

In summary, while LLM tools may not be able to fully replace human experts in abstract screening, their ability to transform the screening process is both undeniable and imminent. With continued advances in technology and ongoing refinement, these tools will play a critical role in the future of abstract screening, ushering in a new era of efficiency and effectiveness.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13643-024-02609-x.

Supplementary Material 1: Table S1. TRIPOD + AI checklist.

Acknowledgements
The authors thank the reviewers and the Editor for their insightful comments and suggestions, which have greatly improved the quality of this work.

Authors' contributions

XT conceived the paper. ML and XT devised the study. ML, JS, and XT analyzed the datasets. ML, JS, and XT wrote the manuscript. All authors edited and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials

The datasets analyzed during the current study are available in the SYNERGY repository, https://github.com/asreview/synergy-dataset. We have created a GitHub repository https://github.com/mikeli380/LLMAbstractScreening to host our code (R/Python), abstracts, and intermediate results so that interested readers can replicate our findings.

Declarations

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details
[1]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. [2]Department of Mathematics and Statistics, University of North Carolina at Greensboro, Greensboro, NC 27402, USA. [3]Department of Biostatistics and Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA.

Received: 20 December 2023 Accepted: 11 July 2024
Published online: 21 August 2024

References

1. Higgins JPT, et al. Cochrane handbook for systematic reviews of interventions. 2019.
2. Moher D, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 2009;6(7). https://doi.org/10.1371/journal.pmed.1000097.
3. Sackett DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. 1996. Clin Orthop Relat Res. 2007;455. https://doi.org/10.1136/bmj.312.7023.71.
4. Lavis JN, Posada FB, Haines PA, Osei E. Use of research to inform public policymaking. Lancet. 2004;364(9445). https://doi.org/10.1016/S0140-6736(04)17317-0.
5. Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up?. PLoS Med. 2010;7(9). https://doi.org/10.1371/journal.pmed.1000326.
6. Polanin JR, Pigott TD, Espelage DL, Grotpeter JK. Best practice guidelines for abstract screening large-evidence systematic reviews and meta-analyses. Res Synth Methods. 2019;10(3). https://doi.org/10.1002/jrsm.1354.
7. Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. Selective publication of antidepressant trials and its influence on apparent efficacy. N Engl J Med. 2008;358(3). https://doi.org/10.1056/nejmsa065779.
8. Edwards P, Clarke M, DiGuiseppi C, Pratap S, Roberts I, Wentz R. Identification of randomized controlled trials in systematic reviews: accuracy and reliability of screening records. Stat Med. 2002;21(11). https://doi.org/10.1002/sim.1190.
9. Wang Z, Nayfeh T, Tetzlaff J, O'Blenis P, Murad MH. Error rates of human reviewers during abstract screening in systematic reviews. PLoS ONE. 2020;15(1). https://doi.org/10.1371/journal.pone.0227742.
10. O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. Syst Rev. 2015;4(1). https://doi.org/10.1186/2046-4053-4-5.
11. OpenAI. Introducing ChatGPT Plus. OpenAI Blog; 2023. https://openai.com/blog/chatgpt-plus. Accessed 8 Nov 2023.
12. Google DeepMind. Introducing PaLM 2. Google Blog; 2023. https://blog.google/technology/ai/google-palm-2-ai-large-language-model/. Accessed 8 Nov 2023.
13. Meta AI Research. Meta and Microsoft introduce the next generation of Llama. Facebook AI Blog; 2023. https://ai.meta.com/blog/llama-2/. Accessed 8 Nov 2023.
14. Claude.ai. Claude AI API reference. 2024. https://docs.anthropic.com/en/api/getting-started. Accessed 4 Jun 2024.
15. Wallace BC, Small K, Brodley CE, Lau J, Trikalinos TA. Deploying an interactive machine learning system in an evidence-based practice center: Abstrackr. 2012. https://doi.org/10.1145/2110363.2110464.
16. Alshami A, Elsayed M, Ali E, Eltoukhy AEE, Zayed T. Harnessing the power of ChatGPT for automating systematic review process: methodology, case study, limitations, and future directions. Systems. 2023;11(7). https://doi.org/10.3390/systems11070351.
17. Qureshi R, Shaughnessy D, Gill KAR, Robinson KA, Li T, Agai E. Are ChatGPT and large language models 'the answer' to bringing us closer to systematic review automation? Syst Rev. 2023;12(1). https://doi.org/10.1186/s13643-023-02243-z.
18. Xian Y, Lampert CH, Schiele B, Akata Z. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. IEEE Trans Pattern Anal Mach Intell. 2019;41(9). https://doi.org/10.1109/TPAMI.2018.2857768.
19. Moreno-Garcia CF, Jayne C, Elyan E, Aceves-Martins M. A novel application of machine learning and zero-shot classification methods for automated abstract screening in systematic reviews. Decis Anal J. 2023;6. https://doi.org/10.1016/j.dajour.2023.100162.
20. Ferdinands G, et al. Performance of active learning models for screening prioritization in systematic reviews: a simulation study into the Average

Time to Discover relevant records. Syst Rev. 2023;12(1). https://doi.org/10.1186/s13643-023-02257-7.
21. Boetje J, van de Schoot R. The SAFE procedure: a practical stopping heuristic for active learning-based screening in systematic reviews and meta-analyses. Syst Rev. 2024;13(1). https://doi.org/10.1186/s13643-024-02502-7.
22. Yang Y, Loog M. A benchmark and comparison of active learning for logistic regression. Pattern Recognit. 2018;83. https://doi.org/10.1016/j.patcog.2018.06.004.
23. De Bruin RJ, Ma Y, Ferdinands G, Teijema J, Van de Schoot. SYNERGY - open machine learning dataset on study selection in systematic reviews. DataverseNL; 2023.https://dataverse.nl/dataset.xhtml?persistentId=doi:10.34894/HE6NAQ.
24. Millard SJ, Weston-Green K, Newell KA. The Wistar-Kyoto rat model of endogenous depression: a tool for exploring treatment resistance with an urgent need to focus on sex differences. Prog Neuropsychopharmacol Biol Psychiatry. 2020;101. https://doi.org/10.1016/j.pnpbp.2020.109908.
25. Zawia NH, Mattia CJ, Bondy SC. Differential effects of difluoromethylornithine on basal and induced activity of cerebral ornithine decarboxylase and mRNA. Neuropharmacology. 1991;30(4). https://doi.org/10.1016/0028-3908(91)90058-J.
26. Shea BJ, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ. 2017;358. https://doi.org/10.1136/bmj.j4008.
27. Bannach-Brown A, Liao J, Wegener G, Macleod M. Understanding in vivo modelling of depression in non-human animals: a systematic review protocol. Evid Based Preclin Med. 2016;3(2). https://doi.org/10.1002/ebm2.24.
28. Meijboom RW, Gardarsdottir H, Becker ML, ten Wolde S, Egberts TCG, Giezen TJ. Incidence of and reasons and determinants associated with retransitioning from biosimilar etanercept to originator etanercept. BioDrugs. 2021;35(6). https://doi.org/10.1007/s40259-021-00501-x.
29. Menon JML, Struijs F, Whaley P. The methodological rigour of systematic reviews in environmental health. Crit Rev Toxicol. 2022;52(3). https://doi.org/10.1080/10408444.2022.2082917.
30. Collins GS, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open. 2021;11(7). https://doi.org/10.1136/bmjopen-2020-048008.
31. Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. Am J Epidemiol. 1995;141(3). https://doi.org/10.1093/oxfordjournals.aje.a117428.
32. Hui SL, Walter SD. Estimating the error rates of diagnostic tests. Biometrics. 1980;36(1). https://doi.org/10.2307/2530508.
33. Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. Biometrics. 2001;57(1). https://doi.org/10.1111/j.0006-341X.2001.00158.x.
34. Collins LM, Lanza ST. Latent class and latent transition analysis: with applications in the social, behavioral, and health sciences. 2010.
35. McCutcheon AL. Latent class analysis. Newbury Park: Sage; 1987.
36. Rillig MC, Ågerstrand M, Bi M, Gould KA, Sauerland U. Risks and benefits of large language models for the environment. Environm Sci Technol. 2023;57(9). https://doi.org/10.1021/acs.est.3c01106.
37. Sagi O, Rokach L. Ensemble learning: a survey. Wiley Interdiscip Rev Data Min Knowl Discov. 2018;8(4). https://doi.org/10.1002/widm.1249.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.