GPT API Models Can Function as Highly Reliable Second Screeners of Titles and Abstracts in Systematic Reviews: A Proof of Concept and Common Guidelines

ABSTRACT

Independent human double screening of titles and abstracts is considered a critical step to ensure the quality of systematic reviews and meta-analyses herein. However, double screening is a costly as well as resource-intensive procedure that slows the review process, ultimately excluding many researchers from using it. To alleviate this issue and potentially increase the reliability of systematic reviews and meta-analyses, we evaluated the use of OpenAI's GPT (generative pre-trained transformer) API (application programming interface) models as alternative second screeners of titles and abstracts. For this purpose, we developed a new benchmark scheme for interpreting the performances of automated screening tools and conducted three large-scale classification experiments on three different kinds of systematic reviews. This included a published and an ongoing review with few well-defined inclusion criteria and one ongoing review with a complex set of inclusion/exclusion criteria. For the latter, we introduced and used multi-prompt screening, that is making one prompt per inclusion/exclusion criteria, to accommodate the complexity of the review. Overall, we found that the GPT API models perform on par or even better than common human screening performance in terms of detecting relevant studies, even in very complex social science review settings. To support future reviews, we develop a reproducible workflow and tentative guidelines for when (and not) reviewers can use GPT API models for title and abstract screening. Our aim is ultimately to make a framework for using GPT API models acceptable as independent second screeners within state-of-the-art reviews. To standardize the application, we present the R package AIscreenR.

KEYWORDS: *title and abstract screening, OpenAI's GPT API models, systematic review, screening benchmarks, AIscreenR*

[CHECK DETAILS HERE: https://onlinelibrary.wiley.com/page/journal/17592887/homepage/forauthors.html]

HIGHLIGHTS

What is already known

- OpenAI's GPT API models have shown promising performance in terms of working as a second screener of titles and abstracts within various scientific fields.
- Automating screening tools can ease the burden of title and abstract screening
- Automating screening tools most often cannot detect/classify all relevant studies, which in turn, can induce the so-called 'artificial screening biases'

What is new

- We show that OpenAI's GPT API models can function as a highly reliable second screener in social science reviews with better recalls than presented in previous evaluations and on par with human performance.
- We introduce the concept of multi-prompt screening and show that this approach can make GPT API models work as reliable second screeners even in very complex review settings.
- We develop empirical benchmarks to make reliable comparisons between the screening performance of humans and automated screening tools.
- We provide general guidelines for how and when GPT API models can (and cannot) safely be used as independent second screeners of titles and abstracts.
- We present and validate the AIscreenR package to ensure standardized conduct of title and abstract screening with (OpenAI's) GPT API models (and in theory with other models such as Claude 2).

Potential impact for Research Synthesis Methods readers

- Changing the duplicate screening of title and abstract screening in systematic reviews
- Increasing the reliability of large-scale systematic reviews
- Making substantial and reliable reductions of human labor in systematic reviews
- Providing a new guideline for reviewers on when and when not to use AI screening tools
- Standardizing screening with prompt-based LLMs

1 INTRODUCTION

Systematic reviews are essential tools for informing policy, research, and practice. Hence, it is all-important that systematic reviews adhere to the highest scientific standards. Yet systematic reviews are time-consuming, potentially hindering a timely transfer of usable knowledge. Distinct from other types of reviews, systematic reviews are defined as the process of collecting, assessing, and synthesizing findings from (ideally all) relevant scientific studies using explicit and replicable research methods (Gough et al., 2017; Hou & Tipton, 2024). A critical first step to ensure the quality of systematic reviews and meta-analyses herein involves detecting all eligible references related to the literature under review (Polanin et al., 2019). This entails searching all pertinent literature databases relevant to the given review, most often resulting in thousands of title and abstract records that need to be screened. Manual screening hereof can indeed be a time-consuming and tedious task. However, overlooking relevant studies at this stage can be consequential, leading to substantially biased results if the missed studies are systematically different from the detected ones. In fact, this can be seen as a special case of publication/selection bias (Hedges, 1992; Rothstein et al., 2005), which threatens the internal validity of systematic reviews (Shadish et al., 2002). Therefore, independent human double-screening is considered to be the 'golden standard' to hinder a biased selection of relevant studies (Guo et al., 2024; Higgins et al., 2019; Stoll et al., 2019; Wang et al., 2020). This is further supported by previous research suggesting that screeners on average tend to miss between 3% to 24% of all eligible studies depending on the level of content knowledge, which most often has a substantial impact on the final quantitative results (Buscemi et al., 2006; Waffenschmidt et al., 2019). In medicine, this number is in some cases even higher when using student screeners (Ng et al., 2014). Nonetheless, duplicate screening of all identified titles and abstracts is a costly and resource-intensive procedure, possibly requiring several months of skilled, full-time human labor (Campos et al., 2023; Hou & Tipton, 2024; Shemilt et al., 2016). Consequently, many reviewers refrain from using duplicate screening methods due to low budgets or narrow time limits, for instance (Pacheco et al., 2023). Alternatively, reviewers make too narrow searches to keep the number of records down to a manageable size which again seriously increases the risk of overlooking relevant studies (Van De Schoot et al., 2021). Over time all these issues will only grow in size since the complexity of identifying all relevant studies increases with the rapid growth in the number of scientific publications (Bornmann et al., 2021; O'Mara-Eves et al., 2015). Thus, it can be considered an economically inefficient and

unsustainable use of human resources only to rely on (duplicate) human screening of titles and abstracts in future systematic reviews[1] (Shemilt et al., 2016), and changes are needed to maintain a high quality of large-scale systematic reviews.

A possible solution, and an alternative to human double-screening, is to use (semi-)automated screening tools based on text-mining and/or machine-learning algorithms to act either as a second screener, a course-grained classifier, or to sort citation records in prioritized order (Cohen et al., 2006; Gartlehner et al., 2019; O'Mara-Eves et al., 2015; Van De Schoot et al., 2021). The use of automated screening tools is considered invaluable in supporting living reviews and has shown a promising ability to reduce the screening workload by 30% to 70% (O'Mara-Eves et al., 2015; Perlman-Arrow et al., 2023). However, a clear disadvantage of these substantial workload savings is that it is expected that they will always result in missing at least 5%-10% of all eligible references since "a 100% recall rate with a stochastic algorithm is generally considered unattainable" (Hou & Tipton, 2024, p. 3). This seems to create a screening paradox which might be one of the main reasons why many reviewers tend to mistrust the application of machine-learning tools (O'Connor et al., 2019). While trying to reduce selection biases caused by single screening, automated screening potentially introduces a novel type of publication/selection bias defined by König et al., (2023) as the 'artificial screening bias' (ASB).

An additional challenge is that most automated screenings are based on supervised and active learning methods. This means that they need to be trained on a large enough set of in- and excluded references to perform adequately which in turn can be a time-consuming task, as well. Moreover, when automation tools are used for prioritized screening, it is most often unknown when it is safe to stop screening with regard to finding all or close to all eligible references. Albeit, various stopping rules have been proposed, the adequacy of these is sensitive to a range of factors such as the length of the database, the prevalence of relevant studies, and the balance between relevant and irrelevant records (Campos et al., 2023; König et al., 2023; Van De Schoot et al., 2021).

To date, many automated screening tools have been thoroughly evaluated (Burgard & Bittermann, 2023; Kugley et al., 2016). From these evaluations, the overall picture is that they are generally not capable of replacing an independent human second screener without a significant risk

[1] But already now, we see that in some applications of systematic reviews, the number of records needed to be screened way exceeds what can be considered an economically efficient and sustainable use of human resources, either due to very broad terms needed to be added to search string to cover all relevant studies (see e.g., Thomsen et al., 2022) or due to a broad aim of the review as is often the case with scoping review and evidence and gap maps (see e.g., Bondebjerg, Filges, et al., 2023).

of omitting a substantial number of eligible studies[2] (Gartlehner et al., 2019; O'Mara-Eves et al., 2015; Olorisade et al., 2016; Rathbone et al., 2015). By using the level of automation heuristic (c.f. Table 1) developed by O'Connor et al. (2019), it can be said that current automated tools generally fail to function at the highest levels of automation (i.e., Level 3 and Level 4) where they make credible independent deterministic screening decisions. Instead, the vast majority of tools are predominately used to conduct Level 2 tasks such as sorting citation records in prioritized order from highest to lowest probability of being relevant to the review (Kugley et al., 2016; O'Connor et al., 2019; Olofsson et al., 2017). If considerable time savings should be realized in future reviews, it is regarded as all-important that automated tools at least rise to Level 3 of automation (Jonnalagadda et al., 2015; Tsafnat et al., 2014).

TABLE 1. Levels of automation for human-computer interactions*

Level	**Task**
Level 4	Tools perform tasks to eliminate the need for human participation in the task altogether, e.g., fully automated article screening decision about relevance made by the automated system.
Level 3	Tools perform a task automatically but unreliably and require human supervision or else provide the option to manually override the tools' decisions, e.g., duplicate detection algorithms and software, linked publication detection with plagiarism algorithms and software.
Level 2	Tools enable workflow prioritization, e.g., prioritization of relevant abstracts; however, this does not reduce the work time for reviewers on the task but does allow for compression of the calendar time of the entire process.
Level 1	Tools improve the file management process, e.g., citation databases, reference management software, and systematic review management software.

*Adopted from O'Connor et al. (2019)

A possible solution to bridge the gap between Levels 2 and 3 of automation[3] is to use the newly developed large language models (LLM), such as the generative pre-trained transformer (GPT) models introduced by OpenAI. The first evaluations of using OpenAI's GPT API (application programming interface) models for screening of medical and software engineering titles and abstracts have generally yielded promising results with recall and specificity measures in most instances on

[2] To overcome/reduce this issue, a new tentative guideline termed SAFE has been developed in which it is suggested to use multiple machine learning algoritmes in order to detect all relevant references in the bulk of records (Boetje & van de Schoot, 2024). However, we do not considered this framework to have been thouroughly enough testing yet to know if the SAFE procedure allows reviewers to detect all relevant studies with the machine learning algoritms including in screening softwares such as ASReview.

[3] We do not consider the level 4 of automation to be the ideal case since we consider human-in-the-loop operation to be state-of-the-art at the time of writing.

par with human performance but always on par with or superior to classical machine-learning tools (Guo et al., 2024; Syriani et al., 2023).

Although previous applications and evaluations of using OpenAIs GPT models for title and abstract screening (henceforth TAB screening) represent a vital first step for validating the use of GPT models as independent second screeners in systematic reviews, many questions are left unanswered. Most pressing, it is still unclear if and how the GPT models can be implemented in systematic reviews in a standardized and reliable manner. In contrast to many well-established automated screening algorithms, there exists no recommended workflow and guideline for how to conduct such screenings, including how to make reliable prompts. Even more critically, no software[4] has yet been developed to support and standardize the setup of this screening approach. Therefore, a major aim of this paper is partly to develop a heuristical workflow for how to conduct TAB screening with GPT API models and partly to present the R package AIscreenR (version 0.0.1). Our goal is to develop an easy-to-implement framework that draws on commonly accessible RIS file data typically used with standard review software such as Covidence and EPPI-reviewer, etc. This might increase the chances of ensuring user deployment and acceptance since complex implementation is often considered to be a major impediment to the wider application of automated screening tools (O'Connor et al., 2019).

Furthermore, there has not yet been laid any solid foundation on which evidence institutions (such as Cochrane and the Campbell Collaboration) can accept and recommend the use of such tools per se. According to the Campbell Collaboration, for them to accept the incorporation of automation tools in their reviews "*requires (a) functioning tech (b) proof that it is functioning appropriately (c) the tech embodied in usable products (d) agreed guidelines for appropriate use (e) training (f) ongoing support.*" (Campbell Collaboration, 2023). Therefore, the overarching goal of this paper is to construct a framework in which TAB screening with GPT API models can be said to meet requirements set forth by the evidence institutions. In the following part, we briefly explicate how we aim to build this framework.

Concerning requirement *(a)*, we cannot as such fulfill it since the GPT API models we draw upon in this paper are closed-source applications with black-box algorithms. That is our suggested framework is only viable as long as given firms provide access to their LLMs. However, our suggested framework and codes can readily be remodeled to work with other API models, such as models from Claude 2 or Mistral AI where the request body takes the same arguments as OpenAI's

[4] To our knowledge, GPT models has so far only be implemented in the EPPI Reviewer software with the aim to support automated data extraction from full texts (see EPPI-Centre, 2024) and not for TAB screening purposes.

GPT API models. Therefore, our setup aims to be agnostic to the given provider of the given LLM. In theory, our approach can be implemented together with LLMs such for instance Mistral open-source LLMs that can be downloaded locally by the users. We, therefore, understand a "functioning tech" to point, in our case, to the broader family of LLM models, which we believe will be around in some or another form for many years.

A key part of fulfilling Campbell's requirement *(b)*, and not compromising the quality of future systematic reviews, is to show that the GPT API models are not significantly inferior to human screening performance (O'Connor et al., 2019). Thus to make a reliable assessment of this, we developed empirical screening benchmarks to which the GPT API screening performance can be compared. We consider this as the only reliable way to assess whether a given recall is good or bad. Say, for example, that if humans on average tend to miss 20%-25% of all relevant studies during the title and abstract screening phase, then it might be misleading to infer that GPT models with a recall of 0.75% imply that GPT cannot be used as an individual second screener. To construct such a benchmark scheme we mapped the human screening performance of 22 large-scale systematic reviews; 17 Campbell Systematic Reviews, and five systematic reviews conducted by the Norwegian Institute of Public Health (NIPH). Thereafter, we conducted three large-scale classifier experiments, where we showed that OpenAI's GPT API models can conduct TAB screening with a performance *at least* on par with human performance relative to our developed benchmarks, even when applied in a very complex review setting. In this framework, we introduce the concept of multi-prompt screening that resembles the common way humans usually screen titles and abstracts (Valentine, 2009). These experiments further aim to show that GPT API models are perfectly viable as independent second screeners in social science reviews.

We aim to fulfill requirement *(c)* by developing the AIscreenR software. A side-effect of conducting the above-mentioned classifier experiments, mentioned under requirement *(b)*, was further to ensure that the AIscreenR package works reliably.

Then, to fulfill requirements *(d)* and *(e),* we develop a heuristic for how to test the performance of one's developed prompt(s) and screening as well as assess under what conditions TAB screening with the GPT API models can (and cannot) be accepted to be used as an independent second screener in systematic reviews. We inform these guidelines by the empirical human screening benchmarks developed under requirement *(b)* as well. Since we are working with pre-*trained* models, requirement *(e)* is not as such necessary in our case. Instead, the performance of the prompt(s) used for screening needs to be *tested* and compared against human performance measures before credible TAB

screening can be initiated. We return to this point when we show how to develop reliable prompts for TAB screening in later sections. Finally, to accommodate requirement *(f)* we have developed the AIscreenR package as an open-source software so that others in the review community can readily contribute to the development and ongoing support of the software. With the exposition sketched above, we hope to make the uptake of such tools more acceptable in future state-of-the-art systematic reviews. This goes without saying that our approach represents a final solution. Our aim is merely to show one way in which GPT API models can be used for TAB screening in large-scale systematic reviews that can inspire future applications of TAB screening with LLMs.

The remainder of the paper proceeds as follows: In Section 2 we review previous evaluations of using OpenAI's GPT models for TAB screening tasks in systematic reviews and reflect on our contributions. In Section 3 we describe the metrics we applied to evaluate the screening performance of the GPT API models and human screeners, respectively. In Section 3, we further develop screening performance benchmarks to assess the performance of the GPT API models. In Section 4, we present our classifier experiments, including our prompt engineering and data underlying these experiments. The results of these experiments are also presented in this section. In section 5, we deduce tentative guidelines for when we think reviewers are 'good to go' in terms of using OpenAI's GPT API models as an independent second screener. Moreover, we flesh out how we think reliable prompts can be developed in future reviews. Finally, in Sections 6 to 8, we recapitulate by reflecting on the limitations of our work and the use of OpenAI's LLMs and what should concern future research as well as the implications of our results and recommendations.

2 RELATED WORK

To our knowledge, the first evaluation of the TAB screening performance of OpenAI's GPT API models was performed by Syriani et al. (2023). Based on five ongoing systematic reviews within the field of software engineering, they compared the TAB screening performance of the GPT API model gpt-3.5-turbo-0301[5] relative to five state-of-the-art machine learning algorithms. Hereto they found that OpenAI's GPT API models perform on par with traditional classifier models, and in some instances even better—without any need for (pre-)training. They only found the models to perform badly when applied on datasets where humans had shown a "high conflict ratio". This might simply

[5] This model has been deprecated

indicate that the models perform badly when given unclear inclusion/exclusion criteria—as the humans did too. Syriani et al. (2023) used Python to reach the GPT API models, but they did not build any publicly available software for others to replicate their workflow.

Guo et al. (2024) tested the leverage of OpenAI's GPT-4 API model[6] for TAB screening of medical research literature. They found that the average recall (referred to as the sensitivity of included paper) and specificity when compared to the final decision of two independent human screeners across six clinical reviews was 0.76 and 0.91, respectively. Based on these results, Guo et al. (2024) inferred that the GPT-4 model is proficient in terms of excluding the right studies whereas it is insufficient in finding relevant studies compared to human screeners. Consequently, Guo et al. (2024) concluded that GPT API models should not replace human screening but instead be seen as a support tool guarding against human errors. Guo et al. (2024) used Python to call the API models without providing any general user software.

Gargari et al. (2024) applied the gpt-3.5-turbo-0613 API model to conduct TAB screening in one clinical systematic review. In line with Guo et al. (2024), they found GPT to be better at making correct exclusion decisions relative to detecting relevant studies. Therefore, they also recommended not replacing any human raters with the gpt-3.5 API model. Gargari et al. (2024) reached the API model via Python, and they shared their codes[7] so that others can replicate their workflow. Yet this requires reviewers to be rather skilled in Python coding.

On a related line of research, Alshami et al. (2023), Khraisha et al. (2024), and Issaiy et al. (2024) all investigated the TAB screening performance of using ChatGPT from the internet interface. Alshami (2023) found that using the ChatGPT interface exhibits performance measures similar to the API model. By contrast, Khraisha et al. (2024) and Issaiy et al. (2024) found that using GPT-3.5 and GPT-4 via the ChatGPT interface worked insufficiently compared to human performance. As we will later discuss further, we found a similar pattern when we compared the performance of OpenAI's GPT API models with that of the ChatGPT interface. To be precise, the GPT API models reached from the *v1/chat/completions* endpoint worked significantly better relative to the GPT models embedded in the ChatGPT interface. In fact, we were not able by any means to replicate our results obtained from the API models with the models available in the ChatGPT interface. We, therefore, consider it pivotal that future research clearly distinguishes between OpenAI's GPT models when doing research with them so that the performance of different GPT models is not unnecessarily mixed

[6] It is uncertain what exact model the authors used. We expect it was the gpt-4-0613 API model.
[7] Can be found at https://github.com/mamishere/Article-Relevancy-Extraction-GPT3.5-Turbo

up. In the paper, we narrowly focus on the use of OpenAIs GPT API models reached from the 'v1/chat/completions' endpoint, not to be confused with the GPT models behind the ChatGPT interface or the 'v1/completions' endpoint. On this note, it was unclear what exact model Syriani et al. (2023) and Guo et al. (2024) used during their investigations, whereas Gargari et al. (2024) used the same endpoint as we drew upon in this paper.

2.1 What we do differently

In this paper, we go beyond previous evaluations in multiple ways and show some key advances in using LLMs for TAB screening relative to (but possibly combined with) traditional machine learning tools. Starting with the latter, one advance of using LLMs is that these models do not need to be pre-trained which, in turn, means that these models are not as (if at all) sensitive to imbalance between relevant and irrelevant records or the number of relevant records in the data as classical machine-learning tools (Campos et al., 2023; König et al., 2023). This is so because the GPT models we applied treat each title and abstract individually without any knowledge of previous decisions. Compared with traditional machine learning algorithms, we will also show that the GPT-4 has the ability in some cases to find almost all relevant studies.

In contrast to all the previous evaluations of using GPT API models for TAB screening, we are the first to draw on function calling in the request body (OpenAI, 2024). This allows users to make prompts without the need to explicitly specify how the model shall respond to the request. The specific advance of function calls is that this permits users to make more refined and concise prompts, which, in turn, ensures that users are getting "more reliably (…) structured data back from the model" (OpenAI, 2024). We believe that the use of function calling potentially explains why we in later sections find significantly better recall performances (i.e., the ability to detect relevant studies) of using the GPT API models than previous evaluations. Differently from the previous evaluation, we have built our function calls so that they also allow the model to express its uncertainty relative to merely making binary decisions (i.e., include or exclude) as all previous evaluations have done. That is if the GPT API model, for example, does not have enough information to make a reliable decision, the given title and abstract is added to the pool of included studies. This significantly reduces the models' ability to overlook potentially relevant studies. Moreover, we built two different types of function calls thus that users can both get simple/trinary (i.e., 1 = {include}, 1.1 ={uncertain}, and 0

= {exclude}) and/or descriptive responses back from their screening requests. Getting detailed descriptive responses can be pivotal especially when examining discrepancies between GPT and human screener decisions.

The main difference between this paper and the previous evaluation is further that we aim to make a standardized and user-friendly workflow for how to use GPT API models for TAB screening that are easy to implement in state-of-the-art systematic reviews. We do so by developing the AIscreenR R package and technically quality-assuring it via the conduct of to large-scale classifier experiment. The AIscreenR is built as a flexible software that allows users to conduct multiple screenings simultaneously based on multiple prompts, API models, iterations of the same request, and nucleus samples (i.e., different top_p or temperature values). The software further allows the user to conduct the same request (i.e., asking the exact same question) multiple times to avoid random noise in the model response (especially when using gpt-3.5 models). When this feature is used the final GPT decision is based on the probability of inclusion across the iterated requests. The specific inclusion threshold can be determined by the user. This also allows the users to test model response consistency. Moreover, the software is built so that it draws on multi-core processing, which allows the users to speed up the timing of the screening significantly.

To conduct a fair assessment of GPT's ability to conduct TAB screening relative to humans but also to outline reliable guidelines on when (and not) to use LLMs for TAB screening (which has not previously been done), requires a clear understanding of common human screening performance in systematic reviews. Therefore, to make a better understanding of common human performance and to develop benchmarks that could be held against the screening performance of GPT, we mapped the human screening performance across 17 Campbell systematic reviews and 5 systematic reviews conducted by the Norwegian Institute of Public Health (NIPH). Relative to the previous evaluations, the contribution of this paper is therefore also to put forward a tentative benchmark scheme to which all types of automated screening tools can be compared.

In all the previous evaluations, multiple inclusion/exclusion criteria were added to a single prompt. Yet Gargari et al. (2024) suggested that broader and less specific prompts do not perform well in terms of finding relevant studies. Instead, concisely framed prompts with clear information seem to have a better performance. This could indicate that single-prompt TAB screening is rather restricted to only work within simple and clearly defined reviews where the inclusion of abstracts can be determined by a few inclusion criteria/questions. However, to overcome this issue, we introduce two types of screenings coined as *hierarchical screening* and *multiple-prompt screening*,

respectively, and show that they can make GPT API models work within extremely complex review settings. With these screening approaches, we suggest making one concise prompt per inclusion (and/or exclusion) criterion. In these types of screenings, studies are only considered relevant if included on all or close to all (say 5 out of 6) of the used prompts.

Finally, all previous evaluations were based on medical or natural science reviews, and we add to the generalizability of these results by showing that GPT API models all exhibit promising screening performance in the more wildly social science reviews as well.

3 METHODS

This section describes the metrics that we used partially to evaluate the screening performance of the GPT API models and partially to develop empirical screening benchmarks to hold against the screening performance of the used GPT API models. Moreover, the section describes the data and results we used to develop our suggested screening performance benchmarks.

3.1 Metrics we use to evaluate the performance of the GPT models

To evaluate the screening performance of the GPT API models, we used a range of different metrics. The choice of metrics was primarily informed by the recommendations made by O'Connor et al. (2019) and Syriani et al. (2023). The two main metrics we used to evaluate the performance of the GPT API models were the *recall* (by some defined as the sensitivity) and *specificity* metrics since these are intuitive to understand and interpret and are not sensitive to imbalanced data (i.e., data with a large difference in the proportion between inclusion and exclusion references, as is commonly the case in systematic reviews). The recall metric "represents the proportion of relevant records being correctly classified" (Hou & Tipton, 2024), and can be written as

$$Recall = \frac{TP}{TP + FN} \tag{1}$$

where TP (true positive) represents all the studies that are correctly included, and FN (false negative) is the number of studies falsely excluded. By contrast, the specificity metric "measures the ability to exclude all references that should be excluded" (Syriani et al., 2023), and is given by

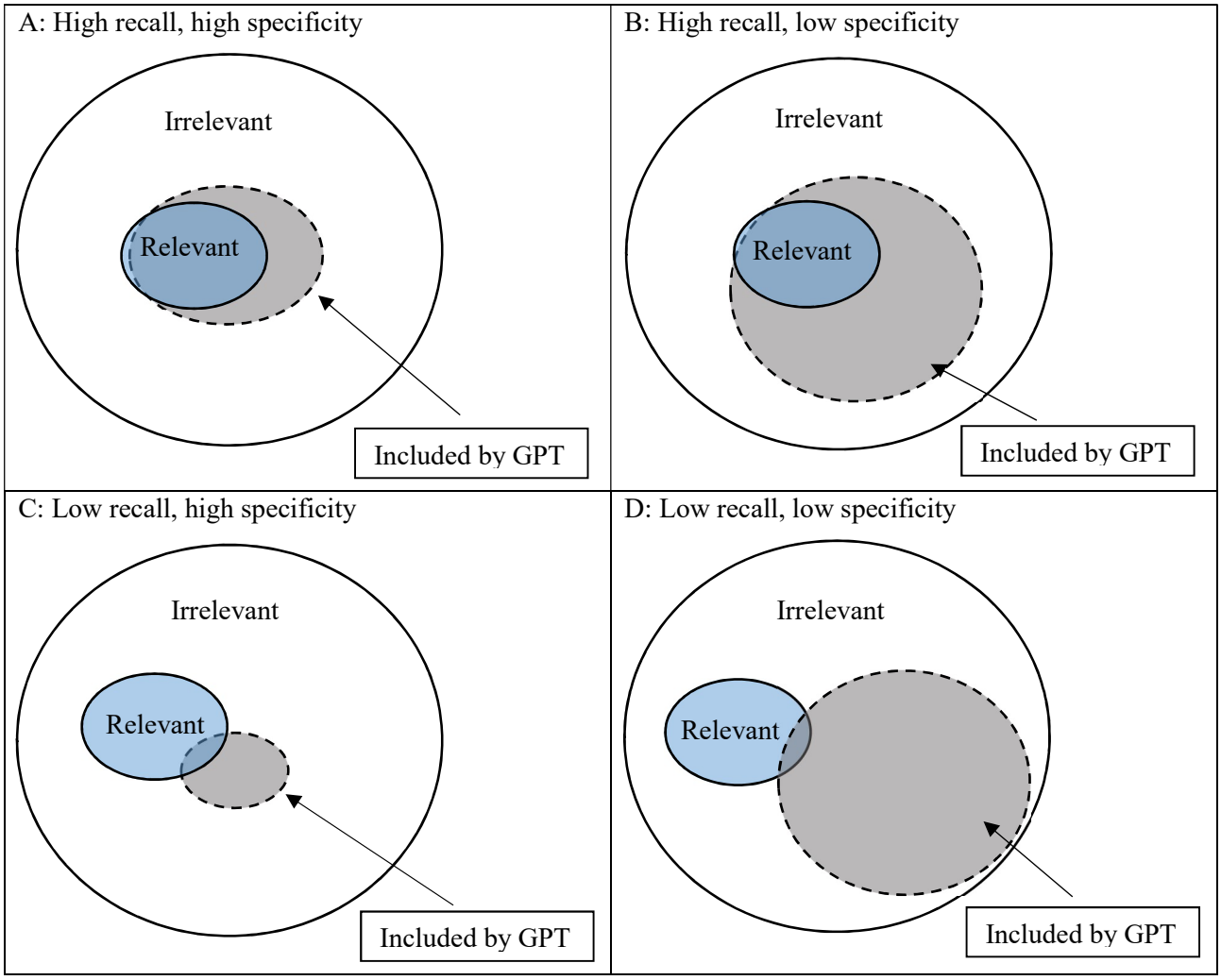
$$Specificity = \frac{TN}{TN + FP} \tag{2}$$

where TN (true negative) represents all the studies that are correctly excluded, and FP (false positive) is the number of studies falsely included. In this regard, we consider the recall measure to be the

absolute most important performance measure in our case since missing relevant studies, that is having a low recall, is the main reason for automated tools to potentially introduce a serious bias in systematic reviews (Hou & Tipton, 2024). Whereas, a low specificity "just" means that reviewers must re-examine the relevancy of a larger share of the total pool of references. If reviewers can be sure that they find all relevant studies but have a specificity of say 50%, this still implies that the reviewer can safely exclude 50% of the irrelevant records, which in most instances can be considered to be a significant reduction in the screening workload. Therefore, we think that automated tools should be accepted as long as they come close to scenarios A and B pictured in Figure 1. That is, they are accepted when high recalls can be made to a large extent independently of the accordingly specificity measure. Yet, this goes without saying low specificity rates should be accepted per se. We will come back to that in the following sections.

For our benchmark development, the TP, TN, FN, and FP conditions were determined by comparing the single human screener decision with the final decision agreed upon between a minimum of two human screeners. In our classifier experiment, the conditions were determined by comparing the GPT decision with the final decision made by a minimum of two independent human screeners.

FIGURE 1: Recall and specificity performances



Note: The blue-colored circles indicate the proportion of relevant title and abstract records; the gray-colored circles represent the proportion of records included by the screener; the white circles represent the proportion of irrelevant records that are correctly excluded by the screener.

The two above metrics concern the inclusion or exclusion performances individually but it might also be desirable to include metrics that incorporate the overall performance across the inclusion and exclusion metrics. A typical issue with such metrics is that they are very sensitive to imbalances in the data. That is for example when the proportion of irrelevant records is much larger relative to the proportion of relevant records, which is most often the case in systematic reviews. To exemplify, if one simply uses the raw agreement metric with imbalanced data then the screening performance will most often be overestimated. For example, assume that you have 10 relevant records per 1000 records, then you could end up reaching a raw agreement of 99% if the given screener just excluded all records. Although the screening performance seems to be high it obviously hides the fact

that the given screener was unable to detect any relevant studies. To overcome this issue, we used two overall metrics that account for imbalances. That is *the balanced accuracy* ($bAcc$) and *the normalized Matthew correlation coefficient* ($nMCC$). The former balances the accuracy of the performance across the recall and specificity metrics and is simply an average of those metrics, and is given by

$$bAcc = \frac{Rec + Spec}{2} \tag{3}$$

The $nMCC$ metric, on the other hand, is considered to be the metric that mostly maximizes the use of the four quantities, *TP, TN, FP,* and *FN* and it has been shown to have better statistical properties than other popular metrics such as the Receiver Operating Characteristic Curve (ROC AUC) (Chicco & Jurman, 2023). It can be calculated as follows.

$$nMCC = \frac{(TP \times TN - FP \times FN)}{2\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} + 0.5 \tag{4}$$

3.2 Human screening performance for benchmark development

In order to make fair comparisons between human and automated screening performances, we consider it pivotal to have a deeper understanding of acceptable human screening performance in high-standard systematic reviews (O'Connor et al., 2019). We believe that many previous evaluations of the performance of automated screening tools overlook the fact that individual human screening is not without significant errors either, and automated screening tools must be evaluated in light of this. If we as a community primarily assess the performance of automated tools and accept the tools with the requirement that they can detect (close to) all relevant studies in all instances or on par with very high human performances, then the tools seem by design to be doomed to fail. Automated screening tools will always err to some degree, as will humans (Waffenschmidt et al., 2019), and the important factor here is to ensure that the difference between the error rates is acceptable. What is acceptable is of course up to discussion but in the next sections, we develop a tentative benchmark scheme for interpreting (acceptable and unacceptable) error rates of screening performances in high-standard systematic reviews.

3.2.1 The data underpinning the benchmark scheme

The data we used to construct this benchmark scheme was based on the human screening performances in 22 high-standard systematic reviews that used independent duplicate human screening. This included 17 Campbell Systematic Reviews and 5 reviews conducted by the Norwegian Institute

of Public Health (NIPH). A descriptive overview of all the included reviews can be found in Table 2, including the imbalance in the given dataset. The included Campbell systematic reviews, represent all Campbell reviews that have been conducted by the Danish Center for Social Science Research in which independent duplicate human screening has been used and tracked. Concretely, this data includes 144,003 title and abstract records, all of which have been double-screened by 46 individual screeners of which 36 were student assistants and/or non-content experts, and 10 were researchers/authors of the given review, respectively. The Campbell reviews were conducted from 2015 to 2024. Since all of the included Campbell reviews drew on assistant (i.e., non-content-expert) screeners, this could potentially downward bias the evaluation metrics for various reasons. For example, assistants might lack sufficient profound content knowledge regarding the topic under review, potentially hindering them from reaching high recall rates. Thus their performances might not necessarily be comparable with the common screening performance of content expert screeners. Hence, we analyzed the Campbell review data separately for assistant/non-expert and researcher/expert screeners. However, relative recall and specificity rate differences between the two types of screeners could also be driven by authority imbalances between the often more senior content expert and the assistant screener, making the performances of the expert screeners look better than they actually were. Therefore, we added the screening performance data from five systematic reviews conducted by NIPH in which all TAB screenings and disagreements were conducted and solved by researchers with specific content knowledge related to the given review. This should, thereby, give a clearer picture of common expert/researcher performances in systematic reviews. This data added 13,825 title and abstract records that had been independently double-screened by 13 individual researchers. The five NIPH reviews were conducted from 2021 to 2024. When analyzing all of the above-presented data, we removed all training data to avoid inflating human disagreements unreliably. In other words, all presented screening performances represent after-training screening performances.

TABLE 2: Description of studies used to develop benchmark scheme

Source **Authors**	**Short title**	$n_{included}/N$	**Ass.**[a]	**Aut.**[b]
Campbell review				
Bøg et al. (2018)	Deployment of personnel to military operations	106/2899	2	-
Bondebjerg et al. (2023)	The effects of small class sizes on students' academic achievement, socioemotional development and well-being in special education	244/11860	4	2
Dalgaard, Bondebjerg, Klokker et al. (2022)	Adult/child ratio and group size in early childhood education or care to promote the development of children aged 0–5 years	258/3667	4	2

Dalgaard, Bondebjerg, Viinholt et al. (2022)	The effects of inclusion on academic achievement, socioemotional development and wellbeing of children with special educational needs	373/14491	5	2
Dalgaard, Filges et al. (2022)	Parenting interventions to support parent/child attachment and psychosocial adjustment in foster and adoptive parents and children	424/13106	3	2
Dalgaard, Jensen et al. (2022)	PROTOCOL: Group-based community interventions to support the social reintegration of marginalised adults with mental illness	557/17614	4	3
Dietrichson et al. (2020, 2021)	Targeted school-based interventions for improving reading and mathematics for students with or at risk of academic difficulties in Grades K-6 [plus 7-12]	2952/15273	6	1
Filges, Dalgaard et al. (2022)	Outreach programs to improve life circumstances and prevent further adverse developmental trajectories of at-risk youth in OECD countries	387/4890	4	-
Filges, Dietrichson et al. (2022)	Service learning for improving academic success in students in grade K to 12	619/6269	4	1
Filges, Montgomery, et al. (2015)	The Impact of Detention on the Health of Asylum Seekers	573/10061	2	-
Filges, Siren et al. (2020)	Voluntary work for the physical and mental health of older volunteers	43/14919	2	0
Filges, Smedslund et al. (2023)	PROTOCOL: The FRIENDS preventive programme for reducing anxiety symptoms in children and adolescents	96/2745	1	1
Filges, Sonne-Schmidt et al. (2018)	Small class sizes for improving student achievement in primary and secondary schools	303/7802	5	1
Filges, Torgerson, et al. (2019)	Effectiveness of continuing professional development training of welfare professionals on outcomes for children and young people	298/5147	1	4
Filges, Verner et al. (2023)	PROTOCOL: Participation in organised sport to improve and prevent adverse developmental trajectories of at-risk youth	158/7021	2	1
Thomsen et al. (2022)	PROTOCOL: Testing frequency and student achievement: A systematic review	627/6239	5	2
NIPH review				
Ames et al. (2024)	Acceptability, values, and preferences of older people for chronic low back pain management	144/425	-	2
Evensen et al. (2023)	Sutur av degenerative rotatorcuff-rupturer [Rotator cuff repair for degenerative rotator cuff tears]	418/2499	-	4
Jardim et al. (2021)	Effekten av antipsykotika ved førstegangspsykose [The effect of antipsychotics on first episode psychosis]	73/3924	-	3
Johansen et al. (2022)	Samværs-og bostedsordninger etter samlivsbrudd [Custody and living arrangements after parents separate]	143/1525	-	4
Meneses Echavez et al. (2022)	Psykologisk debriefing for helsepersonell involvert i uønskede pasienthendelser [Psychological debriefing for healthcare professionals involved in adverse events]	45/5452	-	3

Note: *a*. Ass. denotes student/non-content expert screener; *b* Aut. denote authors of the review

3.2.2 Statistical analysis used to derive benchmarks

All statistical data analyses were conducted using R 4.4.0 (R Core Team, 2022) in RStudio (RStudio Team, 2015). For the main analyses, we used the package metafor (Viechtbauer, 2010), including the sandwich estimators herein (Pustejovsky, 2020). To work with ris-file data, we used the revtools package (Westgate, 2019). All materials behind this article can be accessed at https://osf.io/apdfw/.

From the data presented in the previous section, we estimated all the performance metrics via Equations (1) to (4). The TP, TN, FP, and FN conditions used in these equations were determined by comparing the single human screener decision with the final decision agreed upon between a minimum of two human screeners. When working with proportion metrics such as the ones presented in Equations (1) to (3), it is usually advantageous to transform these metrics into measures that have more appropriate statistical properties. This includes having a sampling distribution that more closely mirrors a normal distribution and a variance component that can more reliably be approximated (Viechtbauer, 2022). Therefore, we used the arcsine transformation (Röver & Friede, 2022; Schwarzer et al., 2019) to calculate sampling variance and confidence intervals for the *recall*, *specificity*, and *balanced accuracy metrics*. For the balanced accuracy metric, we calculated the sampling variance of the transformed measure by using the total number of records as the sample size. We did not use double arcsine transformation (Doi & Xu, 2021) due to the inadequate properties of the back transformation of this measure (Röver & Friede, 2022; Schwarzer et al., 2019). For the *nMCC* metric, we calculated the sampling variance and confidence interval by transforming the correlations to Fisher's z-scores, as typically done in meta-analysis (Borenstein et al., 2009).

To derive the overall average performances across the *recall*, *specificity balanced accuracy metrics*, and the *nMCC* metrics, we fitted two versions of the so-called *correlated-hierarchical effects* (CHE) working models (Pustejovsky & Tipton, 2021). For investigation related to the differential performances between assistant and author screeners, we applied the *subgroup correlated effects* (SCE+) model, whereas we used the CHE-RVE model when analyzing the NIPH performance data. Both types of models account for the multi-level structure of the data with the screener performance measures nested within studies. At the same time, the models account for the correlation between the within-study performance estimates. The sample correlation, ρ, is often entirely or partially unknown and must be imputed. In all the used working models, we assumed $\rho = .7$. To guard against model misspecification both models have incorporated robust variance estimators. The main difference between the two models is that the SCE+ model concerns the use of different weighting schemes but the SCE model is generally recognized as the main working horse for deriving subgroup effects

and conducting reliable contrast tests (Pustejovsky & Tipton, 2021). For differential effects comparisons, we used the HTZ Wald test suggested by Tipon and Pustejovsky (2015). Across both models, we estimated two sources of heterogeneity. That is the variability of the true performance differences within (ω) and between studies (τ). This allowed us to investigate at what level the largest true difference between the human screener performances existed.

3.2.3 Results

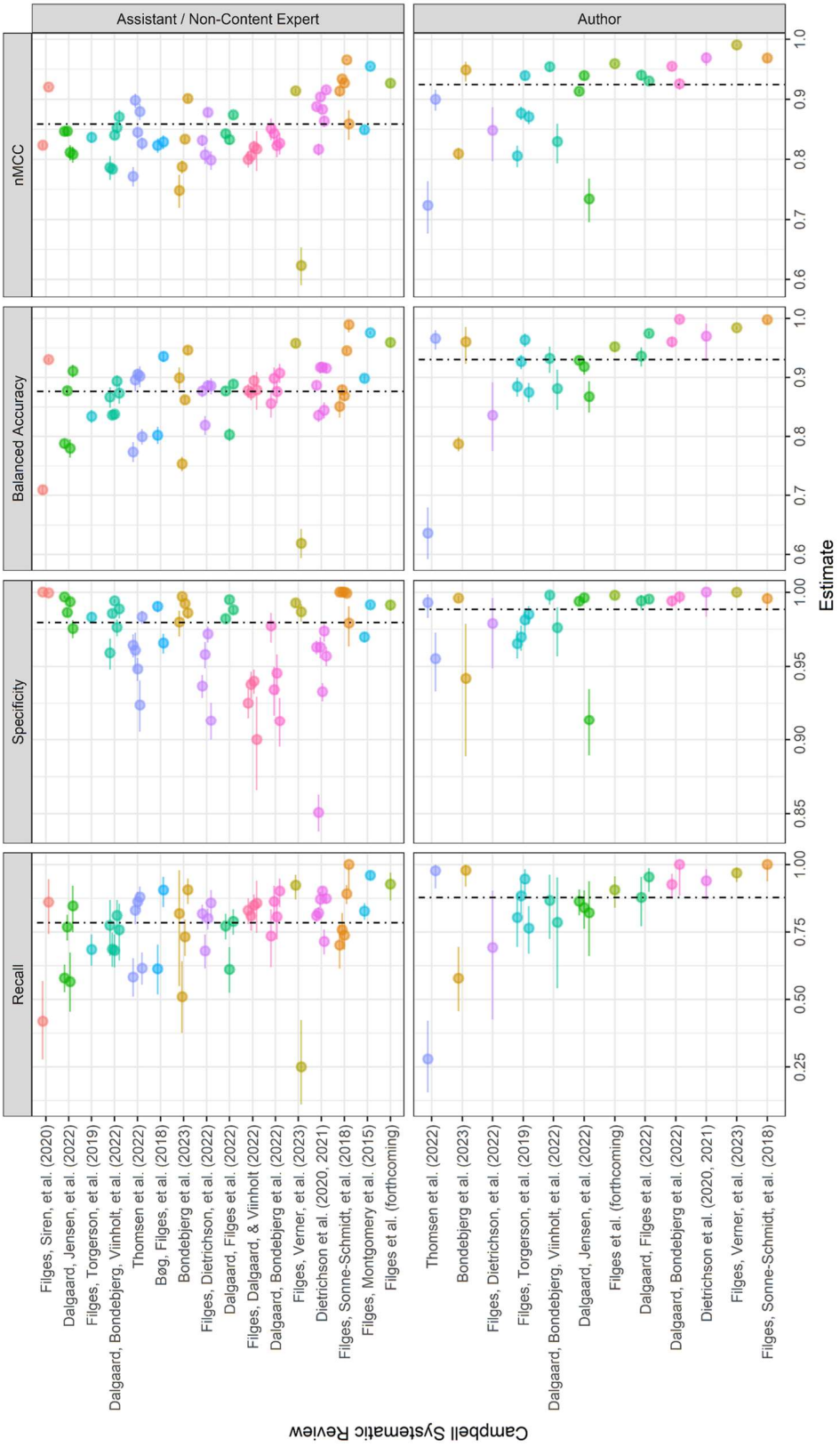
All individual screening performances across the included reviews and how these are distributed around the overall performance means are exhibited in Figures 2 and 3. We found the overall average recall rate for the assistant and author screeners in the included Campbell reviews to be 0.782, 95% *CI*[0.747, 0.817] and 0.881, 95% *CI*[0.823, 0.931], respectively. Hereto, we found the two groups averaged recalls to be statistically distinct from each other with $F(1, 10.3) = 14.58$, $p = .003$. We detected minor substantial variations between the performance measures within studies with $\omega = 0.026$ and $\omega = 0.035$ for the assistant and author screeners, respectively. We were not able to detect any true differences in performances between studies, indicating that the average screening performance seems to be consistent across the Campbell reviews both for assistants and expert screeners. The overall average specificity for assistant screeners was 0.980, 95% *CI*[0.966, 0.990], and for review authors 0.988, 95% *CI*[0.980, 0.995]. We found no statistically significant difference between the two average estimates with $F(1, 13.6) = 2.08$, $p = 0.172$. We did only find very minor non-substantial variation within and between studies with the within-study variability $\omega = 0.004$ as the maximum for author screeners.

For assistant screeners, the average balanced accuracy was 0.874, 95% *CI*[0.857, 0.890], and for authors screeners it was 0.933, 95% *CI*[0.899, 0.961]. We found the difference between the group means to be statistically significant with $F(1, 10.1) = 18.22$, $p = .002$. Finally, the overall *nMCC* was 0.860, 95% *CI*[0.835, 0.882] and 0.925, 95% *CI*[0.880, 0.953] for the assistant and author screeners, respectively. These averages were found to be statistically different with $F(1, 11) = 9.65$, $p = .01$.

Based on these results it might look like that research screeners are substantially better at detecting relevant studies relative to assistant screeners. Yet this difference can be driven by other factors than the actual screening performance of the assistants. Interestingly, when investigating the NIPH data, which was only based on independent researcher-researcher screening comparisons, we found performance patterns closer to the performance of the assistant screeners in the included Campbell review. The overall recall rate in the NIPH data was 0.839, 95% *CI*[0.737, 0.920]. Again, we

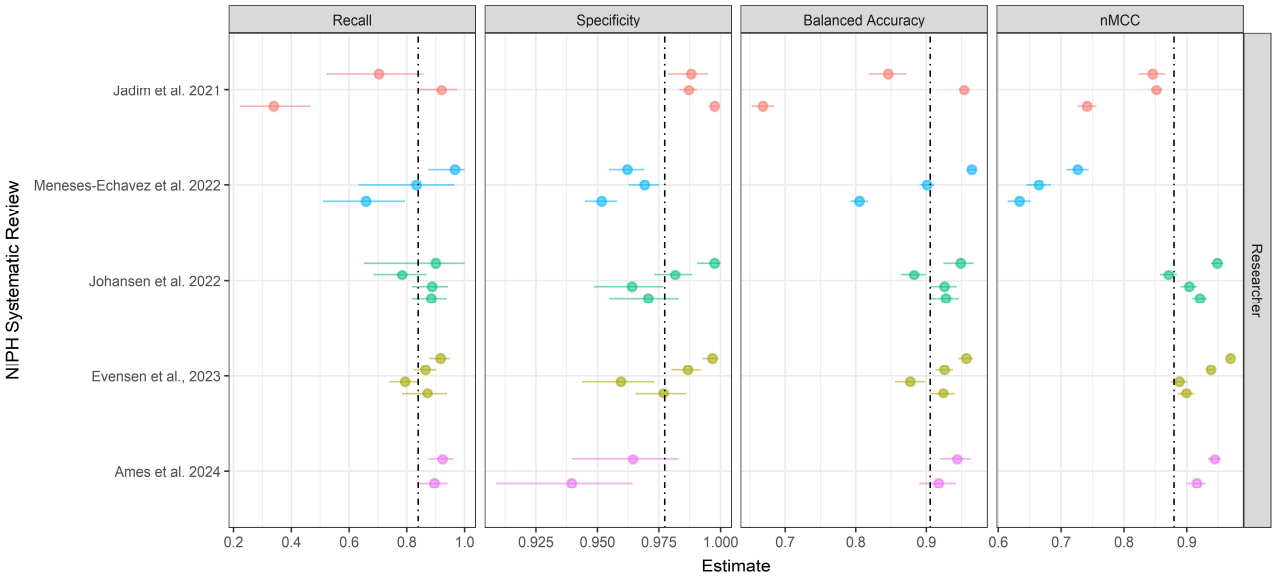
primarily found minor true variation between the screener recall performances within studies with ω = 0.029 and τ = .0. The overall average specificity rate was 0.977, 95% *CI*[0.955, 0.992], with almost no true variability either at the levels of the performance measure or the study. The overall average balanced accuracy and nMCC were 0.905, 95% *CI*[0.859, 0.943] and 0.879, 95% *CI*[0.720, 0.951], respectively.

FIGURE 2. Performance measures within Campbell Systematic Reviews across assistant vs. author screeners. Dashed lines indicate the average estimated via the SCE+ model.



Note: Dashed lines indicate the average estimated via the CHE-RVE model

FIGURE 3. Researcher-researcher screening performance measures within NIPH Systematic Reviews.



Note: Dashed lines indicate the average estimated via the CHE-RVE model

3.2.4 Benchmark scheme

Bearing on the empirical results presented in the previous section, we developed the benchmark scheme presented in Table 3. On this basis, we suggest a course-grained rule of thumb saying that automated tools that can be shown to have recall equal to or above 80% and specificity rates above 95% resemble common human screener performance. There are of course nuances to this broad guideline since we believe that automated tools can also be useful under less restrictive conditions as well. Hold against common human error rates, we consider a recall rate between 0.75 to 0.80 to be acceptable because it closely mirrors the common recall rate of assistant screeners. As a consequence, and in contrast with previous evaluations (Guo et al., 2024), we would not necessarily interpret a recall rate of 0.76 to be so low that it excludes the GPT models to function as an independent second screener. Also, we believe that automated tools can still be viable even when they yield a recall rate between 50-75%. Under such conditions, the automated tool could function as an extra assurance, working as a third screener that forces the duplicate human screeners to double-check close-to-relevant study records. This would enhance the screening, ensuring that the human screeners have not overlooked any relevant records.

As can be seen from the benchmark scheme, we do not necessarily conceive a specificity of 100% to be ideal, since we think it is alright that the GPT API models are over-inclusive. This forces the reviewers to double-check close-to-relevant references which in turn assures that fewer or

no relevant studies are missed. Thus, we think that a specificity rate equal to or above 80% is acceptable as long as the recall rate is also equal to or above 80% (c.f. Figure 1). Consequently, we suggest that automated screening performances that reach recall and specificity rates above 80% should be accepted as independent screeners in high-standard systematic reviews.

Finally, we think that automated tools that yield high recalls can be used to reduce the total amount of title and abstract records needed to be screened even if the specificity rate is below 80%. This would especially be relevant when working with very large amounts of title and abstract records (see an example of this in Shemilt et al., 2014). As a course-grained guideline, we do not consider it viable to use automated tools when they yield performance measures below 0.5.

TABLE 3: Screening performance benchmarks

	Values				
Metric	**.0 < 0.5**	**0.5 < 0.75**	**0.75 < 0.8**	**0.8 < 0.95**	**0.95 <**
Recall	Ineligible performance	Not ideal performance. Only use for extra security as a third screener (Can be used as second screener if resources are scarce since the alternative is worse)	On par with non-content expert screeners. Can be accepted.	On par with common researcher screening performance	Better than common human performance and traditional machine learning tools
Specificity	Ineligible performance	Not ideal performance. Only use to reduce the total number of records if having a high recall.	Low performance. Only use to reduce the total number of records if having a high recall.	Can be accepted if having a high recall rate above 80%	On par with common human screening performance

Note: Red areas indicate conditions under which a TAB screening performance is unacceptability low. Gray areas represent insufficient performance conditions but some applications with these performance measures are still viable. Green shaded areas represent acceptable, on par with common screening performance, or better than human screening, respectively.

With this benchmark scheme, we aim to make a more flexible tool partially for assessing the screening performance of automated tools and partially for assessing which screening tasks can be made under what performance conditions. This allows for more case-specific discussions regarding the adequacy of using GPT API models for TAB screening tasks in systematic reviews.

4 CLASSIFIER EXPERIMENT[8]

In the following section, we present the data and prompts used as well as the results for three large-scale classifier experiments. Differently, from previous research, these classifier experiments aimed to test the performance of GPT API models 1) when applied in social science reviews, 2) when the request body includes the use of function calling, avoiding the need to describe response behavior in the screening prompt(s), and 3) when using multi-prompt screening in complex review settings. A side-effect of conducting these experiments was further to quality assure the AIscreenR package and ensure that the software yields appropriate screening behavior. We considered this test to be all-important if our suggested screening approach shall be scaled up. We narrowed the investigation to only include three experiments since the purpose of this paper is not primarily to show that GPT API models work in all instances across all types of reviews. Instead, we aim to show that if set up adequately these models can function as *highly* reliably independent second screeners. This also means that using GPT API models as a second screener is not always ideal for various reasons. We return to this issue in Section 5.2.

4.1 Data

In classification experiment 1, we tested the performance of AIscreenR in the context of a Campbell systematic review concerning the effects of functional family therapy (FFT) on drug abuse reduction for young people in treatment for nonopioid drugs conducted by Filges et al. (2015). By leveraging a previously published review, we were able to immediately evaluate AIscreenR's performance against the inclusion and exclusion decisions made by two human screeners during the original review. Moreover, the inclusion criteria of the review were rather simple and the intervention represented a well-defined intervention. This made it an ideal initial test case for our proof of concept purposes. Thus, if AIscreenR could not achieve satisfactory performance in this context, it would unlikely be able to do so in the context of more complex reviews. Another interesting feature experiment is that it was based on a highly imbalanced dataset with only 69 relevant records out of 4135 records. That amounts to an approximate inclusion ratio of 17 relevant in 1000 records. This made it an ideal case to test if and to what extent screening with GPT API models was sensitive to data imbalances as is the case with all traditional semi-automated tools (König et al., 2023).

[8] All replication materials behind this experiments can be accessed at https://osf.io/apdfw/.

A critique against classification Experiment 1 is it draws on a published open-access review, meaning OpenAI's GPT models can potentially have been trained on this review. If this is the case, this possibly excludes the opportunity to generalize the results of this experiment to applications where GPT API models are used to conduct screenings on prospective reviews where no previous information has been fed to OpenAI's GPT models. To test and potentially overcome this issue, we conducted a second classification experiment drawing on data from an unpublished/ongoing systematic review. In classification experiment 2, we used screening data from a Campbell systematic review regarding the effects of the FRIENDS preventive programme on anxiety symptoms in children and adolescents conducted by Filges Smedslund et al. (2023).[9] This FRIENDS data in many aspects resembles the FFT data. For example, the inclusion criteria were rather simple and the intervention is well-defined. Moreover, the data is highly imbalanced with 64 relevant records in 2572 records, amounting to an approximate inclusion ratio of 25 relevant per 1000 records.

A fair critique of experiments 1 and 2 is that they both represent very simple TAB screening cases, not necessarily resembling common systematic review structures. To accommodate this issue, we, therefore, conducted a third experiment that aimed to investigate if GPT API models can be used for TAB screening in what we consider to be a very complex review setting. For classification experiment 3, we used screening data from an ongoing Campbell Systematic Review of the effects of different testing frequencies on student achievement (Thomsen et al., 2022). We considered this to be a difficult screening case because the review draws on rather subtle inclusion/exclusion criteria that include notions that are not well-defined. One of the reasons why this particular case is difficult is that the intervention (i.e., student testing) is a type of learning strategy that is ubiquitous in education and used in a variety of ways and for very different purposes. Testing can be used as a formative tool, e.g. to promote retention of academic content, adjust instructional strategies, and uncover student needs for remediation or more intensive support. In most school systems, testing is also used summatively for assigning grades, determining graduation or certification, and for school accountability assessment. Important to note here is that the distinction between formative and summative testing is not clear-cut as tests can serve both purposes simultaneously and can have more or less stakes attached to them, both from a student and from a school perspective. It follows that testing is not a uniform type of intervention, but a multi-facetted phenomenon encompassing a variety of approaches and a heterogeneous terminology (tests are not just called tests, but may also be referred to

[9] We conducted this experiment on the 4th of November 2023. This was before the corresponding protocol where published on the 15th of December 2023.

as e.g. quizzes, progress-monitoring, curriculum-based measures, and retrieval practice). Judging the eligibility of particular interventions therefore requires subject matter familiarity. Therefore, and contrary to Experiment 1, we think that if the GPT API models can achieve satisfactory performances in this context, they would likely be able to do so in most review contexts. The data we used for the experiment consisted of 2000 irrelevant and 100 relevant records randomly sampled from the total pool of 5612 irrelevant and 627 relevant records. We did so to ease the screening since this screening was based on multi-prompt screening, meaning that all title and abstract records were screened with six individual prompts each including one inclusion criterion.

For all datasets, we excluded all study records without an abstract. This excluded 208, 150, and 41 study records for the FFT, FRIENDS, and testing frequency (henceforth TF) data, respectively. For the FRIENDS data, we further deleted 20 titles and abstracts containing a myriad of special symbols, causing the GPT response to return insufficient JSON data from the server.

4.2 Prompt engineering

For Experiments 1 and 2, we engineered prompts so that they included an introduction section describing the general aim of the review followed by the inclusion/exclusion criteria of the review. To exemplify, Textbox 1 exhibits the prompt used for classifier experiment 2.

TEXTBOX 1: Prompt example

"We are screening studies for a systematic literature review. The topic of the systematic review is the effect of the FRIENDS preventive programme on reducing anxiety symptoms in children and adolescents. The FRIENDS programme is a 10-session manualised cognitive behavioural therapy (CBT) programme which can be used as both prevention and treatment of child and youth anxiety. The study should focus exclusively on this topic and we are exclusively searching for studies with a treatment and a comparison group. For each study, I would like you to assess: 1) Is the study about the FRIENDS preventive programme? 2) Is the study estimating an effect between a treatment and control/comparison group?

Short/concise background information

Inclusion/exclusion criteria

Then when given study IDs (if not provided by the user, these are automatically generated), titles, and abstracts, the AIscreenR automatically pastes together the text presented in Textbox 2:

TEXTBOX 2: End of prompt added by AIscreenR

"Now, evaluate the following title and abstract for Study [the study id is inserted here]: -Title: [the study title is inserted here] -Abstract: [the study abstract is inserted here]"

Pasting the prompt together with each title and abstract aims to guard against model drifting/hallucinations. As previously mentioned, we did not add any instruction regarding how the model should respond to our request in the main prompt, as done in previous research evaluations. Instead, we built two JSON functions providing this instruction to the model (OpenAI, 2024). This should theoretically ensure that we get more reliable and standardized responses from the models. The main JSON respond function[10] we built included the instructions presented in Textbox 3:

TEXTBOX 3: Function call text

"If the study should be included for further review, write '1'. If the study should be excluded, write '0'. If there is not enough information to make a clear decision, write '1.1'. If there is no or only a little information in the title and abstract also write '1.1'. When providing the response only provide the numerical decision."

In the initial phase of our prompt engineering, we assumed that the more detailed background information we could add to a single prompt the better the GPT API model would perform. This approach was based on the conception that the model needed to be "trained" with the correct wording. Yet, from our experience, this is a misperception of how this type of model works. As indicated in the GPT acronym, these models are *pre-trained*, meaning that they do not need to be further trained in terms of wording. Instead what they need to work properly are concise (into-the-bone) prompts. Said differently, less is more. It was not that the models were entirely off but the test performance of the models dramatically increased when given more precise prompts with fewer inclusion/exclusion criteria. Therefore, for Experiment 3, we developed and evaluated the concept of multi-prompt screening where each inclusion/exclusion criteria were prompted individually. This approach to a large extent resembles the common way screening guidelines are constructed and used by humans when conducting first-level screening of titles and abstracts (see Valentine, 2009, p. 141).

[10] Find the exact functions here: bit.ly/3Vl0SRp

All engineered prompts used for Experiment 3 are presented in Appendix A. We elaborate further on multi-prompt screening in Section 5.1. [11]

4.2.1 Performance tests

Before initiating the three classifier experiments, we tested the performance of our developed prompts. For the FFT review, we started by testing the prompt on one relevant reference only. Hereto, we refined the prompt until the models consistently included this particular study record. Then, we scaled up the test to include 200 references, including 150 irrelevant and 50 relevant records. This test yielded results very similar to the ones presented in Table 4. Thereafter, we screened all records with the GPT API models to investigate whether the test performances persisted when used in the full sample of records. For both the FRIENDS and TF reviews, we tested the prompts on 150 irrelevant and 50 relevant study records randomly sampled from the total pool of irrelevant and relevant records, respectively. After detecting results very similar to the ones presented later in Table 4, we initiated the full screening.

4.3 Evaluation design

In all three classifier experiments, we evaluated the performance of the GPT API model by using Equations (1) to (3). In this regard, the TP, TN, FN, and FP conditions were determined by comparing the GPT decision with the final decision made by agreement between at least two independent human screeners. Human inclusion at this first level of screening did not necessarily imply that study records were relevant for the final review—merely that they were considered to be relevant for full-text screening. In Experiments 1 and 2, we used the gpt-3.5-turbo-0613 and gpt-4-0613 reached from the 'v1/chat/completions/' endpoint. Since the gpt-3.5 models are generally considered to be less accurate in their response, we repeated the same screening 10 times for each title and abstract when using this model, as also done by Syriani (2023). We did so to test its consistency across the screenings and how it impacted its final inclusion decision. The final inclusion decision of GPT was then based on the probability of inclusion across the repeated requests. In part because the GPT-4 models are considered to be more accurate (meaning that they are more consistent in their responses) and in part because of the cost, we only conducted one screening per title and abstract when calling gpt-4-0613. For Experiment 3 involving multi-prompt screening, we only drew on GPT-4 and the final inclusion decision of GPT was then based on the probability of inclusion across all used prompts. In

[11] Moreover, we show how this can be done in one of the accompaning vignette to the AIscreenR package.

our main analysis of Experiment 3, we included study records if they were included by GPT in at least 5 out of the 6 used prompts. For all experiments, we used invariant top_p and temperature values. That is the default value of 1 for both hyperparameters. As previously mentioned, we interpreted the results of the experiments using the benchmark scheme developed in Section 3 (cf. Table 3).

4.4 Results

All results for the three classifier experiments are presented in Table 4. As can be seen from Table 4, the gpt-4 model yielded recall and specificity rates equal to 89.9% and 93.3% in Experiment 1, which can be considered to be on par with human screening. The gpt-3.5-turbo model was also able to reach human-like screening performances. Yet these results were substantially impacted by the chosen inclusion probability threshold, indicating that these model generally yields rather inconsistent decisions, especially when it comes to detecting relevant studies. Figure 4A shows the decision sensibility of the gpt-3.5 model across inclusion probabilities for the FFT data. When setting the inclusion probability equal to 0.2 (meaning that gpt-3.5 included the study in at least 2 out of 10 screenings), the gpt-3.5 model yielded a recall of 81% and a specificity of 93.7%. However, when setting the inclusion probability equal to 0.5 yielded a performance unacceptably low compared to human screening with a recall below 80%.

When used on the FRIENDS data, the gpt-4 model performed extremely well with performance measures that can considered to exceed common human screening performances. Concretely, it yielded a recall of 98.4% (only missing one relevant study) and a specificity rate of 97.4%. In this regard, the gpt-3.5 model performed closely on par with humans with a recall of 95.3% and specificity of 89.9% when the inclusion probability was set to be 0.7. Yet again, the performance of gpt-3.5 model was highly influenced by the chosen inclusion probability. Figure 4B shows the decision sensibility of the gpt-3.5 model across inclusion probabilities for the FRIENDS data. An impressive fact regarding these results is further that we approximately spend 5-10 minutes engineering the used prompt presented in Textbox 1. In fact, the prompt represents a first trial prompt. Unless we were extremely lucky to hit the right prompt in the first trial, this might indicate that in some applications the GPT API models are not as prompt-sensitive as often expected. Yet we only presume this to be the case in very specific circumstances as is the case here where the screening involves a standardized intervention with a specific name.

Finally, when used on the TF data, the gpt-4 model yielded a recall performance on par with humans with a recall of 80% when studies were included in at least 5 out of 6 prompts. This,

furthermore, exceeded three out of six human recalls within this review (see Thomsen et al. (2022) under column 1 in Figure 2). The model also yielded an acceptable specificity, that is a specificity of ~84%. Relative to human performance, the model in this case was rather over-inclusive. Yet again, we do not necessarily consider this to be disadvantageous, since it provides extra insurance of not overlooking relevant studies, which might be even more important in complex review settings where hard decisions are difficult to make at the first level of screening due to insufficient information in the abstracts. Since these results are underpinned by data from an unpublished review, a side-effect of this experiment is further that it shows that GPT API models can work in complex settings where they have not been pre-trained. Moreover, when studies were included in at least 3 out of 6 prompts, the gpt-4 model was able to reach a recall of 95%, but the specificity was quite low, that is 67%. However, if this approach was used it could potentially, and most importantly safely, reduce the total screening workload by approximately 64%.

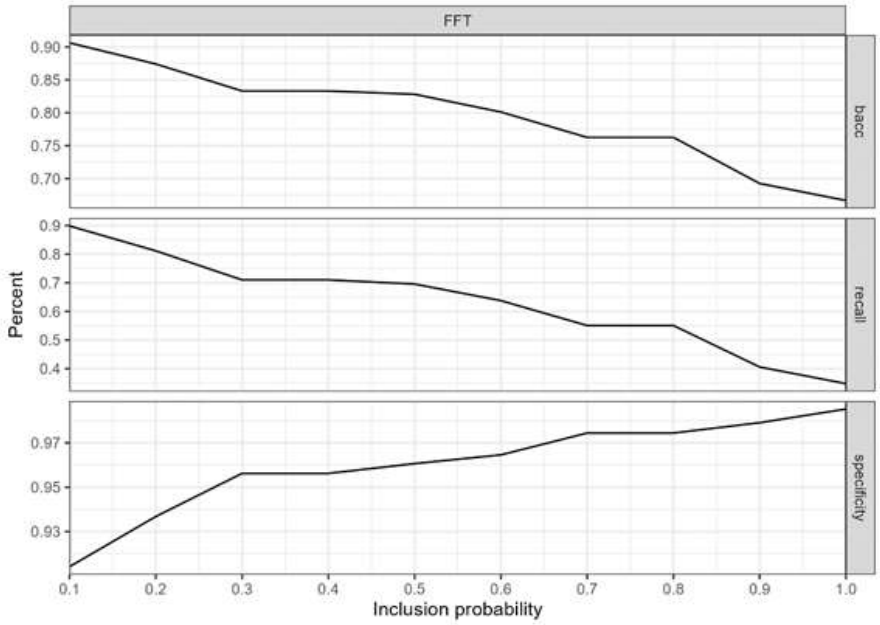
TABLE 4: Results of the two classifier experiments

Review Model	**Reps Per Prompt**	**Recall (%) [TP/(TP + FN)]**	**Specificity (%) [TN/(TN + FP)]**	**Raw agreement (%) [(TP + TN)/*N*][a]**	**bAcc (%)**
FFT					
gpt-3.5-turbo-0613 (incl. prop = .5)	10	69.9 (48/69)	96.1 (3906/4066)	95.6 (3954/4135)	82.8
gpt-3.5-turbo-0613 (incl. prop = .2)	10	81.2 (56/69)	93.7 (3809/4066)	93.5 (3865/4135)	87.4
gpt-4-0613	1	89.9 (62/69)	93.7 (3810/4066)	93.6 (3872/4135)	91.8
FRIENDS					
GPT-3.5-turbo-0613 (incl. prop = .5)	10	95.3 (61/64)	81.3 (1918/2508)	81.6 (2100/2572)	88.3
gpt-3.5-turbo-0613 (incl. prop = .7)	10	95.3 (61/64)	89.9 (2254/2508)	90.0 (2315/2572)	92.6
gpt-4-0613	1	98.4 (63/64)	97.4 (2442/2508)	97.9 (2518/2572)	97.9
TF					
gpt-4-0613 (incl. ≤5 out of 6 prompts)	1	80 (80/100)	83.8 (1676/2000)	83.6 (1756/2100)	81.9
gpt-4-0613 (incl. ≤ 4 out of 6 prompts)	1	89 (89/100)	74.3 (1486/2000)	75 (1575/2100)	81.6
gpt-4-0613 (incl. ≤ 3 out of 6 prompts)	1	95 (95/100)	67 (1340/2000)	68.3 (1435/2100)	81

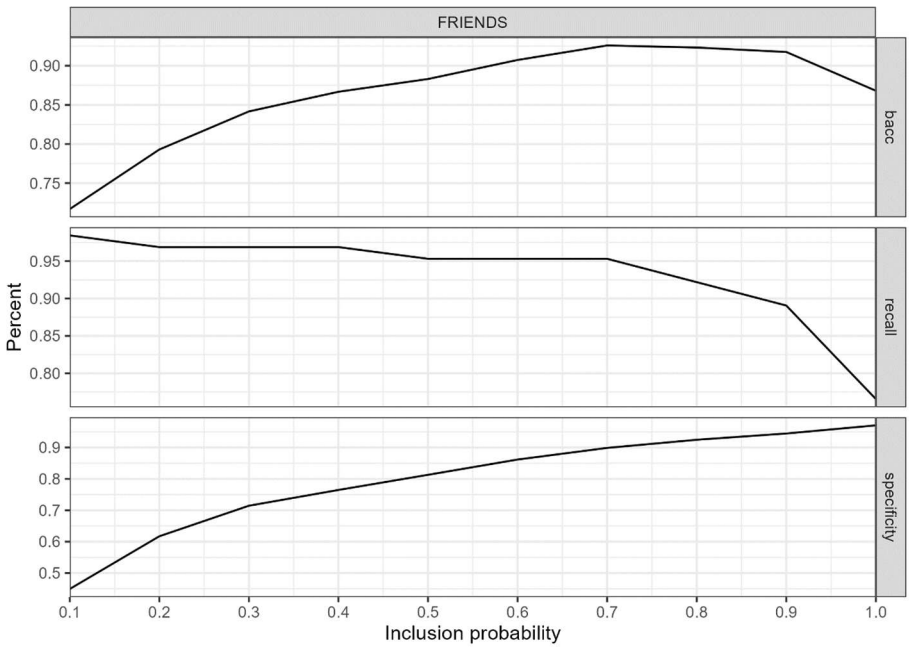
a: N is the total number of references

FIGURE 4: Decision sensibility of the gpt-3.5-0613 model across inclusion probabilities

A



B



To summarise, we derive the following conclusions from the classifier experiments. First, we found that GPT API models can work as highly reliable and independent second screeners with recall performances on par or even better than common human screeners, even in very complex screening settings. This finding contrasts previous evaluations (Gargari et al., 2024; Guo et al., 2024)

finding that the GPT API models mainly have high performances in terms of correctly excluding irrelevant records. This discrepancy might be explained by the fact that we drew on function calling aiming to provide more reliable responses from the GPT API models and that we used multi-prompt screening, circumventing that all inclusion/exclusion criteria were added to a single prompt. Second, and in contrast with the performance of classical semi-automated screening tools (König et al., 2023), we partially found that GPT API models are not sensitive to imbalanced data and partially that the GPT-4 API models are capable of reaching recall rates close to 100%. Third, since we used the AIscreenR software to conduct all classifier experiments, we feel safe to conclude that the software seems to work as expected. Hence, we believe that reviewers can safely use this software in high-standard systematic reviews. Fourth, we found that the GPT API models are not always as prompt-sensitive as suggested in previous evaluations (Gargari et al., 2024). Fifth, we found the GPT-4 API model to be preferable relative to GPT-3.5 since the latter is rather sensitive to the chosen inclusion probability across multiple identical screenings. Based on this finding, we generally recommend not using the GPT-3.5 API models when GPT-4 API models are available. Finally, we found that in some applications, the specificity rate reached by the GPT-4 API model can be seen to be on the lower end compared with human screeners. Yet, we do not find this to be a major issue when having high recall rates (cf. Figure 1) since this can just be seen as an extra opportunity to double-check close-to-relevant studies. Thus, enhancing the change of not overlooking any relevant study records.

Overall, we think that using GPT API models for TAB screening tasks in state-of-the-art systematic reviews has huge potential—also as independent second screeners in complex reviews. Furthermore, we believe that the relevancy of using LLMs will only increase over time as the models improve, which demands a standardized setup to ensure reliable use of these in systematic reviews. In the next section, we, therefore, develop a tentative guideline and workflow for how such screening can be set up in practice.

5 TENTATIVE GUIDELINES AND WORKFLOW

Premised on our developed benchmark scheme, our experience, and the results of the three classifier experiments, we have developed the following tentative guidelines and workflow for when and how GPT API models can be used as independent second screeners of titles and abstracts. All steps in this process are fleshed out in Table 5.

TABLE 5: Workflow for how to conduct TAB screening with GPT API models

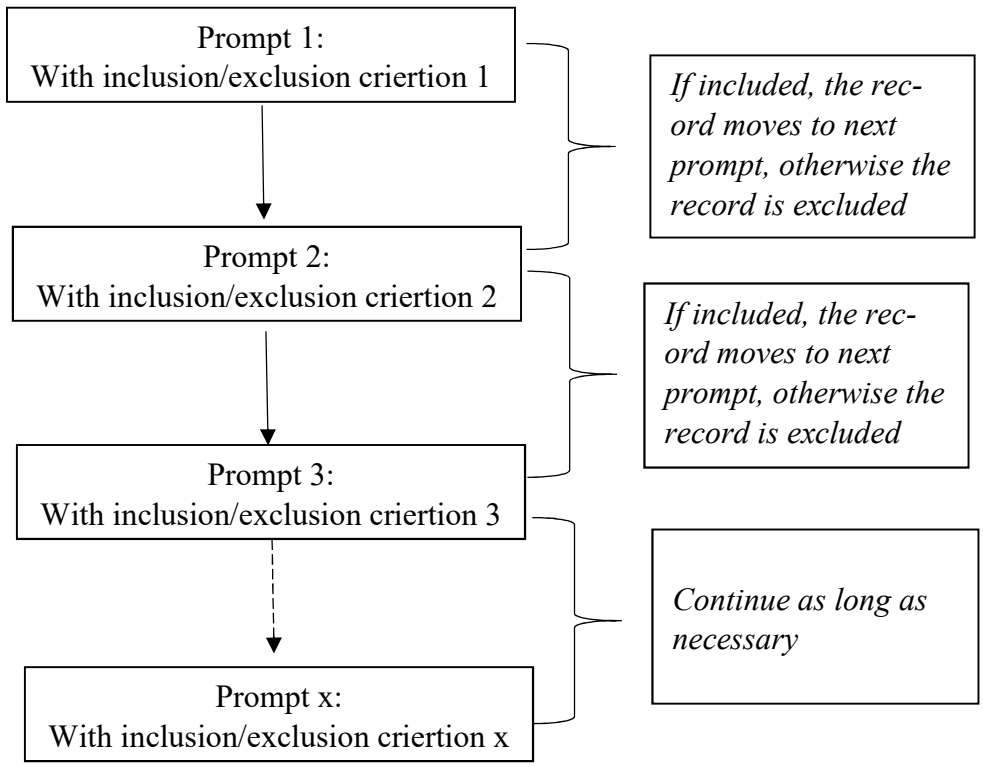
Step	**Reviewer action**
1	Find approximately 10 relevant study records (ideally more).
2	Find a minimum of 200 irrelevant study records (ideally randomly sampled from the pool of records).
3	Construct the test dataset by combining the records from steps 1 and 2.
4	Develop one or multiple prompts and test the(ir) performance.
5	Repeat/refine step 4 until reaching a recall close to 80% or more, and a specificity between 90-100%. If this step cannot be fulfilled, we recommend not to use the GPT API model as a second screener. Thus, human double screening is the ideal solution. Yet, the GPT API model can still be used as a third screener for extra insurance of not missing any relevant studies. In cases where low budgets exclude human duplicate screening, we considered it fair to work with recall performances below 80% since the alternative (i.e., stand-alone single-screening) in these cases is worse.
6	Manually single screen all study records (could be divided into batches of 500-1000 study records). If a GPT API model has shown to be a reliable second screener based on the text data, then this can be done by multiple reviewers/screeeners.
7	Download ris-files individually for included and excluded references. Load this data into R and track the human decision.
8	Run the full TAB screening with the GPT API model. Consider removing all study records without an abstract and human screen those references.
9	Investigate and solve disagreements between the human and automated screening decisions.

Note: See the vignette accompanying the AIscreenR package for a detailed presentation of the to conduct TAB screening with GPT API models in practice.

Before initiating a full-scale TAB screening with GPT API models, we generally recommend thoroughly testing the screening performance of the prompt(s) and GPT API model aimed to be used for the screening until it is ensured that the screening performances pass certain thresholds within the training setting. The first step of the testing procedure involves locating approximately 10 relevant and 200 irrelevant study records including titles and abstracts, respectively. Locating more than 10 relevant study records might be ideal to test if the prompt(s) can detect various types of relevant records. That said, we experienced that using fewer than 10 relevant records could also unveil

a proper recall performance of the prompt and models in more simple screening cases. Consequently, we cannot set this step in stone. When locating irrelevant records, we suggest randomly sampling those from the total pool of records. This aims to ensure that the specificity test rate can be generalized to the full sample of study records. After having collected the training dataset composed of the relevant and irrelevant study records, the next step concerns prompt engineering. A key part of developing well-performing prompts entails making them as concisely written as possible, only feeding them with the absolute most necessary information. Importantly to remember, these models do not need to be trained, as we initially were caused to believe. Thus, if conducting a complex review including many inclusion/exclusion criteria, we suggest conducting what we have coined multi-prompt screening. That is screening with multiple prompts, where each inclusion/exclusion criteria should be prompted individually. All title and abstract records are then screened with all prompts. Alternatively, one could conduct what we define as hierarchical screening where a study record is excluded if it is excluded at any step in the multi-prompt screening. This procedure is depicted in Figure 5[12].

FIGURE 5: *Hierarchical screening*



[12] To support this type of screening, we show how this can be practically executed in the accompanied vignette to the AIscreenR.

If using hierarchical screening, we suggest ordering the prompts so that the prompts excluding the largest body of references appear first and prompts with more specific inclusion/exclusion criteria following thereafter (Brunton et al., 2017). This approach will be more efficient both in terms of money and time. A further side-effect of this approach is that all title and abstract records will be mapped on what exact inclusion/exclusion criteria they were excluded upon. However, a shortage of this screening approach is that it is heavily dependent on the quality of the used prompts. Although more costly, we, therefore, recommend using multi-prompt screening where all title and abstract records are screened with all prompts, since this approach potentially guards against insufficient prompting. Assume for example that one made seven prompts, one for each of the inclusion/exclusion criteria, but one of the prompts wrongly excludes a large share of relevant records at the early stage of the sceening when scaled up from the test setting. Then those studies would be lost in the hierarchical screening suggested in Figure 5. If instead all records had been screened with all prompts then one could avoid the above bias by including all records that were included in 6 out of 7 prompts, as we did in classifier experiment 3, for instance. This approach is less sensitive to the order of the prompts in the hierarchy.

When engineering prompts, we suggest that these should be re-written/refined until they reach recall and specificity rate thresholds of at least 80%. Recall rates between 75% < 80% can also be accepted, but the reviewers should try to increase this performance as much as possible. Lower specificity rates can also be accepted as long as the recall exceeds 80%. If the specificity rate of 80% cannot be reached, then the GPT API models should mainly be used to reduce the total number of study records needed to be screened by two independent reviewers. More importantly, we suggest that if a recall rate of 80% cannot be reached, then the given GPT API model should not be used as an independent second screener. This can only be accepted if the given reviewer lacks financial resources since single-screening is less desirable than using a bad-performing GPT API model as an extra assurance of finding all relevant studies. However, the reviewer must be earnest about this shortcoming of the screening, and we do not think this should be accepted in state-of-the-art reviews. Alternatively, if the thresholds cannot be reached, the GPT API model can still be used as a third screener, again as an extra security for detecting all relevant studies.

When the test has been passed, and the reviewers have decided to leverage the GPT API model as second screener, we suggest that the reviewers screen all study records before initiating the automated screening. Thereby, it is prevented that the human reviewers are impacted by GPT's deci-

sions. In other words, we recommend that decisions on whether GPT API model screening is appropriate in a given review should primarily be made before the main TAB screening has been initiated. An alternative to manually screening all records at once is that reviewers repeat steps 6 to 9 in Table 5 with batches of 500-1000 study records. This would be an adequate way to steer the screening process and to continuously ensure that the given GPT API model performs as expected. Moreover, this reduces the risk of running large screenings that break for some technical reasons, which in turn hinders unnecessary money waste.

When all study records have both been screened by human and automated screeners, reviewers should investigate and solve disagreements. In this regard, it can be advantageous to re-screen all study records where humans and the automated screener disagreed to test the consistency of the automated screening decision but also to get detailed responses for GPT's decisions. For the latter purpose, we mainly recommend using the GPT-4 model since it provides substantially better descriptions of its screening behavior. If the specificity performance of the GPT screener is high (e.g. < 99%), the reviewers can consider just letting all study records that have been included by either human or GPT enter the full-text screening stage. Whether this is viable of course depends on the number of records needed to be screened.

5.1. When not to use GPT API models for TAB screening?

Although we think that GPT API models can have a huge impact on TAB screening in systematic reviews, we can envision at least two cases, beyond when the test performance thresholds are not met, where we find this screening approach to be inappropriate. That is for example when the complexity of the review question(s) or/and inclusion/exclusion criteria is high and the number of reference records needed to be screened is low (e.g. < 2000). In such a situation, it might take longer to construct reliable prompts than instantly initiating the duplicate human screening. In general, we think that when having few records, it is better merely to let humans double-screen all records because it is more time-efficient relative to engineering well-performing prompts. That said, we experienced that we were able to set up a reliable screening with the FRIENDS data within a few hours. Therefore, when having few records and the complexity of the review is low, it can be advantageous to conduct a rapid investigation of whether GPT API model screening is appropriate in the specific case. However, we do not think reviewers should spend/waste too much time on this task in such cases. Table 6 describes under what conditions, we consider it adequate and inadequate to use GPT API models for TAB screening tasks in systematic reviews.

TABLE 6: When to use GPT API models for TAB screening

		Number of studies	
		Low	High
Complexity of the review question(s) and/or inclusion criteria	Low	Questionable whether the time is worth investing in prompt development relative to merely initiating human screening	GPT screening is likely well-suited.
	High	Apply duplicate human screening	GPT screening is potentially well-suited. Consider using hierarchical or multi-prompt screening

6 LIMITATIONS

Although we have strived to make a comprehensive evaluation of the use of GPT API models for TAB screening tasks, our study has some important limitations. First of all, none of our analyses were pre-registered. However, to at least ensure openness and thereby make it possible to replicate our work, we have shared all data, codes, and material behind the analyses conducted in this study. It can be accessed at https://osf.io/apdfw/. A clear limitation of the shared material is that it will not necessarily be possible to make an exact replication of our results since minor model decision deviations can appear from screening to screening. Yet, we still firmly believe that the overall patterns of our results can be replicated. Hereto, it is important to note that humans would properly also change their screening decisions if they had to reiterate their first TAB screening. Therefore, we consider these minor discrepancies to be very human-like. Nonetheless, in future applications, users will be able to add a specific seed to the request body ensuring the reproducibility of the given screenings. Yet this functionality currently only exists in an experimental beta version why we did not use it.

Another clear limitation is that the models we drew on in this paper represent black box and close-end algorithms. Though, we can show that they can be used for TAB screening tasks at the current state of time, and with the current models, we are not able to say anything about why they work. Model dependency is a major issue when working with GPT API models since we do not know how they are trained and how they will develop. This also means the generalizability across different

models and over time is unclear. Consequently, we cannot infer that the results of our experiments are generalizable to other GPT-4 models such as the GPT-4o or the GPT-4-turbo, and, more so, to other models such as the API models from Claude or Mistral AI. On a similar but technical line, it is furthermore extremely demanding and time-consuming to be up-to-date with new models and updates of models as well as how to reach them via the API correctly. Model deprecation is a serious threat to the validity of our suggested approach. For now, the gpt-4-0613 model is stable but we expect that this model will eventually deprecate. Therefore, future research must evaluate whether our results can be exerted with other GPT models such as updated models as well as the GPT-4o and GPT-4-turbo etc. The gpt-3.5-turbo-0613 model that we drew upon is expected to deprecate during 2024, so eventually one cannot replicate the screening we have made with this model. On this note, we think it is pivotal that future research investigates if and how downloadable GPT models perform such as the ones provided by Mistral AI. This would secure a more stable application of using GPT models for TAB screening, supporting a functional technology.

- Black box (but this does not only count for GPT this is often true for human screening as well). Point to alternative models.
- Prizing (only use on a sub-sample of studies) gpt-4o is cheaper, and it could be advantageous to test if our results can be replicated with that model.
- Prompt sensitive
- Convient dataset used to construct the screening performance benchmark scheme.
- Can be time-consuming constructing reliable multi-prompt screening
- Although striving to be user-friendsly the screeening approach we suggested is function-based, meaning that user might be now know or learn how to use R. Thus, it can be advanceous to develop more generic soluation. This could for example be to develop a shiny application
- Evaluate Mistral which provides the possibility of locally downloading their model. This will overcome issues with deprecations and ensure reproducibility over time.
- Shiny app to ease user set-up challenges (O'Connor et al., 2019) to make the workflow more user-friendly.

8 DISCUSSION

- Talk about the interface here – cannot replicate the results on the ChatGPT interface

- Reviewers should not consider screening prioritization methods and GPT screening as two incommensurable methods. Instead, the strength from both should ideally be combined.
- We believe that the GPT-4 models will perform even better when fed with abstracts following a rigorous structure as in medicine.
- When not to use. If you cannot make the prompt work properly or if you screen very few studies.
- We believe that no automated tool should ever be at level 4 – there shall always be a human-in-the-loop to ensure adequate behavior the the screening tools. Consequently, GPT models used in non-systematic to reduce the number of studies needed to be screened should always include safety checks. For example, reviewers should randomly sample 5-10% of the studies excluded by GPT to test for serious flaws in its decision-making. If serious flaws are detected the reviewers must re-test the used prompt(s) or refrain from using the given GPT model.
- More rapid transfer of knowledge from review to policy, research, and practice
- Makes it possible to help to screen in extreme-sized reviews (Shemilt et al., 2014, 2016). Combine with traditional classifier models
- Extra security in low-budget and/or time-limited projects where there is only access to a single screener.
- No need for unnecessary restriction on search string.
- To reduce the environmental impact and reduce the number of references needed to be screen. GPT API models could be used on a subset of studies, for example on all references not examined by humans after using priority screening.
- Future models should use seed to ensure reproducible screening. This is currently only available in the beta version but should be implemented in the software over time.
- Draw on ‘function call’ needs to be updated to work with tools.
- Requires continuous software development.
- When reviewers want to keep duplicate screening, we suggest that the GPT API models can be used as a third screener for extra insurance that all relevant studies are detected.
- The GPT API model seems generally to be over inclusive than human but this just force the reviewers to double-check close-to-relevant studies.

- Environmental impact (embrace the critiques from van Lissa). Combine with traditional machine learning and text-mining tools to reduce the number of records needed to be screened by the GPT API model

ACKNOWLEDGEMENT

Thanks to Jens Dietrichson, Trine Filges, Tiril Borge, Heather Melanie R. Ames, and Christopher James Rose for valuable comments and sharing of screening data. Also thanks to Sofie Elgaard Lisager Jensen and Johan Klejs for testing the AIscreenR software and for valuable inputs to the workflow.

FUNDING STATEMENT

This manuscript was funded by VIVE Campbell, Denmark

DATA AVAILABILITY STATEMENT

To adhere to the reproducibility framework proposed by Olorisade et al. (2017), replicate codes can be found at OSF bit.ly/3spivoG:

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

REFERENCES

* marks studies used for the benchmark development

Alshami, A., Elsayed, M., Ali, E., Eltoukhy, A. E. E., & Zayed, T. (2023). Harnessing the power of ChatGPT for automating systematic review process: Methodology, case study, limitations, and future directions. *Systems*, *11*(7), 351.

Ames, H., Hestevik, C. H., & Briggs, A. M. (2024). Acceptability, values, and preferences of older people for chronic low back pain management; a qualitative evidence synthesis. *BMC Geriatrics*, *24*(1), 1–22. https://doi.org/10.1186/s12877-023-04608-4

Boetje, J., & van de Schoot, R. (2024). The SAFE procedure: a practical stopping heuristic for active learning-based screening in systematic reviews and meta-analyses. *Systematic Reviews*, *13*(1), 81.

Bøg, M., Filges, T., & Jørgensen, A. M. K. (2018). Deployment of personnel to military operations: impact on mental health and social functioning. *Campbell Systematic Reviews*, *14*(1), 1–127. https://doi.org/https://doi.org/10.4073/csr.2018.6

Bondebjerg, A., Dalgaard, N. T., Filges, T., & Viinholt, B. C. A. (2023). The effects of small class sizes on students' academic achievement, socioemotional development and well-being in special education: A systematic review. *Campbell Systematic Reviews*, *19*(3), e1345.

Bondebjerg, A., Filges, T., Pejtersen, J. H., Kildemoes, M. W., Burr, H., Hasle, P., Tompa, E., & Bengtsen, E. (2023). Occupational health and safety regulatory interventions to improve the work environment: An evidence and gap map of effectiveness studies. *Campbell Systematic Reviews*, *19*(4), e1371. https://doi.org/https://doi.org/10.1002/cl2.1371

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis* (1st ed.). John Wiley & Sons.

Bornmann, L., Haunschild, R., & Mutz, R. (2021). Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, *8*(1), 1–15.

Burgard, T., & Bittermann, A. (2023). Reducing Literature Screening Workload With Machine Learning. *Zeitschrift Für Psychologie*.

Buscemi, N., Hartling, L., Vandermeer, B., Tjosvold, L., & Klassen, T. P. (2006). Single data extraction generated more errors than double data extraction in systematic reviews. *Journal of Clinical Epidemiology*, *59*(7), 697–703.

Campbell Collaboration. (2023). *Stepping up evidence synthesis: faster, cheaper and more useful*.

https://www.campbellcollaboration.org/news-and-events/news/stepping-up-evidence-synthesis.html

Campos, D. G., Fütterer, T., Gfrörer, T., Lavelle-Hill, R. E., Murayama, K., König, L., Hecht, M., Zitzmann, S., & Scherer, R. (2023). *Screening Smarter, Not Harder: A Comparative Analysis of Machine Learning Screening Algorithms and Heuristic Stopping Criteria for Systematic Reviews in Educational Research*.

Chicco, D., & Jurman, G. (2023). The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining*, *16*(1), 1–23.

Cohen, A. M., Hersh, W. R., Peterson, K., & Yen, P.-Y. (2006). Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association*, *13*(2), 206–219.

Dalgaard, N. T., Bondebjerg, A., Klokker, R., Viinholt, B. C. A., & Dietrichson, J. (2022). Adult/child ratio and group size in early childhood education or care to promote the development of children aged 0–5 years: A systematic review. *Campbell Systematic Reviews*, *18*(2), e1239. https://doi.org/https://doi.org/10.1002/cl2.1239

Dalgaard, N. T., Bondebjerg, A., Viinholt, B. C. A., & Filges, T. (2022). The effects of inclusion on academic achievement, socioemotional development and wellbeing of children with special educational needs. *Campbell Systematic Reviews*, *18*(4), e1291. https://doi.org/https://doi.org/10.1002/cl2.1291

Dalgaard, N. T., Filges, T., Viinholt, B. C. A., & Pontoppidan, M. (2022). Parenting interventions to support parent/child attachment and psychosocial adjustment in foster and adoptive parents and children: A systematic review. *Campbell Systematic Reviews*, *18*(1), e1209. https://doi.org/https://doi.org/10.1002/cl2.1209

Dalgaard, N. T., Flensborg Jensen, M. C., Bengtsen, E., Krassel, K. F., & Vembye, M. H. (2022). PROTOCOL: Group-based community interventions to support the social reintegration of marginalised adults with mental illness. *Campbell Systematic Reviews*, *18*(3), e1254. https://doi.org/10.1002/cl2.1254

Dietrichson, J., Filges, T., Klokker, R. H., Viinholt, B. C. A., Bøg, M., & Jensen, U. H. (2020). Targeted school-based interventions for improving reading and mathematics for students with, or at risk of, academic difficulties in Grades 7–12: A systematic review. *Campbell Systematic Reviews*, *16*(2), e1081. https://doi.org/10.1002/cl2.1081

Dietrichson, J., Filges, T., Seerup, J. K., Klokker, R. H., Viinholt, B. C. A., Bøg, M., & Eiberg, M. (2021). Targeted school-based interventions for improving reading and mathematics for students with or at risk of academic difficulties in Grades K-6: A systematic review. *Campbell Systematic Reviews*, *17*(2), e1152. https://doi.org/10.1002/cl2.1152

Doi, S. A., & Xu, C. (2021). The Freeman–Tukey double arcsine transformation for the meta-analysis of proportions: Recent criticisms were seriously misleading. *Journal of Evidence-Based Medicine*, *14*(4), 259–261. https://doi.org/https://doi.org/10.1111/jebm.12445

EPPI-Centre. (2024). *Automated data extraction using GPT-4*. https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=3921

Evensen, L. H., Kleven, L., Dahm, K. T., Hafstad, E. V., Holte, H. H., Robberstad, B., & Risstad, H. (2023). *Sutur av degenerative rotatorcuff-rupturer: en fullstendig metodevurdering [Rotator cuff repair for degenerative rotator cuff tears: a health technology assessment]*. https://www.fhi.no/publ/2023/sutur-av-degenerative-rotatorcuff-rupturer/

Filges, T., Andersen, D., & Jørgensen, A.-M. K. (2015). Functional Family Therapy (FFT) for Young People in Treatment for Non-opioid Drug Use: A Systematic Review. *Campbell Systematic Reviews*, *11*(1), 1–77. https://doi.org/https://doi.org/10.4073/csr.2015.14

Filges, T., Dalgaard, N. T., & Viinholt, B. C. A. (2022). Outreach programs to improve life circumstances and prevent further adverse developmental trajectories of at-risk youth in OECD countries: A systematic review. *Campbell Systematic Reviews*, *18*(4), e1282. https://doi.org/https://doi.org/10.1002/cl2.1282

Filges, T., Dietrichson, J., Viinholt, B. C. A., & Dalgaard, N. T. (2022). Service learning for improving academic success in students in grade K to 12: A systematic review. *Campbell Systematic Reviews*, *18*(1), e1210. https://doi.org/https://doi.org/10.1002/cl2.1210

Filges, T., Montgomery, E., Kastrup, M., & Jørgensen, A.-M. K. (2015). The Impact of Detention on the Health of Asylum Seekers: A Systematic Review. *Campbell Systematic Reviews*, *11*(1), 1–104. https://doi.org/https://doi.org/10.4073/csr.2015.13

Filges, T., Siren, A., Fridberg, T., & Nielsen, B. C. V. (2020). Voluntary work for the physical and mental health of older volunteers: A systematic review. *Campbell Systematic Reviews*, *16*(4), e1124. https://doi.org/https://doi.org/10.1002/cl2.1124

Filges, T., Smedslund, G., Eriksen, T., & Birkefoss, K. (2023). PROTOCOL: The FRIENDS preventive programme for reducing anxiety symptoms in children and adolescents: A systematic review. *Campbell Systematic Reviews*, *19*(4), e1374.

https://doi.org/https://doi.org/10.1002/cl2.1374

Filges, T., Sonne-Schmidt, C. S., & Nielsen, B. C. V. (2018). Small class sizes for improving student achievement in primary and secondary schools: A systematic review. *Campbell Systematic Reviews*, *14*(1), 1–107. https://doi.org/10.4073/csr.2018.10

Filges, T., Torgerson, C., Gascoine, L., Dietrichson, J., Nielsen, C., & Viinholt, B. A. (2019). Effectiveness of continuing professional development training of welfare professionals on outcomes for children and young people: A systematic review. *Campbell Systematic Reviews*, *15*(4), e1060. https://doi.org/https://doi.org/10.1002/cl2.1060

Filges, T., Verner, M., Ladekjær, E., & Bengtsen, E. (2023). PROTOCOL: Participation in organised sport to improve and prevent adverse developmental trajectories of at-risk youth: A systematic review. *Campbell Systematic Reviews*, *19*(2), e1321. https://doi.org/https://doi.org/10.1002/cl2.1321

Gargari, O. K., Mahmoudi, M. H., Hajisafarali, M., & Samiee, R. (2024). Enhancing title and abstract screening for systematic reviews with GPT-3.5 turbo. *BMJ Evidence-Based Medicine*, *29*(1), 69 LP – 70. https://doi.org/10.1136/bmjebm-2023-112678

Gartlehner, G., Wagner, G., Lux, L., Affengruber, L., Dobrescu, A., Kaminski-Hartenthaler, A., & Viswanathan, M. (2019). Assessing the accuracy of machine-assisted abstract screening with DistillerAI: a user study. *Systematic Reviews*, *8*(1), 277. https://doi.org/10.1186/s13643-019-1221-3

Gough, D., Oliver, S., & Thomas, J. (2017). *An introduction to systematic reviews* (2nd ed.). Sage.

Guo, E., Gupta, M., Deng, J., Park, Y.-J., Paget, M., & Naugler, C. (2024). Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study. *J Med Internet Res*, *26*, e48996. https://doi.org/10.2196/48996

Hedges, L. V. (1992). Modeling Publication Selection Effects in Meta-Analysis. *Statistical Science*, *7*(2), 246–255. http://www.jstor.org/stable/2246311

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M. S., Li, T., Page, M., & Welch, V. (2019). *Cochrane handbook for systematic reviews of interventions* (2nd ed.). Wiley Online Library. https://doi.org/10.1002/9781119536604

Hou, Z., & Tipton, E. (2024). Enhancing recall in automated record screening: A resampling algorithm. *Research Synthesis Methods*, *n/a*(n/a). https://doi.org/https://doi.org/10.1002/jrsm.1690

Issaiy, M., Ghanaati, H., Kolahi, S., Shakiba, M., Jalali, A. H., Zarei, D., Kazemian, S., Avanaki,

M. A., & Firouznia, K. (2024). Methodological insights into ChatGPT's screening performance in systematic reviews. *BMC Medical Research Methodology*, *24*(1), 78. https://doi.org/10.1186/s12874-024-02203-8

Jardim, P. S. J., Borge, T. C., & Johansen, T. B. (2021). *Effekten av antipsykotika ved førstegangspsykose: en systematisk oversikt [The effect of antipsychotics on first episode psychosis]*. https://fhi.no/publ/2021/effekten-av-antipsykotika-ved-forstegangspsykose/

Johansen, T. B., Nøkleby, H., Langøien, L. J., & Borge, T. C. (2022). *Samværs-og bostedsordninger etter samlivsbrudd: betydninger for barn og unge: en systematisk oversikt [Custody and living arrangements after parents separate: implications for children and adolescents: a systematic review]*. https://www.fhi.no/publ/2022/samvars--og-bostedsordninger-etter-samlivsbrudd-betydninger-for-barn-og-ung/

Jonnalagadda, S. R., Goyal, P., & Huffman, M. D. (2015). Automating data extraction in systematic reviews: a systematic review. *Systematic Reviews*, *4*(1), 78. https://doi.org/10.1186/s13643-015-0066-7

Khraisha, Q., Put, S., Kappenberg, J., Warraitch, A., & Hadfield, K. (2024). Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Research Synthesis Methods*, *n/a*(n/a). https://doi.org/https://doi.org/10.1002/jrsm.1715

König, L., Zitzmann, S., Fütterer, T., Campos, D. G., Scherer, R., & Hecht, M. (2023). *When to stop and what to expect—An Evaluation of the performance of stopping rules in AI-assisted reviewing for psychological meta-analytical research*.

Kugley, S., Wade, A., Thomas, J., Mahood, Q., Jørgensen, A.-M. K., Hammerstrøm, K., & Sathe, N. (2016). Searching for studies: A guide to information retrieval for Campbell. *Campbell Systematic Reviews*, *13*(1), 1–73. https://doi.org/10.4073/cmg.2016.1

Meneses Echavez, J. F., Borge, T. C., Nygård, H. T., Gaustad, J.-V., & Hval, G. (2022). *Psykologisk debriefing for helsepersonell involvert i uønskede pasienthendelser: en systematisk oversikt [Psychological debriefing for healthcare professionals involved in adverse events: a systematic review]*. https://www.fhi.no/publ/2022/psykologisk-debriefing-for-helsepersonell-involvert-i-uonskede-pasienthende/

Ng, L., Pitt, V., Huckvale, K., Clavisi, O., Turner, T., Gruen, R., & Elliott, J. H. (2014). Title and Abstract Screening and Evaluation in Systematic Reviews (TASER): a pilot randomised controlled trial of title and abstract screening by medical students. *Systematic Reviews*, *3*, 1–8.

O'Connor, A. M., Tsafnat, G., Thomas, J., Glasziou, P., Gilbert, S. B., & Hutton, B. (2019). A question of trust: can we build an evidence base to gain trust in systematic review automation technologies? *Systematic Reviews*, *8*(1), 1–8.

O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., & Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic Reviews*, *4*(1), 1–22.

Olofsson, H., Brolund, A., Hellberg, C., Silverstein, R., Stenström, K., Österberg, M., & Dagerhamn, J. (2017). Can abstract screening workload be reduced using text mining? User experiences of the tool Rayyan. *Research Synthesis Methods*, *8*(3), 275–280.

Olorisade, B. K., Brereton, P., & Andras, P. (2017). Reproducibility of studies on text mining for citation screening in systematic reviews: Evaluation and checklist. *Journal of Biomedical Informatics*, *73*, 1–13. https://doi.org/https://doi.org/10.1016/j.jbi.2017.07.010

Olorisade, B. K., de Quincey, E., Brereton, P., & Andras, P. (2016). A critical analysis of studies that address the use of text mining for citation screening in systematic reviews. *Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering*, 1–11.

OpenAI. (2024). *Function calling*. https://platform.openai.com/docs/guides/function-calling

Pacheco, R. L., Riera, R., Santos, G. M., Sá, K. M. M., Bomfim, L. G. P., da Silva, G. R., de Oliveira, F. R., & Martimbianco, A. L. C. (2023). Many systematic reviews with a single author are indexed in PubMed. *Journal of Clinical Epidemiology*, *156*, 124–126.

Perlman-Arrow, S., Loo, N., Bobrovitz, N., Yan, T., & Arora, R. K. (2023). A real-world evaluation of the implementation of NLP technology in abstract screening of a systematic review. *Research Synthesis Methods*, *14*(4), 608–621.

Polanin, J. R., Pigott, T. D., Espelage, D. L., & Grotpeter, J. K. (2019). Best practice guidelines for abstract screening large-evidence systematic reviews and meta-analyses. *Research Synthesis Methods*, *10*(3), 330–342. https://doi.org/https://doi.org/10.1002/jrsm.1354

Pustejovsky, J. E. (2020). *clubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections* (0.5.5). cran.r-project.org. https://cran.r-project.org/web/packages/clubSandwich/index.html

Pustejovsky, J. E., & Tipton, E. (2021). Meta-analysis with robust variance estimation: Expanding the range of working models. *Prevention Science*, *23*(1), 425–438. https://doi.org/10.1007/s11121-021-01246-3

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/

Rathbone, J., Hoffmann, T., & Glasziou, P. (2015). Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers. *Systematic Reviews*, *4*(1), 1–7.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). Publication bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Wiley Online Library.

Röver, C., & Friede, T. (2022). Double arcsine transform not appropriate for meta-analysis. *Research Synthesis Methods*, *13*(5), 645–648. https://doi.org/https://doi.org/10.1002/jrsm.1591

RStudio Team. (2015). *RStudio: Integrated development for R*. RStudio, Inc., Boston, MA. https://www.rstudio.com/

Schwarzer, G., Chemaitelly, H., Abu-Raddad, L. J., & Rücker, G. (2019). Seriously misleading results using inverse of Freeman-Tukey double arcsine transformation in meta-analysis of single proportions. *Research Synthesis Methods*, *10*(3), 476–483. https://doi.org/https://doi.org/10.1002/jrsm.1348

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (2nd ed.). Cengage Learning, Inc.

Shemilt, I., Khan, N., Park, S., & Thomas, J. (2016). Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews. *Systematic Reviews*, *5*, 1–13.

Shemilt, I., Simon, A., Hollands, G. J., Marteau, T. M., Ogilvie, D., O'Mara-Eves, A., Kelly, M. P., & Thomas, J. (2014). Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. *Research Synthesis Methods*, *5*(1), 31–49. https://doi.org/https://doi.org/10.1002/jrsm.1093

Stoll, C. R. T., Izadi, S., Fowler, S., Green, P., Suls, J., & Colditz, G. A. (2019). The value of a second reviewer for study selection in systematic reviews. *Research Synthesis Methods*, *10*(4), 539–545. https://doi.org/10.1002/jrsm.1369

Syriani, E., David, I., & Kumar, G. (2023). Assessing the Ability of ChatGPT to Screen Articles for Systematic Reviews. *ArXiv Preprint ArXiv:2307.06464*.

Thomsen, M. K., Seerup, J. K., Dietrichson, J., Bondebjerg, A., & Viinholt, B. C. A. (2022). PROTOCOL: Testing frequency and student achievement: A systematic review. *Campbell*

Systematic Reviews, *18*(1), e1212. https://doi.org/https://doi.org/10.1002/cl2.1212

Tipton, E., & Pustejovsky, J. E. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics*, *40*(6), 604–634. https://doi.org/10.3102/1076998615606099

Tsafnat, G., Glasziou, P., Choong, M. K., Dunn, A., Galgani, F., & Coiera, E. (2014). Systematic review automation technologies. *Systematic Reviews*, *3*(1), 74. https://doi.org/10.1186/2046-4053-3-74

Valentine, J. C. (2009). Judging the quality of primary research. *The Handbook of Research Synthesis and Meta-Analysis*, *2*, 129–146.

Van De Schoot, R., De Bruin, J., Schram, R., Zahedi, P., De Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., & Ferdinands, G. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, *3*(2), 125–133.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Viechtbauer, W. (2022). *Miller (1978)*. https://www.metafor-project.org/doku.php/analyses:miller1978?s[]=proportion

Waffenschmidt, S., Knelangen, M., Sieben, W., Bühn, S., & Pieper, D. (2019). Single screening versus conventional double screening for study selection in systematic reviews: a methodological systematic review. *BMC Medical Research Methodology*, *19*(1), 132. https://doi.org/10.1186/s12874-019-0782-0

Wang, Z., Nayfeh, T., Tetzlaff, J., O'Blenis, P., & Murad, M. H. (2020). Error rates of human reviewers during abstract screening in systematic reviews. *PloS One*, *15*(1), e0227742.

Westgate, M. J. (2019). revtools: An R package to support article screening for evidence synthesis. *Research Synthesis Methods*, *10*(4), 606–614. https://doi.org/https://doi.org/10.1002/jrsm.1374

Appendix A: Multi-prompt screening

TEXTBOX A1

PROMPT 1: "We are conducting a systematic review, which examines the potential effects of different frequencies or intensities in testing on the academic achievement or testing-related anxiety of primary and secondary school students.
We want to include studies with quantitative measures. For each study, we would like you to assess:
1) Does the study report quantitative measures?"

PROMPT 2: "We are conducting a systematic review, which examines the potential effects of different frequencies or intensities in testing on the academic achievement or testing-related anxiety of primary and secondary school students.
Only investigations performed in a school setting on children or students (ages 4-18 years old) are relevant for this review. This means that experiments performed in laboratories must be excluded, because we are only interested in real school settings and educational systems. For each study, we would like you to assess:
1) Does the intervention take place within a school setting?"

PROMPT 3: "We are conducting a systematic review, which examines the potential effects of different frequencies or intensities in testing on the academic achievement or testing-related anxiety of primary and secondary school students.
We only want to include studies that investigate children or students attending either primary or secondary school, this means from kindergarten until grade 12. In other words, we are looking for studies where the participants are students 4-18 years old. For each study, we would like you to assess:
1) Are the participants in the study children or students attending either primary or secondary school, this means from kindergarten until grade 12."

PROMPT 4: "We are conducting a systematic review, which examines the potential effects of different frequencies or intensities in testing on the academic achievement or testing-related anxiety of primary and secondary school students.
The study must entail testing students or children. The testing can be standardized and non-standardized tests as well as formative assessments and summative tests, and high-stakes and low-stakes exams. This also include repeated testing, interim assessment testing, class quizzes, multiple choice testing, progress monitoring assessments or measures, curriculum-based measurement or assessments, retrieval practice measures or assessments, etc. For each study, we would like you to assess:
1) Does the study report on tests or testing of students or children?"

TEXTBOX A1 (Continued)

PROMPT 5: "We are conducting a systematic review, which examines the potential effects of different frequencies or intensities in testing on the academic achievement or testing-related anxiety of primary and secondary school students.
We like to include randomized controlled trials (RCT), field experiments, quasi-experimental studies (QES), or observational studies, which use a control/comparison research design to examine effects. This means that the study must compare at least two groups of students or children. Such studies can have many labels and the different designs can have different notations. The most common sub-categories of randomised controlled trials and quasi-experimental studies are: individual randomised assignment, cluster randomised assignment, stratified/blocked random assignment, pseudo-randomisation, matching cohort studies, difference-in-differences, regression-discontinuity designs, instrumental variable designs, propensity score matching, case-control studies, etc. Studies employing a within-subject design are also eligible for inclusion. For each study, we would like you to assess:
1) Is the study a randomized controlled trial (RCT), a field experiment, a quasi-experimental study, an observational study, or a study employing a within-subject design?"

PROMPT 6: "We are conducting a systematic review, which examines the potential effects of different frequencies or intensities in testing on the academic achievement or testing-related anxiety of primary and secondary school students.
In the review, we would like to include studies that measure students' academic achievement. In this review, we do not restrict measures of academic achievement to specific subjects. For each study, we would like you to assess:
1) Does the study report on measures of academic achievement or academic skills?"

Textbox A1 presents all the prompts we engineered and used to conduct the third classifier experiment. When added to the AIscreenR, each of the above six prompts was pasted together with the text present in Textbox 2 in the main paper.