GPT API Models Can Function as a Highly Reliable Second Screener of Titles and Abstracts in Systematic Reviews

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 a time- and 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 an alternative second screener of titles and abstracts in systematic reviews. 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 to be included. To support future reviewers, 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 the uptake of using GPT API models acceptable as independent second screeners within reviews facilitated by evidence institutions such as Cochrane and Campbell Collaboration. To standardize this application using GPT API models for title and abstract screening tasks, 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 develop empirical benchmarks to make reliable comparisons between AI and human screening performances.
- We provide general guidelines for how and when GPT models safely can used
- We present and validate the R package AIscreenR 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 double screening workflow of title and abstract screening in systematic reviews
- Increasing the reliability of large-scale systematic reviews
- Substantial and reliable reduction of human labor in systematic reviews
- Provides 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 be a time-consuming and tedious task. However, overlooking relevant studies in this phase can be consequential, potentially 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 the fact that previous research suggests 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, potentially 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. Alternatively, reviewers make too narrow searches to keep the number of records down to a manageable size which again heavily 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

Commented [MHV1]: Find examples

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 bias defined by König et al., (2023) as the 'artificial screening bias' (ASB).

An additional challenge is that most automated screening 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. 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). The overall picture is that they are generally not capable of replacing an independent human second screener without a significant risk of omitting a substantial number of eligible

[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).

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 (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 rise to at least 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 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 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 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 target 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 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 21 large-scale systematic reviews; 16 Campbell Systematic Reviews, and five systematic reviews conducted by the Norwegian Institute of Public Health (NIPH). Thereafter, we conducted two large-scale classification 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.

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 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 and clearer in future reviews. This goes without saying that our approach represents the final solution. Our aim is just 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 evaluation of using OpenAI's GPT models for TAB screening tasks in systematic reviews and reflect on what contributions our work provides. In Section 3 we describe the metrics we applied to evaluate the screening performance of GPT models and humans, respectively. In Section 3, we also develop benchmark measures that can be used to assess the screening performance of the GPT API models. In Section 4, we focus partly on how we developed prompts and partly on how we think appropriate prompts can be developed to ensure reliable TAB screening. In this regard, we also describe the advance of using function calling with the GPT API models to ensure reliable response messages. In Section 5, we present the data used to conduct the two large-scale classifier experiments and the results of these experiments. In section 6, 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. Finally, in Sections 7 to 9, 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 screening performance of OpenAI's GPT API models to be used for TAB screening 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 3.5-turbo-0301 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" which might just indicate that the models perform badly when given unclear inclusion criteria—as would humans do. Syriani et al. (2023) used Python to reach the GPT API models. Yet they did not build any publicly available software for others to replicate this workflow.

Guo et al. (2024) tested the leverage of OpenAI's GPT-4 API model[5] for TAB screening of medical research literature. They found that the average recall (referred to as the sensitivity of included paper) across six clinical reviews was 0.76 and the average specificity (referred to as the sensitivity of excluded paper) was 0.91. Based on these results, Guo et al. (2024) infer that the GPT-4 model is proficient in terms of excluding the right studies whereas it is insufficient in finding relevant studies. Consequently, Guo et al. (2024) conclude 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) did also reach the API models via Python without providing any 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 recommend not replacing any human raters with the GPT-3.5 API model. Gargari et al. (2024) reach the API model via Python, and they shared their codes[6] so that others can replicate their workflow.

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, that is 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 so that the performance of different GPT models is not unnecessarily mixed 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. 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 draw upon in this paper.

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

2.1 What we do differently

This paper goes beyond the above-mentioned evaluations in multiple ways and shows 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 to find almost all relevant studies when well prompted.

In contrast to all the previous evaluations of using GPT API models for TAB screening, we are the first to draw on the 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 a request. The advance here 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). The use of function calling can potentially explain why we in later sections find a better recall performance (i.e., the ability to detect relevant studies) of using the GPT API models than previous evaluations by Guo et al. (2024) and Gargari et al. (2024). Differently from the previous evaluation, we have built our function calls so that they also allow the model to express its uncertainty relative to just making binary decisions (include or exclude) as all previous evaluations have done. That is if the GPT API model 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 = {exlcude}) 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 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 large-scale systematic reviews. We do so by developing the AIscreenR R package and technically quality-assured 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 using multiple prompts, API models, iterations of the same request, and nucleus samples

(i.e., different top_p values). We allow the user to conduct the same request (i.e., asking the 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 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 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 16 Campbell systematic reviews and 5 systematic reviews conducted by the Norwegian Institute of Public Health (NIPH). The contribution of this paper is therefore to put forward a tentative benchmark scheme to which all types of AI screening tools can be compared.

Previous research (Gargari et al., 2024) suggests that broader and more complex prompts do not perform well in terms of finding relevant studies. Instead, concisely framed prompts with clear information perform better. This might 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. To overcome this issue, we suggest using multiple prompts separately containing each inclusion criterion when using GPT models for TAB screening in broad and complex reviews, as we often find within the social sciences. We coin this procedure as hierarchical TAB screening.

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

Meanwhile, the current evaluations were either premised on the original **GPT-3.5-0301 models that will soon deprecate** or did not draw on up-to-date features of the newest GPT models such as function calling (OpenAI, 2024). Moreover, **it is unclear if these findings generalize to social science reviews** in which the scientific abstracts are less structured. Therefore, one of the major

aims of this paper is to evaluate the use and performance of OpensAI's GPT API (application programming interface) models in **social science reviews**. Hereto, we confirm that OpenAI's GPT API models can function as a highly reliable second screener with recalls (i.e., the ability to detect relevant studies) similar or superior to human performance.

3 METHODS

To evaluate and develop quantitative benchmarks, we used a range of different metrics. The choice of metric was informed by the recommendations made by Syriani et al. (2023) and O'Connor et al. (2019). These metrics are presented below.

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

The two main metrics we used to evaluate the performance of the GPT API models were the recall and specificity metrics since these are intuitive to understand and interpret and are not sensitive to imbalanced data (i.e., data with a large discrepancy between inclusion and exclusion references). The recall "represents the proportion of relevant records being correctly classified" (Hou & Tipton, 2024), and can be written as

$$Rec = \frac{TP}{TP + FN}$$

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, specificity "measures the ability to exclude all references that should be excluded" (Syriani et al., 2023), given by

$$Spec = \frac{TN}{TN + FP}$$

where TN (true negative) represents all the studies that are correctly excluded, and FP (false positive) is the number of studies falsely included. The recall metric can be considered the most important metric since it can seriously bias a review if the screener excludes references that should have been included. [FIND FURTER REASONS IN HOU & TIPTON] Whereas, a low specificity "just" means that reviewers must re-examine a larger share of the reference. This goes without saying that reviewers should accept low specificity rates. We will come back to that in the following sections.

We applied to overall assessment metric deduced from the above measure: mention imbalanced data

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

In our simulation, the TP, TN, FN, and FP conditions are determined by comparing the GPT decision with the final decision made by a minimum of two independent human screeners. For benchmark development, the conditions are determined by comparing the single screener decision with the final decision agreed upon between a minimum of two human screeners. This approach is suggested by O'Connor et al. (O'Connor et al., 2019)

Mention how to calculate variance and confidence intervals. Viectbauer and Research synthesis methods. (Röver & Friede, 2022; Schwarzer et al., 2019)

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

Mention the nMCC model and formula and why it is prefered above the receiver operating characteristic Curve (ROC AUC) (Chicco & Jurman, 2023)

Insert WSS (Campos et al., 2023)

$$WSS = 1 - \frac{TP + FP}{N}$$

3.2 Human screening performance for comparison

Give us and idea of what human screening performances that are accepted with in evidence institutions such as the Campbell collaboration.

To grasp a better understanding of the AI performance. (O'Connor et al., 2019) map how humans perform.

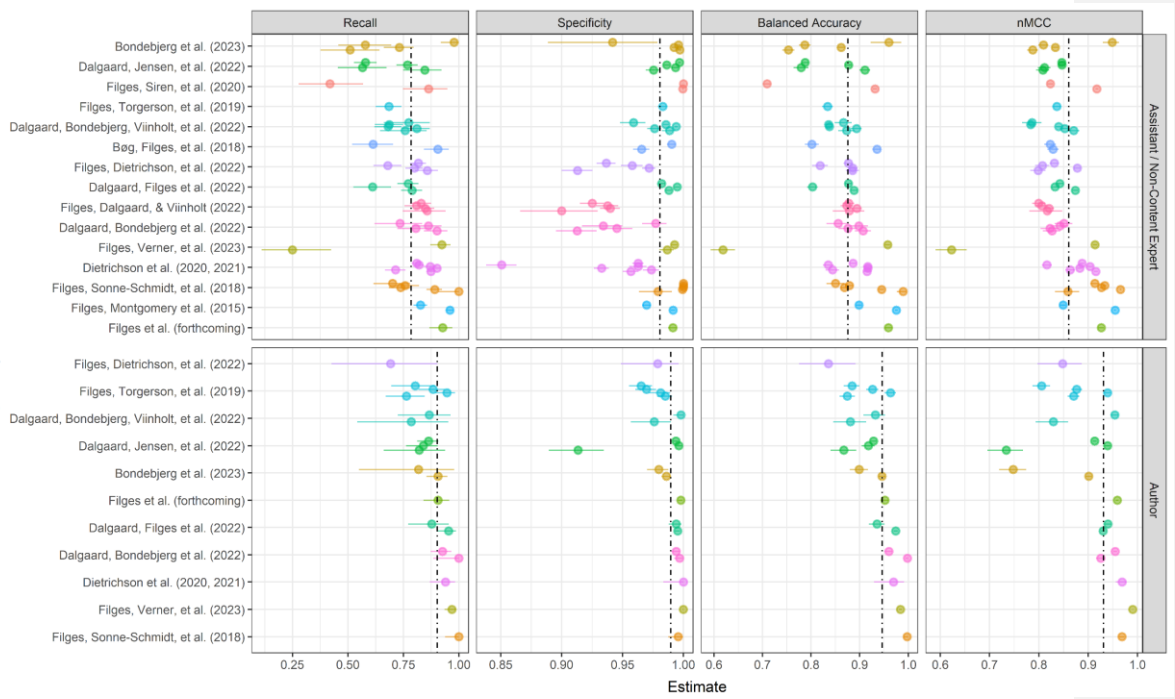
We think it is more fair to compare the performance of the GPT models

Deleted all training data.

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	244/1160	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, Torg-erson, 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
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



IGURE 1. Performance measures within Campbell Systematic Reviews across assistants vs. authors. Dashed line indicate the verage estimated via the SCE+ model.

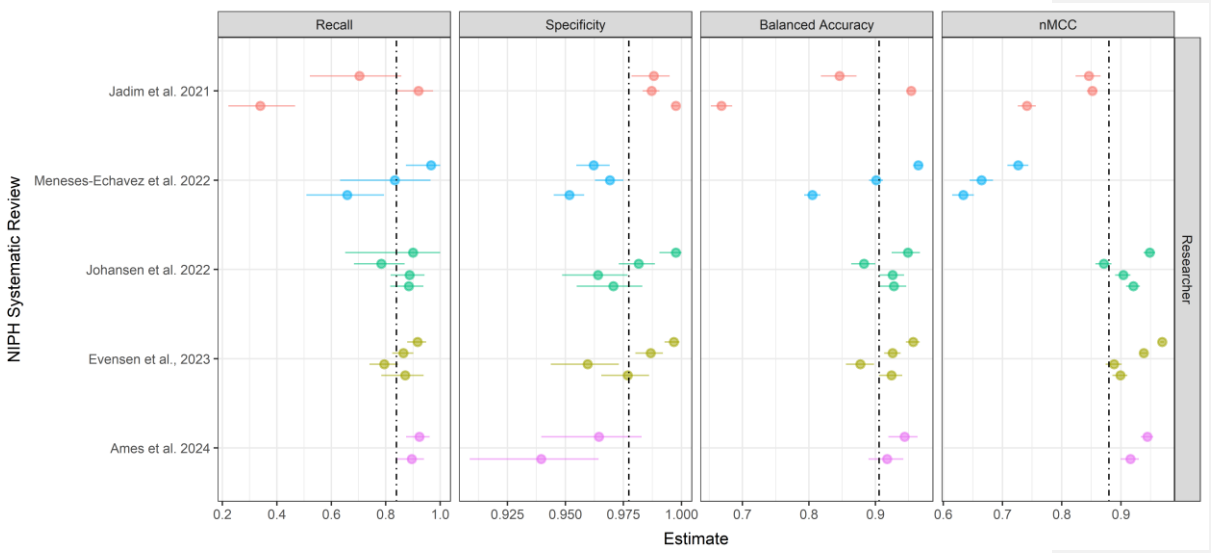


FIGURE 2. Researcher-researcher screening performance measures within NIPH Systematic Reviews. Dashed line indicate the average estimated via the CHE model.

Mention the authority and deeper content knowledge of the main author which might cause the recall to increase when review author screen with student assistants. Therefore to compare screenings with more equal relations, we analyze data from sixe systematic reviews conducted by the Norwegian Institue of Public Health (NIPH).

Imbalance is not a problem with GPT models cf. FRIENDS

Does not need to be trained. Only initial testing is needed.

4 PROMPT DEVELOPMENT AND FUNCTION CALLING

5 NUMERICAL STUDY

Simulation data

FRIENDS and FTT, only citation records with abstracts

What type of reviews.

How many titles and abstracts used.

Prompt engineering

[insert prompt example]

The simulation results

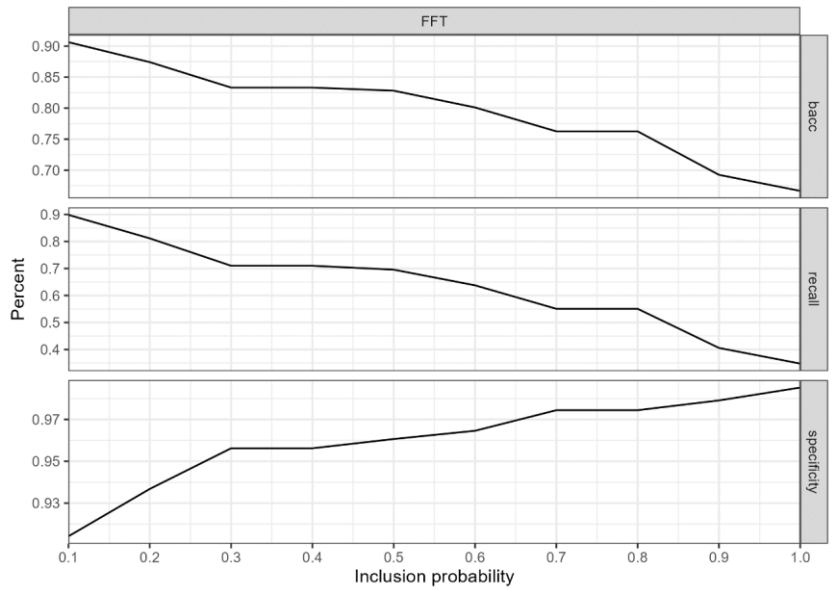
Review Model	Reps	Recall (%) [TP/(TP + FN)]	Specificity (%) [TN/(TN + FP)]	Raw aggrement (%) [(TP + TN)/*N*][a]	bAcc (%)	WSS
FFT						
GPT-3.5-turbo-0613 (incl. prop = .5)	10	71 (49/69)	95.6 (3888/4066)	95.2 (3937/4135)	83.3	94.5
GPT-3.5-turbo-0613 (incl. prop = .3)	10	81.2 (56/69)	93.7 (3809/4066)	93.5 (3865/4135)	87.4	92.4
GPT-4-0613	1	89.9 (62/69)	93.7 (3810/4066)	93.6 (3872/4135)	91.8	92.3
FRIENDS						
GPT-3.5-turbo-0613 (incl. prop = .5)	10	96.9 (62/64)	76.5 (1930/2511)	77.1 (1992/2575)	86.7	75
GPT-3.5-turbo-0613 (incl. prop = .7)	10	95.3 (61/64)	89.8 (2256/2511)	90.0 (2317/2575)	92.6	87.7
GPT-4-0613	1	98.4 (63/64)	97.4 (2455/2511)	97.4 (2518/2585)	97.9	95

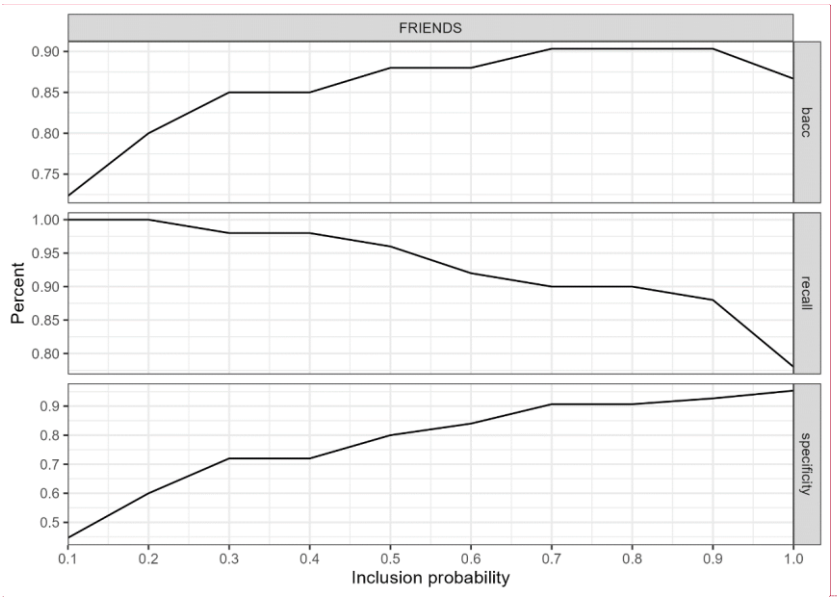
a: *N* is the total number of references

Commented [MHV2]: Go back to WSS when all simulation results are re-screened.

Commented [MHV3]: Extra security

GPT AS SECOND SCREENER





Concise text more important than information-dense prompt.

GPT-3.5-turbo is sensitive to the number of times a reference is included across the 10 iterations. If 3.5 models are used then this most efficient threshold must be determined in the test phase.

Commented [MHV4]: Check this - – this results seem to vary from the table

Commented [MHV5]: Check this – this results seem to vary from the table

Due to costs, we have not investigated the performance of GPT-4 with 10 iterations. As soon as the cost get close to the current cost of GPT-.3.5 models, users could considered screening all titles and abstracts with 10 iterations. For now suggest just to re-screening all references where humans and GPT disagree.

In contrast with priority screening methods (Hou & Tipton, 2024), the gpt models do have the potential to find more than 95% of the relevant study cf. FRIENDS.

A side-effect of this simulation was also to validate the performance of the AIscreenR software. Especially the use of function calling.

Student screening evaluation (Ng et al., 2014)

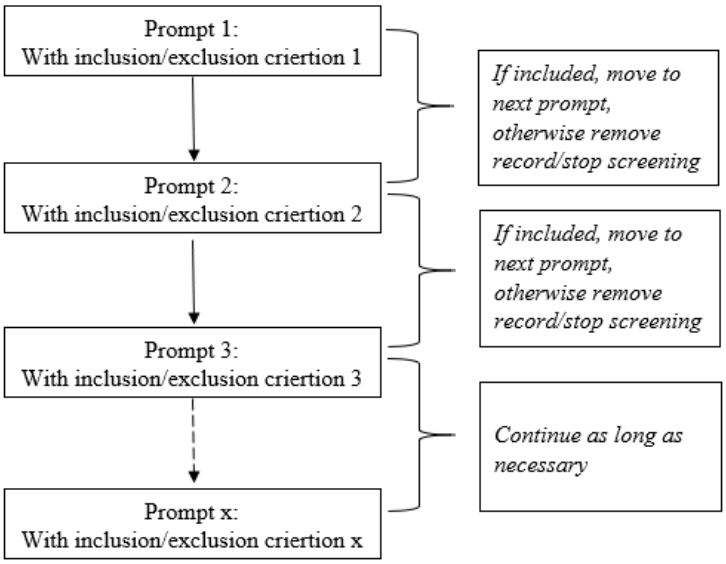
6 TENTATIVE GUIDELINES

80% recall and 95% specificity.

Workflow and short package presentation

Testing, not training. Less is more.

Figure x: Hierarchical screening



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

7 LIMITATIONS

- Black box (but this does not only count for GPT this is often true for human screening as well)
- Different performance across model updates
- Function tech? We have no control over the existence of OpenAI
- Environmental impact (embrace the critiques from van Lissa)

8 FUTURE RESEARCH

- The use of hierarchical prompting in complex reviews. Simple prompts instead of long onces
- 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.

9 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.
- Forces review times to make very narrow searches due to lack of ressources to conduct the title and abstract screening rigorously (Guo find in ICloud)
- 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)

- 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.

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, Johan Klejs, and Frederikke Lykke Witthöft Schytt 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

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

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

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., 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*.

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

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

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

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

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

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.

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.