**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 resource-intensive procedure that decelerates the review process, ultimately excluding many researchers from using it. To alleviate this issue and potentially increase the reliability of systematic reviews, we evaluated the use of OpenAI’s GPT (generative pre-trained transformer) API (application programming interface) models as an alternative to human second screeners of titles and abstracts. To make a comprehensive evaluation of this screening approach, we developed a new benchmark scheme for interpreting the performances of automated screening tools against common human screening performances in high-quality systematic reviews, and we conducted three large-scale classifier experiments on three systematic reviews with different levels of complexity typically encountered in high-quality reviews. Across all experiments, we show that the GPT API models can perform on par with or even better than typical human screening performance in terms of detecting relevant studies while showing high exclusion performance as well. Hereto, we introduce the use of multi-prompt screening, that is making one concise prompt per inclusion/exclusion criteria in a review, and show that this can make GPT screening work in highly complex review settings. To support future reviews, we develop a reproducible workflow and tentative guidelines for when reviewers can or cannot use GPT API models as independent second screeners of titles and abstracts. Finally, to standardize and scale up this application, we present the R package AIscreenR. Our aim is ultimately to make GPT API models acceptable as independent second screeners within high-quality systematic reviews.

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

**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.
* Traditional automated screening tools can ease the burden of title and abstract screening.
* Traditional automated screening tools most often cannot detect/classify all relevant studies, which in turn, can induce ‘artificial screening biases’.

**What is new**

* We show that OpenAI’s GPT API models can function as highly reliable second screeners in social science reviews with recall performances at least on par with human screeners.
* 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 highly complex review settings.
* We develop a new benchmark scheme based on typical human screening performances from high-standard systematic reviews. This scheme is meant partially to make judgments about acceptable and unacceptable screening performances in high-quality reviews and partially to make reliable comparisons between the screening performance of humans and automated screening tools in general.
* We provide general guidelines for how and when GPT API models can and cannot be used as independent second screeners of titles and abstracts.
* We present and validate the AIscreenR package to allow for standardized conduct of title and abstract screening with (OpenAI’s) GPT API models.

**Potential impact for Research Synthesis Methods readers**

* Changing duplicate screening of title and abstract screening in systematic reviews.
* Increasing the reliability of large-scale systematic reviews.
* Making substantial reductions of human labor in systematic reviews without compromising the reliability of reviews.
* Providing a new guideline for reviewers on when and when not to use automated screening tools in high-quality systematic reviews.
* Standardizing title and abstract screening with prompt-based Large Language Models (LLMs).

**1 INTRODUCTION**

Systematic reviews are essential 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 titles and abstracts to be screened for relevance. 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, threatening the internal validity of systematic reviews (Brunton et al., 2017; Hedges, 1992; Rothstein et al., 2005; Shadish et al., 2002). In effect, independent human double-screening is considered the ’gold standard’ to hinder a biased selection of studies (Guo et al., 2024; Higgins et al., 2019; Stoll et al., 2019; Wang et al., 2020).

Duplicate screening of all identified titles and abstracts is, however, a resource-intensive procedure, often 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, for example, due to low budgets or narrow time limits (Pacheco et al., 2023). Alternatively, reviewers narrow their searches to keep the number of records down to a manageable size, again increasing the risk of overlooking relevant studies (Van De Schoot et al., 2021). Over time these issues will only grow in magnitude as 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). As such, it can be considered an economically inefficient and unsustainable use of human resources to rely solely on duplicate human screening of titles and abstracts in future systematic reviews (Shemilt et al., 2016).

An alternative to human double-screening is to use (semi-)automated screening tools either based on text-mining or machine-learning algorithms to act either as a second screener, a course-grained classifier, or to sort citation records in a prioritized order, thereby allowing for a more efficient screening (Cohen et al., 2006; Gartlehner et al., 2019; O’Mara-Eves et al., 2015; Van De Schoot et al., 2021). These tools have the potential to make substantial reductions in human screening workloads in systematic reviews (O’Mara-Eves et al., 2015; Perlman‐Arrow et al., 2023). However, a clear disadvantage of these workload savings is that the tools may always be expected to miss a part 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 for reviewers 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, referred to by König et al. (2023) as ‘artificial screening bias’.[[1]](#footnote-1)

Most evaluations of traditional automated screening tools also yield the general conclusion that these tools are not yet capable of replacing an independent human second screener without a significant risk of omitting a substantial number of eligible studies (Burgard & Bittermann, 2023; Gartlehner et al., 2019; Kugley et al., 2016; O’Mara-Eves et al., 2015; Olorisade et al., 2016; Rathbone et al., 2015). By using the level of automation heuristic, 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 where they make credible independent, deterministic screening decisions.

A potential solution to elevate automated screening tools to the highest levels of automation is to use large language models (LLM), such as the generative pre-trained transformer (GPT) models that have recently been introduced by OpenAI. Initial evaluations have generally yielded some promising results with regard to using OpenAI’s GPT API (application programming interface) models for screening titles and abstracts.

To our knowledge, Syriani et al. (2023) were the first team to evaluate these models for screening tasks, specifically comparing the GPT-3.5-turbo-0301 model to state-of-the-art machine learning algorithms in systematic reviews within software engineering.[[2]](#footnote-2) They found the model's performance to be comparable with traditional classifier models and most often better. In another study, Guo et al. (2024) tested the use of OpenAI’s GPT-4 API model for title and abstract (henceforth TAB) screening of medical research literature.[[3]](#footnote-3) Across six clinical reviews, when evaluating the model’s inclusion and exclusion decisions against the final decisions of two independent human screeners, GPT-4 was found to have average recall (i.e., the proportion of relevant records being correctly classified as relevant. See the statistical definition in next section) and specificity (i.e., the proportion of irrelevant records being correctly classified as irrelevant. See the statistical definition in next section) values of 0.76 and 0.91, respectively. Based on this, the authors concluded that GPT-4 was effective in excluding irrelevant studies but less reliable in identifying relevant ones, and therefore recommended its use as a support tool but not as a replacement for human screeners. Similar conclusions were reached by Gargari et al. (2024) after testing the use of the gpt-3.5-turbo-0613 API model for TAB screening in a clinical systematic review.[[4]](#footnote-4)

On a related line of research, Alshami et al. (2023), [Khraisha](https://onlinelibrary.wiley.com/authored-by/Khraisha/Qusai) et al. (2024), and Issaiy et al. (2024) explored the ChatGPT web-browser interface for TAB screening, with mixed results, indicating performance similar to the API models but generally insufficient compared to human reviewers.

Although previous applications and evaluations of using OpenAIs GPT models for TAB screening represent a vital first step in validating these models as independent second screeners in systematic reviews, many questions remain unanswered. In this respect, 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, no common workflow and guidelines exist for how to conduct GPT-based TAB screenings, including how to make reliable prompts. Also, no software has yet been developed to support and standardize the setup of GPT-based TAB screenings. This study therefore aims 1) to test and validate the TAB screening performance of GPT API models, 2) to develop a heuristical workflow for how to conduct TAB screening with GPT API models, and 3) to present the R package AIscreenR (Vembye, 2024). Our goal is to develop an easy-to-implement framework that works in combination with standard review software.

The remainder of the paper proceeds as follows: In Section 2 we ask how well a GPT model needs to perform to be accepted as an independent second screener of tables and abstracts in a high-quality systematic literature review. To answer this question, we analyze human screening performances across 22 high-quality systematic reviews. We use this evaluation as the basis for developing a novel, empirically informed benchmark scheme for interpreting acceptable and unacceptable screening performances in high-quality systematic reviews. Next, in Section 3, we present three classifier experiments to systematically evaluate the use of GPT API models for TAB screening. This includes presentations of the prompt engineering and data underlying these experiments as well as the results of the experiments. In Section 4, we deduce tentative guidelines for when it is acceptable and unacceptable to use GPT API models as independent second screeners. Moreover, we elaborate on how we think reliable prompts can be developed in future reviews. Finally, in sections 5-7, we recapitulate by reflecting on the limitations of our work, the prospect of using LLMs for TAB screening in high-quality systematic reviews, and what should concern future research as well as the implications of our results and recommendations.

**2 HOW MUCH ERROR CAN WE ACCEPT IN HIGH-STANDARD SYSTEMATIC REVIEWS?**

Before adopting automated tools, such as GPT API models, as independent TAB screeners for systematic reviews, we need to ensure that these tools are not inferior to human screeners to avoid compromising the quality of the review (O’Connor et al., 2019). To allow for assessments of this, in this section, we will develop a novel, empirically informed benchmark scheme for interpreting acceptable and unacceptable TAB screening performance in high-quality reviews. We start by presenting the performance metrics that we use to develop our benchmarks. Next, we present and analyze data on human screening performances across 22 state-of-the-art systematic reviews and use this as a basis for our benchmark scheme.

**3.1 Evaluation metrics**

In existing literature, a wide range of different metrics has been used to evaluate TAB screening performances in the context of systematic literature reviews. Our choice of metrics has been informed by the recommendations of O’Connor et al. (2019) and Syriani et al. (2023). As such, the most central performance metrics in our analyses are the *recall* (sometimes referred to as the sensitivity) and *specificity* metrics since these are intuitive to understand and interpret and are not sensitive to imbalanced data, that is when data contains a large difference in the proportion between inclusion and exclusion references, which is commonly found in systematic reviews (Brunton et al., 2017). The recall (defined above) can be written as

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

where (true positive) represents all the studies that are correctly included, and (false negative) is the number of studies that are falsely excluded. By contrast, the specificity metric (defined above) is given by

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

where (true negative) represents all the studies that are correctly excluded, and (false positive) is the number of studies falsely included.

For our purpose, we consider the recall measure to be the most important performance measure since missing relevant studies (i.e., having a low recall) is the main reason for automated tools to potentially introduce bias in systematic reviews (Hou & Tipton, 2024). A low specificity does not induce bias, it just means that reviewers have to 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, 0.5, this still implies that reviewers can confidently exclude 50% of the irrelevant records, which would in most cases still be considered a significant reduction in screening workload. As such, we think that automated tools should be accepted as long as they come close to scenarios A and B pictured in Figure 1 (i.e., tools should be accepted when high recalls can be reached, to a large extent independently of the accompanied specificity value). We will come back to this aspect in the following sections.

In addition to the recall and specificity metrics, which concern the inclusion or exclusion performances individually, it may also be relevant to look at summary metrics that incorporate the overall performance across the inclusion and exclusion metrics. A typical issue with such summary metrics is that they tend to be sensitive to imbalances in the data. For instance, if one simply uses the raw agreement metric in a context with imbalanced data, then the screening performance will most often be overestimated. Assume that you have 10 relevant records per 1000 records. You could then reach a raw agreement of 99% simply by excluding all records (i.e., without identifying any relevant record). To overcome this issue, we used two overall metrics that account for imbalances*; the balanced accuracy* and *the normalized Matthew correlation coefficient* . The balanced accuracy metric balances the accuracy of the performance across the recall and specificity metrics and is simply an average of those metrics given by

|  |  |  |
| --- | --- | --- |
|  |  | (3) |

The metric is considered as the metric that maximizes the use of the four quantities, *TP, TN, FP,* and *FN* can be calculated as follows

|  |  |  |
| --- | --- | --- |
|  |  | (4) |

This metric has, moreover, been shown to have better statistical properties than other popular metrics such as the Receiver Operating Characteristic Curve (ROC AUC) (Chicco & Jurman, 2023). Therefore, we drew on this metric relative to the ROC AUC.

FIGURE 1: Recall and specificity performances

|  |  |
| --- | --- |
| A: High recall, high specificity  Irrelevant  Relevant    Included by GPT | B: High recall, low specificity  Irrelevant  Included by GPT  Relevant |
| C: Low recall, high specificity  Irrelevant  Relevant  Included by GPT | D: Low recall, low specificity  Relevant  Included by GPT  Irrelevant |

*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.

**3.2 Typically human screening performances in high-quality systematic reviews**

In order to make fair comparisons between human and automated screening performances, we consider it important to have a deep understanding of acceptable *human* screening performance in high-standard systematic reviews (O’Connor et al., 2019). We consider this comparison as the only reliable way to assess whether a given recall is good or bad. For example, 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. Hereto, we think it is important to acknowledge that individual human screening is not without significant errors 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 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. The overall purpose of this scheme is to aid and establish a common guideline for evaluating automated screening results for systematic reviews across different fields of research.

*3.2.1 Data*

As the empirical basis of our benchmark scheme, we analyze human screening performances in 22 high-standard systematic reviews that used independent duplicate human screening. This includes 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 having 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 independently double-screened by at least two individuals. A total of 46 individual screeners participated in this process, 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 instance, assistants might be expected to have limited content expertise with regard to the topic(s) under review, which might potentially lower their performance. Thus, their performances might not necessarily be on par with the typical screening performance of content expert screeners. Hence, we analyze the Campbell review data separately for assistant/non-expert and researcher/expert screeners. This, as well, is not without problems. As such, differences in performance levels between the two types of screeners may not only reflect different levels of content expertise but 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 are. We therefore add screening performance data from five systematic reviews conducted by NIPH. In these reviews, all TAB screenings were conducted by researchers with a high level of content expertise related to the given review. This should give a clearer picture of common expert/researcher performances in systematic reviews. The added NIPH data includes 13,825 title and abstract records that have been independently double-screened by a total of 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. 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** |  | **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](https://onlinelibrary.wiley.com/doi/10.1002/cl2.1212) | 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 Analysis*

To analyze screening performances in the 22 above-described systematic reviews, 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 metafor package (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/>.

Using the data presented in section 3.2.1, we estimated all the performance metrics via Equations (1) to (4). The , , , and conditions used in these equations were determined by comparing each single human screener decision with the final decision reached by 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.[[5]](#footnote-5) 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. 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 in terms of *recall*, *specificity,* *balanced accuracy*, and *nMCC* metrics across the included studies, we fitted two versions of the so-called *correlated-hierarchical effects* (CHE) working models (Pustejovsky & Tipton, 2021). For the investigation regarding 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, , is often entirely or partially unknown and must be imputed. In all the used working models, we assumed . To guard against model misspecification both models have incorporated robust variance estimators. The main difference between the two models is that they draw on slightly different weighting schemes but the SCE model is generally recognized as the main working engine for deriving subgroup effects and conducting reliable contrast tests (Pustejovsky & Tipton, 2021). For differential effects comparisons, we used the HTZ Wald test suggested by Tipton and Pustejovsky (2015). Across both models, we estimated two sources of heterogeneity; the variability of the true screener performances within () and between studies (). 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 their distribution around the overall performance means are illustrated in Figures 2 and 3. We find the overall average *recall* value 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 find the two groups’ average recal values to be statistically different from each other with *F*(1, 10.3) = 14.58, *p*  = .003. We detect minor substantial variations between the performance measures within studies with = 0.026 and = 0.035 for the assistant and author screeners, respectively. We are unable to detect any true differences in performance levels between studies, thereby indicating consistent average screening performances across the Campbell reviews, both for assistants and expert screeners. The overall average *specificity* for assistant screeners is 0.980, 95% *CI*[0.966, 0.990], and for review authors 0.988, 95% *CI*[0.980, 0.995]. We find no statistically significant difference between the two average estimates with *F*(1, 13.6) = 2.08, *p* = 0.172. We do only find very minor, non-substantial variation within and between studies with = 0.004 as the maximum for author screeners.

Moving on to the *balanced accuracy* metric, we find average performance levels of 0.874, 95% *CI*[0.857, 0.890] among assistant screeners and 0.933, 95% *CI*[0.899, 0.961] among author screeners. We find the difference between the group means to be statistically significant with *F*(1, 10.1) = 18.22, *p* = .002. Finally, the overall *nMCC* values are 0.860, 95% *CI*[0.835, 0.882] and 0.925, 95% *CI*[0.880, 0.953] for the assistant and author screeners, respectively. The group averages are found to be statistically different from eachother with *F*(1, 11) = 9.65, *p* = .01.

Based on these results it appears that compared to students, researcher (i.e., content expert) screeners are substantially better at detecting relevant studies. Yet, as noted, this difference may be driven by factors other than mere screening quality (e.g., authority relations with a potential to inflate the researchers’ performance relative to that of the assistants). Interestingly, when analyzing the NIPH data, which is in all cases based on independent researcher-researcher screening comparisons, we find performance levels closer to those of the assistant screeners in the Campbell reviews. As such, the average recall value in the NIPH data is 0.839, 95% *CI*[0.737, 0.920]. Again, we find minor true variation between the screener recall performances within studies with = 0.029 and = .0. The overall average specificity value of the NIPH researchers is 0.977, 95% *CI*[0.955, 0.992], with almost no true variation between screeners and between studies. The average balanced accuracy and *nMCC* values of the NIPH researchers are 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

**3.3 Benchmarking**

Bearing on the empirical results presented in the previous section, we developed the screening benchmark scheme presented in Table 3. We do thus suggest—as a coarse-grained rule of thumb—that if a GPT API model (or other automated screening tool) can reach a recall level equal to or above 0.8 and a specificity level above 0.95, then the performance can be said to resemble common human screener performance in the context of high-quality systematic reviews, meaning that the model can safely be used as an independent second screener.

There are nuances to this guideline as we believe that automated tools may be useful under less restrictive conditions as well. Held against common human error rates, we consider a recall between 0.75 to 0.80 to be acceptable as such recalls are similar to those accepted from assistant screeners today. Consequently, and in contrast with previous evaluations (Guo et al., 2024), we would not necessarily interpret a recall value of 0.76 to be too low for a GPT API model to function as an independent second screener. We even believe that recall values between, say, 0.5-0.75 should disqualify automated screening tools from playing a role in TAB screening. Under such conditions, we would warn against using the tools as independent screeners but they could function as an extra assurance, working as a third screener that forces the human screeners to double-check close-to-relevant study records. This would enhance the screening, lowering the risk of humans overlooking any relevant records.

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 | Low 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 | Low 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 value above 0.8 | On par with common human screening performance |

*Note*: Red areas indicate conditions under which the TAB screening performance is unacceptability low. Gray areas represent insufficient performance conditions but some applications with these performance measures might still be viable. Green areas represent acceptable screening performances on par with or better than human screening.

As can be seen from Table 3, we do not necessarily conceive a specificity value of 1 to be the ideal, since we do rather want our automated screenings to be over-inclusive than over-exclusive. Thus, a low specificity value merely forces human screeners to double-check a larger number of potentially relevant references, which in turn lowers the risk of relevant studies being missed. As such, we think that a specificity value equal to or above 0.8 is acceptable as long as the recall value is equal to or above 0.8 (c.f. Figure 1) as well. We, therefore, suggest that automated screening performances reaching recall and specificity values above 0.8 should be accepted as independent screeners in high-quality systematic reviews. Finally, we think that automated tools that yield high recalls may be used to reduce the total number of title and abstract records needed to be screened, even if the specificity value is below 0.8. 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).

We think that in most cases, it will not be viable to use automated tools with performance levels (recall and specificity values) below 0.5. However, in Table 3, we use graduated shades of red, where the light red color for specificity values below 0.5 indicates that we cannot reject that that there might be cases where automated screenings could be usefulm even with specificity values below 0.5 (as long as the recall is high). As such, in extreme-sized reviews, even a 30% screening reduction may represent substantial workload savings, amounting to multiple days of work saved. The number of title and abstract records does, however, need to be very high in order for it to make sense to conduct automated screenings with such low specificity levels.

With our benchmark scheme, we aim to make a more flexible tool partially for assessing the screening performance of automated tools in general 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 or other automated tools for TAB screening tasks in systematic reviews, avoiding trivial for and against discussions. Furthermore, we will use this benchmark scheme for interpreting our conducted classifier experiment that we present in the next section.

**4 CLASSIFIER EXPERIMENTS**

In this section, we present the data and prompts used as well as the results of three large-scale classifier experiments.[[6]](#footnote-6) Differently from previous research, these classifier experiments aim to test the performance of OpenAI’s GPT API models 1) when applied in social science reviews and 2) when using multi-prompt screening in complex review settings. A side-effect of conducting these experiments is further to quality assure the AIscreenR package (Vembye, 2024) for automated TAB screenings using GPT API models. We consider this test all-important in order to be able to scale up our suggested screening approach. The three experiments each represent different levels of complexity in terms of inclusion criteria. The main purpose of the paper is not to show that GPT API models work in all instances. Instead, we aim to show that if configured adequately, these models *can* function as highly reliably independent second screeners across various types of systematic review questions. This also means that using GPT API models as a second screener is not always ideal for various reasons, cf. the proposed guidelines in Section 3.3 and our discussion in Section 5.1. We return to this issue in Section 5.1.

**4.1 Data**

In Experiment 1, we tested the performance of GPT API models in the context of a Campbell Systematic Review, conducted by Filges et al. (2015), on the effects of functional family therapy (FFT) on drug abuse reduction for young people in treatment for nonopioid drugs. By leveraging a previously published review, we were able to immediately evaluate the GPT API models’ performances 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 (FFT) represented a well-defined intervention. This made it an ideal initial test case for proof of concept purposes. If the GPT API models could not achieve satisfactory performance in this context, they would unlikely be able to do so in the context of more complex reviews. Another interesting feature of this experiment is that it was based on a highly imbalanced dataset with only 69 of 4135 records being relevant, amounting to an approximate inclusion ratio of 17 relevant studies per 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).

A possible critique against Experiment 1 is that it draws on a published open-access review, meaning that OpenAI’s GPT models can potentially have been trained on the review. If this is the case, may reduce the generalizability of the experiment’s results to applications where GPT API models are used for prospective reviews where the screening results have not been fed to the models. To address this issue, we conducted a second classifier experiment drawing on data from an unpublished/ongoing systematic review. Thus, in Experiment 2, we used screening data from a Campbell systematic review of the effects of the FRIENDS preventive programme on anxiety symptoms in children and adolescents conducted by Filges, Smedslund et al. (2023).[[7]](#footnote-7) The FRIENDS data in many aspects resembles the FFT data, with inclusion criteria being rather simple and the intervention being well-defined. Moreover, the data is highly imbalanced with only 64 of 2572 records being relevant, amounting to an approximate inclusion ratio of 25 relevant studies per 1000 records.

Finally, a fair critique of Experiments 1 and 2 is that they both involve ‘easy’ TAB screening tasks with interventions being more well-defined and inclusion criteria being more simple than is often the case in systematic reviews. To address this, we conducted a third classifier experiment to investigate if GPT API models can be used for TAB screening in review settings that are more complex. For Experiment 3, we used screening data from an ongoing Campbell Systematic Review of the effects of testing frequencies on students’ academic achievement (Thomsen et al., 2022). Compared to the screenings in Experiments 1 and 2, we considered this a more difficult screening case as the review’s inclusion criteria are more subtle, including notions that are not well-defined. As such, the intervention (student testing) is a type of learning strategy that is ubiquitous in education and is used in a variety of ways for many 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. In effect, 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, e.g., as quizzes, progress-monitoring, curriculum-based measures, and retrieval practice). Judging the eligibility of particular studies therefore requires a great deal of subject matter familiarity. Therefore, and contrary to Experiments 1 and 2, we think that if GPT API models can achieve satisfactory performances in this context, they would likely be able to do so in most review contexts. In this experiment, our data consists of 2000 irrelevant and 100 relevant records randomly sampled from the total pool of 5612 irrelevant and 627 relevant records, respectively. We did so to optimize our use of resources as this screening was carried out as a multi-prompt screening, with each title and abstract being evaluated against six individual prompts, each corresponding to a specific inclusion criterion.

In all three experiments, we excluded study records that did not have an abstract. This excluded 208, 150, and 41 study records in the FFT, FRIENDS, and testing frequency (henceforth TF) data, respectively. In 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 to include an introduction section describing the general aim of the review followed by the inclusion criteria of the review. To exemplify, Textbox 1 exhibits the prompt used for 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?*

Inclusion criteria

Short/concise background   
information

Next, when given study IDs,[[8]](#footnote-8) titles, and abstracts, the AIscreenR automatically integrates this information in the prompt, using the text in Textbox 2:

*“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]”*

TEXTBOX 2: End of prompt added by AIscreenR

By pasting the full prompt together with each title and abstract, we aim to guard against model drifting/hallucinations. We did not add any instruction regarding how the model should respond to our request in the main prompt, as done in previous studies (REFS). Instead, we relied on function calling and built two JSON functions (one function call yielding simple trinary results and another yielding descriptive screening responses) with instructions to be provided to the model. According to OpenAI, this should result in more reliable and standardized responses from the models (OpenAI, 2024). The main JSON respond function[[9]](#footnote-9) 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 was 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. What they need to work properly are concise prompts and thus, test performances increased dramatically when given more precise prompts with fewer inclusion/exclusion criteria at a time. As such, for Experiment 3, we developed and evaluated the concept of multi-prompt screening where each inclusion 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). All engineered prompts used for Experiment 3 are presented in Appendix A. We elaborate further on multi-prompt screening in Section 5.[[10]](#footnote-10)

*4.2.1 Prompt calibration*

Before initiating each classifier experiment, we tested and refined our prompts on a subet of the title and abstract records. For the FFT review, we started by testing our prompt on a single relevant reference, and we refined the prompt until the models consistently included this study record. Then, we scaled up the test to include 150 irrelevant and 50 relevant records. Using our proposed benchmark criteria from section 3.3, this test yielded satisfactory results, demonstrating an ability to reach a recall above 0.8 and a specificity above 0.9, and thus, we moved on to screening all records using the GPT API models to investigate if the models’ 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. Again, we initiated the full screening after finding the GPT models to yield satisfactory screening performances.

**4.3 Evaluation design**

In all three classifier experiments, we evaluated the performance of the GPT API models by using Equations (1) to (3) from section 3.1. The , , , and 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 models, reached from the ‘v1/chat/completions/’ endpoint. Since the GPT-3.5 models are generally less consistent in their responses than GPT-4, 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 the model’s consistency across screenings and to assess how this impacted its final inclusion decisions. The final inclusion decision of GPT-3.5 was then based on the probability of inclusion across the repeated requests. In part because the GPT-4 models are more consistent in their responses, and in part because of the higher costs of using these models, we only conducted one screening per title and abstract when calling gpt-4-0613. For Experiment 3, which involved multi-prompt screening, we only drew on GPT-4 and the final inclusion decision of the model was then based on the probability of inclusion across all used prompts. In our main analysis of Experiment 3, we included study records if they were included by the GPT model in at least 5 of the 6 used prompts. For all experiments, we used invariant top\_p and temperature values, using the default value of 1 for both hyperparameters.

**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 values of 0.899 and 0.933 in Experiment 1, which, using the benchmark scheme from Section 3, can be considered to be on par with typical human screening performances in high-quality systematic reviews. The gpt-3.5-turbo model was also able to reach human-like screening performances. Yet these results varied substantially depending on the chosen inclusion probability threshold, reflecting the model’s higher level of inconsistency in screening decisions, especially when it comes to detecting relevant studies (cf. Table 4’s recall column). Figure 4A shows the decision sensibility of the GPT-3.5 model across different inclusion probability thresholds in the context of Experiment 1. When using an inclusion probability threshold of 0.2 (meaning that studies were coded as relevant if the GPT model included it in 2 or more of the 10 screenings), the GPT-3.5 model yielded a recall of 0.81 and a specificity of 0.94. However, when using an inclusion probability threshold of 0.5, the performance became unacceptably low compared to human screening, with a recall of only 0.69.

When used on the FRIENDS data, the GPT-4 model performed extremely well with performance measures exceeding common human screening performances. Specifically, it yielded a recall of 0.98 (only missing one relevant study) and a specificity value of 0.97. When using an inclusion probability threshold of 0.699, the GPT-3.5 model performed very well as well, with a recall of 0.95 and specificity of 0.90. Yet, again, as can be seen in Figure 4B, the screening performance of GPT-3.5 varied by a lot depending on the chosen inclusion probability threshold. An impressive fact regarding these results is further that we approximately spent 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 would be theoretically 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 (Experiment 3), the GPT-4 model yielded a recall of 0.80 when including studies that were included by the model in at least 5 out of 6 prompts. This is on par with typical human screening performances (cf. our benchmark scheme from Section 3) and does even exceed three out of six human recalls within this review (cf. Thomsen et al. (2022) under column 1 in Figure 2). Moreover, the model yielded an acceptable level of specificity of ~0.84. Relative to human performances, the model in this case was rather over-inclusive. Yet again, we do not necessarily consider this to be disadvantageous, since it reduces the risk of overlooking relevant studies, which might be even more important in complex review settings where exclusion decisions may more often be difficult to make at the first level of screening due to insufficient information in the abstracts. As can be seen in Table 4, when setting a more inclusive threshold, coding studies as relevant if the GPT model included it in at least 3 of the 6 prompts, the model was able to reach a recall of 0.95, but the specificity was then quite low, that is 0.67, leaving a rather high number of title and abstract records to be double-checked by human screeners.

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)/]a** | **bAcc** |
| *FFT* |  |  |  |  |  |
| gpt-3.5-turbo-0613  (incl. prop .5) | 10 | 0.699  (48/69) | 0.961  (3906/4066) | 0.956  (3954/4135) | 82.8 |
| gpt-3.5-turbo-0613  (incl. prop .2) | 10 | 0.812  (56/69) | 0.937  (3809/4066) | 0.935  (3865/4135) | 87.4 |
| gpt-4-0613 | 1 | 0.899  (62/69) | 0.937  (3810/4066) | 0.936  (3872/4135) | 91.8 |
| *FRIENDS* |  |  |  |  |  |
| GPT-3.5-turbo-0613  (incl. prop .5) | 10 | 0.953  (61/64) | 0.813  (1918/2508) | 0.816  (2100/2572) | 88.3 |
| gpt-3.5-turbo-0613  (incl. prop .7) | 10 | 0.953  (61/64) | 0.899  (2254/2508) | 0.900  (2315/2572) | 92.6 |
| gpt-4-0613 | 1 | 0.984  (63/64) | 0.974  (2442/2508) | 0.979  (2518/2572) | 97.9 |
| *TF* |  |  |  |  |  |
| gpt-4-0613  (incl. 5 out of 6 prompts) | 1 | 0.80  (80/100) | 0.838  (1676/2000) | 0.836  (1756/2100) | 81.9 |
| gpt-4-0613  (incl. 4 out of 6 prompts) | 1 | 0.89  (89/100) | 0.743  (1486/2000) | 0.75  (1575/2100) | 81.6 |
| gpt-4-0613  (incl. 3 out of 6 prompts) | 1 | 0.95  (95/100) | 0.67  (1340/2000) | 0.683  (1435/2100) | 81 |

*a*: is the total number of references

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

|  |
| --- |
| A |
| B |

*Note*: The inclusion probability threshold on the x-axis indicates the required number of times the given title and abstract record needed to be included over the 10 repeated requests in order to be coded as relevant. For example an inclusion probability threshold of 0.1 means that the record was coded as relevant if the GPT model included it in 1 or more out of 10 requests.

To summarise, we find that GPT API models can work as highly reliable and independent second screeners with recall performances on par or better than common human screeners, even in highly 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, in our third classifier experiment, we used multi-prompt screening instead of including all inclusion criteria in a single prompt. Part of our comparably high performance might also be due to the fact that we instructed the models to include title and abstreact records with insufficient information (cf. Textbox 3 in Section 4.2). We note that based on our tests, the GPT-4 API model seems to be preferable relative to GPT-3.5 since the latter is rather sensitive to the chosen inclusion probability threshold, as illustrated in Figure 4. Based on this finding, we generally recommend not using the GPT-3.5 API models when GPT-4 API models are available. Moreover, in cases where researchers have to rely on GPT-3.5, different inclusion probability thresholds should be considered. Overall, we think that using GPT API models for TAB screening tasks in high-quality 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. This 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 screenings 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 guideline 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 screenings using 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 entire 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 or above 0.8, and a specificity equal to or above 0.8 (ideally between 0.9 and 1). If this step cannot be fulfilled, we recommend *not* using the GPT API model as an independent 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 relevant studies. In cases where low budgets exclude human duplicate screening, we consider it acceptable to work with recalls below 0.8 as the alternative (i.e., stand-alone single-screening) is worse. |
| 6 | Manually single screen all study records (could be divided into batches of 500-1000 study records). |
| 7 | Download ris-files individually for included and excluded references. Load this data into R and track the human decisions. |
| 8 | Run the full TAB screening using 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*: For a detailed presentation of how, in practice, to conduct TAB screening using GPT API models, see the vignette accompanying the AIscreenR package (Vembye, 2024).

Before initiating a full-scale TAB screening using GPT API models, we generally recommend thoroughly testing and validating the screening performance of the prompt(s) and GPT API model(s) aimed to be used for the screening to ensure 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)/model(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 prompts 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, thereby ensuring that the specificity test value can be generalized to the full sample of study records. If any relevant studies are detected among the randomly sampled study records, these can just be added to the pool of relevant 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. The models do not need to be trained and should therefore in general only be fed with a minimum of information. 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 criteria is 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 considered irrelevant if it is excluded at any step in the multi-prompt screening. This procedure is depicted in Figure 5.[[11]](#footnote-11)

FIGURE 5: *Hierarchical screening*

Prompt 1:   
With inclusion/exclusion criertion 1

*If included, the record moves to next prompt, otherwise the record is excluded*

Prompt 2:   
With inclusion/exclusion criertion 2

*If included, the record moves to next prompt, otherwise the record is excluded*

Prompt 3:   
With inclusion/exclusion criertion 3

*Continue as long as necessary*

Prompt x:   
With inclusion/exclusion criertion x

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 criteria follow thereafter (as suggested by 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 criteria they were excluded upon. However, a shortage of this screening approach is that it is strongly 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 six prompts, one for each of the inclusion criteria, but one of the prompts wrongly excluded a large share of relevant records at an early stage of the screening when scaled up from the test setting. Then those studies would be lost in the hierarchical screening in Figure 5. If instead all records had been screened with all prompts, then one could overcome the above problem by coding records as relevant if they were included in 5 out of 6 prompts, as we did in Experiment 3 above.

When engineering prompts, we suggest that these should be refined until they reach recall and specificity values of at least 0.8 (cf. our benchmark scheme in Section 3). Recall values between 0.75 and 0.8 may also be accepted, but reviewers should try to increase this performance as much as possible. Lower specificity values may be accepted as long as the recall exceeds 0.8. If a specificity value of 0.8 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. We suggest that if a recall of 0.8 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. In this case, single-screening is still less desirable than using a low-performing GPT API model as an extra ‘pair of eyes’ to increase one’s chances 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 high-quality reviews. If the thresholds cannot be reached, the GPT API model can still be used as a third screener, again providing 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 a second screener, we suggest that the reviewers screen all study records before initiating the automated screening. Thereby, the human reviewers are not impacted by GPT’s decisions. In general, we recommend that decisions on whether GPT API model screening is appropriate in a given review should be made before the main TAB screening has been initiated or after the human screening has been conducted. An alternative to manually screening all records at once is to 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 revolutionary 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) and/or inclusion criteria is high *and* the number of reference records needed to be screened is low (e.g., less than 2000). In such a situation, it might take longer to engineer reliable prompts than it would take to instantly initiate 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 quickly set up a reliable screening with the FRIENDS data (i.e., in Experiment 2 above). Therefore, if the complexity of a review’s inclusion criteria is low, it may be advantageous to conduct a rapid investigation of whether GPT API model screening is appropriate in the specific case, even if the number of records is not high. However, we do not think reviewers should spend too much time on this task in such cases. Table 6 visualizes the conditions under which 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** | | |
| **Complexity of the review question(s) and/or inclusion criteria** |  | Low | High |
| 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/>. It should be noted that – even if running the exact same codes as we used in our tests – the screening performances might not be identical to ours since, as illustrated in our analyses, the screening decisions of the GPT API models (like humans) are not 100 % consistent across screenings. Yet, we still firmly believe that the overall patterns of our results can be replicated, and we gladly invite readers to test this hypothesis. In future applications, reviewers will be able to set a specific seed (currently a beta argument) to the request body ensuring the reproducibility of the given screening. We did not use this functionality since it was not developed at the time when we ran our experiments but we consider it to be a helpful feature for future applications.

Another clear limitation is that the models we drew on are based on black box, closed-source algorithms. We have demonstrated that current GPT API models – if configured adequately – are capable of doing TAB screening tasks, but we are unable to say *why* the models work. Model dependency is a major issue when working with GPT API models since we do not know how they are trained and/or will develop. This also means that the generalizability across different models and across time is unclear. Consequently, we cannot infer that the results of our experiments are generalizable to other GPT models, such as the GPT-4o model, the GPT-4-turbo model, or the API models from Claude or Mistral AI. From a scientific point of view, and to increase the transparency of GPT API screening, it is, therefore, important that future research revolve around investigating the performances of local, open-source, and downloadable models such as the ones provided by Mistral AI. That said, we do think it is important to note that most human duplicate screenings are also black box operations that are hardly entirely replicable, and we believe the GPT API models should be judged in light of this. This goes without saying that we should not strive to make screenings replicable and reproducible since this would clearly increase the transparency of high-quality reviews. Yet, we just do not think that the black box argument should be a major reason for not drawing on GPT API models for TAB screening in high-quality reviews.

On a similar, but technical line, it is demanding and time-consuming to keep up with new model developments and updates as well as how they are reached via the API. 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 as with previous models. Already, the original function calling arguments have been deprecated and moved to the tools argument in the request body. Therefore, future research must evaluate whether our results can be exerted with other GPT API models such as updated models as well as the GPT-4o and GPT-4-turbo, etc. Likewise, the GPT-3.5-turbo-0613 model that we drew upon is expected to deprecate during 2024, so eventually one will not be able to replicate the screening we have made with this model. On this note, we again think it is pivotal that future research investigates if downloadable GPT models can perform on par with OpenAI’s GPT API models. This would secure a more stable applicability of using GPT models for TAB screening, supporting the functionality of this technology.

Even though, the use of GPT API models as second screeners can be considered more efficient than using a human second screener, reviewers should be aware that it still can induce significant costs to one’s project, especially when working with GPT-4 models and multi-prompt screenings. To exemplify, we spent approximately $220 making the 12,600 screening requests (2100 references x 6 prompts) for Experiment 3. Even though prices will certainly drop over time,[[12]](#footnote-12) we recommend using the models carefully. In some applications, it might be advantageous to combine traditional classifier tools with GPT API models to reduce the total cost. In extreme-size reviews (i.e., 100,000< references), reviewers could e.g. consider combining priority screening/classifier modeling with the GPT API screening. The GPT API screener could then, for example, be used as an extra guardian, checking the performance of one’s selected stopping rule (Boetje & van de Schoot, 2024; Campos et al., 2023; König et al., 2023), e.g., by screening a subsample of, say, 1000 references on the wrong side of the set threshold. Then all references on the right side of the threshold of the stopping rule could be screened by at least one human and the GPT screener together. This is just one idea of how a more cost-efficient screening could be set up in large-scale reviews, and reviewers could play around with other solutions as well. Hereto, it is important to stress that we do not think traditional automated tools and GPT API models should be considered competing tools. Instead, they should be used together to overcome each other's disadvantages.

The screening approach that we suggest is limited by its prompt dependency, meaning that this screening approach is in theory rather sensitive to the prompt(s) made by the user. This can potentially complicate the use of the GPT API screenings as it can be time-consuming to build well-performing prompts. Reviewers must, therefore, always thoroughly consider whether the use of GPT API screening is resource-efficient in the given review case. A key purpose of our paper is, thus, also to guide reviewers on when GPT API screening might *not* be appropriate, which would be the case if prompt performances are not on par with human screening (cf. our benchmark scheme in Section 3), or if the time needed to reach satisfactory performance exceeds the time required for a human second screener to independently screen the titles and abstracts (cf. Table 6).

Although, with the AIscreenR, we have strived to make a user-friendly setup for GPT API screening, a limitation is that our screening approach is function-based, meaning that reviewers need to have (or at least acquire) some minor R coding skills. In the future, the screening approach may be embedded in a shiny app or in existing screening tools (similar to what has been done with the data extraction GPT API tool in the EPPI-Reviewer (EPPI-Centre, 2024)), thereby making it easier to use without prior R coding skills. A tempting solution to accommodate user-friendliness on the short run is to copy our approach to the ChatGPT interface. Importantly, however, we have *not* been able to reach satisfactory screening performances by using the ChatGPT internet interface. Specifically, the GPT API models reached from the ‘v1/chat/completions’ endpoint worked significantly better relative to the GPT models embedded in the ChatGPT interface. Consequently, we consider it pivotal that future research clearly distinguishes between OpenAI’s GPT models when doing research with them, so that different GPT model performances are not unnecessarily mixed up. In this paper, we have narrowly focused 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. It is, therefore, also unknown how our results generalize to the performance of using the ‘v1/completions’ endpoint as a second screener. Further, it should also be noted that comparisons between our results and the results of prior evaluations are restricted by the fact that we do not know what exact model endpoints were used by Syriani et al. (2023) and Guo et al. (2024).

Finally, several caveats should be mentioned regarding the data underlying the benchmark scheme that we have developed for interpreting screener performances in high-quality reviews. For instance, it is based on screener performances in a convenient sample of systematic reviews, possibly restricting the generalizability of the estimated average screening performance measures. We believe that the analyzed Campbell and NIPH screening performances provide key insights regarding what is currently being accepted from humans in high-standard reviews, and they indicate that the screener performances seem to be comparable across distinct disciplines. Yet, future research may usefully investigate typical screener performances more systematically and across various research fields, such as medicine and the social sciences to make even more refined screening guidelines.

**8 CONCLUSION AND DISCUSION**

Human duplicate TAB screening in systematic reviews is time-consuming, requiring a substantial amount of human labor which decelerates the review process and thereby the dissemination of evidence for practice, research, and policy. In this study, we have shown that OpenAI’s GPT API models can function as highly reliable second screeners even in complex review settings, making it possible to substitute one human in the duplicate screening process and reallocate human resources. Our findings suggest that, when configured correctly, GPT API models can perform on par with or even surpass human screeners with regard to finding relevant studies. Moreover, we found that the GPT-4 model outperforms the GPT-3.5-turbo model, and we therefore recommend primarily using the GPT-4 model for TAB screening. Moreover, we found that GPT API models *can* yield specificity values that are on par with humans, but in some applications appear to be slightly over-inclusive (i.e., they yield lower specificity values than typical human screeners). We do, however, not consider this a problem as long as the models obtain high recall values since low specificity values do not induce a bias—they just force human reviewers to double-check a higher number of records.

Our results contrast previous research in which GPT API models were found to perform well in terms of specificity but less well in terms of recall (Gargari et al., 2024; Guo et al., 2024). While it would be premature—based on our data—to make hard conclusions about the reasons for the higher performance (in terms of recall) in our classifier experiments, some differences are worth noting in terms of the workflow used by us compared to prior evaluations. First, as noted earlier, contrary to prior evaluations of GPT API models for TAB screening, we relied on function calling (OpenAI, 2024), thereby improving the models’ response consistency. Second, we instructed the models to account for uncertainties when TAB screenings were based on limited information. Third, in Experiment 3 (our most complex case), instead of adding all inclusion/exclusion criteria to the same prompt, we introduced and used multi-prompt screening, using one concise prompt per inclusion/exclusion criteria in the review.

Based on our findings, we believe TAB screening with GPT API models can revolutionize the way duplicate title and abstract screening is conducted in high-quality systematic reviews since these models show an ability to work at very high levels of automation (c.f. O’Connor et al., 2019), thereby making it possible – in some cases – to replace human *second* screeners. To do so, however, a standardized screening approach is needed in order to make it scalable and acceptable in high-quality reviews. Therefore, we also developed a reproducible workflow and tentative guidelines for when such screenings can (and cannot) be accepted in high-quality reviews. To further support automated GPT API model-based screenings, we developed the AIscreenR R package (Vembye, 2024). Among other things, this allows reviewers to draw on features such as function calling (i.e., making prompts without the need to explicitly specify how the model shall respond to the screening request) as well as multi-core processing, something that speeds up the screening significantly. After having used the AIscreenR software to conduct our classifier experiments, we feel confident to conclude that the software works as expected. Hence, we believe that reviewers can confidently use this software in their high-quality systematic reviews as well.

A key part of setting up a reliable GPT API screening is to thoroughly validate the performances of one’s screening prompt(s) and (if relying on repeated screenings or multi-prompt screenings of each title and abstract record) inclusion probability threshold before making any full-scale screening. For such assessments, we developed a new, empirically informed benchmark scheme for interpreting acceptable and unacceptable screening performance in high-quality reviews. Based on typical human screening performances in 22 high-standard systematic reviews, we suggest that if an automated screening can yield a recall value above 0.8, it should be acknowledged as being on par with typical human performances and can be confidently used as an independent second screener. In addition, we suggest that a specificity value equal to or above 0.8 should be accepted in high-standard reviews as long as the recall is equal to above 0.8 as well since a low specificity does not induce bias.

It is important to note that no matter how much effort is invested in developing good prompts, GPT API models—like humans—can err and, therefore, it is of vital importance that GPT API screening is combined with other traditional screening techniques, such as forward and backward citation tracking, to ensure that potentially missed studies re-enter the review. In that regard, GPT-based screenings are not different from screenings conducted by humans. Although our recommendations allow for minor errors, we generally recommend not to use GPT API screening if reviewers cannot reach satisfying recall and specificity values. In a similar vein, we never think a GPT API model should be used as a stand-alone screener. There must always be a human in the loop, meaning that humans must always take the role of the first screener of titles and abstracts in high-quality systematic reviews.

With this paper, we have strived to make the foundation on which evidence organizations (such as Cochrane and Campbell Collaboration) and review journals can assess the use of TAB screening with GPT API models. According to the Campbell Collaboration, the acceptance of using 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). These requirements have played a key part in this paper, and we have used them as the main pillars to build the suggested screening framework. Concretely, we have aimed to accommodate requirement *(a)* by building our framework and codes so that they can readily be remodeled to work with other API models than OpenAI’s. This means that our setup aims to be agnostic to the given provider of the given LLM and will be viable as long as reviewers have public access to LLM models. We aimed to support Campbell’s requirement *(b)* by developing the new benchmark scheme and by showing that GPT API screening is perfectly appropriate in high-quality reviews, whereas the development of the AIscreenR package and the quality tests hereof were meant to accommodate Campbell’s requirement *(c)*. Moreover, to fulfill requirement *(f)*, we built the AIscreenR package as open-source software, allowing others (e.g., the Evidence Synthesis Hackathon, Campbell Collaboration, or the EPPI-Reviewer team) to readily contribute to the development and ongoing support of the software. Finally, we developed our workflow and guidelines to underpin requirements *(d)* and *(e)*. Requirement *(e)* is as such not necessary in our case since we are working we *pre-*trained models. 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.

However, some caveats and limitations follow our work. First, we agree with Schoot et al. (2021) that transparency and reproducibility represent the highest scientific standards. Yet, OpenAI’s GPT API models are based on black box algorithms. Nonetheless, we do not believe that this argument should prevent reviewers from using OpenAI’s GPT API models for TAB screening since human screening decisions most often represent black-box operations as well. Nonetheless, we consider it all-important that future research investigates the performance of alternative open-source GPT models. A side-effect of such research would further be that the costs of using GPT models may be substantially reduced, which can be a major barrier to using GPT-4 models for TAB screening at the current point in time. These models are still rather expensive (in absolute terms, not compared to hiring a human screener). In addition, a more general challenge, when using GPT API models, is that it requires a substantial amount of software maintenance to keep up to date with the newest model developments. Therefore, it requires continuous software development, for this screening approach to be viable which, in turn, will probably require collaborations in the research community to ensure the stability of the software over time.

Although this study is not without limitations, we believe that the implications of this work are rather extensive beyond what we have presented and possibly can imagine. First, using well-functioning automated tools renders the possibility for reviewers not to make unnecessary restrictions on their search string to steer the number of study records, which, in turn, increases the likelihood of finding all or close to all relevant studies for the review in the given databases. Moreover, it makes it possible to screen literature for extreme-sized reviews (Shemilt et al., 2014, 2016) that would otherwise have been considered unmanageable and/or unremunerative for humans. Second, this approach can be all-important in elevating the quality of reviews conducted by single researchers restricted by resources such as limited budgets and/or time. Third, we believe that a huge potential exists in combining traditional automated tools and GPT modeling. For example, GPT API models could play a key role in validating a decided stopping rule (Campos et al., 2023; König et al., 2023) whereto it could partly be used to screen records close to the stopping rule on the wrong side, and partly be used to more precisely detect relevant studies on the right side of a given stopping rule, thereby reducing the risk of relevant studies being overlooked. Combining traditional tools and GPT screening could furthermore reduce the cost of using GPT API models since it reduces the number of titles and abstracts needed to be screened by the GPT API models. Another application could also be that GPT API models are used together with prioritization resampling algorithms such as the one suggested by Hou and Tipton (2024), thereby making it possible to reach recall values closer to 1, which is generally considered unattainable when using stochastic algorithms. Fourth, even if reviewers prefer to use duplicate human screening, we think that using a GPT API model as a third screener would be valuable since it can guard against missing relevant studies due to human screener drifting.

To recapitulate, we believe that using GPT API models can change duplicate TAB screening in high-quality reviews across all kinds of scientific disciplines. In fact, we envision that the GPT-4 models will perform even more adequately when used on more structured abstracts as typically found in medicine. We think TAB screening is an ideal use case where artificial intelligence (AI) can meaningfully take on rigid human labor, and where no legal issues arise. Even more edifying, GPT API model screening can ensure a more rapid transfer of usable knowledge to research, practice, and policy, which ultimately underpins the core rationalefor doing systematic reviews.

**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. Also thanks to Terri Pigott for valuable discussions.

**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](https://bit.ly/3spivoG" \t "_blank):

**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.

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.

Brunton, J., Stansfield, C., Caird, J., & Thomas, J. (2017). Finding relevant studies. In D. Gough, S. Oliver, & J. Thomas (Eds.), *An introduction to systematic reviews* (2nd ed., pp. 93–122). Sage.

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

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/

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/

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.

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

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.

Vembye, M. H. (2024). *AIscreenR: AI screening tools for systematic reviews.* (R package version 0.0.1). https://mikkelvembye.github.io/AIscreenR/

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

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.

1. To alleviate 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, this framework has not yet been thouroughly enough tested to know if the SAFE procedure allows reviewers to detect all relevant studies using machine learning algoritms included in screening softwares, such as ASReview (Van De Schoot et al., 2021). [↑](#footnote-ref-1)
2. Note, that they used the gpt-3.5-turbo-0301 model which has been deprecated. [↑](#footnote-ref-2)
3. It is uncertain what exact model the authors used. We expect it was the gpt-4-0613 API model. [↑](#footnote-ref-3)
4. Syriani et al. (2023), Guo et al. (2024), and Gargari et al. (2024) did all use Python to access the GPT API models. While the former two did not share replication materials, Gargari et al. (2024) shared their codes, thereby allowing others to replicate their workflow (though, requiring rather advanced Python coding skills). [↑](#footnote-ref-4)
5. 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). [↑](#footnote-ref-5)
6. All replication materials behind this experiments can be accessed at <https://osf.io/apdfw/>. [↑](#footnote-ref-6)
7. We conducted this experiment on the 4th of November 2023. This was before the corresponding protocol where published on the 15th of December 2023. [↑](#footnote-ref-7)
8. If not provided by the user, study IDs are automatically generated when using the AIscreenR. [↑](#footnote-ref-8)
9. Find the exact functions here: [bit.ly/3Vl0SRp](https://bit.ly/3Vl0SRp" \t "_blank) [↑](#footnote-ref-9)
10. Moreover, we show how this can be done in one of the accompanying vignettes to the AIscreenR package (Vembye, 2024). [↑](#footnote-ref-10)
11. To support this type of screening, we show how this can be practically executed in one of the accompanied vignettes to the AIscreenR (Vembye, 2024). [↑](#footnote-ref-11)
12. For now, it might e.g. be beneficial for researchers to investigate the performance of GPT-4o or GPT-4-turbo since these models are significantly cheaper than the GPT-4 model we used. [↑](#footnote-ref-12)