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

Mikkel Helding Vembye[1], Julian Christensen[1], Anja Bondebjerg Mølgaard[1], & Frederikke Lykke Witthöft Schytt[1]

[1] VIVE - The Danish Center for Social Science Research

Author Note

Mikkel Helding Vembye, https://orcid.org/0000-0001-9071-0724, Department of Quantitative Methods, VIVE. Julian Christensen, https://orcid.org/0000-0002-4596-6998, Department of Governance and Management, VIVE. Anja Bondebjerg Mølgaard, https://orcid.org/0000-0002-2825-4921, Department of Quantitative Methods, VIVE. Frederikke Lykke Witthöft Schytt, Department of Quantitative Methods, VIVE. This research was funded by VIVE Campbell. All authors declare no conflict of interest.

Supplementary materials for this article are available at OSF https://osf.io/apdfw/ and GitHub https://github.com/MikkelVembye/screen_benchmarks. The AIscreenR R package presented in the article can be assessed at https://mikkelvembye.github.io/AIscreenR/. A preprint of this article is available at https://osf.io/preprints/osf/yrhzm.

We would like to thank Jens Dietrichson, Trine Filges, Terri Pigott, 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.

Correspondence regarding this article should be addressed to Mikkel H. Vembye, Soeren Frichs Vej 36 G, 8230 Aabyhoej, Denmark. E-mail: mihv@vive.dk

Abstract

Independent human double screening of titles and abstracts is a critical step to ensure the quality of systematic reviews and meta-analyses herein. However, double screening is a resource-demanding procedure that decelerates the review process. To alleviate this issue, we evaluated the use of OpenAI's GPT API models as an alternative to human *second* screeners of titles and abstracts. We did so by developing a new benchmark scheme for interpreting the performances of automated screening tools against common human screening performances in high-quality systematic reviews and conducting three large-scale experiments on three psychological systematic reviews with different levels of complexity. Across all experiments, we show that the GPT API models can perform on par with and in some cases 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 it can be a valuable tool to use for screening 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. Moreover, we present the R package AIscreenR to standardize and scale up the suggested application. Our aim is ultimately to make GPT API models acceptable as independent *second* screeners within high-quality systematic reviews, such as the ones published in *Psychological Bulletin*.

Keywords: title and abstract screening, OpenAI's GPT API models, systematic review, screening benchmarks, Large Language Models (LLM)

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

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 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).

Independent human double-screening of all identified titles and abstracts is, however, a resource-demanding procedure, often requiring several months of skilled, full-time human labor (Campos et al., 2024; Hou & Tipton, 2024; Shemilt et al., 2016). Consequently, many reviewers refrain from using duplicate screening methods (see Pacheco et al., 2023 for examples in medicine), for instance, due to low budgets or narrow time limits. Alternatively, reviewers narrow their searches

to keep the number of records down to a manageable size, which again increases the risk of overlooking relevant studies (Van De Schoot et al., 2021). Over time these issues will only grow in size 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 based on text-mining or machine-learning algorithms to act either as a second screener, a coarse-grained classifier, or to sort citation records in a prioritized order (Cohen et al., 2006; Gartlehner et al., 2019; O'Mara-Eves et al., 2015; Van De Schoot et al., 2021). These tools can make substantial reductions in human screening workloads in systematic reviews (König et al., 2023; O'Mara-Eves et al., 2015; Perlman-Arrow et al., 2023). However, most evaluations of traditional automated screening tools 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 alleviate this issue and 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 title and abstract (henceforth TAB) screening. 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.[1] 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 TAB screening of medical research literature.[2] 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) and *specificity* (i.e., the proportion of irrelevant records being correctly classified as irrelevant) values of .76 and .91, respectively. Based on these results, the authors concluded that GPT-4 was effective in excluding irrelevant studies but less reliable in identifying relevant ones, and therefore recommended using it as a support tool but not as a full replacement of the second human screener. 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.[3]

On a related line of research, Alshami et al. (2023), Khraisha 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 how big

[1] Note that they used the gpt-3.5-turbo-0301 model which has been deprecated and is not longer available at the server.
[2] It is uncertain what exact model the authors used. We expect it was the gpt-4-0613 API model.
[3] 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).

error rates can be accepted when working with automated tools and how these tools' performances compare to typical human screening performances in high-quality systematic reviews. It is further unclear if screening performances can be improved by drawing on newly developed API features such as function calling and fine-tuning (OpenAI, 2024c, 2024b). In addition, it is 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 to 1) build a generic benchmark scheme to improve assessments about what constitutes acceptable screening behaviors of automated tools in high-quality systematic reviews, 2) test and validate the TAB screening performance of GPT API models using novel model features, 3) develop a heuristic workflow and guidelines for how and when to conduct TAB screening with GPT API models, and 4) present the R package AIscreenR (Vembye, 2024).

The remainder of the paper proceeds as follows. In the next section, we investigate how well a GPT model needs to perform to be accepted as an independent second screener of titles and abstracts in high-quality systematic reviews. To answer this question, we analyze human screening performances across 22 high-quality systematic reviews and use this investigation as the basis for developing a novel, empirically informed benchmark scheme for interpreting acceptable and unacceptable screening performances in high-quality systematic reviews. In the following section, we present three classification experiments to evaluate the screening performances of GPT API models relative to human screeners. This includes presentations of the prompt engineering and data underlying these experiments as well as the results of the experiments. In the subsequent section, we deduce tentative guidelines for when we consider it acceptable and unacceptable to use GPT API models as independent second screeners. In this section, we also elaborate on how we think reliable prompts

can be developed in future reviews and present a standardized workflow for how to incorporate GPT screening in high-quality reviews. In the final sections, 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.

What are acceptable human error rates 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 systematic reviews (O'Connor et al., 2019). To allow for assessments of this, in this section, we develop a novel, empirically informed benchmark scheme for interpreting acceptable and unacceptable TAB screening performance in high-quality systematic 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 that are used as the basis for our benchmark scheme.

Transparency and openness

All statistical data analyses were conducted using R 4.4.0 (R Core Team, 2022) in RStudio (RStudio Team, 2015). For the main analyses behind the benchmark scheme, we used the `metafor` package, version 4.6.0 (Viechtbauer, 2010), including the sandwich estimators herein (Pustejovsky, 2020). RIS file data was handled by using the `revtools` package, version 0.4.1 (Westgate, 2019), and we used the `ggplot2`, version 3.5.1 (Wickham, 2016) for visualization. Code and data for replicating the investigation behind the benchmark scheme are available at https://github.com/MikkelVembye/screen_benchmarks. All replication materials behind the experiments presented in later sections can be accessed at https://osf.io/apdfw/. This study has not been preregistered.

Evaluation metrics

In the existing literature, a wide range of different metrics has been used to evaluate TAB screening performances in the context of systematic 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, meaning that they are not sensitive to the data containing a large difference in the proportion of relevant and irrelevant references, which is commonly found in systematic reviews (Brunton et al., 2017). To be formal, the recall can be written as

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

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

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

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

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). By contrast, a low specificity does not induce bias, it just means that reviewers have to re-examine the relevance 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, .5, this still implies that reviewers can confidently exclude 50% of the irrelevant records, which in most cases would be a significant reduction in the screening workload. Therefore, we think tools should be accepted when high recalls can be reached, to a large extent

independently of the accompanied specificity value. These scenarios are depicted in the Figures 1A and 1B. We will come back to this 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 common issue with such summary metrics is that they tend to be sensitive to imbalances in the data. For instance, 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 *the balanced accuracy* ($bAcc$) metric, which accounts for imbalanced data. Thus, the balanced accuracy metric balances the accuracy of the performance across the recall and specificity metrics and is simply an average of those metrics, formally given by

$$bAcc = \frac{Recall + Specificity}{2} \quad (3)$$

We could have calculated other relevant metrics as well, such as the normalized Matthew correlation coefficient (Chicco & Jurman, 2023). However, we will mainly focus on the metrics presented in Equations 1-3 since these are most intuitive in their interpretation. Yet, we have shared all data behind all of our analyses, thereby allowing readers to add further estimations if necessary.

Typical human screening performances in high-quality systematic reviews

In order to make valid comparisons between human and automated screening performances, we consider it important to have an impression of the human screening performance typically accepted in high-standard systematic reviews (O'Connor et al., 2019). We consider this comparison a reliable way to assess whether a given recall is good or bad. If, for example, humans on average tend to miss 20%-25% of all relevant studies during the title and abstract screening phase, as some

previous research has suggested (Buscemi et al., 2006; Waffenschmidt et al., 2019),[4] then it might be misleading to infer that a GPT model with a recall of .75 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. Automated screening tools will probably 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. To assess this difference, the next section presents a tentative benchmark scheme for interpreting acceptable and unacceptable screening performances in high-standard systematic reviews.

Data used for benchmarking

As the empirical basis of our benchmark scheme, we analyzed human screening performances in 22 high-standard systematic reviews that used independent duplicate human screening. This included 17 Campbell Systematic Reviews and five reviews conducted by The Norwegian Institute of Public Health (NIPH). An overview of all the included reviews can be found in Table 1, including information on the imbalance in the given dataset. The included Campbell Systematic Reviews represent all Campbell Systematic Reviews that have been conducted by VIVE (the Danish Center for Social Science Research) in which independent duplicate human screening has been used and tracked. This data includes 144,003 title and abstract records. A total of 46 individual screeners participated in screening the studies, of which 36 were student assistants and/or non-content experts, and 10 were researchers/authors of the given review. The Campbell Systematic Reviews were either conducted from 2015 to 2024 or represent ongoing reviews.[5]

[4] In medicine, the number of missed studies may be even higher, especially when relying on student screeners (Ng et al., 2014).

[5] In four of the Campbell Systematic Reviews the first level of title and abstract screening has been conducted but the final review has not yet been published. We, therefore, refer to the protocol of these four reviews in Table 1.

Since all of the included Campbell Systematic Reviews drew on assistant (i.e., non-content-expert) screeners, this could potentially downward bias the performance metrics for various reasons. For instance, assistant screeners have limited content expertise regarding the topic(s) under review, which might potentially lower their performance. In effect, their performances might not necessarily be on par with the typical screening performance of content expert screeners. Hence, we analyzed the Campbell review data separately for assistant/non-expert and researcher/expert screeners.

Moreover, 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 were. We, therefore, added 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 artificially inflating human disagreements.

Statistical analysis of the human screening performances

We estimated all the performance metrics via Equations (1) to (3). The TP, TN, FP, and FN 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 within in the given review.

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 all metrics.[6] 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.

To derive the overall average performances in terms of *recall*, *specificity,* and *balanced accuracy* 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 being nested within studies. At the same time, the models account for the correlation between the within-study performance estimates. The sample correlation, ρ, is often entirely or partially unknown and must be imputed. In all the used working models, we assumed $\rho = .7$.

To guard against model misspecification both models have incorporated robust variance estimators. The main difference between the two models is that 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).

[6] We did not use double arcsine transformation (Doi & Xu, 2021) due to the inadequate properties of the back transformation of this measure (Röver & Friede, 2022; Schwarzer et al., 2019).

For both models, we estimated two sources of heterogeneity: the variability of the true screener performances within (ω) and between studies (τ). This allowed us to investigate at what (if any) level the largest true difference between the human screener performances existed.

Typical human screening performance results

All individual screening performances across the included reviews and their distribution around the overall performance means are illustrated in Figures 2 and 3. In this section, we primarily comment on the results for the recall and specificity measures, since these are the main metrics we use to derive and develop the benchmark scheme. All results can be found in the background material.

Across the included Campbell Systematic Reviews, we found the overall average *recall* value for the assistant and author screeners to be .782, 95% *CI*[.747, .817] and .881, 95% *CI*[.823, .931], respectively. Hereto, we found the two groups' average recall values to be statistically different from each other with $F(1, 10.3) = 14.58$, $p = .003$. We detected minor substantial variations between the performance measures within studies with $\omega = 0.026$ and $\omega = 0.035$ for the assistant and author screeners, respectively. We were unable to detect any true differences in performance levels between studies, indicating consistent average screening performances across the Campbell reviews, both for assistants and expert screeners. The overall average *specificity* for assistant screeners was .980, 95% *CI*[.966, .990], and for review authors .988, 95% *CI*[.980, .995]. We found no statistically significant difference between the two average estimates with $F(1, 13.6) = 2.08$, $p = 0.172$. We did only find very minor non-substantial variation within and between studies with $\omega = 0.004$ as the maximum for author screeners. Moving on to the *balanced accuracy* metric, we found average performance levels of .874, 95% *CI*[.857, .890] among assistant screeners and .933, 95% *CI*[.899, .961] among author screeners. We found the difference between the group means to be statistically significant with $F(1, 10.1) = 18.22$, $p = .002$.

From these results, it appeared that compared to students, researcher (i.e., content expert) screeners are substantially better at detecting relevant studies. Yet, as previously noted, this difference may be driven by factors other than mere screening quality (e.g., authority relations with the potential to inflate the researchers' performance relative to that of the assistants). Interestingly, when analyzing the NIPH data, which was in all cases based on independent researcher-researcher screening comparisons, we found performance levels closer to those of the assistant screeners in the Campbell Systematic Reviews. The average recall value in the NIPH data was .839, 95% *CI*[.737, .920]. Again, we only found minor true variation between the screener recall performances within studies with $\omega = 0.029$ and $\tau = .0$. The overall average specificity value of the NIPH researchers was .977, 95% *CI*[.955, .992], with only minor non-substantial true variation between screeners and between studies. The average balanced accuracy value of the NIPH researchers was .905, 95% *CI*[.859, .943].

Benchmark scheme

Bearing on the empirical results presented in the previous section, we developed the screening benchmark scheme presented in Table 2. On this basis, we suggest—as a coarse-grained rule of thumb—that if an automated tool can reach a recall rate of at least .75 and specificity rates above .95, they can be said to resemble common *second*-screener performances in the context of high-quality systematic reviews. Consequently, and in contrast with previous evaluations (Guo et al., 2024), we would not necessarily interpret a recall value of .76 to be too low for a GPT API model to function as an independent second screener.

As can be seen from the benchmark scheme in Table 2, we do not necessarily conceive a specificity of 1 (i.e., 100%) to be ideal, since we would rather have our automated screenings 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 .80 is acceptable as long as the recall value is equal to or above .75 (c.f. Figure 1) as well. We, therefore, suggest that automated screening performances reaching recall of at least .75 and specificity above .80 should be accepted as independent screeners in high-quality systematic reviews. Also, 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 .80. This would especially be relevant when working with very large amounts of title and abstract records (see an example of this in Shemilt et al., 2014).

Finally, we believe that automated tools can also be useful under less restrictive conditions. We even believe that recall values between, say, .5-.75 should not 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 quality, lowering the risk of humans overlooking any relevant records.

Yet, to be clear, we generally think that it will not be viable to use automated tools with performance levels (recall and specificity values) below .5. However, in Table 2, we use graduated shades of red, where the light red color for specificity values below .5 indicates that we cannot entirely reject that there might be cases where automated screenings could be useful even with specificity values below .5 (as long as the recall is high). For example, in extreme-sized reviews, even a 30% workload reduction might save multiple days of human labor. The number of title and abstract records does, however, need to be very high for this approach to be viable.

With the benchmark scheme presented in Table 2, 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 and binary 'for and against' discussions. Furthermore, we will use this benchmark scheme for interpreting our conducted classifier experiment that we present in the next section.

Classification experiments: How do GPT API models perform in light of the benchmark scheme?

In this section, we present the data and prompts used as well as the results for three large-scale classifier experiments that we have conducted to evaluate the screening performance of OpenAI's GPT API models. Differently from previous research, these classifier experiments aimed to test the performance of GPT API models 1) when applied in psychological reviews, 2) when using function calling (OpenAI, 2024c), that is using JSON functions ensuring structured responses, and 3) when using multi-prompt screening in complex review settings. We set up the experiment so that each of the three experiments represented different levels of complexity in terms of inclusion criteria and contained different challenges to overcome when conducting TAB screening using GPT API models. The main purpose of the experiments 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. We return to this issue in a later section. A side-effect of conducting these experiments was further to ensure the quality of the AIscreenR package (Vembye, 2024) for automated TAB screenings using GPT API models. We considered this test to be an important step in ensuring a scalable screening approach.

Data used for classifier experiments

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 FFT intervention is well-defined. 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 be unlikely to do so in the context of more complex reviews. This experiment 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.

A potential weakness in Experiment 1 is that it is assumed that the OpenAI's GPT models have been trained on publicly available text data from the internet, available from 2021 and back, which could in theory have inflated the screening performances in Experiment 1 caused by the GPT models being trained on this particular open-access review and its protocol. If this was the case, it may reduce the generalizability of the experiment's results to other review cases where the models have not been trained on any relevant data. 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 regarding the effects of the FRIENDS preventive programme on anxiety symptoms in children and adolescents conducted by Filges, et al. (2023).[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

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

only 64 of 2572 records being relevant, amounting to an approximate inclusion ratio of 25 relevant studies per 1000 records.

As noted, Experiments 1 and 2 can both be said to involve rather simple TAB screening tasks and without further investigation, it is unclear to what extent their results can be generalized to more complex review settings with more complex inclusion and exclusion criteria as is often the case in psychological reviews. A challenge in making GPT screening work in complex review settings is that it likewise requires reviewers to make a broader and more complex prompt. Hereto, Gargari et al. (2024) suggested that long and broad prompts may not perform well in terms of finding relevant studies. To address this challenge, we conducted a third experiment. In this experiment, we introduced and tested the performance of multi-prompt screening (i.e., sometimes referred to as prompt chaining), that is making one concise prompt per inclusion/exclusion criteria in a review (instead of adding all inclusion and exclusion criteria to the same prompt), to test if this screening approach would yield better performance in a complex review setting.

To emulate a complex psychological review setting, we used screening data from an ongoing Campbell Systematic Review of the effects of testing frequencies on students' academic achievement (Thomsen et al., 2022), which has been a common review topic in psychology (c.f. Yang et al., 2021). Compared to the screenings in Experiments 1 and 2, we considered Experiment 3 a more difficult screening case, since the inclusion criteria of this review were more complex, including concepts that are not well-defined. As such, the intervention (student testing) is a type of learning strategy that is ubiquitous in education and psychology and is used in a variety of ways for many different purposes.[8] 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,

[8] 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.

but may also be referred to, e.g., as quizzes, progress-monitoring, curriculum-based measures, and retrieval practice). Assessing the eligibility of particular studies therefore requires a great deal of subject matter familiarity, and contrary to Experiments 1 and 2, we think that if GPT API models can achieve satisfactory performances in this context, they will likely be able to do so in most review contexts. In this experiment, the data consisted 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.

Across 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.

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.

Next, when given study IDs,[9] titles, and abstracts, the AIscreenR automatically integrates this information in the prompt, using the text in Textbox 2.

By pasting the prompt together with each title and abstract, we aim to guard against model drifting/hallucinations. Different from prior evaluation studies, we did not add any instruction regarding how the model should respond to our request in the main prompt. Instead, we relied on

[9] If not provided by the user, study IDs are automatically generated when using the AIscreenR.

function calling and built two JSON functions (one function call yielding simple trinary results and another yielding descriptive screening responses) with instructions on how the models should respond to our requests.[10] According to OpenAI, this should result in more reliable and standardized responses from the models (OpenAI, 2024c). The main JSON respond function we built included the instructions presented in Textbox 3.

For Experiment 3, involving six inclusion criteria, we tested two different prompt engineering strategies: *multi-prompt/prompt-chaining* and *single-prompt* screening. In the former approach, six short prompts were made, each containing one inclusion criterion only. In the latter, we added all inclusion criteria to a single prompt. This allowed us to test how GPT performances are impacted by different prompt strategies, and particularly whether multi-prompt screening can make GPT screenings viable in settings where the usefulness of GPT screenings has been questioned in previous research (Gargari et al., 2024).

All engineered prompts used for Experiment 3 were initiated by a short introduction to the review followed by a description of the given inclusion criterion. As with the prompts used in Experiments 1 and 2, the prompts used in Experiment 3 were all pasted together with the text present in Textbox 2 and drew the function call from Textbox 3. The specific prompts can be found in the Supplementary Material Textbox S1 and S2. Moreover, we elaborate further on multi-prompt screening below.

Performance testing

Before initiating each classification experiment, we tested and refined our prompts on a subset 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 GPT models consistently included this

[10] The exact wording of each function can be found at bit.ly/3Vl0SRp

study record. Then, we scaled up the test to include 150 irrelevant and 50 relevant records. When the test yielded satisfactory results, demonstrating an ability to reach a recall above .75 and a specificity above .9, we moved on to screening all records using the GPT API models to investigate if the models' test performances persisted when used on 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 API models to yield satisfactory screening performances. It should be noted that in our multi-prompt screening of title and abstract records for the TF review, we found that the best screening performances yielded by coding studies as relevant if they were included in at least five of the six prompts (instead of requiring them to be included by all six prompts). Therefore, we applied this threshold in the full screening. We allowed for this flexible inclusion threshold to further account for the uncertainty in the models' decision when limited information appears in an abstract.

Evaluation design

In all three classifier experiments, we evaluated the performance of the GPT API models by using Equations (1) to (3). In this respect, the TP, TN, FN, and FP conditions were determined by comparing the GPT decision with the final decision made by agreement between at least two independent human screeners. Human inclusion at this first level of screening did not necessarily imply that study records were relevant for the final review—merely that they were considered to be relevant for full-text screening. In Experiments 1 and 2, we used the gpt-3.5-turbo-0613 and gpt-4-0613 models, reached from the 'v1/chat/completions/' endpoint. Since the GPT-3.5 models are generally less consistent in their responses relative to GPT-4, we repeated the same screening request 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[11] of using these models, we only conducted one screening per title and abstract when calling GPT-4. We will present the result for the GPT-3.5 model but our main focus is on the performance of GPT-4 since the GPT-3.5 model has been deprecated[12] and expired September 13, 2024.

For Experiment 3, which involved multi-prompt screening, we only drew on GPT-4, and the final inclusion decision was then based on the probability of inclusion across all used prompts.

For all experiments, we used invariant *top_p* and *temperature* values, using the default value of 1 for both hyperparameters.

Results of the classification experiments

All results for the three classifier experiments are presented in Table 3. As can be seen from Table 3, the GPT-4 model yielded recall and specificity values of .899 and .933 in Experiment 1, which, held against the benchmark scheme presented in Table 2, can be considered to be on par with typical researcher screener performances in high-quality systematic reviews. The GPT-3.5 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 3's recall column). When using an inclusion probability threshold of .2 (meaning that a study was coded as relevant if the GPT-3.5 model included it in two or more of the ten screenings), the GPT-

[11] Note: After we conducted the experiment OpenAI has presented the GPT-4o-mini which is cheaper than the GPT-3.5 models that we used. All of OpenAI's new GPT API models are configured differently than the GPT-3.5 and GPT-4 model, we draw upon. Therefore, we have not evaluated these models yet since they need to be coded anew.

[12] Deprecation "refer[s] to the process of retiring a model or endpoint. When we announce that a model or endpoint is being deprecated, it immediately becomes deprecated. All deprecated models and endpoints will also have a shut down date. At the time of the shut down, the model or endpoint will no longer be accessible." (OpenAI, 2024a)

3.5 model yielded a recall of .81 and a specificity of .94. However, when using an inclusion probability threshold of .5, the performance became unacceptably low compared to human screening, with a recall of only .69. A full overview of the impact of the inclusion probability threshold on the performance metrics for Experiment 1 can be found in Supplementary Material Figure S1A.

When used on the FRIENDS data, the GPT-4 model yielded performances exceeding common human screening performances. Specifically, it yielded a recall of .98 (only missing one relevant study) and a specificity value of .97. When using an inclusion probability threshold of .7, the GPT-3.5 model also performed well, with a recall of .953 and specificity of .899. Yet, again, the screening performance of GPT-3.5 hinged on the chosen inclusion probability threshold. A full overview of the impact of the inclusion probability threshold on the performance metrics for Experiment 2 can be found in Supplementary Material Figure S1B.

Finally, when used on the TF data (Experiment 3), the GPT-4 model yielded a recall of .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. the benchmark scheme in Table 2) and did exceed three out of six human recalls within this review (see Thomsen et al. (2022) under column 1 in Figure 2). Moreover, the model yielded an acceptable level of specificity of ~.84 when based on multi-prompt screening. Relative to human performances, the model in this case can be said to be rather over-inclusive. As such, 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.

When using a single prompt to screen the TF data, we found a human-like recall of .9. Compared to human performances, the single-prompt screening was rather overinclusive with a specificity of .743. This performance resembled the results of the multi-prompt screening when using a

threshold where title and abstract records were coded as relevant if included by the GPT model in at least four of the six prompts.

As can be seen in Table 3, when using an even more inclusive threshold, coding studies as relevant if included by the GPT model in at least three of the six, the GPT-based reached a recall of .95, but with a low specificity of .67, leaving a rather high number of title and abstract records to be double-checked by human screeners. Yet, if this approach was used it could still reliably reduce the total screening workload.

To summarise, we find that GPT API models can work as highly reliable and independent second screeners with recall performances on par with 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) suggesting 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 used different prompting strategies. A part of our comparably high performance might also be caused by the fact we instructed the models to include title and abstract records with very little information (cf. Textbox 3). 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. 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 at the initial stage of the screening.

We found that in some applications, the specificity rate reached by the GPT-4 API model can be seen to be on the lower end compared with human screeners. Yet, we do not find this to be a major issue when having a high recall rate (cf. Figure 1) since this can just be seen as an extra opportunity to double-check close-to-relevant studies, thus enhancing the chance of not overlooking any relevant study records.

Based on our results, we cannot firmly conclude that multi-prompt screening is significantly better than single-prompt screening in complex review settings. Yet, it is a more flexible approach that can reduce the over-inclusiveness of GPT models, while still yielding sufficient recalls on par with typical human second screeners. Although we cannot reject that single-prompt screening might be viable in complex review settings with many inclusion criteria, we think multi-prompt screening is more appropriate to use in complex review settings. When using multi-prompt screening, all titles and abstracts will be mapped on the exact reasons for exclusion, increasing the transparency of the review and making it easier to decode what factors made the GPT model work or not. Furthermore, as we discuss in the next section, the use of multi-prompt screening can have additional advances in the prompt development and testing phases, making it a useful tool adding to the screening toolbox.

Overall, we think there is a huge potential for GPT API models to be used for TAB screening tasks in high-quality systematic reviews—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 a reliable use of LLMs in systematic reviews. In the next section, we therefore develop a tentative workflow and guidelines for how such screenings practically can be set up in a standardized manner.

Tentative workflow and guidelines

Premised on our developed benchmark scheme, our experience, and the results of the three classifier experiments, we have developed the following tentative workflow and guidelines for when and how GPT API models can be used as independent second screeners of titles and abstracts. All steps in this process are presented in Table 4.

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) until it is ensured that the screening performances pass certain thresholds within the given review context. The first step of the testing procedure involves locating approximately 10 relevant and 150 irrelevant titles and abstracts, respectively. Locating more than 10 relevant study records might be ideal to test if the prompt(s) and 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. Thus, we cannot set this step in stone. When locating irrelevant records, we suggest randomly sampling those from the total pool of records, thereby increasing the chances that the specificity test value can be generalized to the full sample of study records.

After having collected the testing 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[13] and should therefore in general only be fed with a minimum of information.

When conducting complex reviews, prompt engineering gets more complicated. Specifically, we experienced that it can be rather difficult to decipher what exact text part(s) of a prompt makes it yield insufficient performances when adding multiple inclusion criteria to a single prompt. To overcome this issue, we suggest that one can draw on the multi-prompt screening strategy, as evaluated in the previous section, where each inclusion criterion is prompted individually. All title and abstract records are then screened with all prompts. It should, thereby, be easier to evaluate what exact inclusion criterion questions and sentences yield (in)sufficient performances. If reviewers opt to use multi-prompt screening for the full screening, the inclusion probability threshold should be tested and decided at this point, meaning that it should be decided how many of the multiple prompts a title and abstract need to be included in to be considered relevant.

[13] A new feature has been developed på OpenAI that allows user to fine-tune/train model to do specific task (OpenAI, 2024b). Therefore, future application might potentially involve model traning as well.

When engineering prompts, we suggest that these should be refined until reaching a recall of .75 and a specificity of at least .8 (cf. the benchmark scheme in Table 2). Lower specificity values may be accepted as long as the recall exceeds .75. However, if a specificity value of .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 human screeners. We suggest that if a recall of .75 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 the 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.

When the test threshold has been passed, and the reviewers have decided to leverage the GPT API model as a second screener, we suggest that the human reviewers screen all study records before initiating the automated screening. This prevents human reviewers from being impacted by GPT's decisions. An alternative to screening all records at once is to repeat steps 6 to 9 in Table 4 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 been screened by both the human and automated screener, 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.

When not to use GPT API models for TAB screening?

Although we think that GPT API models have the potential to revolutionize 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 references 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., Experiment 2). 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 5 visualizes the conditions under which we consider it adequate vs. inadequate to use GPT API models for TAB screening tasks in systematic reviews.

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 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 represent black-box and closed-source algorithms. While we have demonstrated that current GPT API models—if configured adequately—are capable of doing TAB screening tasks, 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 and GPT-4-turbo models, and, more so, to other models such as the API models from Claude (ANTHROPIC, 2024) or Mistral AI (2024). From a scientific point of view, and to increase the transparency of GPT API screening, it is, therefore, important that future research revolves around investigating the performances of local, open-source, and downloadable models.

That said, we do think it is important to note that most human duplicate screenings also represent black box operations that are hardly replicable, and we believe that the GPT API models should be judged in light of this. This is not to say that we should not strive to make screenings replicable and reproducible since this would increase the transparency of high-quality reviews. Yet, we just do not think that the black box argument should be a major reason for abandoning GPT API

models for TAB screening in high-quality reviews. Moreover, model sensitivity is the exact reason why we have developed the benchmark scheme so that if the GPT models eventually appear to underperform, this scheme serves as a means to ensure that biased screenings do not enter high-quality reviews.

On a similar, but technical line, it is rather 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 that we used have been deprecated (but can still be used) and moved to the tools argument in the request body. Therefore, future research must evaluate whether our results can be achieved 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 has expired so one cannot replicate the screenings we 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. Although prices have already dropped,[14] we recommend using the models carefully. In some applications, it might be advantageous to combine traditional classifier

[14] For now, it might e.g. be beneficial for researchers to investigate the performance of GPT-4o-mini, GPT-4o, or GPT-4-turbo since these models are significantly cheaper than the GPT-4 model we used. Moreover, OpenAI has developed batch APIs, allowing the user to lower the cost by 50% if they can wait up till 24 hours to get answers to its requests.

tools with GPT API models to reduce the total cost. In extreme-size reviews (i.e., > 100,000 references), reviewers could consider combining priority screening/classifier modeling with the GPT API screening. The GPT API screener could, for example, then be used as an extra guardian, checking the performance of one's selected stopping rule (Boetje & van de Schoot, 2024; Campos et al., 2024; König et al., 2023), either by screening a subsample of, say, 1000 references on the wrong side of the set threshold or by randomly sampling 1000 references from the pool of studies considered to be irrelevant. 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. In this respect, 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. the benchmark scheme in Table 2), 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 5).

Although we have strived to build a user-friendly setup for GPT API screening in the AIscreenR package, 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 in the short run is to copy our approach to the ChatGPT web browser interface. 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, future research must be aware of these model deviations, avoiding the performance of different models being mixed up.

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. First of all, it is based on screener performances deduced from a convenience sample of systematic reviews, possibly restricting the generalizability of the estimated average screening performance measures. Even so, we believe that the screening performance measures provide key insights regarding what human screening standards are currently being accepted in high-standard reviews. Although our results indicate that the human screener performances seem to be comparable across distinct disciplines, future research may usefully investigate typical screener performances more systematically and across various research fields within psychology and the social sciences to make even more refined screening guidelines.

Discussion

Independent 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 paper, we evaluated the use of OpenAI's GPT API models to conduct title and abstract screening to reduce human labor in systematic reviews and found that 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, potentially speeding up the review process.

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. 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.

Based on our findings, we believe TAB screening with GPT API models can change the way duplicate title and abstract screening is conducted in high-quality systematic reviews, making it possible to replace human *second* screeners. However, this necessitates a standardized screening approach 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 increase the user-friendliness of our suggested approach, we developed the AIscreenR R package (Vembye, 2024).

With this paper, we have strived to make a foundation on which evidence organizations (such as Cochrane and Campbell Collaboration) and review journals (such as *Psychological Bulletin*) can assess and potentially accept 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 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 can be 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, we developed our work-flow and guidelines to underpin requirements *(d)* and *(e)*. Requirement *(e)* is as such not necessary in our case since we are working with *pre*-trained models.[15] 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. Finally, to fulfill requirement *(f)*, we built the AIscreenR package as open-source software, allowing others (e.g., the Evidence Synthesis Hackathon, Campbell Collaboration, the EPPI-Reviewer team, etc.) to contribute to the development and ongoing support of the software.

Although this study is not without limitations, as mentioned in the previous section, 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

[15] As mentioned in footnote 13, model training can be come a key part of future GPT API screening applications.

(Shemilt et al., 2014, 2016) that would otherwise have been considered unmanageable and/or unremunerative for humans to initiate. Second, this approach can potentially elevate 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., 2024; König et al., 2023) whereto it could be used to screen records close to the stopping rule on the wrong side, 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. 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.

Conclusion

To recapitulate, we believe that using GPT API models can radically change duplicate TAB screening in high-quality reviews across all kinds of scientific disciplines. In fact, we envision that the GPT-4 models can perform even more adequately when used on more structured article 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 rationale for doing systematic reviews.

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. https://doi.org/10.3390/systems11070351

*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

ANTHROPIC. (2024). *Claude 2*. https://claude.ai/new

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. https://doi.org/10.1186/s13643-024-02502-7

*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/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. https://doi.org/10.1002/cl2.1345

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. https://doi.org/10.1057/s41599-021-00903-w

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*. https://doi.org/10.1027/2151-2604/a000509

Buscemi, N., Hartling, L., Vandermeer, B., Tjosvold, L., & Klassen, T. P. (2006). Single data extraction generated more errors than double data extraction in systematic reviews. *Journal of Clinical Epidemiology*, *59*(7), 697–703. https://doi.org/10.1016/j.jclinepi.2005.11.010

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. (2024). Screening smarter, not harder: A comparative analysis of machine learning screening algorithms and heuristic stopping criteria for systematic reviews in educational research. *Educational Psychology Review*, *36*(19). https://doi.org/10.1007/s10648-024-09862-5

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. https://doi.org/10.1186/s13040-023-00322-4

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. https://doi.org/10.1197/jamia.M1929

*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/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/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/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/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/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/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/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/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/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/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/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*. 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*. 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. *Open Science Framework*. https://doi.org/10.31234/osf.io/ybu3w

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/

Mistral. (2024). *La Plateforme*. https://console.mistral.ai/

Ng, L., Pitt, V., Huckvale, K., Clavisi, O., Turner, T., Gruen, R., & Elliott, J. H. (2014). Title and Abstract Screening and Evaluation in Systematic Reviews (TASER): A pilot randomised controlled trial of title and abstract screening by medical students. *Systematic Reviews*, *3*, 1–8. https://doi.org/10.1186/2046-4053-3-121

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. https://doi.org/10.1186/s13643-019-1062-0

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. https://doi.org/10.1186/2046-4053-4-5

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. (2024a). *Deprecations*. https://platform.openai.com/docs/deprecations

OpenAI. (2024b). *Fine-tuning*. https://platform.openai.com/docs/guides/fine-tuning

OpenAI. (2024c). *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. https://doi.org/10.1016/j.jclinepi.2023.01.007

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. https://doi.org/10.1002/jrsm.1636

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/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. https://doi.org/10.1186/s13643-015-0067-6

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/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/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. https://doi.org/10.1186/s13643-016-0315-4

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/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/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

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. https://doi.org/10.1038/s42256-020-00287-7

Vembye, M. H. (2024). *AIscreenR: AI screening tools for systematic reviews.* (GitHub version 0.0.0.9999). 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

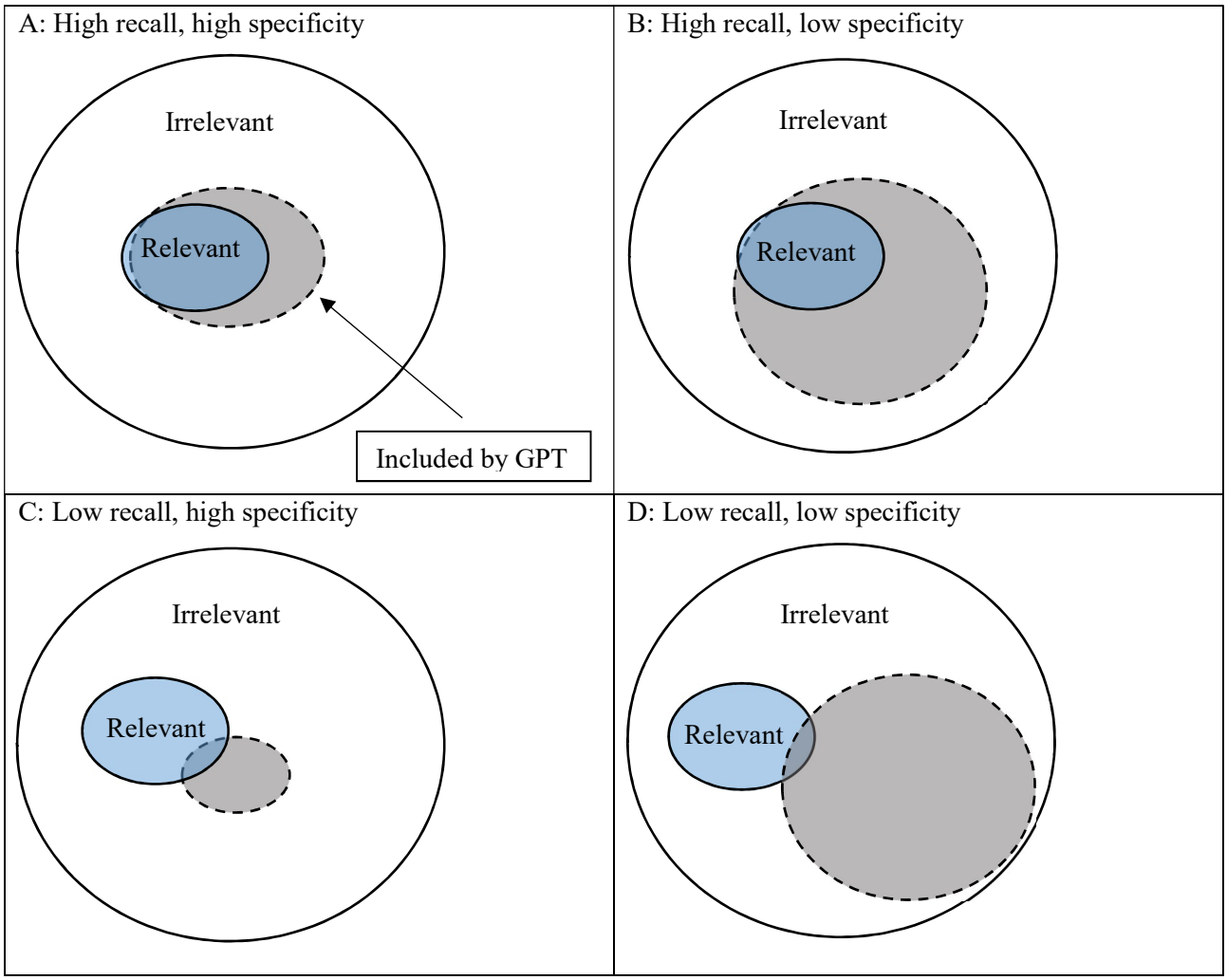
Waffenschmidt, S., Knelangen, M., Sieben, W., Bühn, S., & Pieper, D. (2019). Single screening versus conventional double screening for study selection in systematic reviews: a methodological systematic review. *BMC Medical Research Methodology*, *19*(1), 132. https://doi.org/10.1186/s12874-019-0782-0

Wang, Z., Nayfeh, T., Tetzlaff, J., O'Blenis, P., & Murad, M. H. (2020). Error rates of human reviewers during abstract screening in systematic reviews. *PloS One*, *15*(1), e0227742. https://doi.org/10.1371/journal.pone.0227742

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/10.1002/jrsm.1374

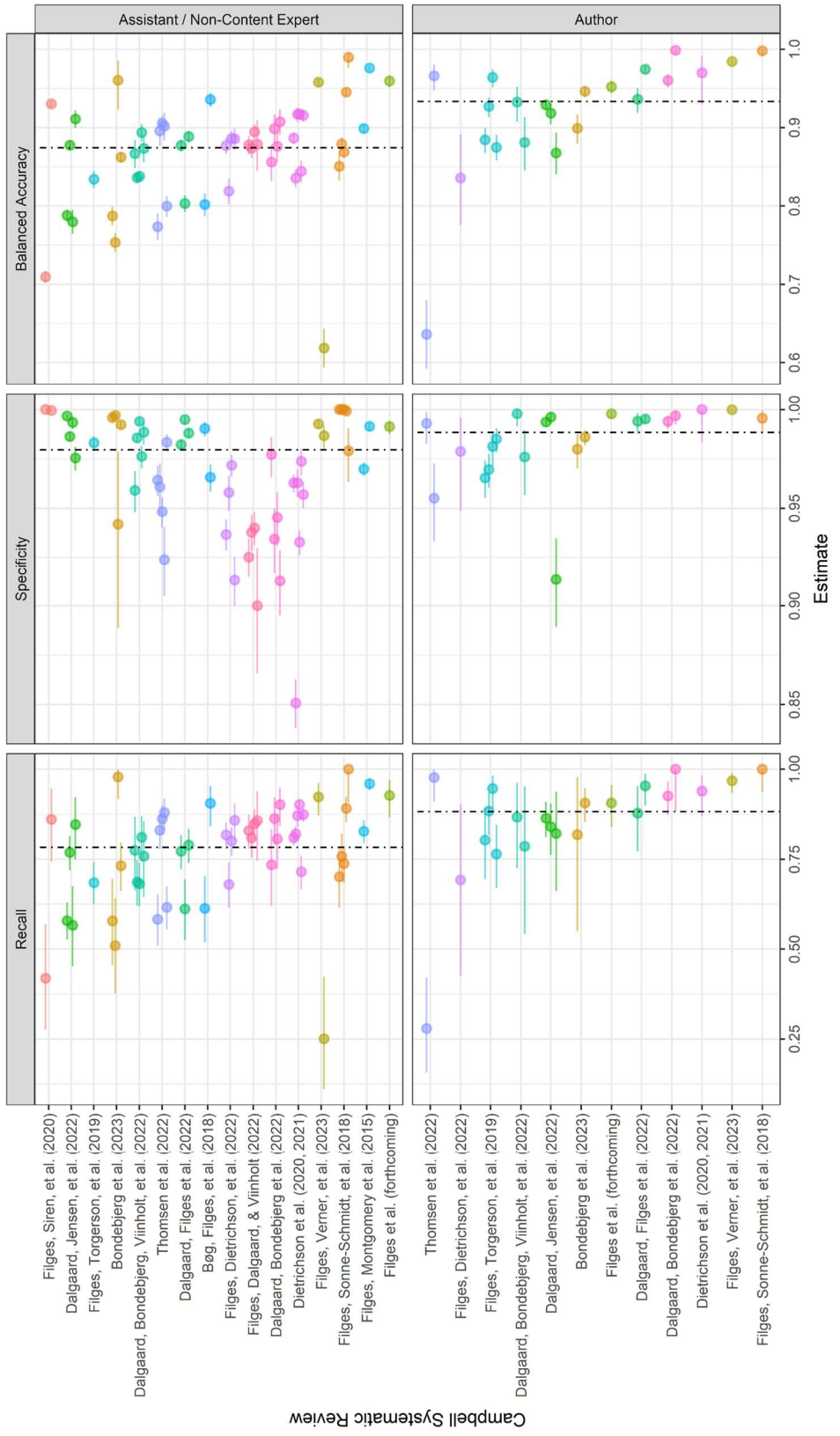
Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. https://cran.r-project.org/web/packages/ggplot2/index.html

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. In *Psychological Bulletin* (Vol. 147, Issue 4, pp. 399–435). American Psychological Association. https://doi.org/10.1037/bul0000309



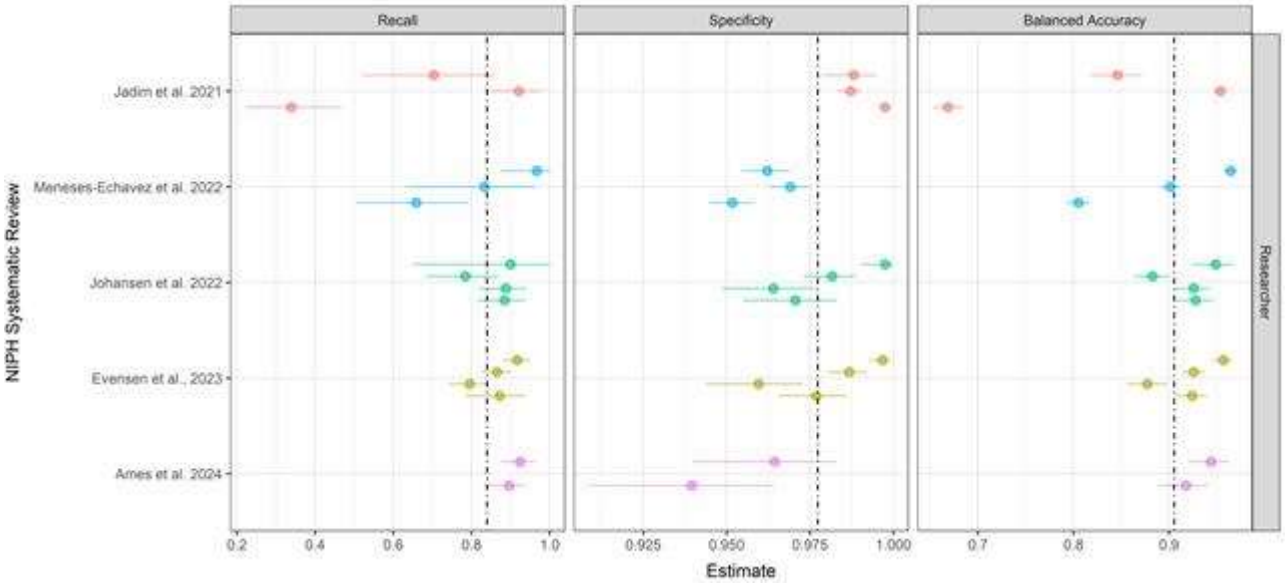
Note: The blue-colored circles indicate the proportion of relevant title and abstract records; the gray-colored circles represent the proportion of records included by the screener; the white circles represent the proportion of irrelevant records that are correctly excluded by the screener.

Figure 1. Recall and specificity performances



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

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.

Source Authors	Short title	$n_{included}/N$	Ass.[a]	Aut.[b]
Campbell review				
Bøg et al. (2018)	Deployment of personnel to military operations	106/2899	2	-
Bondebjerg et al. (2023)	The effects of small class sizes on students' academic achievement, socioemotional development and well-being in special education	244/11860	4	2
Dalgaard, Bondebjerg, Klokker et al. (2022)	Adult/child ratio and group size in early childhood education or care to promote the development of children aged 0–5 years	258/3667	4	2
Dalgaard, Bondebjerg, Viinholt et al. (2022)	The effects of inclusion on academic achievement, socioemotional development, and wellbeing of children with special educational needs	373/14491	5	2
Dalgaard, Filges et al. (2022)	Parenting interventions to support parent/child attachment and psychosocial adjustment in foster and adoptive parents and children	424/13106	3	2
Dalgaard, Jensen et al. (2022)	PROTOCOL: Group-based community interventions to support the social reintegration of marginalised adults with mental illness	557/17614	4	3
Dietrichson et al. (2020, 2021)	Targeted school-based interventions for improving reading and mathematics for students with or at risk of academic difficulties in Grades K-6 [plus 7-12]	2952/15273	6	1
Filges, Dalgaard et al. (2022)	Outreach programs to improve life circumstances and prevent further adverse developmental trajectories of at-risk youth in OECD countries	387/4890	4	-
Filges, Dietrichson et al. (2022)	Service learning for improving academic success in students in grade K to 12	619/6269	4	1
Filges, Montgomery, et al. (2015)	The Impact of Detention on the Health of Asylum Seekers	573/10061	2	-
Filges, Siren et al. (2020)	Voluntary work for the physical and mental health of older volunteers	43/14919	2	0
Filges, Smedslund et al. (2023)	PROTOCOL: The FRIENDS preventive programme for reducing anxiety symptoms in children and adolescents	96/2745	1	1
Filges, Sonne-Schmidt et al. (2018)	Small class sizes for improving student achievement in primary and secondary schools	303/7802	5	1
Filges, Torgerson, et al. (2019)	Effectiveness of continuing professional development training of welfare professionals on outcomes for children and young people	298/5147	1	4
Filges, Verner et al. (2023)	PROTOCOL: Participation in organised sport to improve and prevent adverse developmental trajectories of at-risk youth	158/7021	2	1
Thomsen et al. (2022)	PROTOCOL: Testing frequency and student achievement: A systematic review	627/6239	5	2
NIPH review				
Ames et al. (2024)	Acceptability, values, and preferences of older people for chronic low back pain management	144/425	-	2
Evensen et al. (2023)	Sutur av degenerative rotatorcuff-rupturer [Rotator cuff repair for degenerative rotator cuff tears]	418/2499	-	4

Jardim et al. (2021)	Effekten av antipsykotika ved førstegangspsykose [The effect of antipsychotics on first episode psychosis]	73/3924	-	3
Johansen et al. (2022)	Samværs-og bostedsordninger etter samlivsbrudd [Custody and living arrangements after parents separate]	143/1525	-	4
Meneses Echavez et al. (2022)	Psykologisk debriefing for helsepersonell involvert i uønskede pasienthendelser [Psychological debriefing for healthcare professionals involved in adverse events]	45/5452	-	3

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

Table 1. Description of studies used to develop benchmark scheme.

	Values				
Metric	**.0 < .5**	**.5 < .75**	**.75 < .8**	**.8 < .95**	**.95 ≤ 1**
Recall	Ineligible performance	Low performance. Only use for extra security as a *third* screener (Only use if resources are scarce since the alternative is worse)	On par with typical human second screener performance. Can be accepted.	On par with common researcher screening performance	Better than common human performance and traditional automated screening tools
Specificity	Ineligible performance	Low performance. Only use to reduce the total number of records if having an acceptable high recall.	Low performance. Only use to reduce the total number of records if having an acceptable high recall.	Acceptable if having a high recall value above .75	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.

Table 2. Screening performance benchmarks.

Review Model	Reps	Recall TP/(TP + FN)	Specificity TN/(TN + FP)	Raw agreement (TP + TN)/N[a]	bAcc
FFT					
gpt-3.5-turbo-0613 (incl. prop $\leq$ .5)	10	.699 (48/69)	.961 (3906/4066)	.956 (3954/4135)	.828
gpt-3.5-turbo-0613 (incl. prop $\leq$.2)	10	.812 (56/69)	.937 (3809/4066)	.935 (3865/4135)	.874
gpt-4-0613	1	.899 (62/69)	.937 (3810/4066)	.936 (3872/4135)	.918
FRIENDS					
gpt-3.5-turbo-0613 (incl. prop $\leq$ .5)	10	.953 (61/64)	.813 (1918/2508)	.816 (2100/2572)	.883
gpt-3.5-turbo-0613 (incl. prop $\leq$.7)	10	.953 (61/64)	.899 (2254/2508)	.900 (2315/2572)	.926
gpt-4-0613	1	.984 (63/64)	.974 (2442/2508)	.979 (2518/2572)	.979
TF					
gpt-4-0613 (incl. $\leq$5 out of 6 prompts)	1	.800 (80/100)	.838 (1676/2000)	.836 (1756/2100)	.819
gpt-4-0613 (incl. $\leq$ 4 out of 6 prompts)	1	.890 (89/100)	.743 (1486/2000)	.75 (1575/2100)	.816
gpt-4-0613 (incl. $\leq$ 3 out of 6 prompts)	1	.950 (95/100)	.670 (1340/2000)	.683 (1435/2100)	.810
gpt-4-0613 (all criteria in one prompt)	1	.91 (91/100)	.741 (1483/2000)	.749 (1574/2100)	.825

a: N is the total number of references

Table 3. Results of the three classifier experiments.

Step	Reviewer action
1	Find approximately 10 relevant study records (ideally more).
2	Find a minimum of 150 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. If finetuning is used, the fine-tuned model should be trained at this point.
4	Develop one or multiple prompts and test the(ir) performance. If repeated requests are used an inclusion probability threshold must also be decided at this point.
5	Repeat/refine step 4 until reaching a recall close to or above .75, and a specificity equal to or above .8. 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 .75 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 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).

Table 4. Workflow for how to conduct TAB screening using GPT API models.

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

Table 5. When to use GPT API models for TAB screening

"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.

Short background information

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

Textbox 1. Prompt example

"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

"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."

Textbox 3. Function call text