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

**ABSTRACT**

Independent human double screening of titles and abstracts is considered a critical step to ensure the quality of systematic reviews and meta-analyses herein. However, double screening is a costly as well as a time- and resource-intensive procedure that slows the review process, ultimately excluding many researchers from using it. To alleviate this issue and potentially increase the reliability of systematic reviews and meta-analyses, we evaluated the use of OpenAI’s GPT API models as an alternative second screener of titles and abstracts in large-scale systematic reviews. Overall, we found that GPT models perform similarly or even better than human screeners in terms of detecting relevant studies to be included. To support future reviewers, we develop a reproducible workflow and tentative guidelines for when reviewers can use GPT models for title and abstract screening. For this purpose, we present the R package AIscreenR.

**KEYWORDS:** *title and abstract screening, ChatGPT, systematic review, meta-analysis, screening benchmarks, AIscreenR*

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

**HIGHLIGHTS**

**What is already known**

* OpenAI’s GPT API models have shown promising performance in terms of working as a second screener of titles and abstracts within clinical and software literature.
* Automating screening tools can ease the burden of title and abstract screening
* Yet, automating screening tools most often cannot detect/classify all relevant studies, which in turn, induces the so-called ‘artificial screening biases’

**What is new**

* We show that OpenAI’s GPT API models can function as a highly reliable second screener in social science reviews with better recalls than presented in previous evaluations.
* We develop general benchmarks to compare the performance of AI screening tools with human screening.
* We provide general guidelines for how and when GPT models safely can used
* We present and validate the R package AIscreenR to ensure standardized conduct of title and abstract screening with OpenAI’s GPT models.

**Potential impact for Research Synthesis Methods readers**

* Changing the double screening workflow of title and abstract screening in systematic reviews
* Increasing the reliability of systematic reviews
* Substantial reduction of human labor in systematic reviews
* Standardizing screening with GPT API models

**1 INTRODUCTION**

Systematic reviews are essential tools for informing policy, research, and practice. Hence, it is all-important that systematic reviews adhere to the highest scientific standards. Yet systematic reviews are time-consuming, potentially hindering a timely transfer of usable knowledge. Distinct from other types of reviews, systematic reviews are defined as the process of collecting, assessing, and synthesizing findings from (ideally all) relevant scientific studies using explicit and replicable research methods (Gough et al., 2017; Hou & Tipton, 2024). A critical first step to ensure the quality of systematic reviews and meta-analyses herein involves detecting all eligible references related to the literature under review (Polanin et al., 2019). This entails searching all pertinent literature databases relevant to the given review, often resulting in thousands of title and abstract records that need to be screened. Manual screening of a large number of titles and abstracts can be a time-consuming and tedious task, indeed. However, overlooking relevant studies in this phase can be consequential, leading to substantially biased results. In fact, this can be seen as a special case of publication bias[[1]](#footnote-1) (Hedges, 1992; Rothstein et al., 2005). Therefore, independent human double-screening is considered to be the ’golden standard’ to hinder a biased selection of relevant studies (Guo et al., 2024; Higgins et al., 2019; Wang et al., 2020). Previous research suggests that screeners tend to miss between 3% to 24% of all eligible studies, which most often impacts the final results substantially (Buscemi et al., 2006; Waffenschmidt et al., 2019). Yet, duplicate screening of all identified titles and abstracts is a costly and resource-intensive procedure, potentially requiring several months of skilled, full-time human labor to complete (Campos et al., 2023; Hou & Tipton, 2024; Shemilt et al., 2016). Consequently, many reviewers refrain from using duplicate screening methods due to low budgets or narrow time limits, for instance. Alternatively, reviewers make too narrow searches to keep the number of records down to a manageable size (Van De Schoot et al., 2021). Over time, these issues will only grow in size since the complexity of identifying all relevant studies increases with the rapid growth in the number of scientific publications (Bornmann et al., 2021; O’Mara-Eves et al., 2015). Thus, it can be considered an economically inefficient and unsustainable use of human resources only to rely on human screening of titles and abstracts in future systematic reviews (Shemilt et al., 2016).

An alternative to human double-screening is to use (semi-)automated screening tools either based on text-mining or machine-learning algorithms to act as the second screener, a course-grained classifier, or to sort citation records in prioritized order (Cohen et al., 2006; Gartlehner et al., 2019; O’Mara-Eves et al., 2015; Van De Schoot et al., 2021). The use of automated screening tools is considered invaluable in supporting living reviews and has shown a promising ability to reduce the screening workload by 30% to 70% (O’Mara-Eves et al., 2015; Perlman‐Arrow et al., 2023). However, a clear disadvantage of substantial workload savings is that they will always result in missing at least 5%-10% of all eligible references since ”a 100% recall rate with a stochastic algorithm is generally considered unattainable” (Hou & Tipton, 2024, p. 3). This creates a screening paradox. While trying to reduce selection biases caused by single screening, automated screening potentially introduces a novel type of publication bias defined by König et al., (2023) as the ‘artificial screening bias’ (ASB). A further challenge with automated screening tools is that most of them are based on supervised and active learning methods. This means that they need to be trained on a large enough set of in- and excluded references to perform adequately which in turn can be a time-consuming task as well. In addition, when automation tools are used for prioritized screening, it is most often unknown when it is safe to stop screening with regard to finding all or close to all eligible references. Albeit, various stopping rules have been proposed, the adequacy of these is sensitive to a range of factors such as the length of the databases and the proportion of included studies (Campos et al., 2023; König et al., 2023; Van De Schoot et al., 2021).

To date, many automated screening tools have been thoroughly evaluated (Burgard & Bittermann, 2023). The overall picture is that they are generally not capable of replacing an independent human second screener without significant risk of omitting a substantial number of eligible studies (Gartlehner et al., 2019; O’Mara-Eves et al., 2015; Olorisade et al., 2016; Rathbone et al., 2015). By using the level of automation heuristic (c.f. Table 1) developed by O’Connor et al. (2019), it can be said that current automated tools generally fail to function at the highest levels of automation (i.e., Level 3 and Level 4) where they make trust-worthy independent, deterministic screening decisions. Instead, the vast majority of tools are predominated used to conduct Level 2 tasks such as sorting citation records in prioritized order from highest to lowest probability of being relevant to the review (O’Connor et al., 2019; Olofsson et al., 2017). If considerable time savings should be realized, it is regarded as pivotal that automated tools rise to Level 3 or 4 of automation.

**Table 1.** Levels of automation for human-computer interactions\*

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

\*Adopted from O’Connor et al. (2019)

A possible solution to bridge the gap between Levels 2 and 3 of automation is to use the newly developed large language models (LLM), such as the generative pre-trained transformer [(Chat)GPT] models introduced by OpenAI. The first evaluations of using GPTs for screening of medical, environmental, and software engineering titles and abstracts have generally yielded promising results (Alshami et al., 2023; Guo et al., 2024; Syriani et al., 2023). Meanwhile, the current evaluations were either premised on the original GPT-3.5-0301 models that will soon deprecate or did not draw on up-to-date features of the newest GPT models such as function calling (OpenAI, 2024). Moreover, it is unclear if these findings generalize to social science reviews in which the scientific abstracts are less structured. Therefore, one of the major aims of this paper is to evaluate the use and performance of OpensAI’s GPT API (application programming interface) models in social science reviews. Hereto, we confirm that OpenAI’s GPI API models can function as a highly reliable second screener with recalls (i.e., the ability to detect relevant studies) similar or superior to human performance.

Although the above findings represent a vital first step to validating the use of GPT as an independent second screener, it is neither enough to ensure user deployment and acceptance (O’Connor et al., 2019) nor that evidence institutions (such as Cochrane and the Campbell Collaboration) will accept and recommend the use of such tools. According to Campbell Collaboration the incorporation of automation tools “*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). Consequently, the main aim of this paper is to build a framework in which title and abstract screening with GPT models meet these standards. This work will furthermore aim to align with the recommendation put forward by O’Connor et al. (2019) for what it takes to gain trust in systematic review automation technologies.

Since the GPT API models we draw on are closed-source applications with black-box algorithms, fulfilling *(a)* is out of our hands. Yet, to prove the working the GPT API model *(b)*, we conducted two large-scale classification experiments. A key part of fulfilling *(b)* and not compromising the quality of future systematic reviews, is also to show that GPT is not inferior to human screen performance (O’Connor et al., 2019). Therefore, we mapped the human screening performance of 21 large-scale systematic reviews and held overall average human screening performances against the performance of the used GPT models. These measures were further used to develop common and flexible guidelines [i.e., complying with *(d)*] for when it can be considered safe to apply the GPT API models as an independent screening screener and when it is not. This also includes guidelines for how reviewers can set up tests [i.e., complying with *(e)*] to make their screening prompt(s) work(s) as expected in terms of identifying the most relevant studies. To standardize and comply with the requirements in point *(c)*, we have developed the R package AIscreenR. A key part of making user-friendly software is to make sure that software aligns with common workflows in systematic reviews (O’Connor et al., 2019). Therefore, we have built the AIscreenR so that it takes in the same input (i.e., ris-file data) as common screening software such as Covidence and EPPI-reviewer. Furthermore, we have built the AIscreenR as open-source software so that others in the review community can easily contribute to the development and ongoing support of the software [and thereby accommodate *(f)*]. With this exposition, we hope to make the uptake of such tools more acceptable in future reviews.

The remainder of the paper proceeds as follows: In

Section 2Previous research, what we add, Section 3 methods and human screening performance, Section 4 simulation/classifier experiment. Section 5 Section 6 Section 7 and Section 8

**2 RELATED WORK**

Syriani et al. (Syriani et al., 2023) test the performance relative to other machine-learning models.

**2.1 What we do differently**

* Use the newest models with function calling. We are the first to present results for the GPT-4 model.

”*Function calling allows developers to more reliably get structured data back from the model.*” (<https://openai.com/blog/function-calling-and-other-api-updates>)

* Instead of comparing GPT to other machine learning models, we develop benchmark for comparing human and AI performance.
* Develop new software (AIscreenR) to standardize the title and abstract screening with GPT.
* Multi-core process to increase the time used on screening.
* Draw on function calling an incorporates uncertain decisions. “Function calling allows developers to more reliably get structured data back from the model” (<https://openai.com/blog/function-calling-and-other-api-updates>)

**3 METHODS**

All metrics presented below were chosen based on the recommendations made by Syriani et al. (2023) and O’Connor et al. (2019).

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

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

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

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

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

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

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

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

Insert WSS (Campos et al., 2023)

**3.2 Human screening performance for comparison**

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

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Source**  **Authors** | **Short title** |  | **Ass.** | **Aut.** |
| *Campbell review* |  |  |  |  |
| Bøg et al. (2018) | Deployment of personnel to military operations | 106/2899 | 2 | - |
| Bondebjerg et al. (2023) | The effects of small class sizes on students’ academic achievement, socioemotional development and well‐being in special education | 244/1160 | 4 | 2 |
| Dalgaard, Bondebjerg, Klokker et al. (2022) | Adult/child ratio and group size in early childhood education or care to promote the development of children aged 0–5 years | 258/3667 | 4 | 2 |
| Dalgaard, Bondebjerg, Viinholt et al. (2022) | The effects of inclusion on academic achievement, socioemotional development and wellbeing of children with special educational needs | 373/14491 | 5 | 2 |
| Dalgaard, Filges et al. (2022) | Parenting interventions to support parent/child attachment and psychosocial adjustment in foster and adoptive parents and children | 424/13106 | 3 | 2 |
| Dalgaard, Jensen et al. (2022) | PROTOCOL: Group‐based community interventions to support the social reintegration of marginalised adults with mental illness | 557/17614 | 4 | 3 |
| Dietrichson et al. (2020, 2021) | Targeted school-based interventions for improving reading and mathematics for students with or at risk of academic difficulties in Grades K-6 [plus 7-12] | 2952/15273 | 6 | 1 |
| Filges, Dalgaard et al. (2022) | Outreach programs to improve life circumstances and prevent further adverse developmental trajectories of at-risk youth in OECD countries | 387/4890 | 4 | - |
| Filges, Dietrichson et al. (2022) | Service learning for improving academic success in students in grade K to 12 | 619/6269 | 4 | 1 |
| Filges, Montgomery, et al. (2015) | The Impact of Detention on the Health of Asylum Seekers | 573/10061 | 2 | - |
| Filges, Siren et al. (2020) | Voluntary work for the physical and mental health of older volunteers | 43/14919 | 2 | 0 |
| Filges, Smedslund et al. (2023) | PROTOCOL: The FRIENDS preventive programme for reducing anxiety symptoms in children and adolescents | 96/2745 | 1 | 1 |
| Filges, Sonne-Schmidt et al. (2018) | Small class sizes for improving student achievement in primary and secondary schools | 303/7802 | 5 | 1 |
| Filges, 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 |
| *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 |

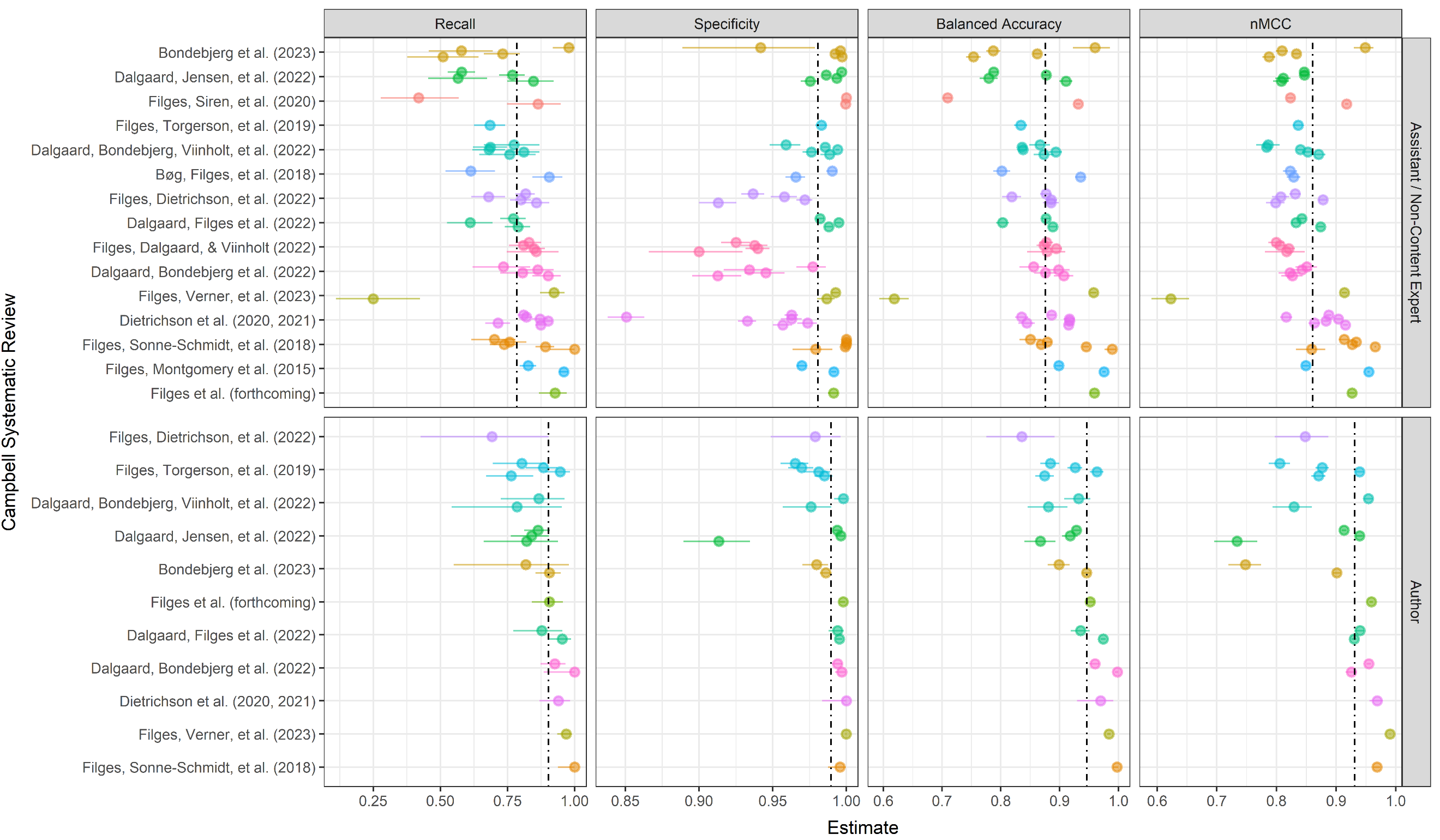


FIGURE 1. Performance measures within Campbell Systematic Reviews across assistants vs. authors. Dashed line indicate the average estimated via the SCE+ model.



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

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

Imbalance is not a problem with GPT models cf. FRIENDS

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

**4 NUMERICAL STUDY**

*Simulation data*

FRIENDS and FTT, only citation records with abstracts

*Prompt engineering*

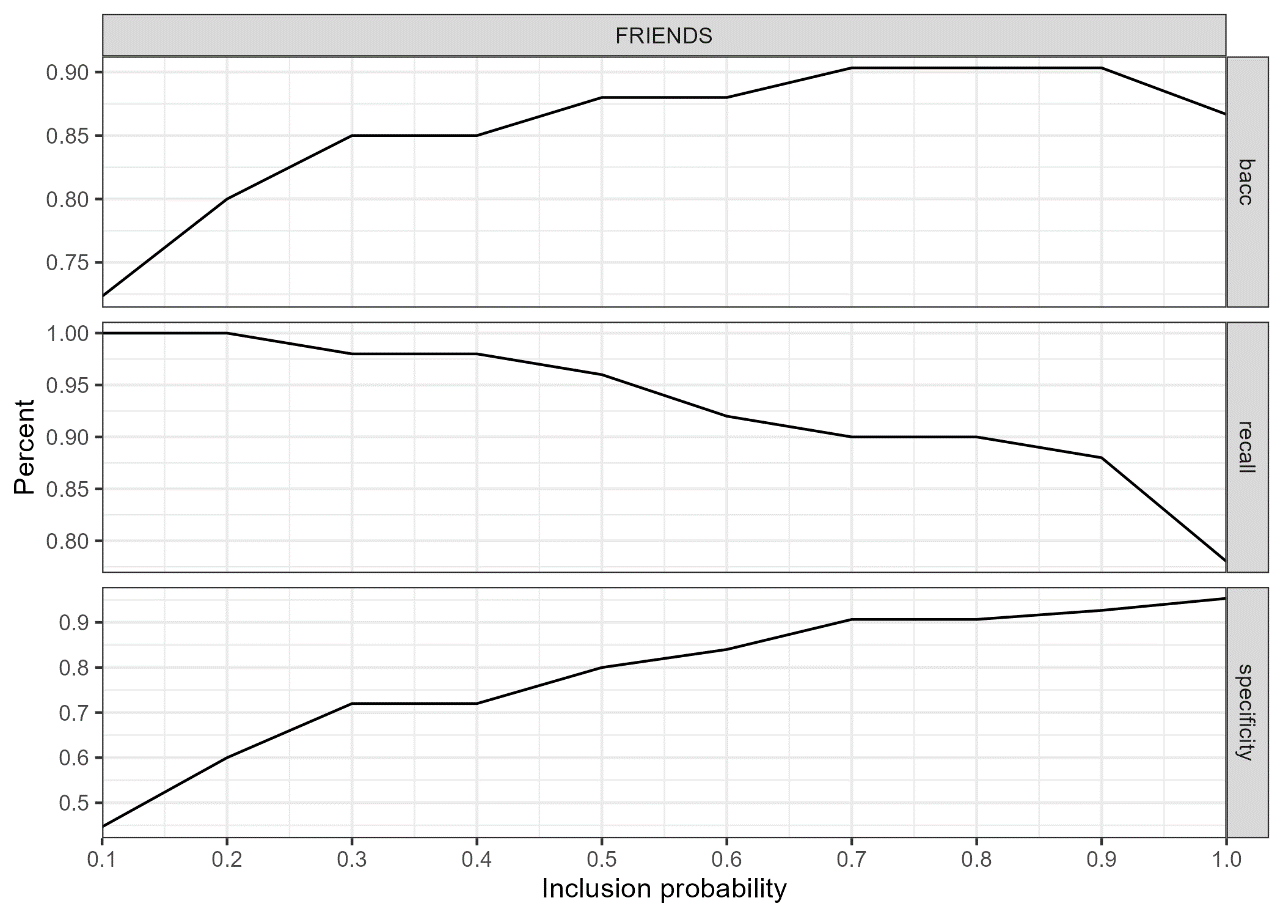
[insert prompt example]

*The simulation results*

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

*a*: is the total number of references





Concise text more important than information-dense prompt.

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

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

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

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

Student screening evaluation (Ng et al., 2014)

**5****TENTATIVE GUIDELINES**

*80% recall and 95% specificity.*

*Workflow and short package presentation*

Testing, not training. Less is more.

**6 LIMITATIONS**

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

**7 FUTURE RESEARCH**

* The use of hierarchical prompting in complex reviews. Simple prompts instead of long onces
* Shiny app to ease user set-up challenges (O’Connor et al., 2019) to make the workflow more user-friendly.

**8 DISCUSSION**

* Talk about the interface here – cannot replicate the results on the ChatGPT interface
* Reviewers should not consider screening prioritization methods and GPT screening as two incommensurable methods. Instead, the strength from both should ideally be combined.
* Forces review times to make very narrow searches due to lack of ressources to conduct the title and abstract screening rigorously (Guo find in ICloud)
* We believe that the GPT-4 models will perform even better when fed with abstracts following a rigorous structure as in medicine.
* When not to use. If you cannot make the prompt work properly or if you screen very few studies.
* We believe that no automated tool should ever be at level 4 – there shall always be a human-in-the-loop to ensure adequate behavior the the screening tools. Consequently, GPT models used in non-systematic to reduce the number of studies needed to be screened should always include safety checks. For example, reviewers should randomly sample 5-10% of the studies excluded by GPT to test for serious flaws in its decision-making. If serious flaws are detected the reviewers must re-test the used prompt(s) or refrain from using the given GPT model.
* More rapid transfer of knowledge from review to policy, research, and practice
* Makes it possible to help to screen in extreme-sized reviews (Shemilt et al., 2014, 2016)
* Extra security in low-budget and/or time-limited projects where there is only access to a single screener.
* No need for unnecessary restriction on search string.

**ACKNOWLEDGEMENT**

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

**FUNDING STATEMENT**

This manuscript was funded by VIVE Campbell, Denmark

**DATA AVAILABILITY STATEMENT**

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

**CONFLICT OF INTEREST STATEMENT**

The authors declare no conflict of interest.

**REFERENCES**

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

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

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

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

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

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

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

Campbell Collaboration. (2023). *Stepping up evidence synthesis: faster, cheaper and more useful*. https://www.campbellcollaboration.org/news-and-events/news/stepping-up-evidence-synthesis.html

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Filges, T., Smedslund, G., Eriksen, T., & Birkefoss, K. (2023). PROTOCOL: The FRIENDS preventive programme for reducing anxiety symptoms in children and adolescents: A systematic review. *Campbell Systematic Reviews*, *19*(4), e1374. https://doi.org/https://doi.org/10.1002/cl2.1374

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

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

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

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

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

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

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

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

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

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/

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

1. Rothstein et al. (2005) defines this type of publication bias as an availability bias. [↑](#footnote-ref-1)