­**ChatGPT 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 a pivotal step to ensure the reliability of systematic reviews. Yet, double screening is a costly as well as a time- and resource-intensive procedure that slows the review process which ultimately excludes many researchers from using it. Therefore, we evaluated the use of ChatGPT as an alternative second screener of titles and abstracts in large-scale systematic reviews. To access the performance of ChatGPT, we develop benchmarks to compare the screening performance between humans and any given AI screener based on conflict rates estimates from 16 large-scale Campbell Systematic Reviews conducted over the last 10 years by the Danish VIVE Campbell group. In contrast to the typical conflict rate between human screeners, we find that ChatGPT can function, as a highly reliable second screener, with a substantially higher recall (i.e., fewer false excluded references) than humans. To support future reviewers, we develop a reproducible workflow and tentative guidelines for when you should be able to trust in the screening performance of ChatGPT. For this purpose, we present the R package AIscreenR.

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

**HIGHLIGHTS**

**What is already known**

* ChatGPT and related GPT APIs have been shown to work as a second screener of titles and abstracts within clinical and software literature.

**What is new**

* We develop new benchmarks for comparing the AI screening performance relative to humans.
* We show that ChatGPT can be a highly reliable second screener in the social science review as well.
* We present from most recent gpt-models.
* We present the R package AIscreenR to ensure standardized conduct of title and abstract screening with ChatGPT.

**Potential impact**

* Changing double screening title and abstract screening in systematic reviews
* Standardizing screening with ChatGPT

**INTRODUCTION**

An all-important step to ensure the quality of systematic reviews involves detecting all relevant references related to the literature under review. Usually, this involves independent human double screening of all references detected in relevant databases and literature with two human screeners. This procedure has shown pivotally since less experienced single screeners tend to miss around 13% of relevant studies (with 3% for experienced screeners), mostly changing the main review findings1. Yet, double-screening is a costly and resource-intensive procedure which ultimately excludes many researchers from using it. An alternative to human double-screening is to use automated tools to act as the second screener 2 (Gartlehner et al., 2019; van de Schoot et al. 2021). Previous evaluations of existing tools find that most automated tools fail to reliably act as/imitating a human second screener. Meanwhile, it is still less known how well or if the newly developed large-language models (LLMs), such as ChatGPT, can work as a reliable second screener, especially within social science reviews.

“*Deployment and user acceptance: 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](https://www.campbellcollaboration.org/news-and-events/news/stepping-up-evidence-synthesis.html))

We focus on proving (b) and developing as well as providing software and user guidelines to fulfill (c) and (d).

*Previous research*

Syriani et al. 3 test the performance relative to other machine-learning models.

*What we do differently*

* Use 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.

*Metrics we use to evaluate the performance of the gpt-model*

The two main metrics we used to evaluate the performance of ChatGPT were the recall and specificity metrics. The “recall measures the ability to include all references that should be included”3 and is given by

where (true positive) represents all the studies that are correctly included, and (false negative) is the number of studies falsely excluded. Contrary, the “specificity measures the ability to exclude all references that should be excluded”3, that is

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. 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 GPT’s decision with the final decision made by two independent human screeners. For benchmark development, the conditions are determined by comparing the single screener decision with the final decision agreed upon between the two human screeners.

Mention how to calculate variance and confidence intervals.

*Human performance vs. AI performance*

To grasp a better understanding of the AI performance.

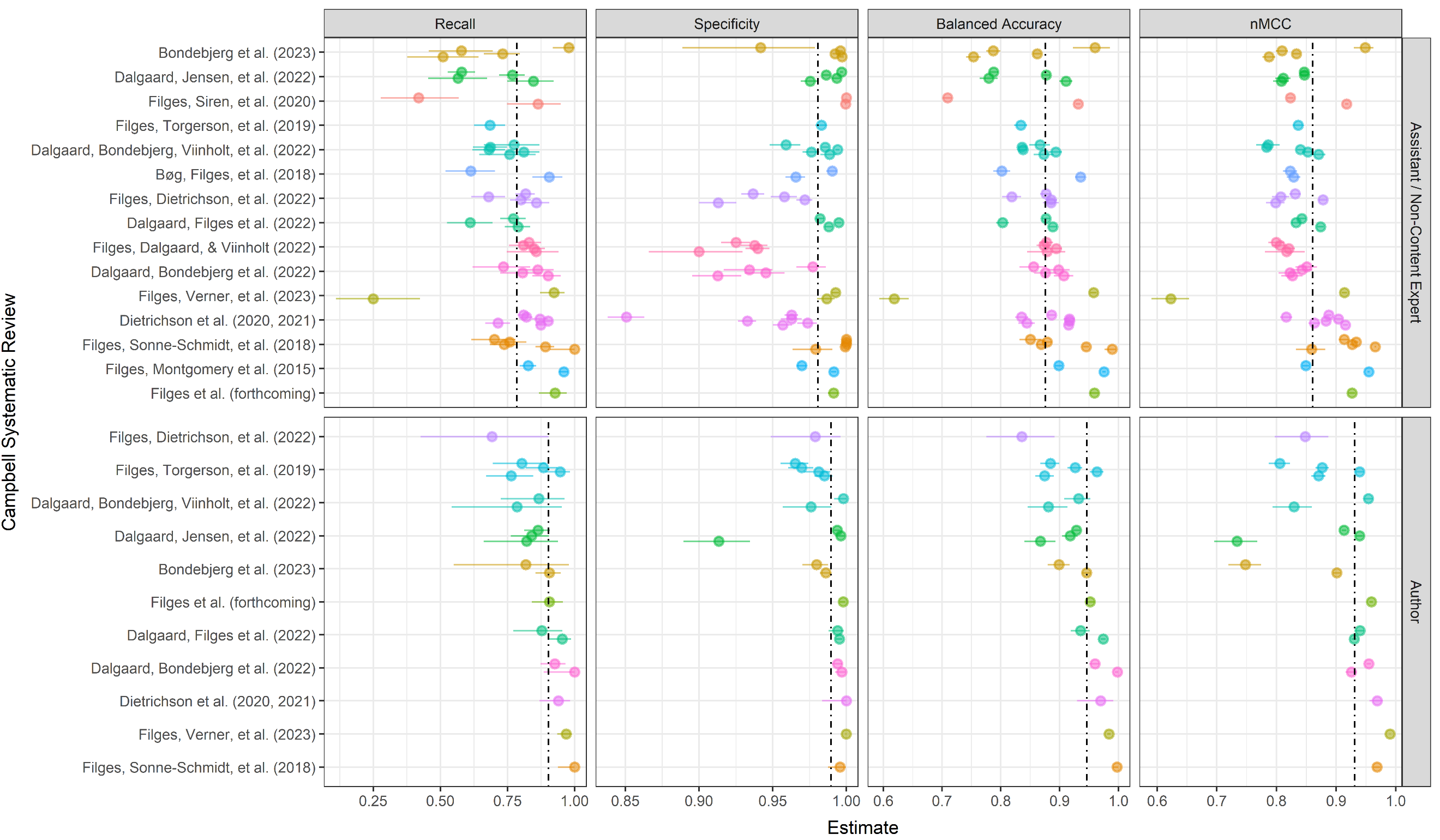


FIGURE 1. Performance measures within Campbell Systematic Reviews across assistants vs. authors. Dashed line indicate the average estimated via the PESCE+ model.

Campbell Systematic Reviews that we use 4–19

*Simulation data*

*The simulation results*

This includes conflict rates across xx references from xx Campbell Systematic Reviews, two reviews from Review of Educational Research

*Tentative guidelines*

*Workflow and short package presentation*

*Deficits of using ChatGPT*

* Black box (but this does not only count for GPT this is often true for human screening as well)

*Discussion*

* *Talk about interface here*

**REFERENCES**

1. Waffenschmidt, S., Knelangen, M., Sieben, W., Bühn, S. & Pieper, D. Single screening versus conventional double screening for study selection in systematic reviews: a methodological systematic review. *BMC Med. Res. Methodol.* **19**, 132 (2019).

2. Gartlehner, G. *et al.* Assessing the accuracy of machine-assisted abstract screening with DistillerAI: a user study. *Syst. Rev.* **8**, 277 (2019).

3. Syriani, E., David, I. & Kumar, G. Assessing the Ability of ChatGPT to Screen Articles for Systematic Reviews. *arXiv Prepr. arXiv2307.06464* (2023).

4. Dietrichson, J. *et al.* 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 Syst. Rev.* **16**, e1081 (2020).

5. Dietrichson, J. *et al.* 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 Syst. Rev.* **17**, e1152 (2021).

6. Filges, T. The FRIENDS preventive programme for reducing anxiety symptoms in children and adolescents: A systematic review. *Campbell Syst. Rev.* (2023).

7. Dalgaard, N. T., Bondebjerg, A., Viinholt, B. C. A. & Filges, T. The effects of inclusion on academic achievement, socioemotional development and wellbeing of children with special educational needs. *Campbell Syst. Rev.* **18**, e1291 (2022).

8. Filges, T., Dalgaard, N. T. & Viinholt, B. C. A. Outreach programs to improve life circumstances and prevent further adverse developmental trajectories of at-risk youth in OECD countries: A systematic review. *Campbell Syst. Rev.* **18**, e1282 (2022).

9. Filges, T., Dietrichson, J., Viinholt, B. C. A. & Dalgaard, N. T. Service learning for improving academic success in students in grade K to 12: A systematic review. *Campbell Syst. Rev.* **18**, e1210 (2022).

10. Filges, T., Verner, M., Ladekjær, E. & Bengtsen, E. PROTOCOL: Participation in organised sport to improve and prevent adverse developmental trajectories of at-risk youth: A systematic review. *Campbell Syst. Rev.* **19**, e1321 (2023).

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

12. Dalgaard, N. T., Bondebjerg, A., Klokker, R., Viinholt, B. C. A. & Dietrichson, J. 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 Syst. Rev.* **18**, e1239 (2022).

13. Dalgaard, N. T., Flensborg Jensen, M. C., Bengtsen, E., Krassel, K. F. & Vembye, M. H. PROTOCOL: Group‐based community interventions to support the social reintegration of marginalised adults with mental illness. *Campbell Syst. Rev.* **18**, e1254 (2022).

14. Filges, T., Sonne‐Schmidt, C. S. & Nielsen, B. C. V. Small class sizes for improving student achievement in primary and secondary schools: A systematic review. *Campbell Syst. Rev.* **14**, 1–107 (2018).

15. Filges, T., Siren, A., Fridberg, T. & Nielsen, B. C. V. Voluntary work for the physical and mental health of older volunteers: A systematic review. *Campbell Syst. Rev.* **16**, e1124 (2020).

16. Filges, T., Montgomery, E., Kastrup, M. & Jørgensen, A.-M. K. The Impact of Detention on the Health of Asylum Seekers: A Systematic Review. *Campbell Syst. Rev.* **11**, 1–104 (2015).

17. Bøg, M., Filges, T. & Jørgensen, A. M. K. Deployment of personnel to military operations: impact on mental health and social functioning. *Campbell Syst. Rev.* **14**, 1–127 (2018).

18. Filges, T. *et al.* Effectiveness of continuing professional development training of welfare professionals on outcomes for children and young people: A systematic review. *Campbell Syst. Rev.* **15**, e1060 (2019).

19. Dalgaard, N. T., Filges, T., Viinholt, B. C. A. & Pontoppidan, M. Parenting interventions to support parent/child attachment and psychosocial adjustment in foster and adoptive parents and children: A systematic review. *Campbell Syst. Rev.* **18**, e1209 (2022).