­**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 and ultimately excludes many researchers from using it. To overcome this tedious burden, we evaluated the use of ChatGPT as a second screener of titles and abstracts in large-scale systematic reviews. Hereto, we developed benchmarks to compare the screening performance between humans and any given AI screener based on conflict rates estimates across xx independently double-screened references from 15 large-scale Campbell Systematic Reviews conducted over the last 10 years at the Danish Center for Social Science Research (VIVE). 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 what we consider to be a reproducible workflow and tentative guidelines for when you should be able to trust in the screening performance of ChatGPT, and we present the R package AIscreenR.

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

**HIGHLIGHTS**

**What is already known**

* 1
* 2
* 3
* 4

**What is new**

* 1
* 2
* 3
* 4

**Potential impact**

* 1
* 2

**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, excluding 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*

*What we do differently*

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

*Human performance vs. AI performance*

Campbell Systematic Reviews that we use 3,4,13–17,5–12

*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*

*Discussion*

* *Talk about interface here*

18

**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. 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).

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 K-6: A systematic review. *Campbell Syst. Rev.* **17**, e1152 (2021).

5. 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).

6. 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).

7. 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).

8. 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).

9. 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).

10. 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).

11. 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).

12. 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).

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

14. 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).

15. 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).

16. 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).

17. 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).

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