­**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. In the hope of overcoming 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 xx large-scale systematic reviews conducted over the last 10 years by various researchers from the Danish Center for Social Science Research (VIVE). In contrast to the typical conflict rate between human screener, 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. Finally, to implement this procedure we have developed the R package AIscreenR and we specify 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.

**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 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), which in most cases substantially changes 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 unknown how well or if the newly developed large-language models (LLMs) such as ChatGPT can work and possibly emulate a human 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*

*Simulation data*

*The simulation results*

This includes conflict rates across xx references from xx Campbell Systematic Reviews, two reviews from Review of Educational Research, and one BMJ review.

*Tentative guidelines*

*Workflow and short package presentation*

*Deficits of using ChatGPT*

*Discussion*

* *Talk about interface here*

3

**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).

**SUPPORTING INFORMATION**R codes for replication of all examples provided in this paper are available on the Open Science Framework at INSERT