­**ChatGPT Can Function as a Highly Reliable Second Screener of Titles and Abstracts in Systematic Reviews**

*Abstract*

Evaluate the GPTs screening performance on three Campbell Systematic Reviews.

Develop benchmarks for performance comparison between humans and the AI.

Present a potential reproducible workflow for how the screen title and abstracts with ChatGPT and guidelines for when you should be able to trust in the screening performance of 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 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.

*Previous research*

*What we do differently*

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

*The simulation results*

*Human performance vs. AI performance*

*Rule of thumps*

*Workflow and shot package presentation*

*Deficits*

*Discussion*

* *Talk about interface here*

3

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).