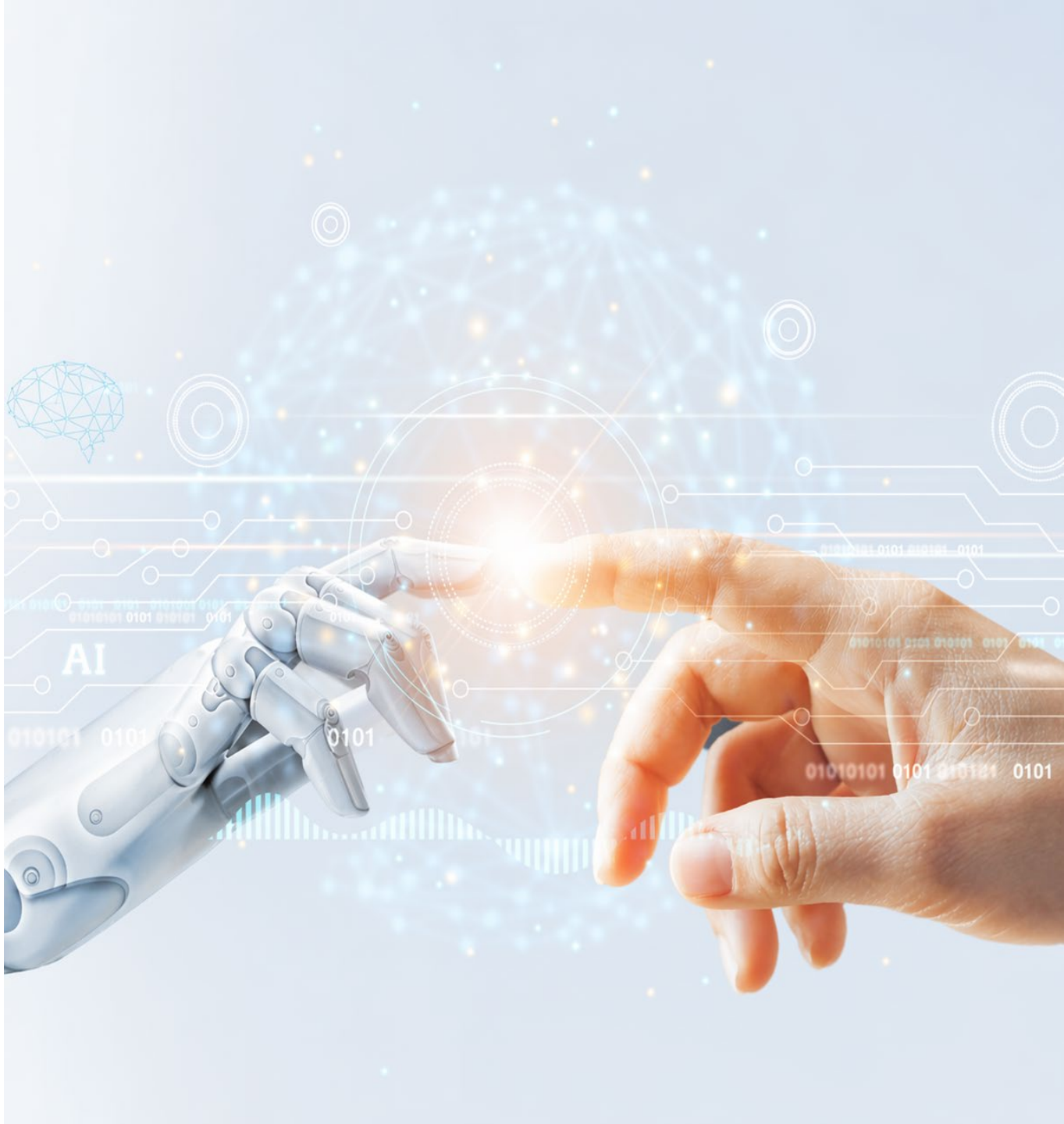
How to Use Generative Pretrained Transformer (GPT) Models as Reliable Second Screeners of Titles and Abstracts in High-Quality Systematic Reviews [workshop]

Mikkel H. Vembye, PhD.
Researcher
The Danish Center for Social Science Research

IRM 2026, Cologne, April 24

VIVE



Workshop objectives

- Introduction to title and abstract screening with GPT (API) models.
- Discuss quality assessment and reliable implementation of AI screening.
- Discuss future direction for using AI for literature screening.
- Showcase the R package AIscreenR.

Find all material behind the presentation at: https://github.com/MikkelVembye/IRM2026

Link to the AIscreenR homepage: https://mikkelvembye.github.io/AIscreenR/

VIVE

Why use AI for screening in systematic reviews?

Improve quality

- Reduces missed studies and human error. While double-screening is the gold standard, it is costly and not always feasible.
- Enables broader, less restrictive database searches

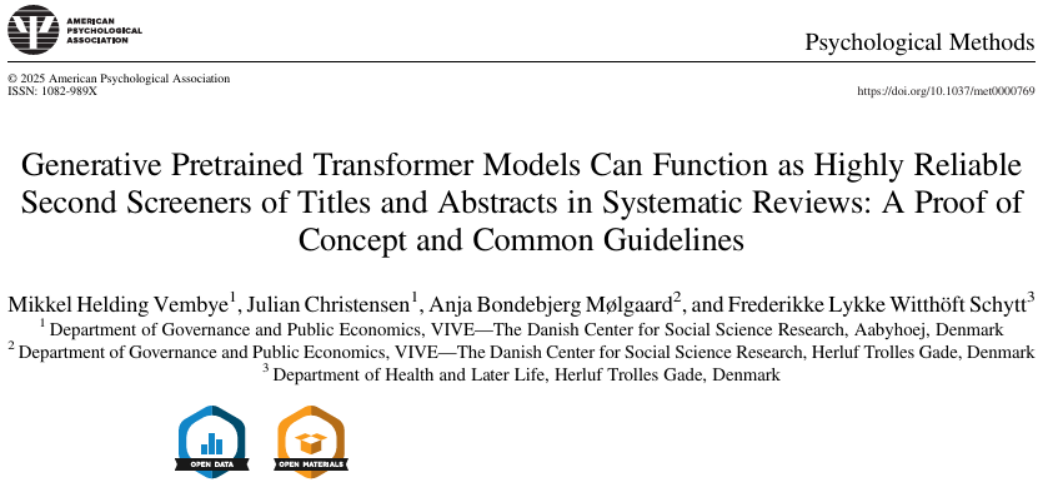
Increase efficiency

- Reduce manual and tedious workload and accelerate the review process
- Some reviews remain incomplete due to the sheer volume of studies—and this challenge is growing as databases expand.

What we have tested and developed

- We have tested the use of OpenAI's **GPT** (Generative Pre-trained Transformer) **API** (Application Programming Interface) **models** to screen titles and abstracts and developed guidelines for how to use them reliably.
- Our results show that GPT API models perform at least on par with human screeners' performances, across many different types of social science reviews.
- To conduct this type of screening, we have developed the R package AIscreenR (Vembye & Olsen, 2026).

Based on this, we suggest that GPT API models can be used as full second screeners in state-of-the-art reviews – let me show you the setup!



AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Methods

© 2025 American Psychological Association
ISSN: 1082-989X

https://doi.org/10.1037/met0000769

Generative Pretrained Transformer Models Can Function as Highly Reliable Second Screeners of Titles and Abstracts in Systematic Reviews: A Proof of Concept and Common Guidelines

Mikkel Helding Vembye[1], Julian Christensen[1], Anja Bondebjerg Mølgaard[2], and Frederikke Lykke Witthöft Schytt[3]
[1] Department of Governance and Public Economics, VIVE—The Danish Center for Social Science Research, Aabyhoej, Denmark
[2] Department of Governance and Public Economics, VIVE—The Danish Center for Social Science Research, Herluf Trolles Gade, Denmark
[3] Department of Health and Later Life, Herluf Trolles Gade, Denmark

AIscreenR: AI screening tools in R for systematic reviewing

The goal of AIscreenR is to use AI tools to support screening processes (including title and abstract screening) in systematic reviews and related literature reviews. At the current stage, the main aim of the `AIscreenR` package is to use generative pre-trained transformer (GPT) models as second screeners of titles and abstracts or alternatively to reduce the number of references needed to be screened by humans. For validation measures and guidance on how to conduct reliable title and abstract screenings with GPT API models, see Vembye et al. (2025).

VIVE

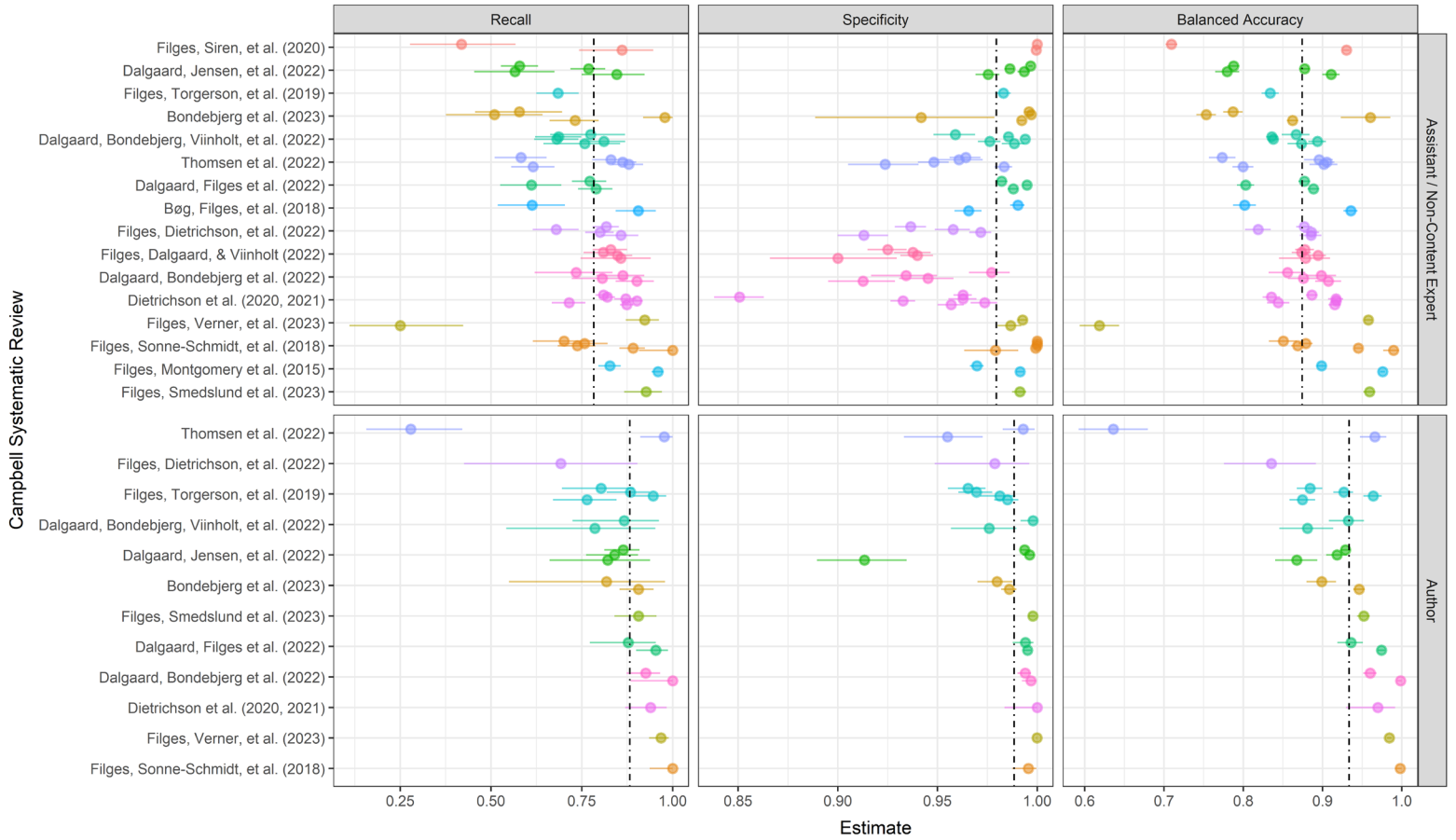
Why use GPT API models and not just ChatGPT?

- Overcomes copy–paste procedures; may reduce hallucination risk.
- Enables systematic comparison of models and prompts.
- Unclear if the ChatGPT setup is agnostic to data imbalance.
- Enables the screening of a very large number of references in a short time frame—potentially up to 30,000 records per minute.
- Supports multi-prompt and repeated, identical screenings, where the inclusion criteria are defined based on the frequency of selection across runs.
- Evidence suggests GPT API models outperform ChatGPT for screening (Alshami et al., 2023; Gargari et al., 2024; Guo et al., 2024; Issaiy et al., 2024; Khraisha et al., 2024; Syriani et al., 2024)..
- Our setup has been scientifically validated.

VIVE

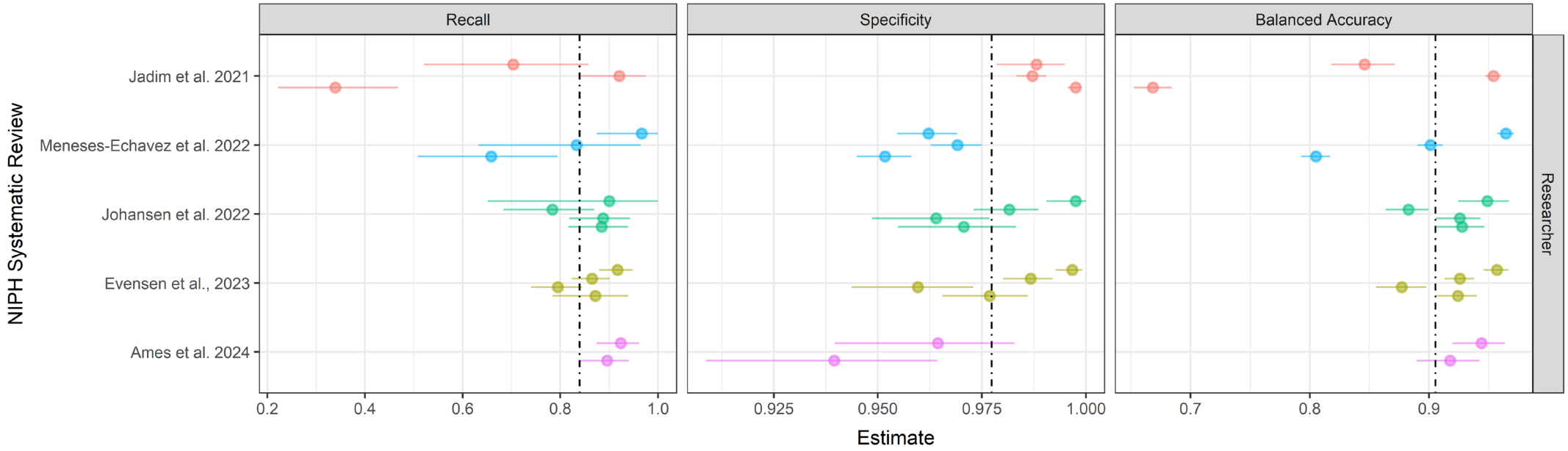
Quality assessing AI screenings

- A precondition for quality assessing AI screening is to know common human screening behavior/performance
- We find that our assistant/student screeners, on average, have a recall = **0.782**, 95% CI[0.747, 0.817]



Quality assessing AI screenings (continued)

- Researchers and/or content-experts have slightly higher recalls.
- We find in our own group researchers have, on average, a recall = **0.881**, 95% CI [0.823, 0.931]
- Using data from the Norwegian Institute for Public Health, comparing researcher-researcher screenings only, we find that the average recall drops to **0.839**, 95% CI [0.737, 0.920].



Quality assessment via benchmarking

- Based on the above-presented results, we have developed a benchmark scheme, which can be used to ensure the quality of title and abstract screenings.
- In general, we recommend that GPT screenings must yield recalls above 75% to be usable in high-quality reviews.
- It is more challenging to set good guidelines for specificity. If recall is high, specificity matters less. It is simply an extra safeguard.

	Values				
Metric	**.0 < .5**	**.5 < .75**	**.75 < .8**	**.8 < .95**	**.95 ≤ 1**
Recall	Ineligible performance	Low performance. Only use for extra security as a *third* screener (Only use if resources are scarce since the alternative is worse)	On par with typical human second screener performance. Can be accepted.	On par with common researcher screening performance	Better than common human performance and traditional automated screening tools
Specificity	Ineligible performance	Low performance. Only use to reduce the total number of records if having an acceptable high recall.	Low performance. Only use to reduce the total number of records if having an acceptable high recall.	Acceptable if having a high recall value above .75	On par with common human screening performance

Note: Red areas indicate conditions under which the TAB screening performance is unacceptability low. Gray areas represent insufficient performance conditions but some applications with these performance measures might still be viable. Green areas represent acceptable screening performances on par with or better than human screening.

Why we consider benchmarking all-important

- Guards against bad and biased screenings that are inferior to human screening.
- Play a key role in testing the accuracy of using AI as a second screener.
- It allows for context-specific assessments of the efficacy of using GPT API models as second screeners.
- As we cannot control model developments, the benchmark scheme ensures that we can monitor model performances over time. Meaning that if we experience that a model suddenly cannot live up to our benchmarks, we can stop using it.
- Avoids the wild-west and ensures standardization of this screening approach.

VIVE

How generalizable is this approach?

We have conducted three large-scale classification experiments with different levels of complexity in terms of the number of inclusion criteria. Here we found that:

- GPT API models can perform on par with or in some cases even better than typical human second screeners in high-quality systematic reviews (Vembye et al., 2025).
- All models yield recalls above 80% in all of our experiments while showing a high exclusion rate as well.
- GPT-4 model can be rather over-inclusive in complex review settings.
- The GPT-4 and GPT-5 models outperform the GPT-3.5 models. We therefore recommend primarily using GPT-4 or GPT-5 models for title and abstract screening.
- Moreover, we find that smaller models such as GPT-4o-mini and GPT-5-mini can be used for conducting reliable screenings. This has been a game-changer in order to reduce the cost of this screening approach.

VIVE

Limitations

- Black-box models (notice: human screening most often represents a black-box operation as well)
- Off-the-shelf method changing over time.
- It can be prompt sensitive
- Potentially large environmental impact
- Hard to keep up with new model developments
- Supporting American and Chinese Big Tech companies.

Advantages

- Equal treatment of all titles and abstracts
- It is most often accurate
- It is fast
- It is cheap
- Can guard against human drifting.
- Extra insurance that you have found all relevant records. It can also be used as a third screener.
- Agnostic to data imbalance
- It’s flexible. That is, it can rapidly be fitted to many different situations.

VIVE

Future research

- Test with local models such as the models on Ollama or from MistralAI. This would freeze the efficacy of this approach and increase its transparency.
- Consider how to combine GPT screenings with traditional (semi)-automated screening tools, such as priority and classifier screening, most efficiently.
- Consider how fine-tuning can support the reliability of the screening approach even further?
- Implementation in standard review tools such as EPPI Reviewer, Covidence, Meta-Reviewer etc.
  - Alternatively, a shiny app could be made to ease user-friendliness.

Future perspectives beyond title and abstract screening

- We need local models to overcome legal issues. Extremely important when conducting full-text screening, data extraction, and risk-of-bias assessment.
- Many countries' licensing agreements with publishers do not allow uploading PDF/full text to remote commercial AIs. Even if they do, it might be problematic due to the European AI Act.
- THE FUTURE IS AI! I project that screening, as we know it today, will be largely eradicated within the next 10 years → I believe that the relevancy of AIscreenR will be short-lived.

VIVE

AIscreenR Demo

Link to the package vignette:

https://mikkelvembye.github.io/AIscreenR/articles/Using-GPT-API-Models-For-Screening.html

Link to R codes behind the presentation:

https://github.com/MikkelVembye/IRM2026/blob/main/Test%20screening.R

VIVE

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. https://doi.org/10.3390/systems11070351
- 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
- Gargari, O. K., Mahmoudi, M. H., Hajisafarali, M., & Samiee, R. (2024). Enhancing title and abstract screening for systematic reviews with GPT-3.5 turbo. *BMJ Evidence-Based Medicine*, *29*(1), 69 LP – 70. https://doi.org/10.1136/bmjebm-2023-112678
- 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
- Issaiy, M., Ghanaati, H., Kolahi, S., Shakiba, M., Jalali, A. H., Zarei, D., Kazemian, S., Avanaki, M. A., & Firouznia, K. (2024). Methodological insights into ChatGPT's screening performance in systematic reviews. *BMC Medical Research Methodology*, *24*(1), 78. https://doi.org/10.1186/s12874-024-02203-8
- Khraisha, Q., Put, S., Kappenberg, J., Warraitch, A., & Hadfield, K. (2024). Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Research Synthesis Methods*. https://doi.org/10.1002/jrsm.1715
- Syriani, E., David, I., & Kumar, G. (2024). Screening articles for systematic reviews with ChatGPT. *Journal of Computer Languages*, *80*, 101287. https://doi.org/10.1016/j.cola.2024.101287
- Tomlinson, B., Black, R. W., Patterson, D. J., & Torrance, A. W. (2024). The carbon emissions of writing and illustrating are lower for AI than for humans. *Scientific Reports*, *14*(1), 3732. https://doi.org/10.1038/s41598-024-54271-x
- Vembye, M. H. & Olsen, T. (2026). *AIscreenR: AI screening tools for systematic reviews.* (0.3.2). CRAN. https://doi.org/10.32614/CRAN.package.AIscreenR
- Vembye, M. H., Christensen, J., Mølgaard, A. B., & Schytt, F. L. W. (2025).Generative Pretrained Transformer Models Can Function as Highly Reliable Second Screeners of Titles and Abstracts in Systematic Reviews: A Proof of Concept and Common Guidelines. *Psychological Methods*. https://doi.org/10.1037/met0000769

Appendix 1 - Assessment Measures

Recall is the proportion of relevant records being correctly classified as relevant, given by

$$Recall = \frac{\{true\ positive\}}{\{true\ positive\} + \{false\ negative\}}$$

Specificity is the proportion of irrelevant records being correctly classified as irrelevant, given by

$$Specificity = \frac{\{true\ negative\}}{\{true\ negative\} + \{false\ positive\}}$$

Appendix 2 – Numerical results

Review Model	Reps	Recall TP/(TP + FN)	Specificity TN/(TN + FP)	Raw agreement (TP + TN)/N[a]	bAcc
FFT					
gpt-3.5-turbo-0613 (incl. prop ≤ .5)	10	.699 (48/69)	.961 (3906/4066)	.956 (3954/4135)	.828
gpt-3.5-turbo-0613 (incl. prop ≤.2)	10	.812 (56/69)	.937 (3809/4066)	.935 (3865/4135)	.874
gpt-4-0613	1	.899 (62/69)	.937 (3810/4066)	.936 (3872/4135)	.918
FRIENDS					
gpt-3.5-turbo-0613 (incl. prop ≤ .5)	10	.953 (61/64)	.813 (1918/2508)	.816 (2100/2572)	.883
gpt-3.5-turbo-0613 (incl. prop ≤.7)	10	.953 (61/64)	.899 (2254/2508)	.900 (2315/2572)	.926
gpt-4-0613	1	.984 (63/64)	.974 (2442/2508)	.979 (2518/2572)	.979
TF					
gpt-4-0613 (incl. ≤5 out of 6 prompts)	1	.800 (80/100)	.838 (1676/2000)	.836 (1756/2100)	.819
gpt-4-0613 (incl. ≤ 4 out of 6 prompts)	1	.890 (89/100)	.743 (1486/2000)	.75 (1575/2100)	.816
gpt-4-0613 (incl. ≤ 3 out of 6 prompts)	1	.950 (95/100)	.670 (1340/2000)	.683 (1435/2100)	.810
gpt-4-0613 (all criteria in one prompt)	1	.91 (91/100)	.741 (1483/2000)	.749 (1574/2100)	.825

a: *N* is the total number of references

Appendix 3: What we also do: further standardization

- *Common guidelines* for when it is (and when it is not) appropriate to use GPT API models for title and abstract screening in high-quality reviews. These guidelines are primarily based on the benchmark scheme.
- *A workflow for how to configure a reliable screening*, including how to test and develop prompts. Hereto we introduce multiple-prompt screening, i.e., making one prompt per inclusion criteria.

According to Campbell Collaboration (2023) using AI in high-quality reviews requires:

(a) functioning tech **[Outside our control]**
(b) proof that it is functioning appropriately: **[Our answer: Experiment results]**
(c) the tech embodied in usable products: **[Our answer: AIscreenR]**
(d) agreed guidelines for appropriate use **[Our answer: The use of benchmark schemes]**
(e) training **[Our answer: Assess the use with test data: Alternatively use fine-tuning]**
(f) ongoing support **[Our answer: Provide AIscreenR as an open-source software]**

In our paper, we strive to accommodate requirements b to f.

VIVE