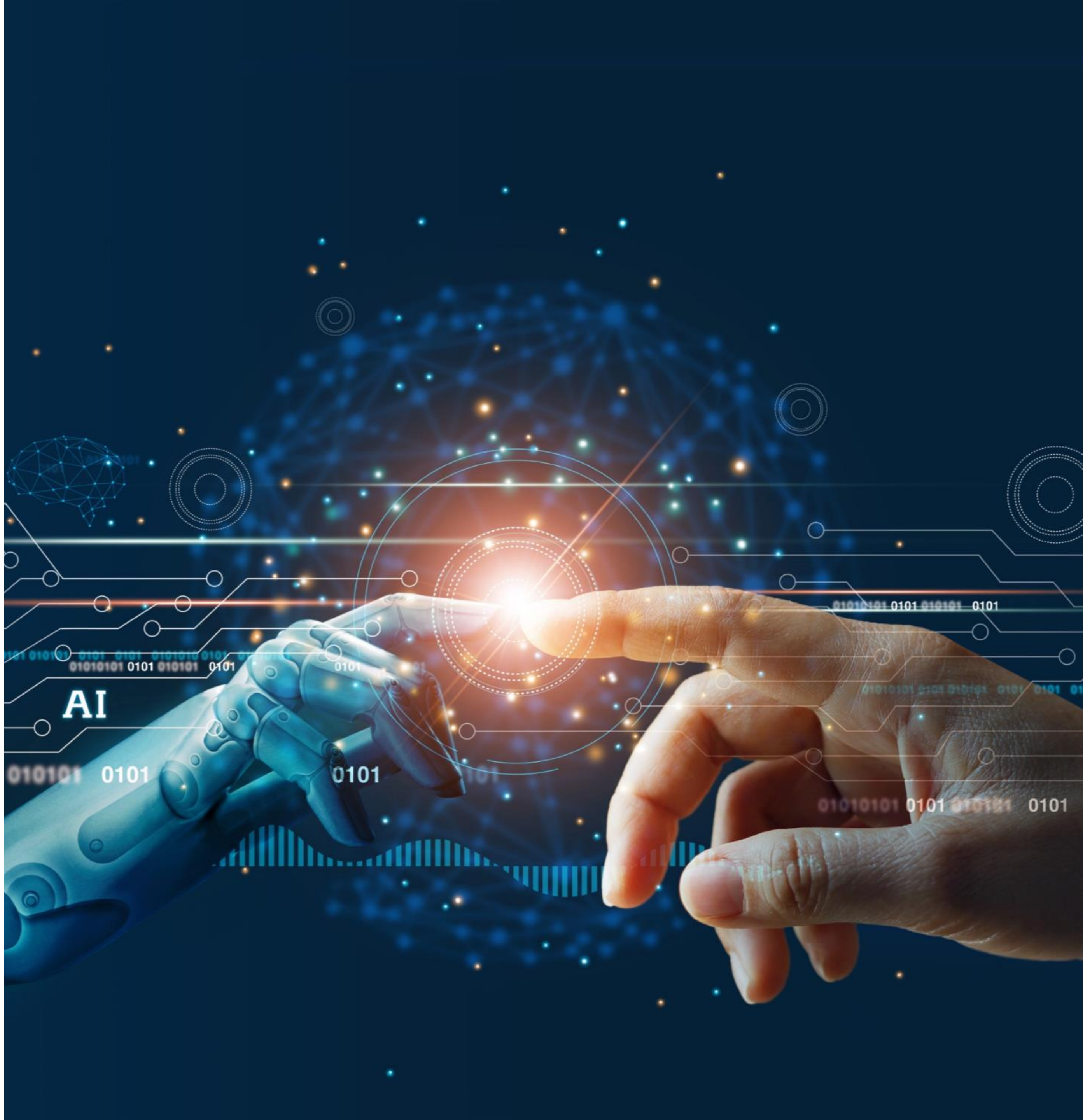
Using GPT API Models as Second Screeners of Titles and Abstracts in High-Quality Systematic Reviews

Mikkel Vembye, PhD

The Danish Center for Social Science Research, VIVE.

SRMA SIG, 2024.11.15

VIVE



Main goals of the presentation

1) To give you a practical understanding of title and abstract screening using GPT API models.

2) To give you an impression of how relatively simple and powerful this screening approach can be.

3) To show how you can quality assess such screenings.

Find all material from this presentation at: https://github.com/MikkelVembye/SRMA-SIG-presentation

Find the paper behind this presentation at: https://osf.io/yrhzm

Why use AI for screening in systematic reviews?

For Quality Reasons: To reduce human errors

- Human screeners overlook relevant studies for various reasons. Therefore, state-of-the-art is to conduct human double-screening. However, this is costly and many research groups cannot afford double-screening.
- Reviewers most often limit their database searches so that they yield a number of studies that has a manageable size for humans to screen. However, this increases the risk of overlooking relevant studies.

For Resource Reasons: To reduce human resources and speed up the review process

- Screening large amounts of references slow the review process as it can take an unmanageable amount of time. In our Campbell reviews, we spend approximately 3–5 months on this process, and even more in our large reviews.
- It is tedious, manual work that can be hard to finish.
- Some reviews cannot be completed because they require screening of too large a number of references. This problem will only grow over time as the number of references in research databases increases.

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. As you will see in a moment, this is NOT the same as using ChatGPT!
- Our results show that GPT API models perform at least on par with human screeners performances, even in very complex reviews with many inclusion criteria. I will come back to these results later in the presentation.
- To conduct this type of screening, we have developed the R package AIscreenR (Vembye, 2024).
- So far, we have not yet encountered a case where we could not reliably use this method, meaning that they show screening behaviors at least on par with human screeners.

Based on this, we suggest that GPT API models can be used as full second screeners in state-of-the-art reviews

Why use GPT API models and not just ChatGPT?

- Avoids copy-paste procedures, and it is easier to bypass model hallucination.
- Makes it easy to test differences between models, prompts, etc.
- You can screen an incredibly large amount of references in a very short time. With a sufficiently powerful computer, you can screen up to 30,000 per minute.
- Multiple identical screenings can be performed, and inclusion criteria can be built based on how many times a reference has been included across these screenings.
- Using gpt-4o-mini is cheaper than subscribing to ChatGPT Plus. With one prompt, you can screen 25,000 references for 1.5 USD.
- Research indicates that GPT API models are better suited for screening than ChatGPT (Alshami et al., 2023; Gargari et al., 2024; Guo et al., 2024; Issaiy et al., 2024; Khraisha et al., 2024; Syriani et al., 2024).

AIscreenR Demo

Link to the package vignette:

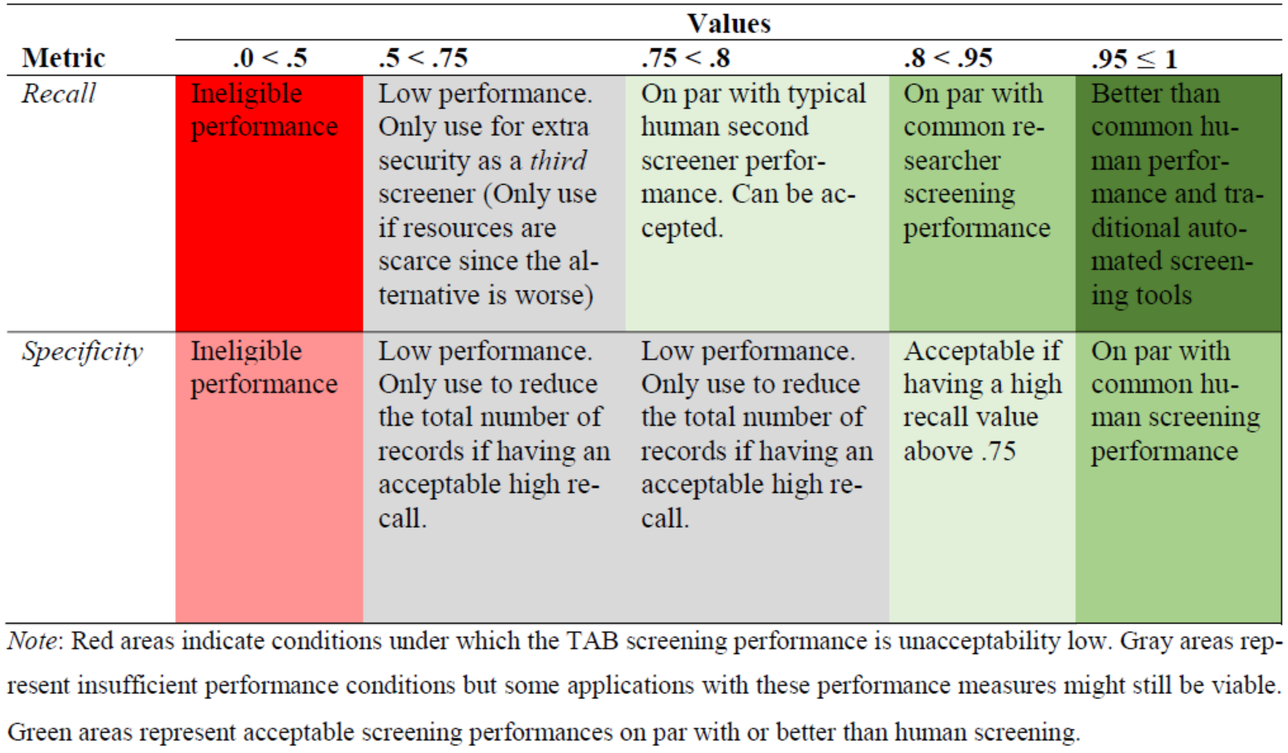
https://mikkelvembye.github.io/AIscreenR/articles/Using-GPT-API-Models-For-Screening.html

Link to R codes behind the presentation:

https://github.com/MikkelVembye/SRMA-SIG-presentation/blob/main/Example%20code.R

Quality assessment via benchmarking

- We have developed a benchmark scheme based on our and our Campbell students' typical screening performances, which can be used to ensure the quality of title and abstract screenings.
- In general, we recommend that GPT screenings should yield recalls above 75% to be usable in high-quality reviews.
- Content-experts/researchers typically have recalls around 83%.
- 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.
- 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.

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. Herein we found that:

- GPT API models can perform on par with or in some cases even better typical human second screeners in high-quality systematic reviews (Vembye et al., 2024). All models yield recalls above 80% in all of our experiments while showing high exclusion rate as well.
- GPT-4 model can be rather over-inclusive in complex review settings.
- The GPT-4 model outperforms the GPT-3.5-turbo model. We therefore recommend primarily using GPT-4 models for title and abstract screening.
- Moreover, our newest results suggest that GPT-4o-mini, which is 200 times cheaper than GPT-4, can perform on par the GPT-4, when using 10 screenings. This is a game-changer in order to reduce the cost of this screening approach.

Limitations

- Black-box models (notice: human screening most often represent black-box operation as well)
- Off-the-shelf method that change over time.
- It can be prompt sensitive
- Potentially large environmental impact. However, this is not easy to fully assess (Tomlinson et al., 2024) but many people have concerns about this.

Advantages

- Equal treatment of all titles and abstracts
- It is fast
- It is cheap
- Can guard against human drifting.
- Extra insurance that you have found all relevant records. Can also be used as a third screener.
- Agnostic to data imbalance
- It's flexible

Future research

- Test with local models such as the models 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 priority and classifier screening most efficiently.
- Considered how can fine-tuning can support the reliability of the screening approach even further?
- Implementation in standard review tools such as Meta-Reviewer, EPPI Reviewer, Covidence, etc.
  - Alternately, a shiny-app could be made to ease user-friendliness.

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. (2024). *AIscreenR: AI screening tools for systematic reviews.* (0.1.0). CRAN. https://doi.org/10.32614/CRAN.package.AIscreenR
- Vembye, M. H., Christensen, J., Mølgaard, A. B., & Schytt, F. L. W. (2024). GPT API Models Can Function as Highly Reliable Second Screeners of Titles and Abstracts in Systematic Reviews: A Proof of Concept and Common Guidelines. *Open Science Framework*. https://doi.org/10.31219/osf.io/yrhzm

Appendix 1 - Assessment Measures

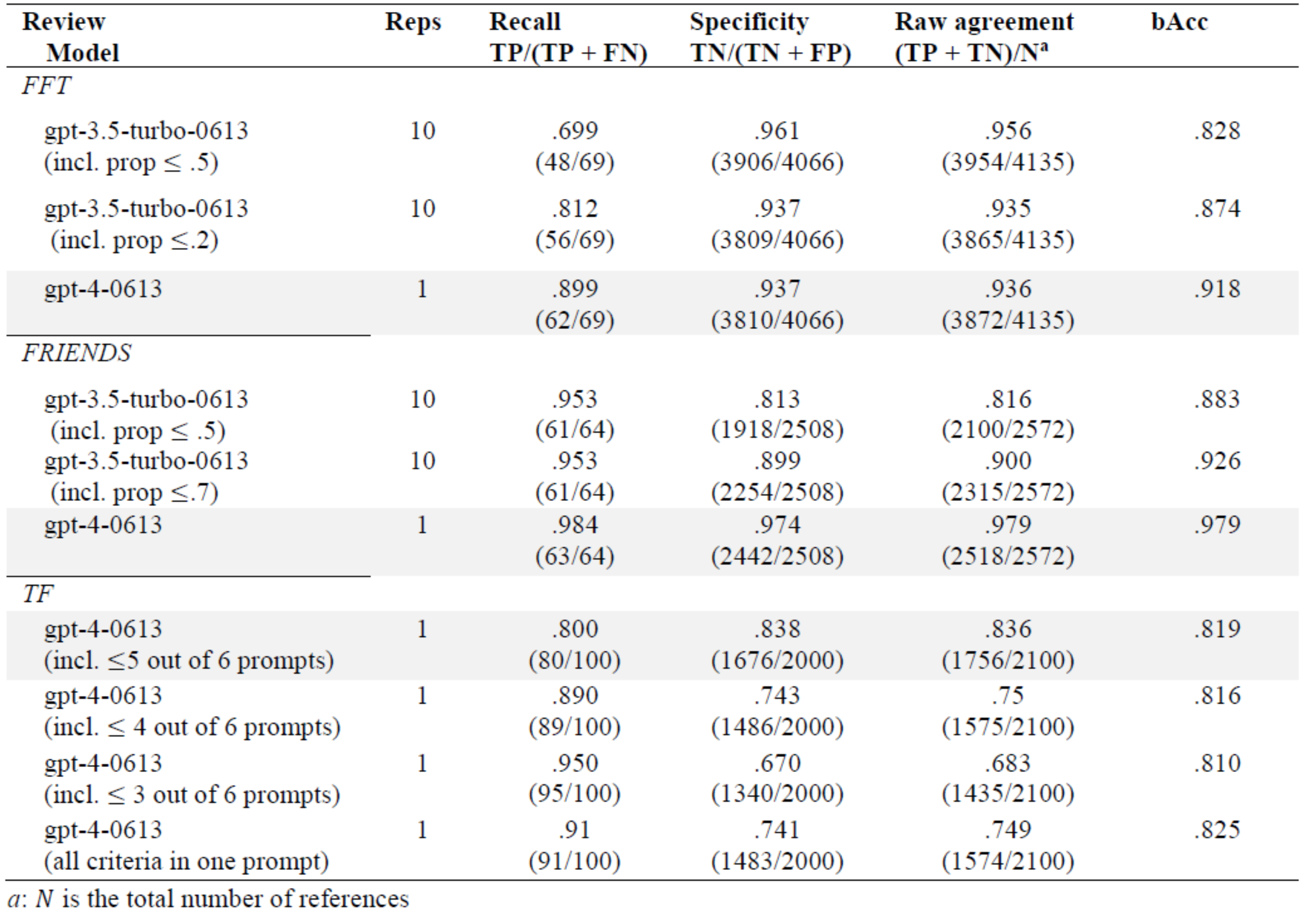
Recall is the proportion of relevant records being correctly classified as relevant, given by

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

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.

Appendix 4: Other possibilities with AIscreenR

- Can be combined with other (semi) automated screening tools such as priority and classifier screening.

- Used to reduce the number of studies needed to be humanly double screened. This can be done by making too over-inclusive prompt on purpose.

- Can be fine tuned to the specific review context