Using GPT API models for title and abstract screening in high-quality reviews

Mikkel Helding Vembye, PhD

INSIA, Prague, 09.09.2024

Common issue in large-scale reviews

Independent human double-screening of titles and abstracts is time-consuming and a resource-dependent procedure which

1. slows the review-process
2. (most often) forces reviewers to make too narrow search strings
3. is costly in terms of skilled human labor
4. makes some topics non-reviewable due to the number of references to screen for relevance

- Overlooking relevant studies at the initial review stage can be consequential, leading to substantially biased results

We suggest to alleviate this issue by substituting the human *second* screener with a GPT (Generative Pre-trained Transformer) API (Application Programming Interface) model.

VIVE

The empirical foundation

- We find 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).
- We conducted three large-scale classification experiments with different levels of complexity in terms of the number of inclusion criteria.
- In simple screening cases, we even find recall close to 100% and with a high specificity values (97.4%), as well.
- In complex review settings, we find the GPT-4 model to yield a recall of 80%.
- Yet, in complex review setting, the GPT-4 model is rather over-inclusive with a specificity of ~84%.
- However, we argue this is not as problem as long as the recall is high (i.e., on par with humans) since a low specificity does not induce any bias to a review.

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

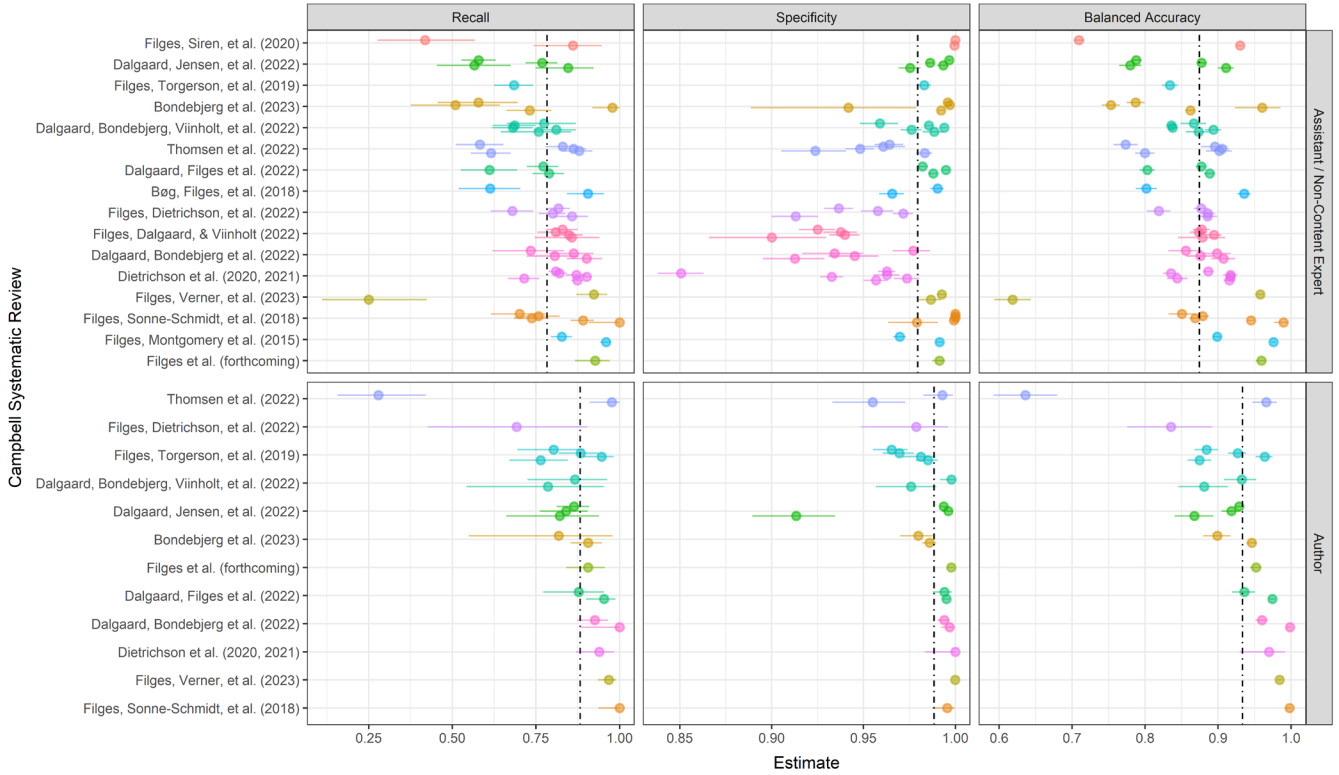
$$Recall = \frac{\{true\ positive\}}{\{true\ positive\} + \{false\ negative\}}$$

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

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

Typical human screening performance

- To make fair comparisons between GPT and human screening, we mapped common human screening performances across 22 high-quality reviews. Hereto, we found the typical *second* screener to have a recall of 78.2%, 95% CI[74.7, 81.7] and specificity of 98%, 95% CI[96.6, 99.0].



Note: Dashed lines indicate the average estimated via the CHE-RVE model. Each point represent an individual screener within the given review

The benchmark scheme

- In light of the typically human screening performances, we developed the following benchmark scheme. The aim is to help assessing screening performances of in general but also to judge when GPT screening is appropriate in high-standard reviews.

	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.

Standardization

To standardize this screening approach, we further developed

- 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.

- The AIscreenR R package (Vembye, 2024). This (among other things) allows the user to screen with multiple prompts and with parallel processing. To exemplify, we have been able to screen 12.000 reference with 6 prompt in less than 30 minutes (prize, $500 USD). Moreover, it includes a vignette with a practical/user-friendly step-by-step tutorial.

Concerns for future research

Some highlights for future research:

- Investigate how our result generalize to other (and cheaper) models, such as the GPT-4o and GPT-4-turbo models as well as models from other companies such as Claude 2 and Mistral AL.

- Test it with local models. This would freeze the efficacy of this approach and increase transparency of this approach

- Consider how best to combine traditional automated screening tools with GPT API screening. For instance can GPT API models play a role in validating stopping rules when using priority screening algorithms?

References:

Vembye, M. H. (2024). *AIscreenR: AI screening tools for systematic reviews.* (GitHub version 0.0.0.9999). https://mikkelvembye.github.io/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 (OSF).* https://doi.org/10.31219/osf.io/yrhzm

Appendix – all 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