We have built the AIscreenR as a flexible software, allowing users to conduct multiple screenings simultaneously based on multiple prompts, API models, iterations of the same request, and nucleus samples (i.e., different top\_p or temperature values). The software further allows the user to send the same request (i.e., repeatedly asking the same question) multiple times to avoid random noise in individual model responses (especially relevant when using GPT-3.5 models). When this feature is used, the final GPT decision is based on the probability of inclusion across the iterated requests, and the specific inclusion threshold can then be determined by the user (see Figure 4 of this paper). This also allows the users to test model response consistency. Moreover, the software has been built to draw on multi-core processing, thereby allowing users to significantly speed up the screening. Finally, we built the package so that reviewers can work with two different types of function calls; one yielding simple/trinary (i.e., 1 = {include}, 1.1 ={uncertain}, and 0 = {exclude}) decisions and/or another yielding descriptive responses to the screening requests. We consider the former to be the main work engine, whereas the latter can be pivotal when examining discrepancies between GPT and human screener decisions.

Although we have tried to accommodate the requirements set forth by evidence organizations, we do not consider our solution to be a final one. Our aim has merely been to show one way in which GPT API models can be used for TAB screening in large-scale systematic reviews that can inspire and be transferred to future applications of TAB screening with all kinds of LLMs

A side-effect of such research would further be that the costs of using GPT models may be substantially reduced, which can be a major barrier to using GPT-4 models for TAB screening at the current point in time. These models are still rather expensive (in absolute terms, not compared to hiring a human screener). Thus, another line for future research could be to investigate the performance of cheaper GPT-4 models, such as GPT-4o and GPT-4-turbo.

*Evidence from previous research about common human performance*:

In fact, research suggests that screeners on average tend to miss between 3% and 24% of all eligible studies depending on the level of content knowledge – often with a substantial impact on the final quantitative results (Buscemi et al., 2006; Waffenschmidt et al., 2019). In medicine, the number of missed studies may be even higher, especially when relying on student screeners (Ng et al., 2014).

An additional challenge is that most automated screening tools are based on supervised and active learning methods. This means that they need to be trained on a large enough set of included and excluded references to perform adequately, which is time-consuming as well. Finally, when automation tools are used for prioritized screening, there is no clear rule for determining when it is safe to stop screening with regard to finding all or close to all eligible references. Although various stopping rules have been proposed, the adequacy of these rules is sensitive to a range of factors, such as the length of the database, the prevalence of relevant studies, and the balance between relevant and irrelevant records (Campos et al., 2023; König et al., 2023; Van De Schoot et al., 2021)

By doing so, we provide a basis for making sense of specific TAB screening performances, which is needed in order to meaningfully assess the fruitfulness of using GPT models (and other LLMs) as independent TAB screeners in systematic reviews.

This seems to create a screening paradox, which might be one of the main reasons for reviewers to mistrust the application of machine-learning tools (O’Connor et al., 2019). While trying to reduce selection biases caused by single screening, automated screening potentially introduces a novel type of publication/selection bias, referred to by König et al. (2023) as ‘artificial screening bias’.[[1]](#footnote-1)

Moving on to the *balanced accuracy* metric, we found average performance levels of 0.874, 95% *CI*[0.857, 0.890] among assistant screeners and 0.933, 95% *CI*[0.899, 0.961] among author screeners. We found the difference between the group means to be statistically significant with *F*(1, 10.1) = 18.22, *p* = .002. Finally, the overall *nMCC* values were 0.860, 95% *CI*[0.835, 0.882] and 0.925, 95% *CI*[0.880, 0.953] for the assistant and author screeners, respectively. The group averages were found to be statistically different from each other with *F*(1, 11) = 9.65, *p* = .01.

The average balanced accuracy and *nMCC* values of the NIPH researchers were 0.905, 95% *CI*[0.859, 0.943] and 0.879, 95% *CI*[0.720, 0.951], respectively.

For the *nMCC* metric, we calculated the sampling variance and confidence interval by transforming the correlations to Fisher’s z-scores, as typically done in meta-analysis (Borenstein et al., 2009).

A side-effect of conducting these experiments was further to quality assure the AIscreenR package (Vembye, 2024) for automated TAB screenings using GPT API models. We consider this test to be important in order to be able to scale up our suggested screening approach.

An impressive fact regarding these results is further that we approximately spent 5-10 minutes engineering the used prompt presented in Textbox 1. In fact, the prompt represents a first trial prompt. Unless we were extremely lucky to hit the right prompt in the first trial, this might indicate that in some applications the GPT API models are not as prompt-sensitive as would be theoretically expected. Yet, we only presume this to be the case in very specific circumstances as is the case here where the screening involves a standardized intervention with a specific name.

FIGURE 4: Decision sensibility of the gpt-3.5-0613 model across inclusion probability thresholds

|  |
| --- |
| A |
| B |

*Note*: The inclusion probability threshold on the x-axis indicates the required number of times the given title and abstract record needed to be included over the 10 repeated requests in order to be coded as relevant. For example an inclusion probability threshold of 0.1 means that the record was coded as relevant if the GPT model included it in 1 or more out of 10 requests.

1. To alleviate this issue, a new tentative guideline termed SAFE has been developed in which it is suggested to use multiple machine learning algoritmes in order to detect all relevant references in the bulk of records (Boetje & van de Schoot, 2024). However, this framework has not yet been thouroughly enough tested to know if the SAFE procedure allows reviewers to detect all relevant studies using machine learning algoritms included in screening softwares, such as ASReview (Van De Schoot et al., 2021). [↑](#footnote-ref-1)