**8 CONCLUSION AND DISCUSION**

Human duplicate title and abstract (TAB) screening in systematic reviews is time-consuming, requiring a substantial amount of human labor which decelerates the review process and thereby the dissemination of important knowledge for practice, research, and policy. In this study, we have shown that OpenAI’s GPT API models can function as highly reliable second screeners even in complex review settings, making it possible to substitute one human in the duplicate screening process and reallocate human resources. Our findings suggest that when configured correctly GPT API models can perform on par with or even surpass human screeners with regard to finding relevant studies. Moreover, we found that the GPT-4 model outperforms the GPT-3.5-turbo model and therefore recommend primarily using the GPT-4 model for GPT API screening. Moreover, we found that GPT API models *can* yield specificity rates that are on par with humans, but in some applications appear to be slightly over-inclusive (i.e., they yield lower specificity rates than typical human screeners). We do, however, not necessarily consider this a deficit as long as the models obtain high recall rates. Thus, low specificity levels do not come with a risk of biased results – they merely force human reviewers to double-check a higher number of records.

Our results contrast previous research in which GPT API models were found to perform well in terms of specificity but less well in terms of recall (Gargari et al., 2024; Guo et al., 2024). While it would be premature – based on our data – to make hard conclusions about the reasons for the higher performance (in terms of recall) in our classification experiments, some differences are worth noting in terms of our workflow and the workflow of prior evaluations. First, as noted earlier, contrary to prior evaluations of GPT API models for TAB screening, we relied on function calling (OpenAI, 2024), thereby improving the models’ response consistency. Moreover, in experiment 3 (our most complex case), instead of adding all inclusion/exclusion criteria to the same prompt, we introduced and used multi-prompt screening, using one concise prompt per inclusion/exclusion criteria in the review. This may further have contributed to the higher levels of recall compared to prior examinations.

Based on our findings, we believe TAB screening with GPT API models can revolutionize the way duplicate title and abstract screening is done in high-quality systematic reviews by making automated tools move up to the highest levels of automation (c.f. O’Connor et al., 2019), where they yield none human-assisted *second* screener decisions. However, this necessitates the need to standardize this screening approach to make it scalable and acceptable in high-quality reviews. Therefore, we also developed a reproducible workflow and tentative guidelines for when such screenings can be accepted in high-quality reviews. To help making automated, GPT API model-based screenings available, we developed the AIscreenR R package, thereby allowing reviewers to draw on features such as function calling (i.e., making prompts without the need to explicitly specify how the model shall respond to the screening request) as well as multi-core processing, something which speeds up the screening significantly.

A key part of setting up a reliable GPT API screening is to thoroughly validate the screening prompt(s) before making any full-scale screening. For such assessments, we developed a new, empirically informed benchmark scheme for interpreting acceptable and unacceptable screening performance in high-quality reviews based on the typical screening performance found in 22 high-standard systematic reviews across 157,828 duplicate human screening decisions. Deduced from this investigation, we suggest that if automated screenings yield a recall rate (i.e., the ability to correctly include relevant studies) above 80%, it should be acknowledged as being on par with typical human performance and can be safely be used as an independent second screener. This recommendation approximately resembles the average recall rate we found in the mapped high-quality, human-conducted reviews. In addition, we suggest that a specificity rate (i.e., the ability to correctly exclude irrelevant studies) equal to or above 80% should be accepted in high-standard reviews as long as the recall is equal to above 80% as well since a low specificity rate does not induce any biases (it merely requires humans to double-check a higher number of records, which is annoying and time consuming but not a threat in terms of biasing the results).

It is important to note that no matter how much effort is invested in developing a good prompt, GPT API models – like humans – can err and therefore, it is of vital importance that GPT API screening is combined with other screening techniques such as forward and backward citation tracking to ensure that potentially missed studies re-enter the review. In that regard, GPT-based screenings are not different from screenings conducted by humans. Although our recommendations allow for minor errors, we recommend not to use GPT API screening, if reviewers can reach satisfying recall and specificity rates. In a similar vein, we never think a GPT API model should be used as a stand-alone screener. There must always be a human in the loop, meaning that humans must always take the role of the first screener of titles and abstracts in high-quality systematic reviews.

Some caveats and limitations follow our work. First of all, we agree with Schoot et al. (2021) that transparency and reproducibility represent the highest scientific standards. Yet, OpenAI’s GPT API models are based on black box algorithms. Nonetheless, we do not believe that this argument should prevent reviewers from using OpenAI’s GPT API models for TAB screening – for instance, human screening decisions most often represent black-box operations as well. However, we consider it all-important that future research investigates the performance of alternative open-source GPT models. 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.

A general challenge when using GPT API models is that it requires a substantial amount of software maintenance to keep up to date with the newest model developments. Therefore, it requires continuous software development, for this screening approach to be viable which, in turn, will probably require collaborations in the research community to ensure the stability of the software over time.

Although this study has some important limitations, we believe that the implications of this work are rather extensive beyond what we have presented. First, using well-functioning automated tools renders the possibility for reviewers not to make unnecessary restrictions on their search string to steer the number of study records, which, in turn, increases the likelihood of finding all or close to all relevant studies for the review. Moreover, it makes it possible to screen literature for extreme-sized reviews (Shemilt et al., 2014, 2016) that would otherwise have been considered unmanageable by humans. Second, this approach can be all-important in elevating the quality of reviews conducted by single researchers restricted by resources such as low budgets and/or time. Third, we believe that a huge potential exists in combining traditional automated tools and GPT modeling. For example, GPT API models could play a key part in validating a decided stopping rule (Campos et al., 2023; König et al., 2023) whereto it could partly be used to screen records close to the stopping rule on the wrong side, and partly be used to more precisely detect relevant studies on the right side of a given stopping rule, thereby reducing the risk of relevant studies being overlooked. Combining traditional tools and GPT screening could furthermore reduce the cost of using GPT API models. Another application could also be that GPT API models are used together with prioritization resampling algorithms such as the one suggested by Hou and Tipton (2024) to come closer to reaching recall rates closer to 100%, which are generally considered unattainable when using stochastic algorithms. Fourth, even if reviewers prefer to use duplicate human screening, we think that using a GPT API model as a third screener would be valuable since it can guard against missing relevant studies due to human screener drifting.

To recapitulate, we believe that using GPT API models can change duplicate TAB screening in high-quality reviews across all kinds of scientific disciplines. In fact, we envision that the GPT-4 models will perform even more adequately when used on more structured abstracts as typically found in medicine. Moreover, we think this is an ideal use case where artificial intelligence (AI) can meaningfully take on tedious human labor, and where no legal issues arise. Even more edifying, GPT API model screening can ensure a more rapid transfer of key knowledge to research, practice, and policy, which ultimately underpins the core rationalefor doing systematic reviews.