Enhancing title and abstract screening for systematic reviews with GPT-3.5 turbo

Omid Kohandel Gargari, Mohammad Hossein Mahmoudi, Mahsa Hajisafarali, Reza Samiee

After conducting a database search, the subsequent phase in the execution of systematic reviews (SRs) involves title and abstract screening. This stage bears significant importance and necessitates the involvement of dedicated and experienced researchers who can exhibit precision and accuracy, particularly when the search yields a substantial number of studies. Besides the qualities of experience and dedication demonstrated by the screeners, several other factors influence the quality of the screening process, such as effective team management, the adoption of a double-screening approach and, notably, the implementation of a well-structured screening design. A screening tool comprises a set of questions that must be addressed by the screeners, and these questions should adhere to the following criteria: (1) they must be objective, (2) they should be single-barreled and (3) they should encompass questions answerable with ‘yes’, ‘no’ or ‘unsure’ responses.2

The domain of large language and transformer models has showcased a promising trajectory of advancement, consistently improving day by day. These models are trained on a vast corpora of text and possess the capability to comprehend and generate human-like text.3 A prominent example within this realm is the Generative Pre-Trained Transformer (GPT) developed by OpenAI, with the latest iteration being GPT-4 at the time of composing this discourse. GPT-4 has exhibited commendable performance across a range of human-related tasks and has surpassed its predecessor, GPT-3.5, in evaluations conducted by the company.4

This single-case study was conceived to assess the performance of GPT 3.5 in the context of title and abstract screening for SRs. To execute this task, a recently published SR titled ‘Light Therapy in Insomnia Disorder: A Systematic Review and Meta-Analysys’ was selected, and the databases were queried using the keywords stipulated in the original paper.5 Two key rationales underpinned the selection of this review: first, it yielded a relatively moderate number of studies, and second, its eligibility criteria were somewhat subjective, and challenging to discern during the screening process, making it a suitable testbed to evaluate GPT-3.5’s capabilities. For instance, this study enrolled patients experiencing sleep difficulties but did not specify the particular types of sleep disorders although reviewers did not face much trouble but models had difficulties with studies that included patients with secondary sleeping troubles like patients with cancer. Furthermore, it was unclear which specific light therapy was chosen for inclusion.

The initial search yielded 330 citations, which were subsequently imported into EndNote X20. An RTF file containing titles and abstracts was generated, followed by its conversion into a more compatible TXT format, thus facilitating further data processing. This transformation laid the foundation for our experimental data set, comprising the research paper titles, abstracts and accompanying metadata. The screening team consisted of three researchers: (1) an expert with screening experience from over 20 SRs, (2) a senior researcher with screening experience from 10 SRs and (3) a junior researcher without any prior screening experience. All researchers possessed a strong command of the English language and a thorough understanding of SR principles.

The senior and junior researchers remained unaware of the study’s objectives. The three screeners were tasked with screening the titles and abstracts in accordance with the criteria outlined in the online supplemental table 1.

To compare the performance of eligibility screening of the human screeners with the performance of GPT-3.5, a range of prompts were devised for GPT by two of the authors (OKG and MMH), these prompts were carefully chosen during several discussion sessions. A prompt is a specific input or instruction given to a language model, to generate a desired output or response. The integration of the OpenAI GPT 3.5 Turbo API played a pivotal role in advancing our research. This powerful tool enabled us to initiate requests to evaluate the pertinence of prompts to individual papers. The binary relevancy results of this interaction were recorded in a structured Pandas data frame that had been prepared in advance. The code for this process is available at the provided link: https://github.com/mamishere/Article-Relevancy-Extraction-GPT3.5-Turbo.

GPT evaluated the eligibility of studies based on the provided prompts, resulting in the creation of numpy arrays containing binary outcomes for each prompt response. These numpy arrays, along with the numpy arrays generated by the researchers, were employed to compute sensitivity, specificity, accuracy and the F1 score for both the researchers and the prompts. The labels
used as the ‘gold standard’ were the studies included in the selected SR.

Prompt 1, which replicated the criteria from the original paper, demonstrated 80% accuracy and 62% sensitivity. In contrast, Prompt 2 broadened the population by focusing on patients with ‘sleep troubles’ instead of providing a specific definition, leading to reduced accuracy, sensitivity, and specificity.

For subsequent prompts, GPT was assigned the role of an ‘Experience Systematic Researcher’ (Prompt 3), which increased accuracy while decreasing sensitivity. Prompt 4 introduced an ‘inclusivity sentence’ to instruct the model to include studies it was uncertain about, prioritising inclusivity, leading to increased sensitivity and reduced specificity.

Prompt 5, which combined the original criteria with the inclusion of an ‘inclusivity sentence’, demonstrated the highest sensitivity, similar to the junior researcher and surpassing the senior researcher. Prompt 6, which omitted segmented criteria in favour of a more descriptive sentence, resulted in a significant reduction in sensitivity, suggesting that GPT responded better to segmented and classic criteria.

Prompt 7 assessed the impact of phrasing by modifying the language of the previous prompt and providing a more detailed description of the inclusivity phrase. This modification substantially increased the model’s sensitivity.

Lastly, Prompt 8 introduced a screening tool with four questions, requiring the model to include a study if the answer to all four questions was ‘Yes’ or ‘Unsure’. Surprisingly, the model performed poorly in this format, potentially due to the technical structure of GPT.

Our results indicate that more specific and detailed prompts elicit better responses, whereas broader and less specific prompts should be avoided, see online supplemental table 2 for performance metrics of 8 prompts three human screeners. We also recommend the incorporation of a sentence prioritising inclusivity to guide the model in including unclear studies.

GPT is a potent tool, and we propose its usage in title and abstract screening for SRs, following the method we have delineated in this case report, alongside other researchers. However, it is imperative to recognise that GPT is not yet fully capable of independently completing this task and should be employed as an assistant to mitigate the risk of overlooking potential studies.

Notably, even the human researchers did not attain exemplary performance, attributable to the inherent challenges posed by the subjective criteria and the absence of clear, objective definitions. We advise researchers wanting to deploy GPT to manually screen a proportion of titles and abstracts, experiment with different prompts and consider the combination of findings, a method unexplored in our study. It is crucial to first establish a clear study objective before designing prompts.

The most formidable challenge encountered in this study pertained to defining the population for the model. For instance, the lack of clarity in the criteria for ‘sleep troubles’ resulted in substantial bias in the model’s performance and significant disparities among researchers. It needs to be acknowledged that the findings of this single-case study are not generalisable, and each study objective necessitates its distinct format. This study serves as an illustrative example and offers guidance for replication with more cases and further research on the topic.

The prompt texts are available in the online supplemental table 1.

References

ORCID iDs
Omid Kohandel Gargari http://orcid.org/0000-0002-8182-0582
Reza Samiee http://orcid.org/0009-0006-4862-1131
### Supplementary table 1. Prompts text

<table>
<thead>
<tr>
<th>Prompt number</th>
<th>Prompt text</th>
</tr>
</thead>
</table>
| 1             | We are trying to conduct a PRISMA systematic review study titled: "Light therapy in insomnia disorder"  
Check the title and abstract if it should be selected for further assessment and full-text screening. Provide you answer only as 0 or 1.  
Our inclusion criteria are the followings:  
Criterion A  
Must have enrolled only patients who had been diagnosed with chronic insomnia disorder. We also include studies in which insomnia diagnosis was not based on proper classification criteria, but was based on “sleep troubles” including at least one insomnia complaint, without any other diagnosis of sleep disorder.  
Criterion B  
Studies assessing the efficacy of light therapy.  
Criterion C  
Publication written in English.  
Criterion D  
Study design could be observational or interventional. |
| 2             | We are trying to conduct a PRISMA systematic review study titled: "Light therapy in insomnia disorder"  
Check the title and abstract if it should be selected for further assessment and full-text screening. Provide you answer only as 0 or 1.  
Our inclusion criteria are the followings:  
Criterion A  
Studies including patients with sleep troubles.  
Criterion B  
Studies assessing the efficacy of light therapy.  
Criterion C  
Publication written in English.  
Criterion D  
Study design could be observational or interventional. |
| 3             | You are an experienced systematic review researcher and you are trying to conduct a |
### PRISMA systematic review and meta-analysis
titled:
"Light therapy in insomnia disorder"

Check the title and abstract if it should be selected for further assessment and full-text screening. Provide you answer only as 0 or 1.

Our inclusion criteria are the followings:

- **Criterion A**
  - Studies including patients with sleep troubles.

- **Criterion B**
  - Studies assessing the efficacy of light therapy.

- **Criterion C**
  - Publication written in English.

- **Criterion D**
  - Study design could be observational or interventional.

---

You are an experienced systematic review researcher and you are trying to conduct a PRISMA systematic review and meta-analysis titled:
"Light therapy in insomnia disorder"

Check the title and abstract if it should be selected for further assessment and full-text screening. Provide you answer only as 0 or 1.

Our inclusion criteria are the followings:

- **Criterion A**
  - Studies including patients with sleep troubles.

- **Criterion B**
  - Studies assessing the efficacy of light therapy.

- **Criterion C**
  - Publication written in English.

- **Criterion D**
  - Study design could be observational or interventional.

If the title and abstract were not clear about criterion A, B or D it is preferred to include the study.

---

Title: "Light therapy in insomnia disorder"

You are conducting a PRISMA systematic review and meta-analysis to assess the efficacy of light therapy in patients with insomnia disorder. Your goal is to identify as many relevant studies as possible to ensure a comprehensive review.

Check the title and abstract if it should be selected
Inclusion Criteria:

Criterion A
Studies involving patients with any form of sleep disturbances or insomnia disorder, regardless of sleep disorder diagnoses.

Criterion B
Studies assessing the efficacy of light therapy.

Criterion C
Publications written in English.

Criterion D
Study design could be observational or interventional.

Your primary objective is to avoid missing any potentially relevant studies. If the title and abstract are not explicit about meeting the inclusion criteria, it is preferred to err on the side of inclusivity and include the study for further assessment.

You are an experienced systematic review researcher and you are trying to conduct a PRISMA systematic review and meta-analysis titled: "Light therapy in insomnia disorder"

Check the title and abstract if it should be selected for further assessment and full-text screening. Provide you answer only as 0 or 1.

You are looking for English original studies working on the effect of light therapy in patients with sleeping problems.
It is important not to miss possible relevant studies.
It is preferred to include studies you are not sure about them.

Prompts made by GPT
English original studies focusing on this topic.

Please be cautious and prioritize inclusivity. If the title and abstract hint at potential relevance, even if the criteria are not explicitly stated, include the study for further assessment.

| You are an experienced systematic review researcher and you are trying to conduct a PRISMA systematic review and meta-analysis titled: |
| "Light therapy in insomnia disorder" |
| Based on the guideline below, check the title and abstract to determine if the study should be selected for further assessment and full-text screening. |
| Provide your answer only as 0 or 1 |
| guideline: |
| 1. Does the title or abstract use English? |
| a. Yes: continue screening |
| b. No: stop screening |
| 2. Does the title or abstract indicate that the study is original and not a review. |
| a. Yes: continue screening |
| b. No: stop screening |
| 3. Does the abstract indicate that patients with sleep troubles were included. |
| a. Yes or Unsure/Unclear: continue screening |
| b. No: stop screening |
| 4. Does the abstract indicate that any type of light therapy was studied? |
| a. Yes, or Unsure/Unclear: continue screening |
| b. No: stop screening |
| Decision: Should this article be included? |
| a. Yes, all 4 screening questions answered Yes or Unclear |
| b. No, at least one answers definitely “No” |

**Your answer should be a binary that determines the final decision output**
<table>
<thead>
<tr>
<th>Metric</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
<th>P7</th>
<th>P8</th>
<th>JUN</th>
<th>SEN</th>
<th>EXP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>0.62</td>
<td>0.56</td>
<td>0.50</td>
<td>0.56</td>
<td>0.69</td>
<td>0.38</td>
<td>0.62</td>
<td>0.38</td>
<td>0.69</td>
<td>0.62</td>
<td>0.81</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.81</td>
<td>0.59</td>
<td>0.69</td>
<td>0.45</td>
<td>0.25</td>
<td>0.67</td>
<td>0.46</td>
<td>0.85</td>
<td>0.73</td>
<td>0.71</td>
<td>0.73</td>
</tr>
<tr>
<td>F1 Score</td>
<td>0.24</td>
<td>0.12</td>
<td>0.13</td>
<td>0.09</td>
<td>0.08</td>
<td>0.09</td>
<td>0.10</td>
<td>0.17</td>
<td>0.19</td>
<td>0.17</td>
<td>0.23</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.80</td>
<td>0.59</td>
<td>0.68</td>
<td>0.46</td>
<td>0.27</td>
<td>0.65</td>
<td>0.46</td>
<td>0.82</td>
<td>0.72</td>
<td>0.70</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Supplementary Table 2: Performance metrics of 8 prompts and three researchers