AUTOMATING LITERATURE SCREENING WITH LARGE LANGUAGE MODELS

Semerikov S. O.[1,2,3,4,5], **Mintii I. S.**[3,1,6,2,5]
[1] Kryvyi Rih State Pedagogical University, Kryvyi Rih, Ukraine
[2] Zhytomyr Polytechnic State University, Zhytomyr, Ukraine
[3] Institute for Digitalisation of Education of the NAES of Ukraine, Kyiv, Ukraine
[4] Kryvyi Rih National University, Kryvyi Rih, Ukraine
[5] Academy of Cognitive and Natural Sciences, Kryvyi Rih, Ukraine
[6] Lviv Polytechnic National University, Lviv, Ukraine

Abstract. Screening research papers for inclusion in a literature review is a time-consuming manual process. We explore automating this process using OpenAI's GPT-3.5 Turbo large language model (LLM). Given text prompts specifying the inclusion/exclusion criteria, the LLM evaluated the abstract of each paper. It is classified into one of four categories: meeting both criteria, violating the first criteria, violating the second criteria, or violating both criteria. Our Python code interfaced with the OpenAI API to pass paper abstracts as prompts to the LLM. For 347 papers, the LLM flagged 173 as meeting the criteria, with 3 additional papers included after accounting for missing abstracts, yielding 176 papers selected for full-text retrieval. A manual review of a sample suggested reasonable accuracy. While further validation is needed, this demonstrates LLMs' potential for accelerating systematic literature reviews.

Keywords: large language models, GPT-3, literature review, automation, screening, inclusion criteria.

Introduction. Conducting a systematic literature review involves screening many research papers to identify those meeting specified inclusion criteria. This manual process is extremely time-consuming and labour-intensive [1]. Recent advances in large language models (LLMs) have demonstrated their capability to understand and reason about natural language [2]. We explore leveraging OpenAI's GPT-3.5 Turbo LLM to automate literature screening based on predefined inclusion/exclusion criteria.

The goal of the work is to explore methods for automating key stages of the systematic literature review process using advanced natural language processing capabilities. Conducting a comprehensive systematic review involves many laborious and time-intensive steps, including:

1. Defining the review scope and inclusion/exclusion criteria.
2. Searching academic databases to identify potentially relevant studies.
3. Screening the titles, abstracts, and full texts to determine which studies meet the criteria.
4. Extracting data from the included studies.
5. Synthesising and analysing the extracted information.

Each of these steps is currently performed primarily through manual human effort. Our work uses recent advances in LLMs and other AI/NLP techniques to automatically or semi-automatically perform as many steps as possible. This can drastically reduce the researcher's time and effort for systematic reviews.

Formulation of the problem. The specific problem we tackle in this paper is automating the screening stage (step 3 above) by using an LLM to evaluate study abstracts against predefined

inclusion/exclusion criteria. This screening stage is a critical bottleneck, requiring reviewing thousands of abstracts individually.

Formally, given:

– A database of N research papers/studies represented by their abstracts $\{A_1, A_2, ..., A_N\}$

– Inclusion criteria $C_1, C_2, ..., C_m$

– Exclusion criteria $E_1, E_2, ..., E_n$

We want a model M that can take each abstract A_i and criteria $\{C_1...C_m, E_1...E_n\}$ as input and output a label L_i indicating if the study meets the criteria or not:

$M(A_i, \{C_1...C_m, E_1...E_n\}) = L_i$,

where L_i can take on one of k values:

– Meets all inclusion criteria and no exclusion criteria

– Violates inclusion criteria C_j

– Violates exclusion criteria E_k

– Violates some inclusion and exclusion criteria

We hypothesise that large language models pre-trained on vast text corpora can effectively perform this classification by understanding the semantic meaning of the abstracts and criteria.

Solving the problem. We implemented a Python script using the OpenAI API to interface with the GPT-3.5 Turbo LLM. The script reads a set of 347 papers with their abstracts, paper IDs, and the two inclusion criteria:

C_1: Publication focuses on teacher training or professional development in AI-powered personalized learning

C_2: Population is students in teacher training programs or teachers

For each paper, the abstract text was provided as a prompt to the LLM along with instructions to analyze whether the criteria were met, violated, or if there was not enough information. The LLM response classified each paper into one of four categories: meeting both criteria ("1 1"), violating criteria C_1 ("0 1"), violating criteria C_2 ("1 0"), or violating both criteria ("0 0").

The LLM classified 173 out of 347 papers as meeting both inclusion criteria. Additionally, 3 papers were incorrectly identified as excluded due to the absence of abstracts in the data source, for a total of 176 papers selected for full-text retrieval after accounting for this error (table 1).

Table 1

Results of screening using LLM

Screening results	Number of papers
Excluded by violation of criteria C_1	44
Excluded by violation of criteria C_2	109
Excluded by violation of both criteria	21
Included	176

We manually reviewed a random sample of 50 papers across all four categories to evaluate the accuracy. For papers meeting the criteria, the accuracy was 84%. More false negatives (16%) were observed than false positives (4%).

Conclusions. Several limitations of this study should be noted. First, the criteria statements were relatively simple semantic criteria about the research focus and population. More complex criteria involving research design, intervention details, statistical analyses, etc., may challenge the current LLM capabilities. However, performance could be improved through prompt engineering or finetuning the model on systematic review data. Additionally, we did not extensively evaluate the LLM's performance across different research domains. The model was given the inclusion/exclusion criteria as prompts without domain-specific training. Customising or finetuning the LLM for each research field may boost effectiveness. Despite these limitations, this proof-of-concept study highlights the immense potential of leveraging advanced AI language models to accelerate evidence synthesis work. As language models advance rapidly, they will likely play an increasingly prominent role in streamlining systematic reviews and other critical research activities.

Promising directions include using LLMs for other systematic review stages like full-text screening, data extraction from studies, analysing the risk of bias and even intelligently synthesising

findings across studies. Exploring how LLMs can semi-automate the systematic review process in a robust, validated pipeline is an exciting area for future research with profound implications for accelerating scientific discovery and evidence-based decision-making.

References

1. Mintii, M.M., 2023. Exploring the landscape of STEM education and personnel training: a comprehensive systematic review. Educational Dimension, 9, pp.149–172. Available from: https://doi.org/10.31812/ed.583
2. Hamaniuk, V.A., 2021. The potential of Large Language Models in language education. Educational Dimension, 5, pp.208–210. Available from: https://doi.org/10.31812/ed.650