Screening Articles in a Qualitative Literature Review Using Large Language Models: A Comparison of GPT4 versus Fine-tuned Open-Source Models Using Expert-annotated Data

Stacie Hudgens[1], Lucy Lloyd-Price[2], Milad Nourizade[3], Claire Burbridge[2], Kristian Thorlund[4]

[1] Clinical Outcomes Solutions Inc., Tucson, AZ, US; [2] Clinical Outcomes Solutions Inc., Folkstone, UK; [3] BioSpark AI Technologies Inc, Vancouver, BC, CA; [4] McMaster University, Hamilton, ON, CA

#CO83

Background

- Clinical Outcome Assessment (COA) landscape reviews begin with the critical activity of identifying proximal and distal concepts important to patients for inclusion of disease-specific patient focused outcomes in clinical trials. They are, however, highly time consuming to conduct by a research team. This, coupled with rapid or fast track clinical trial development timelines, any tool, technique or approach that could substantially reduce the time to completion of a COA landscape review would be of tremendous value.
- Artificial intelligence (AI), and particularly large language models (LLMs) have shown great promise for co-piloting human tasks that involve processing of large amounts of text. This applies to many tasks involved with COA landscape reviews. Generative AI chatbots like ChatGPT have become highly popularized but have a more general “comprehension” and are not trained or optimized for specific tasks. Bespoke models fine-tuned from OpenSource language models may thus prove a better solution.

Objective

- To assess 2 AI models’ performance for literature screening to identify relevant qualitative research that can be used to develop COA conceptual models. We also compared run-time for both models.

Methods

- We manually curated a dataset of PubMed references including n = 1,300 independent study titles and abstracts. These were obtained from 17 previously conducted landscape reviews across oncology, rheumatology, dermatology, and rare diseases. Each citation was annotated for eligibility (Y/N) by population, study design (qualitative), and reporting of candidate concepts (how patients feel or function). Each reference was screened in duplicate and disagreements were resolved via discussion (*comparison to researcher is not presented in this poster*).
- We developed 2 LLM approaches for screening references. First, we iteratively engineered a series of prompts using GPT4 (OpenAI). Second, we fine-tuned an existing open-source biomedical language model, SciFive, using data from the 17 previously conducted landscape reviews. We used 70% of the data for training the fine-tuned model and 30% for test. Both LLMs were set up to predict “Y/N” eligibility by population, study design, and candidate concepts independently. We compared the performance of the 2 LLMs by obtaining Precision (% true positives), Recall (positive prediction value), F1-score (harmonic mean of precision and recall), and accuracy (% correct predictions).

Results

Figure 1. F1-Score and Precision for Population Eligibility

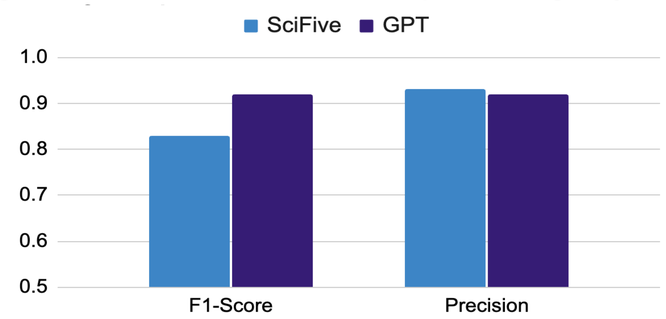


Figure 2. Performance Metrics for Concepts Eligibility Screening

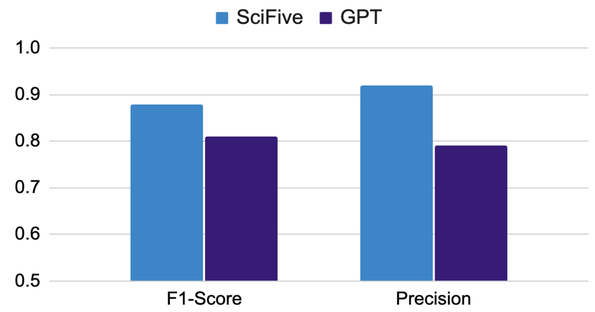
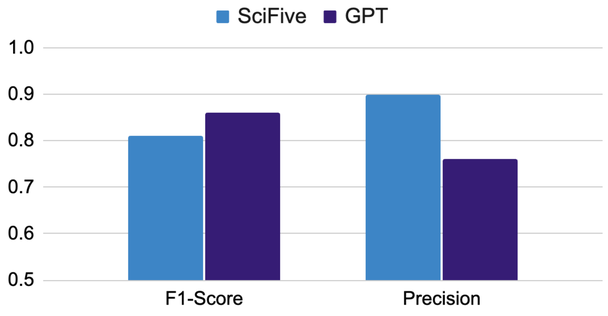
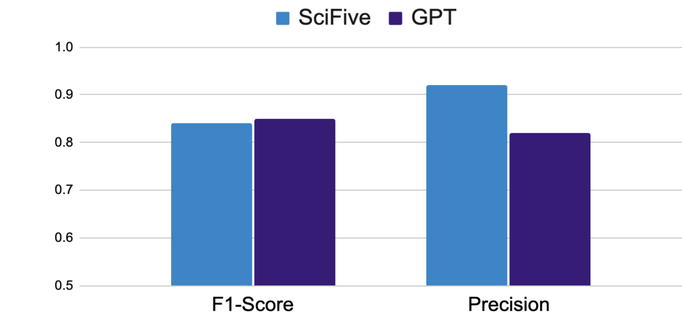


Figure 3. Performance Metrics for Study Design Eligibility Screening



Results

Figure 4. Overall Eligibility Screening Performance



- Both models performed well for assessing relevance by population. The F1-scores of the fine-tuned SciFive model and the GPT4 prompts, respectively, were 0.83 and 0.92, whereas the model precision estimates were 0.93 and 0.92 (see Figure 1).
- For concept reporting the fine-tuned SciFive model outperformed the GPT4 prompts with an F1-score and precision 0.88 and 0.92 versus 0.81 and 0.79 Figure 2).
- The same was true but less pronounced for eligibility by study design, with an F1-score and precision 0.81 and 0.90 versus 0.86 and 0.76 (Figure 3).
- For overall eligibility, the fine-tuned SciFive model outperformed the GPT4 prompts with an F1-score and precision of 0.84 and 0.92 versus 0.85 and 0.82 (Figure 4).
- It took the GPT4 prompts between 10- to 30-minutes to screen 100 abstracts. By contrast, the fine-tuned SciFive model took 1- to 2-minutes on a computer with a Quadro RTX 8000 GPU.

Conclusion

- Both the fine-tuned SciFive model and the GPT4 prompts appear promising. However, given that the former was trained on a limited size data set, fine-tuned open-source models appear to hold more promise than GPT4 prompt that are based on a very large language model's comprehensive, but not specialized, understanding of language.
- The next phase of work will be to compare the AI models to the researcher screening results.

This study was sponsored by COA-AI | Editorial and programming support was provided by COA-AI.

Presented at the International Society for Pharmacoeconomics and Outcomes Research 2024, Atlanta, GA, US, 5-8 May 2024.