Automated title and abstract screening for scoping reviews using the GPT-4 Large Language Model

Introduction

A scoping review is a relatively novel type of literature review that aims to map the key concepts and existing activity within an area of research [@Arksey.2005]. Like systematic reviews, scoping reviews typically use rigorous, transparent [@Pham.2014], and sometimes pre-registered methods for gathering and synthesising evidence, and are increasingly using formal frameworks for both performing and reporting reviews [@Peters.2021]. Scoping reviews can inform future systematic reviews or primary research in the same area [@Sutton.2019]. However, they differ from systematic reviews in focussing on describing the breadth of coverage of the available literature rather than research findings in depth [@Arksey.2005].

Frameworks for performing a scoping review typically involve defining a research area or question, searching bibliographic databases for potentially relevant published material ('sources'), screening these sources to identify those relevant to the area or question, and systematically extracting and reporting data from the sources [@JBI.2015, @Arksey.2005]. The screening stage will usually involve initial screening of source titles and abstracts against pre-determined inclusion and exclusion criteria, followed by screening of the full text of sources, with both steps performed in replicate by at least two human reviewers [@Peters.2020, @Pham.2014]. Because database searches can return many hundreds or thousands of potentially relevant sources, these screening steps can require intensive human effort. Many software methods have been proposed or used to support or partially automate this process, including text mining to prioritise potentially more relevant sources for human screening [@Shemilt.2014, @Howard.2016, @Chai.2021], automated clustering and labelling of sources to support human decision-making [@Stansfield.2013], and 'crowdsourcing' screening to untrained workers via online platforms [@Mortensen.2017]. A similar but more extensive set of methods have been developed and employed for systematic reviews [@Khalil.2022, @Gates.2019] for which the process of source screening is broadly comparable.

Since the release of the first Generative Pre-trained Transformer (GPT) Large Language Model (LLM) by OpenAI in 2018 [@Radford.2018], transformer-based LLMs and the GPT lineage in particular have seen rapid and widespread adoption for a range of tasks. Broadly, these models generate a probabilistically weighted list of tokens (parts of text such as letter combinations and punctuation) which might follow or complete some input text (a 'prompt'), having been trained to do so by processing large human-written corpora. When this generative process is iterated, it allows for a range of applications involving analysis and production of text, such as summarising articles, generating fiction in a specified genre or style, or engaging in conversation with a human user [@OpenAI.2023].

While LLMs are not yet widely used to screen sources for literature reviews, early work suggests they may perform well in this role. Guo et al. [@Guo.2023] reported the use of a GPT-lineage model (they do not specify which, though their published code suggests OpenAI's 'gpt-3.5-turbo' model) to screen 24,307 titles and abstracts from five systematic and reviews one scoping review, achieving pooled sensitivity of 76% and specificity of 91% when compared to human reviewers. Their approach involved giving the model a brief prompt instructing it to take on the persona of a researcher screening titles and abstracts, and to respond with a decision to include or exclude a source, followed by the source's title and abstract as well as the study inclusion and exclusion criteria. Syriani et al. [@Syriani.2023] similarly reported the use of 'gpt-3.5-turbo' to screen titles and abstracts for a systematic review and achieved sensitivities of above 70%. They also systematically evaluated prompts given to the LLM to identify a prompt that performed best at the screening task; their chosen prompt, like that of Guo et al., placed the LLM in the role or persona of an academic reviewer.

Both of these approaches made use of a single, fixed text prompt template, which the LLM then completes with additional text representing its response (the decision to include or exclude a source), a method sometimes

called 'one-shot prompting'. Recent work has identified a number of methods which can be superior to one-shot prompting when using LLMs for tasks that require complex or multi-step reasoning. These methods include chain-of-thought prompting [@Wei.2022], in which a complex task is broken down into a series of intermediate steps that the model completes sequentially, and the tree of thoughts strategy [@Yao.2023] in which multiple chains of thought are generated, compared, and integrated.

In this paper, I introduce a package for the R programming language [@R.2023] called GPTscreenR, and evaluate its performance in screening titles and abstracts for a scoping review. The aim of this package was is to assist and augment rather than replace human reviewers in performing scoping reviews. This paper and the associated package represent four novel developments in the use of LLMs for source screening in literature reviews. Firstly, they provide an open-source software package which can be downloaded and used as well as modified by reviewers. Secondly, they provide the first such application of LLMs specifically for scoping reviews, though pragmatically there is little difference in approach when compared to the use for systematic reviews. Thirdly, they provide to my knowledge the first report on the accuracy of this approach using the most recent iteration of the GPT model lineage, GPT-4. Finally, they incorporate the use of chain-of-thought reasoning in an effort to maximise the accuracy of screening decisions.