30/01/2024, originally posted as a blog post on www.i-deel.org

Title: Around meta-analysis (15): emerging Large Language Models (LLM) tools

by Malgorzata Lagisz (Losia)

Systematic reviews (and meta-analyses based on a systematic review of literature) are extremely time-consuming. Anyone who conducted one in a rigorous and robust way can attest to this fact. Not surprisingly, researchers across disciplines have been looking for using computer algorithms and software to automate and accelerate systematic reviews of academic literature.

Such efforts have brought some success. Algorithms based on text-mining, artificial intelligence (AI), and more specifically machine-learning approaches, are now integrated in some of the popular software dedicated to literature screening (e.g., Rayyan, Abstracr, ASReview) and even data extraction (e.g., RobotReviewer). Other group of algorithms can suggest relevant evidence based on similarities among documents (e.g., ConnectedPapers, and recommendation systems built into major literature search platforms). However, these tools perform well only in a limited set of scenarios and applications, require extensive and expert initial training investment, and many are not freely accessible. For a recent scoping review of diverse types of automation tools, their applications and drawbacks, see Khalil et al. (2022).

It is tempting to think that a recent development of a new generation of AI models and software has better performance and new capabilities. Especially, Large Language Models (LLMs) are trained on large datasets of written language (think ChatGPT and similar models). They can be operated by using user prompts in conversational language, rather than technical programming languages, which makes them user-friendly. Why not ask them to find relevant studies, highlight or summarise relevant information do the screening for you?

Unfortunately, generic LMMs, like ChatGPT are less than ideal for systematic reviews. They tend to hallucinate (invent evidence), are not accurate, and require expert knowledge and careful set up to provide useful output (Qureshi et al. 2023). Among the many likely reasons for the poor performance, one is their probabilistic nature (making decisions based on probabilities of patterns) and the other one is that ChatGPT models are not trained specifically on academic literature. And they were not rally designed to do work for scientists.

Are there LLMs tailored to academic literature and requirements of researchers? In the last months, such tools were rapidly emerging (Sanderson 2023). Since the y are new, there are no rigorous published assessments of their performance. I tried a few out, but cannot provide any concrete data or recommendations yet. I think we should not aim to fully automate any systematic review steps but, instead, we can use such new tools as an “another reviewer” or an alternative approach that supplements and strengthens our existing workflows.

If you are interested in LLMs that look like potentially useful in systematic reviews workflows (and testing how much you can trust them!), here is a short list of suggestions:

- Elicit (elicit.com)
- Scite (scite.ai)
- Typeset (typeset.io)
- Consensus (consensus.app)

References:

Khalil H, Ameen D, Zarnegar A. Tools to support the automation of systematic reviews: a scoping review. J Clin Epidemiol. 2022 Apr;144:22-42. doi: 10.1016/j.jclinepi.2021.12.005. Epub 2021 Dec 8. PMID: 34896236.

Qureshi, R., Shaughnessy, D., Gill, K.A.R. *et al.* Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation?. *Syst Rev* **12**, 72 (2023). https://doi.org/10.1186/s13643-023-02243-z)

Sanderson K. AI science search engines are exploding in number - are they any good? Nature. 2023 Apr;616(7958):639-640. doi: 10.1038/d41586-023-01273-w. PMID: 37069302.