NAPCRG 52nd Annual Meeting — Abstracts of Completed Research 2024.

Submission Id: 6188

Title

Reducing The Effort for Performing Systematic Reviews Using Natural Language Processing And Large Language Models

Priority 1 (Research Category)

Systematic review, meta-analysis, or scoping review

Presenters

Dr. Arya Rahgozar, PhD, Pouria Mortezaagha, Jessie McGowan, Kelly Cobey, Jodi Edwards, PhD, Andrea Tricco, Doug Manuel, Dean Fergusson, PhD, MHA, David Moher

Abstract

Context

Systematic reviews are critical to support future research, investigation, and development of adherence to the reporting guidelines and to identify the gaps but require intensive human synthesis.

Objective

The aim is to evaluate the feasibility of an instructed large language model (LLM) to assist with human synthesis such as temporal topic search and screening by citation in systematic reviews for Brain-Heart-Interconnectome (BHI) to support reporting guidelines, such as CONSORT and SPIRIT.

Study Design and Analysis

We used Lang-Chain framework to implement a Retrieval Augmented Generation (RAG) system including document loader, text-splitter, vectorizer (OpenAI Embedding to facilitate similarity), database and an instructed LLM (GPT3.5-Turbo Model) to respond to human questions automatically regarding systematic reviews given a set of pre-selected papers. We then compared the answers with those generated by a general purpose LLM.

Setting or Dataset

Dataset includes a set of 846 research papers in BHI that claimed to have followed relevant reporting guidelines. We compared the 2 answers to a set of 20 user questions.

Population Studied

Corpus includes a variety of categories in Randomized Clinical Trails (RCT), their assessments, quality, prevalence, outcomes, bias, methods, interventions, and adherence to CONSORT.

Intervention/Instrument

The instrument is an especially trained conversational system that significantly expedites the BHI systematic review development against the reporting guidelines criteria.

Outcome Measures

We used a normalized combination of precision and recall (F1) index to compare and count the corresponding same terms between the two generated answers by the two LLMs.

Results

We qualitatively assessed the two versions of the auto-generated answers to the systematic review questions. The retrained LLM that was instructed to use the selected papers generated more specific answers from the defined citations and known source of selected papers. The F1 index comparison between the two sets of 20 answers showed 30.7% concordance on average, which showed the effects of our controlled source and prompt-engineering.

Conclusion

We showed it was workable to use both a general LLM, and an instructed LLM with a set of specific constrained source of citations of research papers to provide relevant insightful answers to assist with human synthesis of material and hence significantly facilitated the development of BHI systematic reviews.