IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 May 2024, accepted 20 May 2024, date of publication 27 May 2024, date of current version 20 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3405529

APPLIED RESEARCH

Fully Automated Scholarly Search for Biomedical Systematic Literature Reviews

LEANDRA BUDAU AND FAEZEH ENSAN
Department of Electrical, Computer, and Biomedical Engineering, Toronto Metropolitan University, Toronto, ON M5B 2K3, Canada
Corresponding author: Faezeh Ensan (fensan@torontomu.ca)

ABSTRACT Biomedical Systematic Literature Reviews (SLRs) play a fundamental role in evidence-informed healthcare and can serve as actionable insights for researchers and policy-making organizations in the field. In this paper, we focus on the phase of 'study search' in conducting SLRs, i.e., the process of organising a comprehensive search via biomedical databases, such PubMed, in order to obtain all the relevant articles on a certain topic of interest. We introduce FASS-BSLR, a dataset and a benchmark suit to facilitate developing and evaluating fully automated techniques for study search. We also provide and analyze a set of basic methods along with a number of generative models, and report the experiment's results over the introduced dataset. We introduce a simple but effective model based on the resent transformer-based generative model, ChatGPT, for generating Boolean queries over PubMed. Through different experiments, we illustrate that this model is more effective than basic search models, than keyword search over PubMed, and than existing methods for crafting Boolean queries using ChatGPT. We show that the introduced model is even more effective than manual queries in terms of Precision, Recall, NDCG, and MAP in positions 10, and 100, but falls short of the recall that manual queries achieve at position 1000. We also report the retrieval performance of different models when a number of relevant articled have been provided as seed documents. We demonstrate that, when three documents are used as seed articles, the introduced model outperforms manual queries in all metrics except Recall@1000, on which its performance is comparable with the performance attained by manual queries.

INDEX TERMS Systematic literature reviews, technology assisted reviews, boolean query formalization.

I. INTRODUCTION

Biomedical Systematic Literature Reviews (SLRs), which provide a systematically-collected and synthesized body of knowledge on key health and medical issues [1], [2], are a fundamental part of evidence-informed healthcare [3], [4], [5]. As the recent COVID-19 pandemic has clearly shown, evidence-informed health policy-making is critically important, as it allows decision makers to understand and act on the most reliable evidence available to them at the right time [6], [7]. Much of the decisions during the COVID-19 pandemic were driven by over 2,000 SLRs [8] covering a range of issues including but not limited to epidemiology, screening and diagnosis, severity assessment, special populations, and treatment.

Despite the development of a variety of Technology Assisted Review (TAR) methodologies that ease some of the burdens in the process of conducting biomedical SLRs [9], [10], [11], [12], SLRs are still exceedingly expensive (>$100, 000 per study on health-related topics), often quite lengthy (>1 year), and labor-intensive (>1, 100 hours, 5 reviewers) [13], [14], [15]. This is primarily due to the fast-growing and complex nature of published biomedical studies (e.g., 2,400 new scientific papers are published every day on the COVID-19 pandemic alone), and the need for the highest level of quality and rigo in health sciences. These timelines and expenditures highlights the importance of further advancement in developing fully automated techniques that can substantially speeding up the process of conducting biomedical SLRs, while maintaining and even improving the quality of results.

During the last years, machine learning, neural models and the recent advances in neural text processing have

The associate editor coordinating the review of this manuscript and approving it for publication was Abdallah Kassem.

gained attention as a solution for the automatically or semi-automatically development of biomedical SLRs. Existing machine learning techniques target automating different steps of developing SLRs [16], [17], [18], [19] including *study search* [20], *study screening* [21], *data extraction* [22], [23], [24], [25], [26], [27], and *quality assessment* [28]. Despite their differences in objectives, methodologies, and the reported results, all these techniques rely on annotated datasets for training a model that can be appropriately generalized and deployed in real context.

In this paper, we focus on the step of 'study search' in conducting SLRs. Study Search refers to formalising a thorough search through biomedical databases such as PubMed and MedLine to get all the relevant articles for a topic of interest. Here, the main objective is to retrieve relevant and reliable studies at higher ranks, while ensuring all the relevant articles have been returned. Study Search has a significant impact on the effectiveness of later stages of generating SLRs, such as study screening and quality evaluation, which deal with removing irrelevant and unreliable research.

Study Search can be conducted through complex Boolean queries that are constructed by well-trained specialists for medical databases such as PubMed. Alternatively, it can be initialized by posing a textual query to scholarly search engines that are equipped with query analysis and processing techniques. Manual construction of Boolean queries has been recognized as a very costly step in conducting SLRs in terms of time and labor [29]. Further, formalizing textual queries that retrieve all possibly relevant studies is cumbersome and challenging.

In this paper, we provide a dataset and a benchmark that facilitates the design of both Boolean and keyword queries tailored for the PubMed Biomedical scholarly database. PubMed is a free online database of scientific articles and research papers that includes more than 36 million citations for biomedical literature. PubMed is widely recognized as one the leading databses for the field [30]. We evaluate and compare the performance of multiple fully-automated query formalization algorithms over the introduced dataset. We also index the COVID-19 Open Research Dataset, known as CORD-19 [31], which consists of more than 1000,000 scientific articles about the Covid-19 pandemic, and analyze the performance of keyword queries over this data. The objective is to compare the effectiveness of crafting Boolean queries for the PubMed vs. conducting regular searches over indexed data. We also analyze the performance of the methods that use generative language models in order to create Boolean queries for PubMed. We introduce a simple but effective fully-automated method based on ChatGPT, denoted as CGT, that outperform all the existing baselines, and even queries manually crafted by human experts in almost all metrics except Recall@1000, on which manual queries have a supremacy. Additionally, we examine how well generative models retrieve articles when given one to five relevant articles to use as seed documents. The main contributions of this work can be enumerated as follows:

- We release FASS-BSLR, a dataset of 111 biomedical SLRs on the topic of Covid-19, their included studies, a set of one to five relevant articles that can be used as seed documents for more effective search, the Boolean queries generated by PubMed given their title and their keywords, and a set of Boolean queries generated by various generative language model approaches. We also release a subset of FASS-BSLR, denoted as Set-B, that come with manual queries, i.e., queries crafted by human experts for efficient search over PubMed.
- We report results of experiments that craft Boolean queries using the title and the keywords of a SLR over FASS-BSLR. We also report the performance of basic search methods over CORD-19, in order to compare the results of a search over an indexed database that contains all relevant articles with a search over PubMed.
- We introduce a simple but effective method for crafting Boolean queries using ChatGPT, denoted as CGT. We show that CGT outperforms all existing fully-automated models (including other models that use ChatGPT for generating Boolean queries over PubMed) in terms of different performance metrics over FASS-BSLR. We also illustrate that CGT outperforms manual queries on almost all metrics except those that measured at position 1000.
- We analyze the performance of generative models over FASS-BSLR when different number of seed studies are provided. We illustrate that CGT is a powerful model, especially when it is augmented by at most three seed documents.

II. RELATED WORK

In this section, we review recent advancements in automated study search for biomedical SLRs, and provide an overview of the existing datasets.

A. AUTOMATED STUDY SEARCH FOR BIOMEDICAL SLRS

Automated study search has been the topic of numerous works in the literature of technology assisted biomedical reviews. The work presented in [20], [29], [32], [33], and [34] focus on formalizing Boolean queries for biomedical scholarly databases such as PubMed. In [29] and [32], a five-steps framework is introduced to craft Boolean queries for Pubmed. The framework begins by extracting the primary high-level concepts from a textual description of a biomedical research topic and grouping them using an AND operator. It next extracts the noun phrases at lower depths and groups them using an OR operator. The UMLS entities relevant to each Boolean clause are extracted and expanded in the second and the third phases. Entities are mapped to keywords in the fourth phase, and keywords are then further processed for stemming and adding MESH terms in the final step. The work presented in [20], denoted as ChatGPT-PE in the reset of this paper, explores the effectiveness of ChatGPT for generating Boolean queries for the PubMed database. Given a biomedical research topic, it adopts a four step pipeline that 1) asks ChatGPT to generate query terms, 2) asks ChatGPT to classify terms, 3) combines terms in the same category by 'OR', and all Boolean clauses of the categories by 'AND', and finally 4) asks ChatGPT to refine the query by adding more terms. The authors in [33] introduce and use a dataset consists of one million PubMed articles' abstracts and their keywords for fine-tuning pre-trained language models to generate biomedical key-phrases. The framework presented in this work generates key-phrases for a given biomedical research topic, generates UMLS concepts, and applies different clustering methods to group terms and entities. It then use OR operators inside and AND operators across clusters to form a Boolean query. The methods in [35] and [36] introduce automated techniques for generating MESH terms that enrich Boolean queries posed to the biomedical databases.

The other group of works investigate *study search* and its challenges in the context of the Continuous Active Learning (CAL) approach [37], [38]. CAL-based methods employ query formalization techniques to initiate a search in a biomedical database like PubMed, solicit expert feedback on the top-ranked results, then retrain the study screening algorithms using the expanded labelled data in an iterative manner [38]. The study search methodology used in each iteration has a significant impact on how well these methods work because expert feedback may be obtained for a limited number of papers and can be noisy and inaccurate. In [39], a learning to rank method has been proposed that defines and employs a set of manually-crafted features for the similarity between the scholarly articles and the biomedical research topic, e.g. cosine similarity and BM25. The ranking model is re-trained using the newly labeled data in each iterations. The works presented in [40] and [41] focus on locating the last relevant studies, i.e., studies that cannot be found in the initial iterations of a CAL-based method. The method in [41] suggests an automated question-generating mechanism that produces yes-or-no inquiries regarding the expected existence of an entity in missing studies. By answering these questions, human experts direct the study search algorithms to perform more effectively. The work in [40] analyzes noisy answers to the generated questions and their impact on training and the performance of the search algorithms. Finally, the main focus of [42], [43], and [44] is to find an automatic stopping strategy for the CAL interations.

B. EXISTING DATASETS

The existing biomedical study search methods, including those that are reviewed in Section II-A, are mostly evaluated over the CLEF technological assisted reviews (TAR) dataset [45], [46], [47]. This dataset is introduced as part of the Task2 (Technology Assisted Reviews (TAR) in Empirical Medicine in English) of the CLEF eHealth Evaluation Labs, which are hold between 2017–2019. While synergistic, FASS-BSLR complements CLEF-TAR and also distinguishes itself in the following respects: FASS-BSLR introduces a new set of SLRs, mostly from the registered SLRS in the PROSPERO, which is a database for registering systematic reviews in different topics including health and social care.[1] Contrary to the dataset shared by CLEF eHealth, which includes a range of diagnostic test accuracy, intervention and prognosis, FASS-BSLR is only focused on biomedical SLRS. Further, FASS-BSLR provides a range of auxiliary meta-data such as a citation network that can be further utilized in training. Finally, FASS-BSLR provides a benchmark for study search, given a set of verified related articles.

TABLE 1. FASS-BSLR statistics. The articles that have not been indexed by PubMed are excluded.

	#SLR	#Included Studies per SLR		
		min	max	avg
Set A	111	2	65	20.4775
Set B	62	2	60	18.01

III. DATASET DESCRIPTION

In this section, we provide details about the dataset, the manual labeling procedure, and the auxiliary information shared for different scholarly search tasks.

A. FASS-BSLR, SET-A

We share a set of 111 SLRs (denoted as Set-A) on the topic of COVID-19 published between 2020/03/30 and 2023/04/14. For each SLR, we share the DOI, title, abstract, list of authors, publication venue, and the publication year. In addition, we share the same information for the included studies in each SLR. The included studies for each SLR are those articles that are reviewed and examined by human experts and have been included in the SLR. We recruited three research assistant to thoroughly read and analyze the content of each SLR and label all those studies that are included in the final revisions of each SLR. There are a total of 2273 included studies, 1693 of which are unique. We provide trec-style qrel files including the list of relevance judgment for each SLR as a query. Since the main objective of this work is to formalize search queries over the PubMed database, we filter out those articles that are not indexed by PubMed. We categorize all SLRs in 5 folds for the purpose of training and testing.

We also share some auxiliary information for facilitating supervised methods. We share the citation graph of each SLR. Each citation graph includes the DOI, title, abstract, list of authors, publication venue, and the publication year of articles referenced in an SLR article including those included and those just referred but not included in the SLR in a depth of two.

For this task, we provide some auxiliary meta-data for each SLR that can be used for training a more effective method: 1) the name and abstract of 5 articles randomly selected from the included studies to the SLR as the initial set of seed

[1] https://www.crd.york.ac.uk/PROSPERO

TABLE 2. An example a SLR topic, the processed text, and the Boolean query that is generated by PubMed API.

SLR Topic	Corticosteroid use in COVID-19 patients: a systematic review and meta-analysis on clinical outcomes
Processed text	Corticosteroid use COVID-19 patients
PubMed-Title (Boolean Query generated by PubMed API for the title terms)	("adrenal cortex hormones"[MeSH Terms] OR ("adrenal"[All Fields] AND "cortex"[All Fields] AND "hormones"[All Fields]) OR "adrenal cortex hormones"[All Fields] OR "corticosteroid"[All Fields] OR "corticosteroids"[All Fields] OR "corticosteroidal"[All Fields] OR "corticosteroide"[All Fields] OR "corticosteroides"[All Fields]) AND ("covid 19"[All Fields] OR "covid 19"[MeSH Terms] OR "covid 19 vaccines"[All Fields] OR "covid 19 vaccines"[MeSH Terms] OR "covid 19 serotherapy"[All Fields] OR "covid 19 nucleic acid testing"[All Fields] OR "covid 19 nucleic acid testing"[MeSH Terms] OR "covid 19 serological testing"[All Fields] OR "covid 19 serological testing"[MeSH Terms] OR "covid 19 testing"[All Fields] OR "covid 19 testing"[MeSH Terms] OR "sars cov 2"[All Fields] OR "sars cov 2"[MeSH Terms] OR "severe acute respiratory syndrome coronavirus 2"[All Fields] OR "ncov"[All Fields] OR "2019 ncov"[All Fields] OR (("coronavirus"[MeSH Terms] OR "coronavirus"[All Fields] OR "cov"[All Fields]) AND 2019/11/01:3000/12/31[Date - Publication])) AND ("patient s"[All Fields] OR "patients"[MeSH Terms] OR "patients"[All Fields] OR "patient"[All Fields] OR "patients s"[All Fields])

studies; 2) the citation network of these initial related studies with the depth of two (the articles cited by these studies in addition to the articles cited inside the citations) 3) all the articles published by the authors of these studies.

B. FASS-BSLR, SET-B

Set-B is a subset of Set-A for which we have access to the queries formalized by human experts. We recruited a research assistant to examine the attached auxiliary resources to each SLR and look for queries used for study search. We ran these queries through PubMed API and collected top 1000 documents for each SLR. FASS-BSLR-Set-B includes 62 SLRs. There are a total of 1117 included studies, 906 of which are unique. Like Set-A, we filter out those articles that are not indexed by PubMed. We categorize these SLRs into 5 folds for the purpose of training and testing. Table 1 shows some statistics about the provided data.

IV. EXPERIMENTS WITH BASIC BASELINES

In this section, we describe a set of basic methods, their experimental set-up and the experiment's results.

We use the following baselines for formalizing queries over PubMed 1: 1) **PubMed-Title**: We pose the title of each SLR to PubMed through the Entrez API[2] and collect the top 1000 articles returned. For this purpose, we pre-process SLR titles, remove stop words and the words refer to the review or systematic review. We then send the title as a set of words to the PubMed API. PubMed generates a Boolean query for a set of given words. Table 2 shows a sample of a SLR title and the Boolean query generated by the PubMed API for this title. 2) **PubMed-Keywords**: We extract the keyword list created by the authors of each SLR and pose it to the same API as PubMed-Title. We collect the top 1000 PubMed articles returned. For both PubMed-Title and PubMed-Keywords, and all other baselines for them the results are collected from PubMed, we limit the search to retrieve all the articles published before the publication date of each SLR.

In addition to these baselines, we indexed the COVID-19 Open Research Dataset, CORD-19, using Elasticsearch. We then search the pre-processed title of each SLR in the (a) title (b) and in the title and the abstract of the indexed articles using BM25 and LM retrieval algorithms and find the top 1000 articles returned by them. We filter-out those articles that have not been indexed by PubMed. Our hypothesis is to compare the performance of the PubMed search API with the basic search methods, when data can be dumped and indexed. Based on whether title or title and abstract are used for searching, we have four baselines here: 3) **BM25-Title**, and 4) **LM-Title** that search SLR title over the title of articles using BM25 and LM algorithms, respectively and 5) **BM24-TitleAbstract** and 6) **LM-TitleAbstract** that search terms over the title and abstract of articles. For LM, we use Dirichlet smoothing, when μ is set to 1000. For BM25, we use the default setting for the parameters in Elasticsearch.

For Set-B, we also report the performance of the manual queries crafted by human experts denoted as (7) **Manual Queries**.

A. RESULTS

Table 3 shows the performance of the basic baselines in terms of Precision, Recall, NDCG, and MAP in the following positions: 10, 100, and 1000, and F1 score (F-Measure) over the Set-A of the FASS-BSLR benchmark. As you can see in this table, PubMed-Keywords is the best performing method among basic baselines. It has the best Precision@10 (20.7% better than the next best performing method, PubMed-Keywords). While PubMed-Keywords achieved the best Precision@10, BM25-Title is the best performing method when considering recall@1000, which performed 8.3% better than the next runner up, LM-Title, and 73.3% better than PubMed-Keywords. PubMed-Keywords performs the best across all metrics at the position of 10, achieving a recall@10, NDCG@10, and MAP@10 which are 10.3%, 13.9%, and 16.1% better than the runner up method in each metric (LM-TitleAbstract, PubMed-Title, and BM25-TitleAbstract, respectively).

Table 4 shows the performance of all the baselines over Set-B of the FASS-BSLR benchmark. This table also includes the manual query performance, i.e., the performance of Boolean queries constructed by human experts and run through PubMed. As we can see in this table, PubMed-Title is the best performing method at a position of 10 while methods run over CORD-19 typically performed better at a position of 1000. PubMed-Title shows a better performance in regards to precision, recall, and NDCG in the position of 10 than all methods that search over the indexed dump and the manual queries searched over PubMed. This conclusion can guide researchers to use the method of searching the SLR title over PubMed in the initial interactions of study search rather than developing a complex query by hand which may be

[2] https://eutils.ncbi.nlm.nih.gov/entrez/eutils/

TABLE 3. The performance of (a) basic methods, and (b) generative methods in terms of Precision, Recall, NDCG, and MAP in the the positions of 10, 100, and 1000, and F1 score (F- Measure) over FASS-BSLR, Set-A. Bold numbers indicate the highest number in each column.

	Precision	Recall	NDCG	MAP	Precision	Recall	NDCG	MAP	Precision	Recall	NDCG	MAP	F-Measure
	@10				@100				@1000				
(a) Basic Methods													
PubMed-Title	0.0264	0.0128	0.0387	0.0078	0.0086	0.0395	0.0376	0.0097	0.0015	0.0648	0.0467	0.0104	0.0247
PubMed-Keywords	0.0309	**0.0310**	**0.0441**	**0.0130**	0.0099	0.0769	0.0532	**0.0160**	0.0017	0.1208	0.0667	0.0171	0.0350
BM25-Title	0.0200	0.0151	0.0276	0.0073	0.0106	0.0737	0.0455	0.0108	0.0026	0.1611	0.0730	0.0121	0.0088
LM-Title	0.0191	0.0150	0.0240	0.0069	0.0082	0.0575	0.0388	0.0104	0.0024	0.1513	0.0677	0.0116	0.0077
BM25-TitleAbstract	0.0245	0.0268	0.0355	0.0112	0.0120	0.0835	0.0554	0.0155	0.0022	0.1399	0.0738	0.0165	0.0108
LM-TitleAbstract	0.0236	0.0281	0.0329	0.0098	0.0104	0.0794	0.0500	0.0135	0.0021	0.1323	0.0681	0.0146	0.0108
(b) Generative Method													
CGT	**0.0373**	0.0225	0.0398	0.0073	**0.0181**	**0.1109**	**0.0666**	0.0137	**0.0047**	**0.2367**	**0.1101**	**0.0167**	**0.0385**
ChatGPT-PE	0.0145	0.0064	0.0152	0.0019	0.0081	0.0313	0.0228	0.0048	0.0016	0.0648	0.0348	0.0054	0.0158

TABLE 4. The performance of (a) basic methods, and (b) generative methods in terms of Precision, Recall, NDCG, and MAP in the the positions of 10, 100, and 1000, and F1 score (F- Measure) over FASS-BSLR, Set-B. Bold numbers indicate the highest number in each column.

	Precision	Recall	NDCG	MAP	Precision	Recall	NDCG	MAP	Precision	Recall	NDCG	MAP	F-Measure
	@10				@100				@1000				
(a) Basic Methods													
PubMed-Title	0.0250	0.0115	0.0342	0.0061	0.0098	0.0446	0.0373	0.0083	0.0019	0.0799	0.0500	0.0093	0.0259
PubMed-Keywords	0.0317	0.0313	**0.0456**	**0.0143**	0.0108	0.0945	0.0631	0.0175	0.0018	0.1553	0.0805	0.0188	0.0380
BM25-Title	0.0233	0.0197	0.0343	0.0095	0.0117	0.0880	0.0539	0.0134	0.0029	0.1986	0.0873	0.0150	0.0100
LM-Title	0.0183	0.0185	0.0286	0.0103	0.0082	0.0632	0.0455	0.0143	0.0025	0.1837	0.0802	0.0157	0.0082
BM25-TitleAbstract	0.0283	0.0261	0.0435	0.0148	0.0122	0.0900	0.0657	**0.0199**	0.0023	0.1632	0.0887	0.0210	0.0117
LM-TitleAbstract	0.0300	0.0299	0.0422	0.0136	0.0102	0.0797	0.0575	0.0174	0.0020	0.1491	0.0795	0.0187	0.0105
(b) Generative Method													
CGT	**0.0417**	**0.0392**	0.0086	0.0077	**0.0188**	**0.1256**	**0.0713**	0.0143	0.0048	0.2594	0.1173	0.0176	**0.0439**
ChatGPT-PE	0.0083	0.0033	0.0086	0.0012	0.0063	0.0240	0.0169	0.0033	0.0016	0.0725	0.0339	0.0042	0.0136
Manual Queries	0.0150	0.0091	0.0177	0.0040	0.0153	0.0988	0.0523	0.0091	**0.0060**	**0.3581**	**0.1320**	**0.0136**	0.0258

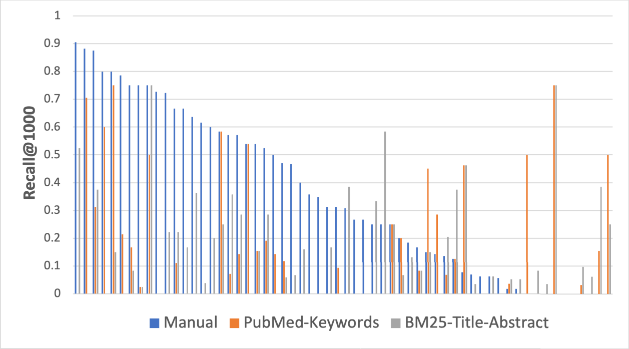


FIGURE 1. Recall@1000 obtained for each of the SLRs in the FASS-BSLR, Set-B dataset obtained from (a) PubMed-Keywords, (b) Bm25-Title Abstract, and (c) the manual queries, where SLRs are sorted based on the recall@1000 achieved by running the manual Boolean queries over PubMed.

more time consuming. Table 4 shows that manually crafted queries outperform all basic methods in terms of recall in the position of 1000. This observation highlights the importance of manually crafted Boolean queries in later interactions of search, when the general objective is to find all the remaining relevant articles.

A key feature of SLR study search, which sets it apart from regular ad-hoc search, is the goal of attaining high recall for biomedical research topics. Figure 1 illustrates the recall obtained for each of the SLRs in the FASS-BSLR, Set-B dataset obtained from (a) the best PubMed-based model (PubMed-Keywords), (b) the best index-based model (Bm25-TitleAbstract), and (c) the manual queries, where SLRs are sorted based on the recall@1000 achieved by running the manual Boolean queries over PubMed. As you can see in this figure, Compared with the PubMed-Keywords, manual Boolean queries result in a better recall for a large number of SLRs (41 out of 62). This observation holds true when comparing manual queries and BM25-TitleAbstract, i.e., 45 out of 62 SLRs show a superior or equal recall under manual queries rather than BM25. This figure also demonstrates that BM25 can be a good search alternative for the biomedical SLR topics for which Boolean queries over PubMed cannot yield satisfactory results.

V. EXPERIMENTS WITH GENERATIVE MODELS

Different information retrieval tasks such as questions answering, document summarization, key-phrase extraction, and ad-hoc retrieval have greatly benefited from recent developments in text generation models [48], [49]. In this section, we analyze the effectiveness of using ChatGPT, one of the most prominent generative models, in creating Boolean queries. For this purpose, we used two different methods, the first of which involved implementing the four-steps framework proposed in [20], denoted as ChatGPT-PE, for generating Boolean queries from each of the SLRs in the introduced dataset. The performance obtained by ChatGPT-PE has been reported in Table 3 for Set-A and in Table 4 for Set-B. As you can see in these tables, ChatGPT-PE typically performs worse than all other basic methods. In terms of Perecision@10, CGT is 45% worse than PubMed-Title and 53% worse than PubMEd-Keywords. It means that using ChatGPT for generating Boolean Queries, the way described in [20], is less effective than posing the

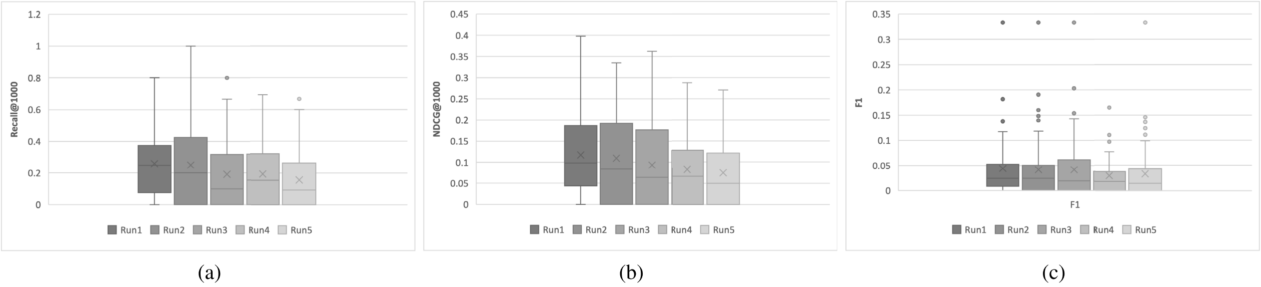


FIGURE 2. F1 (a), NDCG@1000 (b), and Recall@1000 (c) for runs 1-5 of CGT over FASS-BSLR, Set-B.

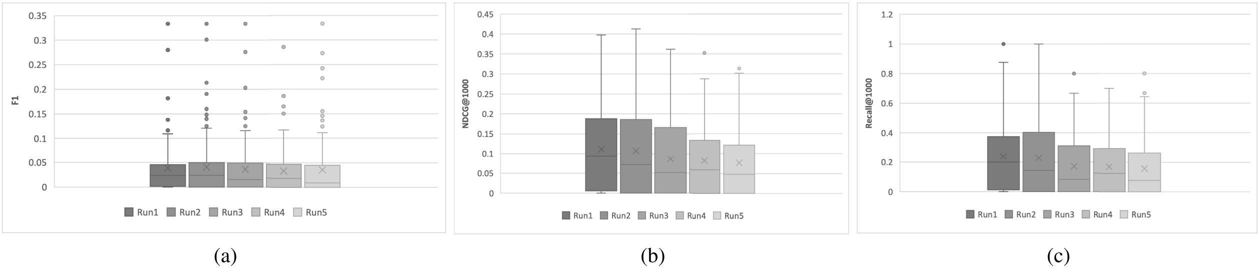


FIGURE 3. F1 (a), NDCG@1000 (b), and Recall@1000 (c) for runs 1-5 of CGT over FASS-BSLR, Set-A.

title or the keywords of the desired biomedical research to PubMed, and Let PubMed API to generate a Boolean query. Similarly, ChatGPT-PE is 50% and 79% worse than PubMed-Title and PubMed-Keywords in terms of Recall@10. This observation has been corroborated across various positions. For Precision@1000 and Recall@1000, ChatGPT-PE is outperformed by PubMed-Keywords (5% and 46%, respectively). It is also outperformed by basic models that search over a dump of all relevant studies, namely BM-Title, LM-Title, BM25-TitleAbstract, and LM-TitleAbstract. At position 10 and compared with ChatGPT-PE, BM25-TitleAbstract achieves a 68% better precision, and almost three times better recall. Similarly, At position 1000, BM25-TitleAbstract achieves a 37% better precision, and almost two times better recall than ChatGPT-PE.

In addition to ChatGPT-PE, we introduce another method, CGT, to use ChatGPT in generating Boolean queries, where only the title of the SLR is used in the ChatGPT prompt. Table 5 shows a sample SLR, a CGT prompt, and a Boolean query generated for that. Here, we ask ChatGPT to provide Boolean queries such that they may be used directly on PubMed's website. We collect the results obtained from each of 5 queries in the order of being generated by ChatGPT. We report the performance of the result obtained from the first query generated by CGT in Tables 3, and 4. As you can see in Table 3, in Set-A, CGT is the best performing method across many metrics including precision@10, recall@1000, NDCG@1000, and F-Measure where it performs 20.7%, 46.9%, 49.2%, and 10% better than the next best performing method. In addition, in Set-A, Table 4, CGT is the best performing method across many metrics including precision@10, Recall@10, Precision@100, Recall@100, NDCG@100, and Precision@1000.

TABLE 5. A sample SLR, a prompt for CGT method, and a Boolean query generated by ChatGPT.

SLR Topic	Corticosteroid use in COVID-19 patients: a systematic review and meta-analysis on clinical outcomes
ChatGPT prompt	Based on the following SLR title, please provide 5 complex pubmed Entrez formatted queries without descriptions, in plain text, such that they may be used directly on Pubmed's website.
CGT (5 Boolean Queries generated by the title based prompt)	(1) (("COVID-19" OR "SARS-CoV-2") AND "Corticosteroids") AND ("Clinical outcomes" OR "Patient outcomes") (2) (("COVID-19" OR "SARS-CoV-2") AND "Corticosteroids") AND ("Mortality" OR "Survival" OR "Death rate") (3)(("COVID-19" OR "SARS-CoV-2") AND "Corticosteroids") AND ("Disease severity" OR "Severe cases" OR "Critical cases") (4) (("COVID-19" OR "SARS-CoV-2") AND "Corticosteroids") AND ("Hospitalization" OR "Admission rate" OR "Critical care") (5)(("COVID-19" OR "SARS-CoV-2") AND "Corticosteroids") AND ("Complications" OR "Morbidity" OR "Adverse events")

According to Table 4, and comparing the manual queries with the generative and basic methods over Set B, we can see that manual queries are not effective at position 10, i.e., on all metrics at the position of 10, generative and basic methods perform better than manual queries. However, at the position of 1000, manual queries perform the best. At the position of 1000, manual queries achieve the best results in precision, recall, and NDCG where they perform 25%, 38%, and 12.5% better than the next best method, CGT. This again highlights the importance of human made, manual queries when attempting to gather all relevant articles for a study.

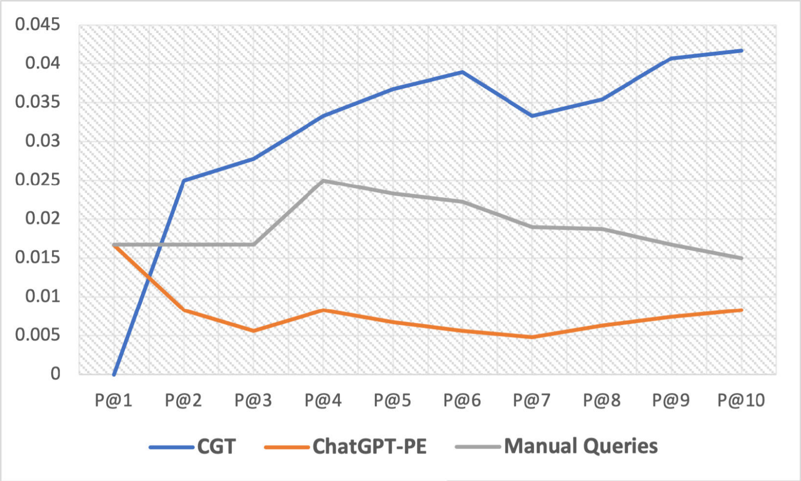


FIGURE 4. Performance of two generative models, ChatGPT- PE, and CGT in terms of Precision in positions 1 to 10, along with the performance of manual queries over FASS-BSLR, Set-B.

TABLE 6. A sample SLR, a prompt for CGT, based on the SLR title and two seed documents, and a Boolean query generated by ChatGPT.

Title: SLR Topic	Clinical Features of COVID-19 and Factors Associated with Severe Clinical Course
SeedTitle1: The title of the Seed document #1	A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster Malaysia
SeedTitle2: The title of the Seed document #2	Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China
ChatGPT prompt	Based on the following SLR title, please provide exactly one complex pubmed Entrez formatted query that encompasses the diverse subjects of the provided SLR title, in plain text, such that it may be used directly on Pubmed’s website:Title+", "+SeedTitle1+", "+SeedTitle2
Boolean Query	(("COVID-19" OR "severe acute respiratory syndrome coronavirus 2" OR "2019-nCoV" AND ("clinical features"OR "clinical course" OR "severe clinical outcome" OR "familial cluster" OR "person-to-person transmission"OR "patient characteristics" OR "Wuhan" OR "China"))

We conducted an experiment of running the prompts of CGT for five times to analyze the variability in responses provided by ChatGPT. Figures 3 and 2 demonstrate the min, max, and median of the F1, NDCG@1000, and recall@1000 measures for runs 1-5 of the prompts of CGT over Set-A, and Set-B, respectively. As you can see, there is very minimal difference in metrics between the first, and the second runs, and these two runs perform slightly better than the other runs. Additionally, the first run has a lightly better minimum values for all the metrics over both detests, which led us to conduct all experiments using only the first run.

We also conduct an experiment to analyze the precision of generative models. Figure 4 shows the result obtained along the precision of manual Boolean queries, CGT queries, and ChatGPT-PE queries for positions 1 to 10. These positions are important especially in CAL-based models in which expert feedback on the top-ranked articles are collected and used for expanding training data. Observably from the figure, CGT is the best performing method on all positions other than the position of 1, where both ChatGPT-PE and the manual queries obtain a better precision. Interestingly, the performance of CGT seems to be improving as the position increases, while the manual query performance only improves from position 1-4, and then steadily decreases from position 4-10

As we mentioned in Section III, For the FASS-BSLR dataset, we provide a set of seed documents for each SLR, that may be exploited in constructing more effective search over biomedical databases. In this section, we analyze the performance of the generative models when seed studies are included into their ChatGPT prompts in creating Boolean queries over PubMed.

In the original ChatGPT-PE method, four prompts were used to firstly obtain the 50 keyphrases and terms from the title and abstract of the SLR, next categorize them, then use the categories to structure a query, and finally incorporate MeSH terms and PICO. To incorporate seed studies into this method, we first obtained the 50 keyphrases and terms from the SLR, and then requested 10 keyphrases and terms from each of the seed studies one at a time. Once all keyphrases and terms were found, the same 3 steps were used to categorize, format, and enhance the query. As for the CGT method, the titles of the seed studies were concatenated to form a new prompt. Table 6 shows the CGT prompt used for creating a Boolean query for a sample SLR using two seed studies.

Tables 7 and 8 show the performance of these two generative models when different number of seed documents are used in their prompts, in terms of Precision, Recall, NDCG, and MAP in the positions of 10, 100, and 1000, and F1 score over Set-A and Set-B, respectively.

VI. EXPERIMENTS WITH SEED STUDIES

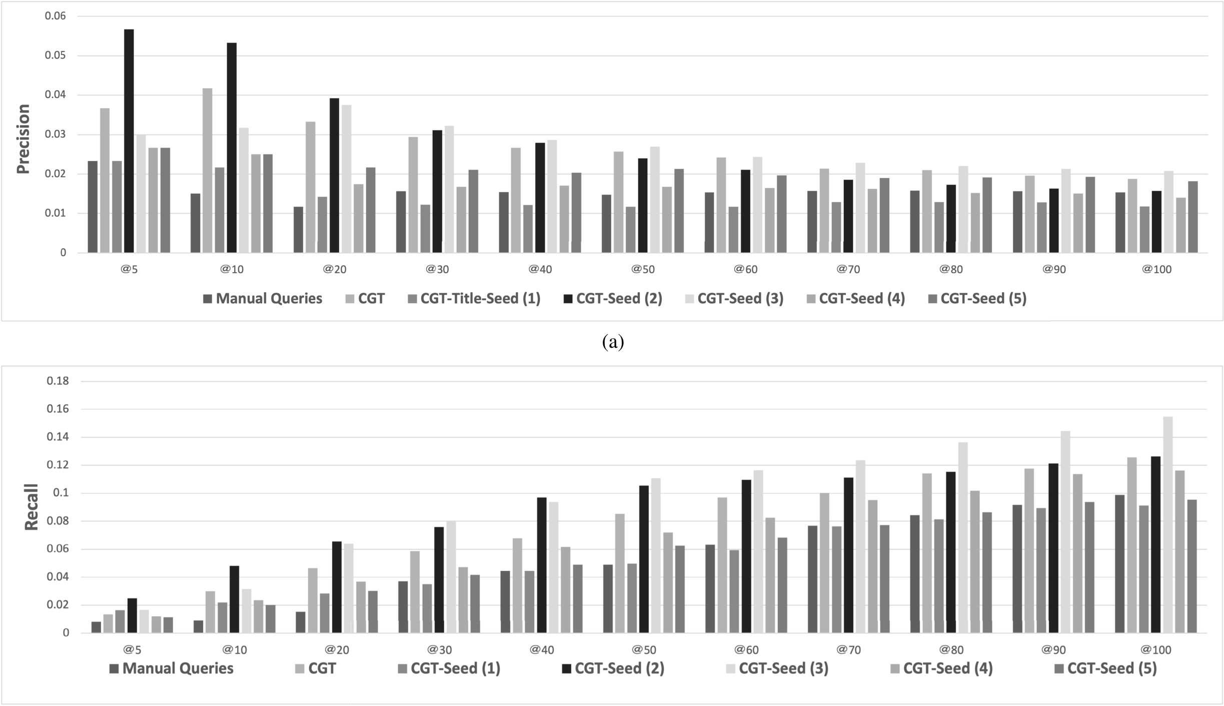
Table 7 and 8 illustrate that CGT with two or three seed documents outperforms all configurations of ChetGPT-PE and all other configuration of CGT with varying numbers of seed studies for every measure at different positions, with the exception of Precision@1000, where they lag slightly behind CGT with the configuration with five seed documents.

TABLE 7. The performance of (a) ChatGPT-PE, and (b) CGT in terms of Precision, Recall, NDCG, and MAP with one to five seed documents over FASS-BSLR, Set-A. Bold numbers indicate the highest number in each column.

	Precision	Recall	NDCG	MAP	Precision	Recall	NDCG	MAP	Precision	Recall	NDCG	MAP	F-Measure
	@10				@100				@1000				
(a) ChatGPT-PE													
One Seed Document	0.0164	0.0121	0.0168	0.0038	0.0084	0.0446	0.0332	0.0097	0.0022	0.0543	0.0583	0.0112	0.0253
Two seed Documents	0.0182	0.0141	0.0456	0.0143	0.0084	0.0593	0.0631	0.0175	0.0020	0.1228	0.0805	0.0188	0.0235
Three Seed Documents	0.0245	0.0229	0.0325	0.0120	0.0088	0.0636	0.0457	0.0152	0.0024	0.1202	0.0666	0.0163	0.0249
Four Seed Documents	0.0291	0.0183	0.0345	0.0072	0.0134	0.0695	0.0489	0.0136	0.0027	0.1340	0.0709	0.0146	0.0315
Five Seed Documents	0.0164	0.0064	0.0161	0.0029	0.0081	0.0412	0.0275	0.0066	0.0023	0.1097	0.0509	0.0075	0.0214
(b) CGT													
One Seed Document	0.0318	0.0304	0.0443	0.0143	0.0136	0.1010	0.0644	0.0194	0.0029	0.1730	0.0885	0.0207	0.0393
Two Seed Documents	**0.0409**	**0.0323**	**0.0475**	**0.0131**	0.0162	0.1129	0.0727	**0.0211**	0.0032	0.1986	0.0995	0.0226	**0.0412**
Three Seed Documents	0.0373	0.0299	0.0416	0.0109	**0.0200**	**0.1316**	**0.0803**	0.0200	0.0059	**0.3138**	**0.1414**	**0.0241**	0.0410
Four Seed Documents	0.0364	0.0261	0.0394	0.0106	0.0175	0.1053	0.0680	0.0175	0.0051	0.2697	0.1219	0.0213	0.0373
Five Seed Documents	0.0255	0.0160	0.0258	0.0059	0.0156	0.0773	0.0479	0.0097	**0.0060**	0.2802	0.1159	0.0143	0.0309

TABLE 8. The performance of (a) ChatGPT-PE, and (b) CGT in terms of Precision, Recall, NDCG, and MAP with one to five seed documents over FASS-BSLR, Set-B. Bold numbers indicate the highest number in each column.

	Precision	Recall	NDCG	MAP	Precision	Recall	NDCG	MAP	Precision	Recall	NDCG	MAP	F-Measure
	@10				@100				@1000				
(a) ChatGPT-PE													
One Seed Document	0.0133	0.0132	0.0143	0.0038	0.0075	0.0597	0.0319	0.0078	0.0020	0.1387	0.0560	0.0089	0.0252
Two seed Documents	0.0200	0.0167	0.0286	0.0076	0.0100	0.0787	0.0481	0.0123	0.0019	0.1343	0.0652	0.0130	0.0287
Three Seed Documents	0.0333	0.0348	0.0431	0.0182	0.0112	0.0964	0.0655	0.0234	0.0028	0.1707	0.0915	0.0248	0.0330
Four Seed Documents	0.0250	0.0204	0.0313	0.0085	0.0138	0.0773	0.0528	0.0173	0.0030	0.1462	0.0766	0.0183	0.0350
Five Seed Documents	0.0100	0.0036	0.0122	0.0017	0.0080	0.0418	0.0261	0.0055	0.0024	0.1161	0.0521	0.0068	0.0181
(b) CGT													
One Seed Document	0.0217	0.0219	0.0301	0.0117	0.0118	0.0914	0.0559	0.0152	0.0028	0.1828	0.0843	0.0167	0.0343
Two Seed Documents	**0.0533**	**0.0481**	**0.0613**	**0.0184**	0.0157	0.1263	**0.0831**	**0.0262**	0.0027	0.1988	0.1051	**0.0273**	**0.0503**
Three Seed Documents	0.0317	0.0317	0.0320	0.0079	**0.0208**	**0.1548**	0.0827	0.0189	0.0059	**0.3300**	**0.1416**	0.0229	0.0407
Four Seed Documents	0.0250	0.0235	0.0312	0.0098	0.0140	0.1162	0.0649	0.0160	0.0042	0.2618	0.1117	0.0185	0.0295
Five Seed Documents	0.0250	0.0201	0.0238	0.0072	0.0182	0.0955	0.0565	0.0118	**0.0062**	0.3154	0.1282	0.0167	0.0373



(a)

(b)

FIGURE 5. The precision (a) and recall (b) values of manual queries and CGT queries with 1-5 seed studies over FASS-BSLR, Set B.

This performance of CGT compared to ChatGPT-PE is expected considering that in the experiments excluding seed studies, CGT consistently performed significantly better than ChatGPT-PE across all metrics. The degradation in results when more seed studies are used can be justified by the fact that the increase in information is narrowing the search scope by too much. These results are useful as they may be able to guide researchers in understanding that even though they may have 5 seed studies for their SLR, the best results can be obtained by only using the two or three seed studies.

As you can see in Table 8, while CGT with 2 seed studies maintains the best results across all metrics at position 10, manual queries obtain 8.5% bester results for recall at the position of 1000. This again highlights the importance of manually crafted Boolean queries in searching for all relevant articles, although the CGT method still achieves comparable results which could make SLR study search more accessible when less budget or time is available.

Figure 5 visualizes the comparative precision and recall values of manual queries and CGT queries with 1-5 seed studies over Set B at positions 5 to 100. When looking at both figures, we can see that CGT with 2 seed studies performs the best from positions 5-20, and CGT with 3 seed studies performs best from positions 30-100. Another interesting result demonstrated in both figures is the comparison in metrics between CGT with no seed studies and CGT with one seed study. Consistently with both recall and precision, CGT with no seeds performs significantly better than CGT with one seed study. This implies that performance is degraded both when too few seed studies are supplied and too many.

VII. CONCLUSION

In this paper, we have focused on the crucial step of 'study search' within the realm of technology assisted review, particularly emphasizing its impact on the effectiveness of subsequent stages such as study screening and quality assessment. We introduced FASS-BSLR, a dataset designed to facilitate the development and evaluation of fully automated techniques for conducting systematic literature reviews in the biomedical field, specifically for PubMed searches. We conducted a comprehensive investigation into the performance of various basic and generative methods over FASS-BSLR. Furthermore, we presented and assessed a novel generative method, CGT, designed for effective searches, and evaluated its performance under different scenarios with varying numbers of relevant studies as seed documents. Our analysis of CGT's performance over FASS-BSLR, considering diverse seed study inputs, reaffirms its effectiveness, particularly when augmented by a limited number of seed documents.

The experimental analysis underscores that although manual queries are less effective than generative and basic methods at higher positions, they surpass automated solutions notably at position 1000. Moreover, the incorporation of seed studies significantly enhances the performance of automatic methods, including the newly introduced CGT model. Notably, CGT with two or three seed documents consistently outperforms all configurations of other generative models across various metrics and positions, even outperforming manual queries at positions 5-100.

Looking ahead, FASS-BSLR holds immense potential for future studies. Researchers can leverage this dataset for fine-tuning language-model-based approaches that generate or optimize Boolean queries for PubMed. Additionally, it can be employed by techniques utilizing citation graphs and auxiliary information to enhance biomedical scholarly searches, ultimately contributing to more efficient and cost-effective approaches in evidence-informed healthcare decision-making.

ACKNOWLEDGMENT

Data and all resources needed to replicate the experiments reported in this paper can be found in https://github.com/TMU-DataLab/FASS-SLR.

REFERENCES

[1] R. Whittemore, A. Chao, M. Jang, K. E. Minges, and C. Park, "Methods for knowledge synthesis: An overview," *Heart Lung*, vol. 43, no. 5, pp. 453–461, Sep. 2014.

[2] M. Egger, G. D. Smith, and D. Altman, *Systematic Reviews in Health Care: Meta-Analysis in Context*. Hoboken, NJ, USA: Wiley, 2008.

[3] V. Smith, D. Devane, C. M. Begley, and M. Clarke, "Methodology in conducting a systematic review of systematic reviews of healthcare interventions," *BMC Med. Res. Methodol.*, vol. 11, no. 1, pp. 1–6, Dec. 2011.

[4] A. C. Tricco, J. Tetzlaff, and D. Moher, "The art and science of knowledge synthesis," *J. Clin. Epidemiol.*, vol. 64, no. 1, pp. 11–20, Jan. 2011.

[5] X. Chen, H. Xie, Z. Li, G. Cheng, M. Leng, and F. L. Wang, "Information fusion and artificial intelligence for smart healthcare: A bibliometric study," *Inf. Process. Manag.*, vol. 60, no. 1, Jan. 2023, Art. no. 103113.

[6] W. Pian, J. Chi, and F. Ma, "The causes, impacts and countermeasures of COVID-19 'infodemic': A systematic review using narrative synthesis," *Inf. Process. Manag.*, vol. 58, no. 6, 2021, Art. no. 102713.

[7] J. Li, "Information avoidance in the age of COVID-19: A meta-analysis," *Inf. Process. Manage.*, vol. 60, no. 1, Jan. 2023, Art. no. 103163.

[8] M. Nassar, N. Nso, M. Alfishawy, A. Novikov, S. Yaghi, L. Medina, B. Toz, S. Lakhdar, Z. Idrees, and Y. Kim, "Current systematic reviews and meta-analyses of COVID-19," *World J. Virol.*, vol. 10, no. 4, p. 182, 2021.

[9] J. Babineau, "Product review: Covidence (systematic review software)," *J. Can. Health Libraries Association/J. de L'Association des Bibliothéques de la Santé du Canada*, vol. 35, no. 2, pp. 68–71, Aug. 2014.

[10] Z. Munn, E. Aromataris, C. Tufanaru, C. Stern, K. Porritt, J. Farrow, C. Lockwood, M. Stephenson, S. Moola, L. Lizarondo, and A. McArthur, "The development of software to support multiple systematic review types: The Joanna Briggs Institute system for the unified management, assessment and review of information (JBI SUMARI)," *JBI Evidence Implement.*, vol. 17, no. 1, pp. 36–43, 2019.

[11] L. Kellermeyer, B. Harnke, and S. Knight, "Covidence and Rayyan," *J. Med. Library Assoc.*, vol. 106, no. 4, p. 580, Oct. 2018.

[12] D. Li, P. Zafeiriadis, and E. Kanoulas, "APS: An active PubMed search system for technology assisted reviews," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2020, pp. 2137–2140.

[13] M. Michelson and K. Reuter, "The significant cost of systematic reviews and meta-analyses: A call for greater involvement of machine learning to assess the promise of clinical trials," *Contemp. Clin. Trials Commun.*, vol. 16, Dec. 2019, Art. no. 100443.

[14] R. Borah, A. W. Brown, P. L. Capers, and K. A. Kaiser, "Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry," *BMJ Open*, vol. 7, no. 2, Feb. 2017, Art. no. e012545.

[15] I. E. Allen, "Estimating time to conduct a meta-analysis from number of citations retrieved," *JAMA, J. Amer. Med. Assoc.*, vol. 282, no. 7, pp. 634–635, Aug. 1999.

[16] T. Muka, M. Glisic, J. Milic, S. Verhoog, J. Bohlius, W. Bramer, R. Chowdhury, and O. H. Franco, "A 24-step guide on how to design, conduct, and successfully publish a systematic review and meta-analysis in medical research," *Eur. J. Epidemiol.*, vol. 35, no. 1, pp. 49–60, Jan. 2020.

[17] J. D. L. Torre-López, A. Ramírez, and J. R. Romero, "Artificial intelligence to automate the systematic review of scientific literature," *Computing*, vol. 105, no. 10, pp. 2171–2194, Oct. 2023.

[18] Á. O. D. Santos, E. S. da Silva, L. M. Couto, G. V. L. Reis, and V. S. Belo, "The use of artificial intelligence for automating or semi-automating biomedical literature analyses: A scoping review," *J. Biomed. Informat.*, vol. 142, Jun. 2023, Art. no. 104389.

[19] H. Almeida, M.-J. Meurs, L. Kosseim, G. Butler, and A. Tsang, "Machine learning for biomedical literature triage," *PLoS ONE*, vol. 9, no. 12, Dec. 2014, Art. no. e115892.

[20] S. Wang, H. Scells, B. Koopman, and G. Zuccon, "Can ChatGPT write a good Boolean query for systematic review literature search?" in *Proc. 46th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.* New York, NY, USA: Association for Computing Machinery, Jul. 2023, pp. 1426–1436.

[21] R. van Dinter, C. Catal, and B. Tekinerdogan, "A decision support system for automating document retrieval and citation screening," *Exp. Syst. Appl.*, vol. 182, Nov. 2021, Art. no. 115261.

[22] Q. Wang, J. Liao, M. Lapata, and M. Macleod, "PICO entity extraction for preclinical animal literature," *Systematic Rev.*, vol. 11, no. 1, pp. 1–12, Sep. 2022.

[23] D. Golinelli, A. G. Nuzzolese, F. Sanmarchi, L. Bulla, M. Mongiovì, A. Gangemi, and P. Rucci, "Semi-automatic systematic literature reviews and information extraction of COVID-19 scientific evidence: Description and preliminary results of the COKE project," *Information*, vol. 13, no. 3, p. 117, Feb. 2022.

[24] S. Liu, Y. Sun, B. Li, W. Wang, F. T. Bourgeois, and A. G. Dunn, "Sent2Span: Span detection for PICO extraction in the biomedical text without span annotations," in *Proc. Findings Assoc. Comput. Linguistics*, 2021, pp. 1705–1715.

[25] S. Farnsworth, G. Gurdin, J. Vargas, A. Mulyar, N. Lewinski, and B. T. McInnes, "Extracting experimental parameter entities from scientific articles," *J. Biomed. Informat.*, vol. 126, Feb. 2022, Art. no. 103970.

[26] S. Goldfarb-Tarrant, A. Robertson, J. Lazic, T. Tsouloufi, L. Donnison, and K. Smyth, "Scaling systematic literature reviews with machine learning pipelines," in *Proc. 1st Workshop Scholarly Document Process.*, 2020, pp. 184–195.

[27] Y. Hu, Y. Chen, R. Huang, Y. Qin, and Q. Zheng, "A hierarchical convolutional model for biomedical relation extraction," *Inf. Process. Manag.*, vol. 61, no. 1, Jan. 2024, Art. no. 103560.

[28] S. Suster, T. Baldwin, J. H. Lau, A. J. Yepes, D. M. Iraola, Y. Otmakhova, and K. Verspoor, "Automating quality assessment of medical evidence in systematic reviews: Model development and validation study," *J. Med. Internet Res.*, vol. 25, Mar. 2023, Art. no. e35568.

[29] H. Scells, G. Zuccon, and B. Koopman, "A comparison of automatic Boolean query formulation for systematic reviews," *Inf. Retr. J.*, vol. 24, no. 1, pp. 3–28, Feb. 2021.

[30] A. Velez-Estevez, I. J. Perez, P. García-Sánchez, J. A. Moral-Munoz, and M. J. Cobo, "New trends in bibliometric APIs: A comparative analysis," *Inf. Process. Manag.*, vol. 60, no. 4, Jul. 2023, Art. no. 103385.

[31] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick, D. Eide, K. Funk, Y. Katsis, R. Kinney, and Y. Li, "CORD-19: The COVID-19 open research dataset," in *Proc. ACL Workshop Natural Lang. Process. COVID-19 (NLP-COVID)*, 2020.

[32] H. Scells, G. Zuccon, B. Koopman, and J. Clark, "Automatic Boolean query formulation for systematic review literature search," in *Proc. Web Conf.*, Apr. 2020, pp. 1071–1081.

[33] M. Pourreza and F. Ensan, "Towards semantic-driven Boolean query formalization for biomedical systematic literature reviews," *Int. J. Med. Informat.*, vol. 170, Feb. 2023, Art. no. 104928. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1386505622002428

[34] W. Hersh, *Information Retrieval: A Health and Biomedical Perspective*. Berlin, Germany: Springer, 2008.

[35] S. Wang, H. Li, H. Scells, D. Locke, and G. Zuccon, "MeSH term suggestion for systematic review literature search," in *Proc. Australas. Document Comput. Symp.*, Dec. 2021, pp. 1–8.

[36] S. Wang, H. Scells, B. Koopman, and G. Zuccon, "Automated MeSH term suggestion for effective query formulation in systematic reviews literature search," *Intell. Syst. with Appl.*, vol. 16, Nov. 2022, Art. no. 200141.

[37] G. V. Cormack and M. R. Grossman, "Multi-faceted recall of continuous active learning for technology-assisted review," in *Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.* New York, NY, USA: Association for Computing Machinery, Aug. 2015, pp. 763–766, doi: 10.1145/2766462.2767771.

[38] G. V. Cormack and M. R. Grossman, "Technology-assisted review in empirical medicine: Waterloo participation in CLEF eHealth 2017," in *Proc. CLEF*, vol. 11, 2017, pp. 1–11.

[39] A. Lagopoulos, A. Anagnostou, A. Minas, and G. Tsoumakas, "Learning-to-rank and relevance feedback for literature appraisal in empirical medicine," in *Proc. Int. Conf. Cross-Language Eval. Forum Eur. Lang.* Springer, 2018, pp. 52–63.

[40] J. Zou and E. Kanoulas, "Towards question-based high-recall information retrieval: Locating the last few relevant documents for technology-assisted reviews," *ACM Trans. Inf. Syst.*, vol. 38, no. 3, pp. 1–35, Jul. 2020.

[41] J. Zou, D. Li, and E. Kanoulas, "Technology assisted reviews: Finding the last few relevant documents by asking yes/no questions to reviewers," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jun. 2018, pp. 949–952.

[42] G. M. Di Nunzio, "A study of an automatic stopping strategy for technologically assisted medical reviews," in *Proc. 40th Eur. Conf. IR Res. (ECIR), Adv. Inf. Retr.*, Grenoble, France. Springer, 2018, pp. 672–677.

[43] G. M. Di Nunzio, "A study on a stopping strategy for systematic reviews based on a distributed effort approach," in *Proc. 11th Int. Conf. CLEF Assoc., Exp. IR Meets Multilinguality, Multimodality, Interact.*, Thessaloniki, Greece. Springer, 2020, pp. 112–123.

[44] N. Hollmann and C. Eickhoff, "Ranking and feedback-based stopping for recall-centric document retrieval," in *Proc. CLEF Workshop*, 2017, pp. 7–8.

[45] E. Kanoulas, D. Li, L. Azzopardi, and R. Spijker, "CLEF 2019 technology assisted reviews in empirical medicine overview," in *Proc. CEUR Workshop*, vol. 2380, 2019, p. 250.

[46] E. Kanoulas, D. Li, L. Azzopardi, and R. Spijker, "CLEF 2018 technologically assisted reviews in empirical medicine overview," in *Proc. CEUR Workshop*, vol. 2125, 2018, pp. 1–34.

[47] L. Goeuriot, L. Kelly, H. Suominen, A. Neveol, A. Robert, E. Kanoulas, R. Spijker, J. Palotti, and G. Zuccon, "CLEF 2017 eHealth evaluation lab overview," in *Proc. 8th Int. Conf. CLEF Assoc., Exp. IR Meets Multilinguality, Multimodality, Interact.*, Dublin, Ireland. Springer, 2017, pp. 291–303.

[48] H. Yuan, Z. Yuan, R. Gan, J. Zhang, Y. Xie, and S. Yu, "BioBART: Pretraining and evaluation of a biomedical generative language model," in *Proc. 21st Workshop Biomed. Lang. Process.*, 2022, pp. 97–109.

[49] G. H. de Rosa and J. P. Papa, "A survey on text generation using generative adversarial networks," *Pattern Recognit.*, vol. 119, Nov. 2021, Art. no. 108098.

LEANDRA BUDAU received the B.Eng. degree in computer engineering with a specialization in software engineering from Toronto Metropolitan University, in 2023, where she is currently pursuing the M.A.Sc. degree in computer and electrical engineering and additionally enhancing their ethical understanding of neural networks through pursuing a professional certificate in Ethics. While this paper marks their inaugural research contribution, they are actively engaged in additional research endeavors focused on *Natural Language Processing* and *Boolean Query Formalization*.

FAEZEH ENSAN received the Ph.D. degree in computer science from the University of New Brunswick, Fredericton, Canada, in 2011. From 2011 to 2019, she was a Research Assistant with The University of British Columbia, and a Data Scientist with the Semantic Technologies Laboratory, Athabasca University, Canada. Since 2019, she has been an Assistant Professor with the Department of Electrical, Computer, and Biomedical Engineering, Toronto Metropolitan University, Toronto, Canada. She has no JBI publications or review work. She has published in several venues, such as *Information Systems Journal*, *Knowledge and Information Systems* journal, *Information Processing and Management*, AAAI, CIKM, WSDM, and *Information Processing and Management*. Also, she was the Program Co-Chair of *Canadian Semantic Web*, in 2009, and edited a subsequent book published by *Canadian Semantic Web: Technologies and Applications* (Springer). She was a Guest Editor of *Information Systems Journal* (Elsevier) and a Guest Editor of *Journal of Biomedical Informatics* (Elsevier).