Can ChatGPT Write a Good Boolean Query for Systematic Review Literature Search?

Shuai Wang
The University of Queensland
Brisbane, Australia
shuai.wang5@uq.net.au

Harrisen Scells
Leipzig University
Leipzig, Germany
harry.scells@uni-leipzig.de

Bevan Koopman
CSIRO
Brisbane, Australia
bevan.koopman@csiro.au

Guido Zuccon
The University of Queensland
Brisbane, Australia
g.zuccon@uq.edu.au

ABSTRACT

Systematic reviews are comprehensive literature reviews for a highly focused research question. These reviews are considered the highest form of evidence in medicine. Complex Boolean queries are developed as part of the systematic review creation process to retrieve literature, as they permit reproducibility and understandability. However, it is difficult and time-consuming to develop high-quality Boolean queries, often requiring the expertise of expert searchers like librarians. Recent advances in transformer-based generative models have shown their ability to effectively follow user instructions and generate answers based on these instructions. In this paper, we investigate ChatGPT as a means for automatically formulating and refining complex Boolean queries for systematic review literature search. Overall, our research finds that ChatGPT has the potential to generate effective Boolean queries. The ability of ChatGPT to follow complex instructions and generate highly precise queries makes it a tool of potential value for researchers conducting systematic reviews, particularly for rapid reviews where time is a constraint and where one can trade off higher precision for lower recall. We also identify several caveats in using ChatGPT for this task, highlighting that this technology needs further validation before it is suitable for widespread uptake.

CCS CONCEPTS

• **Information systems** → **Query suggestion**; • **Human-centered computing** → **Interaction design process and methods**; • **Computing methodologies** → **Natural language generation**; • **Applied computing** → **Health informatics**.

KEYWORDS

Boolean query, Systematic Review, Prompt Engineering, ChatGPT

1 INTRODUCTION

To construct a high-quality systematic review, the researchers of the review examine all evidence related to a research topic; those relevant to the research topic will be selected and further evaluated and synthesised. Boolean queries are then developed to search medical databases to gather all evidence related to the review. Boolean queries provide reproducibility, explainability, and the benefit of filtering out articles irrelevant to the research topic [39], reducing the workload of unnecessary document assessments. However, constructing a high-quality Boolean query is challenging, even for experienced searchers, and the resulting queries are often suboptimal [59]. Automatic Boolean query formulation and refinement methods have been developed to help researchers construct better systematic review Boolean queries [6, 49, 59, 62–66, 77, 78, 80, 80]. With formulation, we refer to the task of creating a Boolean query from scratch, at times relying on one or more "seed" documents, i.e., example studies related to the systematic review. With refinement, we refer to the task of improving an existing Boolean query, e.g., by adding or removing Boolean clauses and operators to yield a higher precision (less irrelevant documents retrieved) while maintaining recall. Although results have been promising and methods have been integrated into tools for assisting in the query creation process [34, 58, 60, 78], automated techniques are still far from able to yield high-quality Boolean queries.

Recent advances in text generation models [21, 35] have improved task-based question-answering effectiveness and provide high-quality responses based on user intentions [35]. One such generative model, ChatGPT[1], is at the time of publication considered the most expressive text-generation model and has achieved state-of-the-art effectiveness across many natural language processing tasks. This paper investigates using ChatGPT to construct systematic review Boolean queries. The following research questions guide our investigation into using ChatGPT:

RQ1: How does ChatGPT compare with current state-of-the-art methods for formulating and refining systematic review Boolean queries?

RQ2: To what extent do the prompts used to generate systematic review Boolean queries impact the effectiveness of the Boolean queries produced by ChatGPT?

RQ3: What is the effect of guiding the query formulation process with ChatGPT through multiple prompts that mimic the process of the current state-of-the-art automated Boolean query generation methods?

RQ4: What are the caveats and potential challenges of using ChatGPT to create systematic review Boolean queries?

---
[1]https://openai.com/blog/chatgpt/

To our knowledge, this is the first such attempt to comprehensively evaluate the effectiveness of using ChatGPT for creating Boolean queries. This includes carefully engineering prompts for ChatGPT and thoroughly evaluating using standardised test collections for systematic reviews.

We find that ChatGPT compares favourably with current state-of-the-art query generation methods. Improvements in precision sometimes come at the expense of a recall, although we show that this may be mitigated with better MeSH term handling and the practice of snowballing (locating more studies via citation networks). We show generating a good prompt is paramount. Providing a sample Boolean query as part of the prompt is beneficial. Multiple prompting interactions with ChatGPT are better than a single interaction. While results are promising, they are also volatile in that effectiveness varies across runs and prompts; MeSH suggestion, in particular, is poor with ChatGPT.

The findings of this research can help systematic review researchers use ChatGPT for systematic review Boolean query construction and refinement, and understand its limitations and caveats.

2 RELATED WORK

Systematic Review Automation. Constructing systematic reviews requires manual effort by trained professionals across several phases.

Phase 1 is where a specific research question is defined; software can help in defining a good question [1].

Phase 2 is where Boolean queries based on the research question are developed. The query defines which (and how many) results are returned, so their quality greatly impacts all later phases. This is why we focus specifically on query generation. Automated methods have been developed for query formulation [13, 34, 55], query refinement [34, 60] and query exploitation [7, 17].

Phase 3 is where all the titles and abstracts retrieved by the query are screened, manually assessing them for relevance. The primary automation approach to help here is screen prioritisation [3–5, 10, 16, 18, 19, 26, 31, 33, 42, 43, 46, 47, 61, 68, 69, 81, 82], which involves ranking the studies according to their likely relevance to the research question. Once ranked, automated approaches can define a 'stopping criteria', after which screening does not continue because all relevant documents are likely to be found. Active learning approaches [15, 44] are often employed during screening to help rank include studies to be screened.

Phase 4 involves extracting specific details from studies relevant to the review; data extraction methods can help here [27, 71].

Phase 5 Phase 5 is the synthesis of all the evidence into a single coherent review document. Synthesis automation [12, 50, 54, 72, 73] have been developed here; they include text thematic analysis [75] and even text generation [51].

Query Formulation and Refinement. Query formulation is the process of deriving a Boolean query, based on the research question, according to a specific set of guidelines. The guideline presents two accepted procedures that exist for developing a query for systematic reviews. These procedures describe the steps one should take when developing a query. The first procedure is called the *conceptual* method [11]. This procedure requires first identifying several high-level concepts from pilot searches or potentially relevant studies known a priori. These high-level concepts are then used to discover synonyms and related keywords. The query is then iteratively refined, a process driven by the expertise of the information specialist. The second procedure is called the *objective* method [24, 67]. The first step of this procedure is to create a small set of potentially relevant studies that will seed the rest of the procedure, e.g., through pilot searchers, as in the conceptual method. Multiple statistical procedures follow, which extract terms from these studies and weakly validate the query. The identified terms must still be added to the query manually, which, like the conceptual method, is driven by the expertise of the information specialist. The exact details of these methods are unimportant for understanding their use in this paper; they should be understood simply as a series of steps one can follow to arrive at a query.

Naturally, these methods require a considerable amount of time (as the information specialist must perform multiple pilot searches and spend time validating the search) [31, 52] and are prone to human error [56, 57]. To this end, Scells et al. [63] investigated automating these two query formulation procedures. The main finding of this line of research was that computationalising these procedures could not match the effectiveness of humans; however, further manual refinement of the automatically generated queries dramatically improved retrieval effectiveness.

Automatic query refinement was developed based on the observed benefit from manual query refinement. Such methods take an initial human authors query and apply a series of transformations (adding terms or clauses) to make the query more effective [34, 60]. In combination with query visualisation [60] tools, these query refinement tools were able to improve the initial query.

Learnings from the query formulation approach drive two clear directions for using ChatGPT to automate query formulation: the first is to allow ChatGPT to generate queries however it sees fit, and the second is to guide ChatGPT by prompting it to follow the instructions of the conceptual or objective procedures. We refer to the first method as *unguided* and the second method as *guided*. Learnings from the query refinement work made us hypothesise that providing an existing query to ChatGPT and asking for a refinement could be beneficial.

Prompt Engineering for Generative Transformer-based Generative Language Models. Prompt engineering is the process of guiding a generative language model to perform a particular task. In some respect, prompt engineering can be seen as a way to program a generative language model through natural language [53]. A popular way of guiding model output through prompt engineering is for text-to-image generative models [37, 48]. This 'zero-shot' or 'few-shot' approach to tasks with generative language models has also achieved state-of-the-art results on several natural language tasks [8]. More recently, prompt engineering has been applied to natural language tasks for medicine, such as question answering [36, 38]. The use of generative language models for query formulation or refinement is relatively under-explored. We are only aware of a single other paper that has published results about using a generative language model for query expansion [14], and no work has examined this method for creating Boolean queries for systematic literature review search.

Table 1: Prompts for unguided prompt query formulation

	Prompt ID	Prompt
Simple	q1	For a systematic review titled "{review_title}", can you generate a systematic review Boolean query to find all included studies on PubMed for the review topic?
Detailed	q2	You are an information specialist who develops Boolean queries for systematic reviews. You have extensive experience developing highly effective queries for searching the medical literature. Your specialty is developing queries that retrieve as few irrelevant documents as possible and retrieve all relevant documents for your information need. Now you have your information need to conduct research on {review_title}. Please construct a highly effective systematic review Boolean query that can best serve your information need.
Detailed	q3	Imagine you are an expert systematic review information specialist; now you are given a systematic review research topic, with the topic title "{review_title}". Your task is to generate a highly effective systematic review Boolean query to search on PubMed (refer to the professionally made ones); the query needs to be as inclusive as possible so that it can retrieve all the relevant studies that can be included in the research topic; on the other hand, the query needs to retrieve fewer irrelevant studies so that researchers can spend less time judging the retrieved documents.
With Example	q4	You are an information specialist who develops Boolean queries for systematic reviews. You have extensive experience developing highly effective queries for searching the medical literature. Your specialty is developing queries that retrieve as few irrelevant documents as possible and retrieve all relevant documents for your information need. You are able to take an information need such as: "{example_review_title}" and generate valid pubmed queries such as: "{example_review_query}". Now you have the information need to conduct research on "{review_title}", please generate a highly effective systematic review Boolean query for the information need.
With Example	q5	You are an information specialist who develops Boolean queries for systematic reviews. You have extensive experience developing highly effective queries for searching the medical literature. Your specialty is developing queries that retrieve as few irrelevant documents as possible and retrieve all relevant documents for your information need. A professional information specialist will extract PICO elements from information needs in a common practice in constructing a systematic review Boolean query. PICO means Patient/ Problem, Intervention, Comparison and Outcome. PICO is a format for developing a good clinical research question prior to starting one's research. It is a mnemonic used to describe the four elements of a sound clinical foreground question. You are able to take an information need such as: "{example_review_title}" and you generate valid pubmed queries such as: "{example_review_query}". Now you have your information need to conduct research on "{review_title}". First, extract PICO elements from the information needs and construct a highly effective systematic review Boolean query that can best serve your information need.

3 BOOLEAN QUERY PROMPT ENGINEERING

In this paper, we investigate the use of ChatGPT to create queries for systematic review literature search. The basic mechanism employed by ChatGPT is to take an input sequence of text (called *prompt*), process it through the model, and output the next token in the sequence. This process is repeated several times to generate a complete response to the prompt. ChatGPT relies upon the Transformer architecture [76] trained on a massive amount of text data, allowing it to learn patterns in the way that language is used. During the generation process, the model uses these learned patterns to generate text that is similar to the text observed at training.

A key component in the use of and interaction with ChatGPT is the design of the prompt used to instruct the model to generate an answer. To instruct ChatGPT to create Boolean queries for systematic review literature search, we iteratively devise a number of prompts of increased complexity, including prompts that rely on example Boolean queries. We also experiment with prompts that guide the generation of the query through multiple interactions (guided prompts), and by leveraging insights from the existing objective method of query formulation [25, 63, 65]. Prompts are devised for two tasks related to Boolean query generation: the formulation task and the refinement task.

3.1 Unguided Prompts for Query Formulation

Unguided query formulation prompt refers to a prompt that instructs to formulate systematic review Boolean queries using the title of the review. Table 1 reports the five prompts we developed for the Boolean query formulation task. They can be classified into three categories: *simple*, *detailed* and *with examples*. A *simple* prompt only uses one sentence to state the task for ChatGPT briefly. This prompt is likely the most common usage for ChatGPT when users are not experts at constructing high-quality prompts. A *detailed* uses a background story that clearly explains what is required from ChatGPT. Finally, a *with example* prompt includes an additional query formulation example to give more context for ChatGPT to understand the task better.

The difference between the two *detailed* prompts 'q2' and 'q3' is the way we describe the background story; for the *with example* prompts 'q4' and 'q5', 'q5' additionally describes a Boolean query formulation strategy to identify PICO elements. PICO refers to Patient/Problem, Intervention, Comparison and Outcome, and it is often used to formulate the research question and generate high-quality systematic review Boolean queries [20]. The intuition for including PICO element extraction in the prompt is to help ChatGPT better understand how the example query is formulated logically.

Table 2: Prompts for unguided prompt query refinement

	Prompt ID	Prompt
Simple	q6	For a systematic review seed Boolean query: "{initial_query}", This query retrieves too many irrelevant documents and too few relevant documents about the information need: "{review_title}", Please correct this query so that it can retrieve fewer irrelevant documents and more relevant documents.
With Example	q7	For a systematic review seed Boolean query: "{example_review_initial_query}" ,This query retrieves too many irrelevant documents and too few relevant documents about the information need: "{example_review_title}", therefore it should be corrected to: "{example_review_refined_query}". Now your task is to correct a systematic review Boolean query: "{initial_query}" for information need "{review_title}", so it can retrieve fewer irrelevant documents and more relevant documents.

For each systematic review topic, we replace "{*review_title*}" with the title of the review. For prompts with an example, we further replace "{*example_review_title*}" with the title of the example review topic and "{*example_review_query*}" with the Boolean query used by that example review. For instance, for the example title *"Thromboelastography (TEG) and rotational thromboelastometry (ROTEM) for trauma-induced coagulopathy in adult trauma patients with bleeding"*, we insert the Boolean query: (Thrombelastography[mesh:noexp] OR (thromboelasto*[All Fields] OR thromblelasto*[All Fields] OR ROTEM[All Fields] OR "tem international"[All Fields] OR (thromb*[All Fields] AND elastom*[All Fields]) OR (rotational[All Fields] AND thromboelast[All Fields])) OR (Thrombelastogra*[All Fields] OR Thromboelastogra*[All Fields] OR TEG[All Fields] OR haemoscope[All Fields] OR haemonetics[All Fields] OR (thromb*[All Fields] AND elastogra*[All Fields])))

3.2 Unguided Prompts for Query Refinement

Unguided query refinement prompts are prompts that provide ChatGPT with a title of a systematic review and a corresponding Boolean query for that review, and instruct the model to produce a modification of that query that leads to higher search effectiveness. These prompts could be used by reviewers to improve a query they have designed, or could be integrated into a more complex automation pipeline. Table 2 contains our query refinement prompts.

We categorise prompts for query refinement into *simple* and *with example*. Similar to the unguided prompts for query formulation, *simple* prompt 'q6' indicates that the prompt is constructed only by briefly describing the task. For the *with example* prompt 'q7', we include an example that instructs ChatGPT what it means to successfully refine the query. When prompting ChatGPT to refine queries, we replace "{*initial_query*}" with the initial Boolean query we want ChatGPT to refine. For prompts that require an example, we further replace {*example_review_title*} with the title of the example review topic; "{*example_review_initial_query*}" with the initial query of the example topic, and "{*example_review_refined_query*}" with the final query refined in the example topic.

We investigate the effectiveness of two types of examples: (1) one high-quality systematic review example (HQE) and (2) an example that is similar or related to the querying topic (RE). To identify a related example, we use a monoBERT architecture to find the closest example from the test collection [45]. Specifically, we concatenate the review title of the querying topic with the title of one potential example review, and we pass it through PubMedBERT, a domain-specific BERT model pre-trained on PubMed abstracts [23]. We get a final classification score which refers to the relatedness of the potential example topic to the querying topic; we select the top-ranked example to include in the prompt during prompt creation.

3.3 Guided Prompts for Query Formulation

We design a multi-step prompt that follows the same logic from one of the current state-of-the-art automated query formulation methods, namely the objective method [63]. Table 3 reports our designed guided prompt. Specifically, we follow a four-step pipeline to generate the Boolean query. In the first step, query terms are identified using a seed study. In the second step, these terms are classified into four categories: terms relating to health conditions (A), terms relating to treatments (B), terms relating to types of study design (C), or others (N/A). For the third step, terms in the same category are combined by 'OR', and final queries are combined by 'AND'. Finally, ChatGPT is asked to refine the query by adding more terms, such as MeSH Terms. A guided prompt can better control the terminology of Boolean query formulation from ChatGPT. It also divides the complex task into multiple, more manageable subtasks.

4 EXPERIMENTAL SETTINGS

Our experiments are conducted using two collections: The CLEF technological assisted reviews (TAR) collections [28–30] and a systematic review collection with seed studies (seed collection) [79]. We use the CLEF TAR 2017 and 2018 collections as they both consist of only diagnostic test accuracy (DTA) systematic reviews. Some topics from CLEF TAR 2017 are duplicated in the 2018 collection, so we obtained 72 review topics after removing these duplicates. The seed collection contains 40 topics along with the seed studies used to guide systematic review researchers during Boolean query formulation. For both collections, each topic includes the title of the review, the Boolean query used for literature retrieval, and the relevance assessments of the retrieved literature. The *with examples* prompts use topic CD010438 from the CLEF TAR collection as the HQE as it has a simple structure for ChatGPT to follow.

The guided prompt method relies on the objective method, which requires at least one seed study to start with query formulation. Thus we did not run guided prompting on CLEF TAR as it does not contain seed studies.

After obtaining Boolean queries from ChatGPT, any incorrectly formatted Boolean queries were removed and generated again for the review topic. To evaluate the generated Boolean queries, we executed them using PubMed's Entrez API to obtain retrieved PubMed IDs [9]. We then used set-based measures such as precision, f-measure, and recall to assess the retrieved PubMed IDs using the abstract-level relevant assessment in the collections.

Table 3: Example designed guided prompt for query formulation.

Step	Prompt	Example Answer
Step1	Follow my instructions precisely to develop a highly effective Boolean query for a medical systematic review literature search. Do not explain or elaborate. Only respond with exactly what I request. First, Given the following statement and text from a relevant study, please identify 50 terms or phrases that are relevant. The terms you identify should be used to retrieve more relevant studies, so be careful that the terms you choose are not too broad. You are not allowed to have duplicates in your list. statement: "Prevalence of Differentiated Thyroid Cancer in Autopsy Studies Over Six Decades: A Meta-Analysis" Text: Ten occult carcinomas of the thyroid gland were found in 274 unselected autopsies at the Pathology Service, Hospital José Joaquín Aguirre, between December 1980 and March 1983. This is the lowest incidence among the most recent published series. The smallest papillary microcarcinomas presented most frequently as nonencapsulated nonsclerosing tumors without inflammatory infiltrate or fibrosis, suggesting that they may represent the early stages of development. Our data show a relatively high and similar frequency of papillary microcarcinomas in surgical and autopsy series, but do not demonstrate risk factors for clinical evolution.	1. Differentiated thyroid cancer 2. Prevalence 3. Autopsy studies 4. Meta-analysis 5. Occult carcinomas 6. Thyroid gland 7. Incidence 8. Environmental factors 9. Etiology 10. Low risk ... 50. Parameters.
Step 2	For each item in the list you created in Step 1, classify it into three categories: terms relating to health conditions (A), terms relating to a treatment (B), terms relating to types of study design (C). When an item does not fit one of these categories, mark it as (N/A). Each item needs to be categorised into (A), (B), (C), or (N/A).	1. (A) Differentiated thyroid cancer 2. (N/A) Prevalence 3. (C) Autopsy studies 4. (C) Meta-analysis 5. (A) Occult carcinomas 6. (A) Thyroid gland 7. (N/A) Incidence 8. (N/A) Environmental factors 9. (N/A) Etiology 10. (N/A) Low risk ... 50. (N/A) Parameters.
Step 3	Using the categorised list you created in step 2, create a Boolean query that can be submitted to PubMed which groups together items from each category. For example: ((itemA1[Title/Abstract] OR itemA2[Title/Abstract] or itemA2[Title/Abstract]) AND (itemB1[Title/Abstract] OR itemB2[Title/Abstract] OR itemB3[Title/Abstract]) AND (itemC1[Title/Abstract] OR itemC2[Title/Abstract] OR itemC3[Title/Abstract]))	(((differentiated[Title/Abstract] OR thyroid[Title/Abstract] OR carcinoma[Title/Abstract] OR papillary[Title/Abstract] OR microcarcinoma[Title/Abstract]) AND (cancer[Title/Abstract] OR incidence[Title/Abstract] OR etiology[Title/Abstract] OR risk[Title/Abstract] OR gender[Title/Abstract] OR hormonal[Title/Abstract] OR nodular[Title/Abstract] OR goiter[Title/Abstract] OR Hashimoto's[Title/Abstract] OR malignancy[Title/Abstract] OR concomitant[Title/Abstract] OR tumor[Title/Abstract] OR infiltrate[Title/Abstract] OR fibrosis[Title/Abstract] OR development[Title/Abstract] OR frequency[Title/Abstract])) AND (autopsy[Title/Abstract] OR surgical[Title/Abstract] OR material[Title/Abstract] OR series[Title/Abstract] OR specimens[Title/Abstract] OR cases[Title/Abstract])).
Step 4	Use your expert knowledge to refine the query, making it retrieve as many relevant documents as possible while minimising the total number of documents retrieved. Also add relevant MeSH terms into the query where necessary, e.g., MeSHTerm[MeSH]. Retain the general structure of the query, however, with each main clause of the query corresponding to a PICO element. The final query still needs to be executable on PubMed, so it should be a valid query.	(((differentiated thyroid cancer[MeSH] OR "differentiated thyroid"[All Fields] OR "thyroid carcinoma"[All Fields] OR "papillary microcarcinoma"[All Fields]) AND (prevalence[All Fields] OR incidence[MeSH] OR "etiology of"[All Fields] OR "risk factors"[All Fields] OR gender[All Fields] OR hormonal[All Fields] OR "nodular goiter"[All Fields] OR "Hashimoto's thyroiditis"[MeSH] OR malignancy[MeSH] OR "concomitant lesion"[All Fields] OR tumor[All Fields] OR infiltrate[All Fields] OR fibrosis[All Fields] OR "early stages of development"[All Fields] OR frequency[All Fields])) AND (autopsy[MeSH] OR surgical[All Fields] OR material[All Fields] OR series[All Fields] OR specimens[All Fields] OR cases[All Fields]))

5 MAIN RESULTS

The following sections contain the results of our unguided and guided query generation experiments.

5.1 Unguided Prompt Query Formulation

Table 4 reports the results of unguided prompt query formulation. These results indicate that queries generated from ChatGPT generally obtain a higher precision compared to the existing state-of-the-art automatic query formulation methods, with a trade-off of lower recall. For F-measure, ChatGPT-generated queries are more effective than both the state-of-the-art and originally authored queries on the CLEF collections. However, they are less effective on the Seed collection. Systematic review literature search generally requires high recall to ensure that all relevant evidence can be found.

All the ChatGPT-generated queries obtain a lower recall than the baseline methods, suggesting that ChatGPT-generated queries may not be suitable for high-recall retrieval, but rather best suited when time is limited; e.g., for rapid reviews [40].

Using different Simple and Detailed prompts (q1–3) only had a minor impact on effectiveness. For CLEF, q2 was statistically significantly better with respect to precision; otherwise, the prompt type did not have a strong effect. However, we found that prompts that include a high-quality systematic review topic as an example are able to significantly outperform those without, shown as a consistently higher F1, F3 and recall. When comparing the effectiveness of two prompts with examples, we found that asking ChatGPT to generate PICO elements before generating its final Boolean query resulted in Boolean queries with a lower recall but higher precision.

Table 4: Unguided prompt query formulation results. *CLEF* **indicates CLEF TAR collection and** *SC* **indicates seed collection. Statistical significant differences (Student's two-tailed, paired t-test with Bonferonni correction, p < 0.05) between q4 and all other methods are indicated by** *.

		Prompts	Precision	F1	F3	Recall
CLEF	Baselines	Original	0.0207*	0.0290*	0.0481*	0.8317*
		Conceptual	0.0015*	0.0027*	0.0101*	0.6997*
		Objective	0.0002*	0.0005*	0.0023*	**0.9128***
	Simple	q1	0.0543	0.0500	0.0590	0.1293*
	Detailed	q2	**0.1166**	0.0654	0.0696	0.1310*
		q3	0.0844	0.0443	0.0497*	0.1175*
	With Example	q4	0.0752	0.0642	**0.0847**	0.5035
		q5	0.0958	**0.0717**	0.0844	0.3335*
SC	Baselines	Original	0.0367	**0.0651***	**0.1099***	**0.7366***
		Conceptual	0.0018			0.4138
		Objective	0.0057			0.5192
	Simple	q1	0.0501	0.0274	0.0298	0.0528
	Detailed	q2	**0.0983**	0.0310	0.0278	0.0394*
		q3	0.0730	0.0329	0.0329	0.0519
	With Example	q4	0.0283	0.0274	0.0374	0.1290
		q5	0.0188	0.0193	0.0271	0.0785

Table 5: Comparison of result for unguided prompt query generation prompt 'q4' when using a different types of examples. *CLEF* **indicates CLEF TAR collection, and** *SC* **indicates seed collection; For each collection, two types of example are used,** $q4 - HQE$ **refers to using one high-quality example, while** $q4 - RE$ **refers to using a related query as an example. Statistical significant differences (p < 0.05) between the two types of examples are indicated by** *.

		Precision	F1	F3	Recall
CLEF	q4-HQE	0.0751	0.0642	0.0872	0.5035
	q4-RE	0.1105(+47.1%)	0.0909(+41.6%)	0.1144(+31.2%)	0.4183(−38.1%)
SC	q4-HQE	0.0283	0.0274	0.0374	0.129
	q4-RE	0.0351(+24.0%)	0.0140(−48.9%)	0.0139(−62.8%)	0.0161*(−87.6%)

Overall, our findings indicate that including a high-quality systematic review query example in the prompt is crucial, while the level of detail in the task description may not have a significant impact.

Next, to assess the impact of query construction example on the effectiveness of generated Boolean queries using ChatGPT, we further test the effectiveness when different types of example is used, as described in section 3.1. Table 5 compares the effectiveness of queries generated using the most relevant topic in the prompt to queries generated using a high-quality example. Using a relevant topic as an example can result in queries with higher precision, but lower recall. ChatGPT will generate different responses for the same prompt. To study the variability of effectiveness, we select the more effective prompt, q4, and we re-run the prompt ten times. Variability

(a) CLEF Precision   **(b) CLEF F1**   **(c) CLEF Recall**

(d) SC Precision   **(e) SC F1**   **(f) SC Recall**
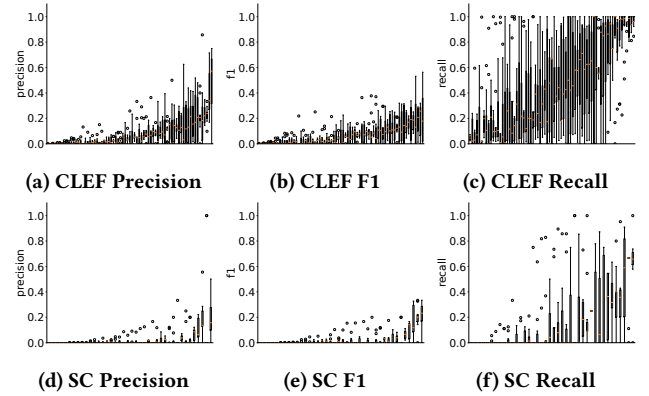
Figure 1: Topic-by-topic variability for the effectiveness of 10 iterative runs in unguided prompt query formulation. *CLEF* **indicates CLEF TAR collection and** *SC* **indicates seed collection.**

Table 6: Result table for unguided prompt query refinement on CLEF TAR collection. For a refinement method, 'q6-Original', 'q6' indicates the prompt used to generate the refined query; 'Original' indicate the seed queries used for ChatGPT to refine. For each query refinement method, statistical significant differences (p < 0.05) between refined prompt and seed queries are indicated by *.

Prompts	Precision	F1	F3	Recall
Original	0.0207	0.0290	0.0481	0.8317
q6-Original	0.0795*	0.0597*	0.0802*	0.5060*
Conceptual	0.0014	0.0027	0.0100	0.6996
q7-conceptual	0.0022	0.0039	0.0069	0.2699*
Objective	0.0002	0.0005	0.0023	**0.9128**
q7-Objective	0.0460*	0.0471*	0.0652*	0.8115*
q4	0.0751	0.0642	0.0872	0.5035
q7-q4	**0.1162**	**0.0772**	**0.0921**	0.3179*

is shown in Figure 1. Recall varied more than precision and F-measures: variance of recall is 12% of its mean value; precision was 7.1%, F-1 6.6% and F-3 7.2%. Our result indicates that the generated queries from the same prompt would mainly differ in the ability to obtain more relevant documents (recall) from the included studies.

5.2 Unguided Prompt Query Refinement

Table 6 suggests that ChatGPT is capable of improving the effectiveness of systematic review literature search queries. ChatGPT for query refinement leads to an increase in precision and F-measure, while obtaining a lower recall. Therefore, it is crucial to first create a query with high recall, and then use ChatGPT to refine.

Using automatically formulated queries from the objective queries obtained the highest effectiveness when ChatGPT was used to refine the queries. Refinement yielded an 11% drop in recall, but considerable gains in precision, F1, and F3.
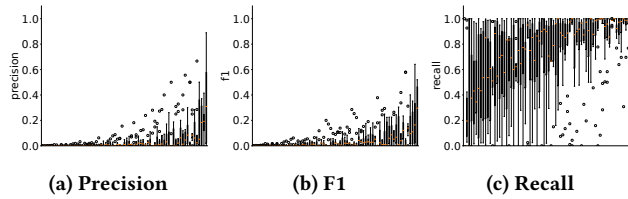
(a) Precision　　　　　(b) F1　　　　　(c) Recall

Figure 2: Topic-by-topic variability for the effectiveness of 10 iterative runs in unguided prompt query refinement.

Table 7: Guided prompt query formulation on Seed Collection, compared with unguided prompt query generation 'q4'; Statistical significant differences (p < 0.05) between guided prompt and unguided prompt is indicated by *.

Prompts	Precision	F1	F3	Recall
q4	0.0284	0.0274	0.0374	0.1290
Guided	0.0993(+249.6%)	0.0492(+79.6%)	0.0565(+51.1%)	0.5171(+301.6%)*

Variability is studied by running the best method, refined-objective run, ten times. We show the variability of the query refinement in Figure 2. There is less variance in query refinement than in query formation (Figure 1). This is understandable, given the query structure is already provided by the seed query, whereas query formulation must be done from scratch from the title of the review.

5.3 Guided Prompt Query Formulation

Table 7 reports the result of guided prompt query formulation. These results suggest that if using a well-chosen seed study, queries generated from guided prompt are more effective than queries generated from unguided prompts.

However, like previous experiments in unguided prompt query generation and refinement, the effectiveness varies considerably across runs; and the effectiveness also depends on the seed study being used. Figure 3 reports the variability of query effectiveness when different seed studies are used to generate queries. Figure 4 reports the variability of effectiveness when the same seed study is used, picked from the best seed study from the first run.

From the variability graph, we see that query generation using guided prompt is not stable across different seed studies. Furthermore, even when the same seed study is used to generate multiple queries, there is a high degree of variability. The range of precision and recall for some topics can span from 0 to 1, especially when the average effectiveness is high.

6 QUERY FAILURE ANALYSIS

We attempt to uncover the characteristics of queries with poor effectiveness, with the intention of identifying ways to improve query generation. To do this, we follow the steps below to select a set of successful and failing queries:

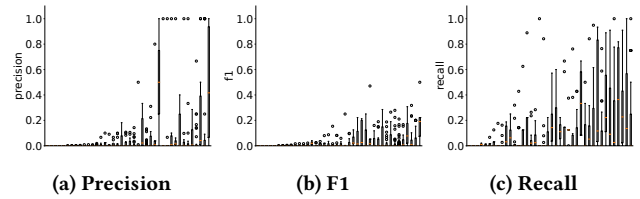(1) Select the best method for query formulation (q4) and query refinement (q7-objective).

(a) Precision　　　　　(b) F1　　　　　(c) Recall

Figure 3: Topic-by-topic variability for the effectiveness of using different seed studies for guided prompt query formulation.

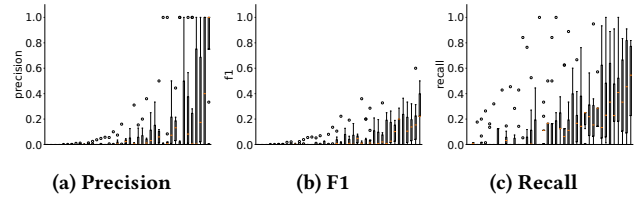(a) Precision　　　　　(b) F1　　　　　(c) Recall

Figure 4: Topic-by-topic variability for the effectiveness of 10 iterative runs using the same seed study in guided prompt query formulation.

(2) From the 10 iterative runs from each task, we extract an 'oracle' result, using recall as an indicator of effectiveness. (Tie broken using precision.)

(3) Next, we compare oracle effectiveness with that of the original query. Successful topics are defined as those where the oracle has higher precision and recall than the original query; failing topics are those where the oracle has lower precision and recall than the original query. The intuition of selecting successful topics and failing topics above is based on the fact that higher recall often means more documents are retrieved from the query, and results in lower precision.

Only the CLEF datasets were used as the Seed collection contains topics that are not all the same type of systematic review; e.g., scoping reviews. Using this method above, we identify seven successful topics[2] and six failed topics[3] for query formulation. For query refinement, we identify three succesfull topics[4] and eleven topics for query refinement[5]. By comparing the failed queries with other queries, we summarise three key findings:

(1) ChatGPT generated queries that are less effective tended to retrieve a large number of results: for comparison between successful queries and failing queries for query formulation, the median ratio for successful queries is 0.30, while for failing queries, this ratio is 10.28. Similarly for query refinement, the median ratio for successful queries and failing queries are 1.39 and 2.74, respectively. This means that failing queries may be categorised by having a large number of OR clauses; thus, it may be possible to detect such queries and introduce mitigation strategies.

---

[2]CD008759, CD009323, CD007427, CD010023, CD010213, CD011515, CD008686
[3]CD010705, CD009591, CD010680, CD010502, CD010296 and CD011126.
[4]CD010023, CD010213, CD007427
[5]CD010657, CD012165, CD009135, CD011926, CD008643, CD011134, CD012083, CD008782, CD009579, CD008691, CD012010.

(2) We also notice that some of the ChatGPT queries contain many incorrect MeSH Terms. For the best-performing prompt for query formulation, on average, three MeSH Terms were generated, and 55% of the MeSH Terms generated were not in the MeSH vocabulary. On the other hand, for the query refinement prompt, only 1.5 MeSH Terms were generated on average per query, with 66% of the MeSH Terms not in the MeSH vocabulary. There is no strong correlation between the ratio of incorrectly generated MeSH Terms and query effectiveness. However, we also acknowledge that this does not mean certain incorrect MeSH Terms do not have a major negative impact on effectiveness. Further experiment is needed (e.g., MeSH Term correction [77, 80]) to check how these MeSH Terms impact query effectiveness and how would effectiveness changes if MeSH Terns are corrected.

(3) Relevance judgements were done using the original query, so it may be biased towards this. In all the failing queries, the number of the unjudged document in the retrieved documents set accounts for more than 94.8% of the retrieved documents, much higher than that from the original queries at 63.7%; On the other hand, the portion of unjudged documents from generated and original queries for other topics was similar, at 56.2% and 55.4%, respectively. It may be that the failing queries simply retrieved many more unjudged documents because they had poor performance. But there is also a possibility that they actually managed to find clusters of documents relevant to topics but never assessed by reviewers. While this is conjecture, it does spark some ideas for solutions to improve such queries. If there are, in fact, clusters of relevant documents amongst a large number of non-relevant ones, then automated classification techniques may help to uncover these [2, 32, 41, 70, 74]. In particular, active learning [15, 44] may be used to selectively assess documents from different clusters to hone in on relevant documents.

7 SUMMARY OF FINDINGS

In light of the empirical results reported in Section 5, we provide a summary of the key findings of our study in terms of the research questions we set to investigate.

7.1 ChatGPT Compared to Existing Methods

The first section addresses our first research question:

RQ1 *How does ChatGPT compare with current state-of-the-art methods for formulating and refining systematic review Boolean queries?*

In terms of automatic query formulation, our results suggest that the use of ChatGPT compares favourably with the current state-of-the-art automated Boolean query generation methods in terms of precision, at the expense of a lower recall.

In terms of automatic query refinement, our results suggest that ChatGPT is effective at refining existing Boolean queries that have been generated by a previous step of automated Boolean query formulation. In particular, the ChatGPT-based refinement of queries

generated using the objective method for Boolean query formulation leads to the most considerable improvements in effectiveness.

In both tasks, improvements are observed in terms of increased precision, i.e. a reduction of irrelevant documents being retrieved by the query. These improvements come at the expense of losses in recall (though for the best query refinement settings, these losses are only marginal). A high recall is often a key requirement for traditional systematic reviews, and thus ChatGPT may not be well suited in these settings. However, we note that typically, in manually created Boolean queries, MeSH terms are used to improve recall, but the queries obtained through ChatGPT contain a large number of incorrect MeSH terms (see Section 6). The addition of a post-processing step that resolves incorrect MeSH terms produced by ChatGPT may alleviate the losses in recall we witnessed in the results. For this, it may be possible to adapt existing methods for MeSH term suggestion [77, 80], which also could be used to further refine the queries produced by ChatGPT by adding more effective MeSH terms the model may have failed to identify. We further note that within the systematic review process, the step of snowballing is designed to further increase recall. Snowballing refers to the recursive analysis of references cited in retrieved documents [22] – in all effects, this process adds to the set to be reviewed studies not retrieved by the Boolean query. Thus, it may be that losses in recall observed when using ChatGPT to create queries may be recovered through snowballing on the identified relevant studies.

7.2 Impact of Prompts on Effectiveness

This section addresses our second research question:

RQ2 *To what extent do the prompts used to generate systematic review Boolean queries impact the effectiveness of the Boolean queries produced by ChatGPT?*

Our results suggest that the type of prompt used has considerable effects on the effectiveness of the queries produced by ChatGPT. We further review the impact of guided prompts in the next research question, and here we only focus on unguided prompts.

In this case, we observe that integrating example Boolean queries into the prompts tend to provide better queries than simple prompts (Table 4). This is particularly the case for the CLEF datasets – while the same does not hold on the seed collection, we note this collection is limited in the number of topics available, so results may be less generalisable. Notably, improvements obtained when including query examples are large for recall (this is also the case on the seed collection), thus partially mitigating the low recall observed throughout the queries produced with ChatGPT. We also observe that the nature of the example query influences results. Specifically, example queries that are semantically close to the systematic review topic for which the new query is being created lead to higher precision. On the other hand, using high-quality but potentially unrelated queries in the prompts, instead, leads to higher recall.

We also have observed that the inclusion in the prompt of instructions to generate PICO elements does lead to considerable losses in recall, and provides improvements in precision only on the CLEF collection.

7.3 Effect of Guided Prompts

This section addresses the following research question:

RQ3 *What is the effect of guiding the query formulation process with ChatGPT through multiple prompts that mimic the process of the current state-of-the-art automated Boolean query generation method?*

We have leveraged the process set out by the existing objective query generation method [63, 65] to design a sequence of prompts that iteratively guide ChatGPT in the process of generating a Boolean query. Our results show that guided prompts lead to higher effectiveness than unguided prompt strategies: improvements are observed for both precision and recall. There are caveats, though, that one needs to be wary about – we discuss these next.

7.4 Caveats and Challenges

RQ4 *What are the caveats and potential challenges of using ChatGPT to create systematic review Boolean queries?*

Our results highlight two main caveats practitioners should be wary of if relying on ChatGPT to create Boolean queries for systematic review literature search: (1) incorrect MeSH terms, (2) high variability in query effectiveness across multiple requests.

We already touched upon the first caveat when analysing RQ1. 55% and over of the MeSH terms generated by ChatGPT are actually not in the MeSH vocabulary, and thus are incorrect. The effect of these incorrect MeSH terms is still unclear, as it is uncertain whether existing methods for automatic MeSH term suggestion could be improved to correct the MeSH terms generated by ChatGPT [77, 80].

The second caveat refers to the results showed in Figures 1- 4, which highlight that if multiple generations are performed with ChatGPT for the same prompt, different queries are generated – and their effectiveness can largely differ. The fact that the output of ChatGPT is non-deterministic given the same prompt. We validated this by issuing our prompts, as well as through other interactions, across separate "chats" to avoid ChatGPT to use the previous input and output as part of the input for the current interaction. suggests that it may use top-k sampling, or a similar technique, for text generation. Top-k sampling involves generating the k most likely next tokens for a given input and then randomly selecting one of those tokens as the next step in the generation process. Nevertheless, aside from the specific technique used for text generation, the non-deterministic nature of ChatGPT and the fact that this has a sensible impact on the effectiveness of the Boolean queries that are generated, pose a threat to the uptake of the methods we investigated in this paper. In fact, users would typically have no or limited insight into the effectiveness of the query that ChatGPT generates and thus may be unable to identify high-yield queries among those generated for the same prompt.

8 LIMITATIONS IN OUR USE OF CHATGPT

In our experiment, we use ChatGPT as a black-box application and test its effectiveness in generating systematic review Boolean queries. One significant limitation of our paper is that details about how ChatGPT works internally are undisclosed, including model architecture and what's included in training data. Therefore, we do acknowledge that ChatGPT may have already seen the original queries for the systematic review topics in our test collection during model training and this might impact the effectiveness of the queries ChatGPT generated. In addition, we are unclear whether interaction history (beyond immediate context of conversation, which the service provider, OpenAI, states being approximately 3000 words or 4000 tokens) may have effect on the generation. Although when executing an experiment for a new query topic we instantiate a new conversation within ChatGPT, we do not know what information the service provider retains for a user and whether they use previous conversations of the user to personalise generation.

Thirdly, our experiments were executed over a two-weeks period in early January 2023. We are unsure whether modifications to the model have been made during this period. OpenAI disclosed the model we used is the January 9 version and no experiment was run once OpenAI updated their model to the Jan 30 version. Yet, we could not control for minor engineering or model updates OpenAI may have made to the model but not disclosed.

9 CONCLUSION

This paper used ChatGPT to formulate and refine Boolean queries for systematic reviews. We designed an extensive set of prompts to investigate these tasks on over 100 systematic review topics. Queries generated by ChatGPT obtained higher precision but lowered recall compared to the queries generated by the current state-of-the-art automatic query formulation methods. We also showed that ChatGPT could generate Boolean queries with even higher effectiveness with a guided prompt. One caveat of our results is that ChatGPT generates different queries even if the same prompt is used, which vary in effectiveness. This issue would need to be resolved before using ChatGPT to generate Boolean queries for systematic review literature search in practice, where reproducibility is paramount. In summary, the paper makes the following contributions:

- To the best of our knowledge, this paper is the first of its kind to investigate the effectiveness of Boolean queries designed for systematic review literature search generated by ChatGPT.
- Several strategies for generating prompts for ChatGPT.
- An extensive evaluation of a large number of systematic review topics and prompts provides insight into how different prompts may impact the effectiveness of the systematic review Boolean queries generated through ChatGPT.
- A query failure analysis provides possible directions for future research on generating better queries with ChatGPT or post-process results from these queries.

Overall, it is still uncertain whether Boolean queries used for systematic review literature search should be developed by transformer-based generative models like ChatGPT. This paper shows the potential for such models, tempered by many issues around variability, robustness and reproducibility. Whatever the outcome, we firmly believe this is a promising and exciting ground for future research.

ACKNOWLEDGMENTS

REFERENCES

[1] 2019. Review Manager Web (RevMan Web).
[2] JJ García Adeva, JM Pikatza Atxa, M Ubeda Carrillo, and E Ansuategi Zengotitabengoa. 2014. Automatic Text Classification to Support Systematic Reviews in Medicine. *Expert Systems with Applications* 41, 4 (2014), 1498–1508.
[3] Amal Alharbi, William Briggs, and Mark Stevenson. 2018. Retrieving and Ranking Studies for Systematic Reviews: University of Sheffield's Approach to CLEF eHealth 2018 Task 2. In *CEUR Workshop Proceedings: Working Notes of CLEF 2018: Conference and Labs of the Evaluation Forum*, Vol. 2125. CEUR Workshop Proceedings.
[4] Amal Alharbi and Mark Stevenson. 2017. Ranking Abstracts to Identify Relevant Evidence for Systematic Reviews: The University of Sheffield's Approach to CLEF eHealth 2017 Task 2.. In *CEUR Workshop Proceedings: Working Notes of CLEF 2017: Conference and Labs of the Evaluation Forum*.
[5] Antonios Anagnostou, Athanasios Lagopoulos, Grigorios Tsoumakas, and Ioannis P Vlahavas. 2017. Combining Inter-Review Learning-to-Rank and Intra-Review Incremental Training for Title and Abstract Screening in Systematic Reviews.. In *CEUR Workshop Proceedings: Working Notes of CLEF 2017: Conference and Labs of the Evaluation Forum*.
[6] Maisie Badami, Boualem Benatallah, and Marcos Baez. 2022. Systematic Literature Review Search Query Refinement Pipeline: Incremental Enrichment and Adaptation. In *International Conference on Advanced Information Systems Engineering*. Springer, 129–146.
[7] Florian Boudin, Jian-Yun Nie, and Martin Dawes. 2010. Clinical Information Retrieval Using Document and PICO Structure. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 822–830.
[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell. 2020. Language Models Are Few-Shot Learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
[9] Kathi Canese and Sarah Weis. 2013. PubMed: the bibliographic database. *The NCBI handbook* 2, 1 (2013).
[10] Jiayi Chen, Su Chen, Yang Song, Hongyu Liu, Yueyao Wang, Qinmin Hu, Liang He, and Yan Yang. 2017. ECNU at 2017 eHealth Task 2: Technologically Assisted Reviews in Empirical Medicine.. In *CEUR Workshop Proceedings: Working Notes of CLEF 2017: Conference and Labs of the Evaluation Forum*.
[11] Justin Clark. 2013. Systematic Reviewing. In *Methods of Clinical Epidemiology*, Gail M. Williams Suhail A. R. Doi (Ed.). Springer.
[12] Justin Clark, Paul Glasziou, Chris Del Mar, Alexandra Bannach-Brown, Paulina Stehlik, and Anna Mae Scott. 2020. A full systematic review was completed in 2 weeks using automation tools: a case study. *Journal of clinical epidemiology* 121 (2020), 81–90.
[13] Justin Michael Clark, Sharon Sanders, Matthew Carter, David Honeyman, Gina Cleo, Yvonne Auld, Debbie Booth, Patrick Condron, Christine Dalais, Sarah Bateup, et al. 2020. Improving the Translation of Search Strategies Using the Polyglot Search Translator: A Randomized Controlled Trial. *Journal of the Medical Library Association* 108, 2 (2020), 195.
[14] Vincent Claveau. 2021. Neural Text Generation for Query Expansion in Information Retrieval. In *IEEE/WIC/ACM International Conference on Web Intelligence*. ACM, ESSENDON VIC Australia, 202–209. https://doi.org/10.1145/3486622.3493957
[15] A.M. Cohen, W.R. Hersh, K. Peterson, and P.Y. Yen. 2006. Reducing workload in systematic review preparation using automated citation classification. *JAMIA* 13, 2 (2006), 206–219.
[16] Aaron M Cohen and Neil R Smalheiser. 2018. UIC/OHSU CLEF 2018 Task 2 Diagnostic Test Accuracy Ranking Using Publication Type Cluster Similarity Measures. In *CEUR Workshop Proceedings: Working Notes of CLEF 2018: Conference and Labs of the Evaluation Forum*, Vol. 2125.
[17] Dina Demner-Fushman and Jimmy Lin. 2007. Answering Clinical Questions with Knowledge-Based and Statistical Techniques. *Computational Linguistics* 33, 1 (2007), 63–103.
[18] Giorgio Maria Di Nunzio, Federica Beghini, Federica Vezzani, and Geneviève Henrot. 2017. An Interactive Two-Dimensional Approach to Query Aspects Rewriting in Systematic Reviews. IMS Unipd at CLEF eHealth Task 2.. In *CEUR Workshop Proceedings: Working Notes of CLEF 2017: Conference and Labs of the Evaluation Forum*.
[19] Giorgio Maria Di Nunzio, Giacomo Ciuffreda, and Federica Vezzani. 2018. Interactive Sampling for Systematic Reviews. IMS Unipd at CLEF 2018 eHealth Task 2.. In *CEUR Workshop Proceedings: Working Notes of CLEF 2018: Conference and Labs of the Evaluation Forum*.
[20] Mette Brandt Eriksen and Tove Faber Frandsen. 2018. The impact of patient, intervention, comparison, outcome (PICO) as a search strategy tool on literature search quality: a systematic review. *Journal of the Medical Library Association: JMLA* 106, 4 (2018), 420.
[21] Roberto Gozalo-Brizuela and Eduardo C Garrido-Merchan. 2023. ChatGPT is not all you need. A State of the Art Review of large Generative AI models. *arXiv preprint arXiv:2301.04655* (2023).

[22] Trisha Greenhalgh and Richard Peacock. 2005. Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. *Bmj* 331, 7524 (2005), 1064–1065.
[23] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. arXiv:arXiv:2007.15779
[24] Elke Hausner, Charlotte Guddat, Tatjana Hermanns, Ulrike Lampert, and Siw Waffenschmidt. 2015. Development of search strategies for systematic reviews: validation showed the noninferiority of the objective approach. *Journal of clinical epidemiology* 68, 2 (2015), 191–199.
[25] Elke Hausner, Siw Waffenschmidt, Thomas Kaiser, and Michael Simon. 2012. Routine development of objectively derived search strategies. *Systematic reviews* 1, 1 (2012), 19.
[26] Noah Hollmann and Carsten Eickhoff. 2017. Ranking and Feedback-Based Stopping for Recall-Centric Document Retrieval.. In *CEUR Workshop Proceedings: Working Notes of CLEF 2017: Conference and Labs of the Evaluation Forum*.
[27] William Hsu, William Speier, and Ricky K Taira. 2012. Automated Extraction of Reported Statistical Analyses: Towards a Logical Representation of Clinical Trial Literature. In *AMIA Annual Symposium Proceedings*, Vol. 2012. 350.
[28] E. Kanoulas, D. Li, L. Azzopardi, and R. Spijker. 2017. CLEF 2017 Technologically Assisted Reviews in Empirical Medicine Overview. In *CLEF'17*.
[29] Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. 2019. CLEF 2019 technology assisted reviews in empirical medicine overview. In *CEUR workshop proceedings*, Vol. 2380.
[30] Evangelos Kanoulas, Rene Spijker, Dan Li, and Leif Azzopardi. 2018. CLEF 2018 Technology Assisted Reviews in Empirical Medicine Overview. In *CLEF 2018 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS*.
[31] Sarvnaz Karimi, Justin Zobel, Stefan Pohl, and Falk Scholer. 2009. The Challenge of High Recall in Biomedical Systematic Search. In *Proceedings of the 3rd International Workshop on Data and Text Mining in Bioinformatics*. 89–92.
[32] S.N. Kim, D. Martinez, L. Cavedon, and L. Yencken. 2011. Automatic Classification of Sentences to Support Evidence Based Medicine. *BMC bioinformatics* 12, 2 (2011).
[33] Grace Eunkyung Lee. 2017. A Study of Convolutional Neural Networks for Clinical Document Classification in Systematic Reviews: Sysreview at CLEF eHealth 2017. In *CEUR Workshop Proceedings: Working Notes of CLEF 2017: Conference and Labs of the Evaluation Forum*.
[34] Hang Li, Harrisen Scells, and Guido Zuccon. 2020. Systematic review automation tools for end-to-end query formulation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2141–2144.
[35] Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2022. A survey of pretrained language models based text generation. *arXiv preprint arXiv:2201.05273* (2022).
[36] Valentin Liévin, Christoffer Egeberg Hother, and Ole Winther. 2023. Can Large Language Models Reason about Medical Questions? arXiv:2207.08143 [cs]
[37] Vivian Liu and Lydia B Chilton. 2022. Design Guidelines for Prompt Engineering Text-to-Image Generative Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–23. https://doi.org/10.1145/3491102.3501825
[38] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: Generative Pre-Trained Transformer for Biomedical Text Generation and Mining. *Briefings in Bioinformatics* 23, 6 (Nov. 2022), bbac409. https://doi.org/10.1093/bib/bbac409
[39] Andrew MacFarlane, Tony Russell-Rose, and Farhad Shokraneh. 2022. Search strategy formulation for systematic reviews: Issues, challenges and opportunities. *Intelligent Systems with Applications* (2022), 200091.
[40] Iain J Marshall, Rachel Marshall, Byron C Wallace, Jon Brassey, and James Thomas. 2019. Rapid Reviews May Produce Different Results to Systematic Reviews: A Meta-Epidemiological Study. *Journal of clinical epidemiology* 109 (2019), 30–41.
[41] Iain J Marshall, Anna Noel-Storr, Joël Kuiper, James Thomas, and Byron C Wallace. 2018. Machine Learning for Identifying Randomized Controlled Trials: An Evaluation and Practitioner's Guide. *Research synthesis methods* 9, 4 (2018), 602–614.
[42] D. Martinez, S. Karimi, L. Cavedon, and T. Baldwin. 2008. Facilitating Biomedical Systematic Reviews Using Ranked Text Retrieval and Classification. In *Proceedings of the 13th Australasian Document Computing Symposium*.
[43] Adamantios Minas, Athanasios Lagopoulos, and Grigorios Tsoumakas. 2018. Aristotle University's Approach to the Technologically Assisted Reviews in Empirical Medicine Task of the 2018 CLEF eHealth Lab.. In *CEUR Workshop Proceedings: Working Notes of CLEF 2018: Conference and Labs of the Evaluation Forum*.
[44] M. Miwa, J. Thomas, A. O'Mara-Eves, and S. Ananiadou. 2014. Reducing systematic review workload through certainty-based screening. *JBI* 51 (2014), 242–253.
[45] Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019. Multi-stage document ranking with BERT. *arXiv preprint arXiv:1910.14424* (2019).

Can ChatGPT Write a Good Boolean Query for Systematic Review Literature Search?

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

[46] Christopher Norman, Mariska Leeflang, and Aurélie Névéol. 2018. LIMSI@ CLEF eHealth 2018 Task 2: Technology Assisted Reviews by Stacking Active and Static Learning.. In *CEUR Workshop Proceedings: Working Notes of CLEF 2018: Conference and Labs of the Evaluation Forum*.

[47] Christopher Norman12, Mariska Leeflang, and Aurélie Névéol. 2017. Limsi@ Clef Ehealth 2017 Task 2: Logistic Regression for Automatic Article Ranking. In *CEUR Workshop Proceedings: Working Notes of CLEF 2019: Conference and Labs of the Evaluation Forum*.

[48] Jonas Oppenlaender. 2022. A Taxonomy of Prompt Modifiers for Text-To-Image Generation. arXiv:2204.13988 [cs]

[49] Mohammadreza Pourreza and Faezeh Ensan. 2022. Towards semantic-driven boolean query formalization for biomedical systematic literature reviews. *International Journal of Medical Informatics* (2022), 104928.

[50] Animesh Prasad, Manpreet Kaur, and Min-Yen Kan. 2018. Neural ParsCit: A Deep Learning Based Reference String Parser. *Journal on Digital Libraries* 19 (2018), 323–337.

[51] John Rathbone. 2017. *Automating systematic reviews*. Ph. D. Dissertation. Bond University.

[52] Scott Reeves, Ivan Koppel, Hugh Barr, Della Freeth, and Marilyn Hammick. 2002. Twelve Tips for Undertaking a Systematic Review. *Medical teacher* 24, 4 (2002).

[53] Laria Reynolds and Kyle McDonell. 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–7. https://doi.org/10.1145/3411763.3451760

[54] Raul Rodriguez-Esteban and Ivan Iossifov. 2009. Figure Mining for Biomedical Research. *Bioinformatics* 25, 16 (2009), 2082–2084.

[55] Tony Russell-Rose and Philip Gooch. 2018. 2dSearch: A Visual Approach to Search Strategy Formulation. In *Proceedings of the 1st Biennial Conference on Design of Experimental Search and Information Retrieval Systems*.

[56] José Antonio Salvador-Oliván, Gonzalo Marco-Cuenca, and Rosario Arquero-Avilés. 2019. Errors in Search Strategies Used in Systematic Reviews and Their Effects on Information Retrieval. *Journal of the Medical Library Association : JMLA* 107, 2 (April 2019), 210–221. https://doi.org/10.5195/jmla.2019.567

[57] Margaret Sampson and Jessie McGowan. 2006. Errors in Search Strategies Were Identified by Type and Frequency. *Journal of Clinical Epidemiology* 59, 10 (2006), 1057–e1.

[58] Harrisen Scells, Connor Forbes, Justin Clark, Bevan Koopman, and Guido Zuccon. 2022. The Impact of Query Refinement on Systematic Review Literature Search: A Query Log Analysis. In *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval*. 34–42.

[59] Harrisen Scells and Guido Zuccon. 2018. Generating Better Queries for Systematic Reviews. In *Proceedings of the 41st annual International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) (SIGIR '18). 10 pages.

[60] Harrisen Scells and Guido Zuccon. 2018. searchrefiner: A Query Visualisation and Understanding Tool for Systematic Reviews. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 1939–1942.

[61] Harrisen Scells, Guido Zuccon, Anthony Deacon, and Bevan Koopman. 2017. QUT Ielab at CLEF eHealth 2017 Technology Assisted Reviews Track: Initial Experiments with Learning to Rank. In *CEUR Workshop Proceedings: Working Notes of CLEF 2017: Conference and Labs of the Evaluation Forum*.

[62] Harrisen Scells, Guido Zuccon, and Bevan Koopman. 2019. Automatic Boolean Query Refinement for Systematic Review Literature Search. In *The World Wide Web Conference*. 1646–1656.

[63] Harrisen Scells, Guido Zuccon, and Bevan Koopman. 2020. A Comparison of Automatic Boolean Query Formulation for Systematic Reviews. *Information Retrieval Journal* (2020), 1–26.

[64] Harrisen Scells, Guido Zuccon, Bevan Koopman, and Justin Clark. 2020. Automatic Boolean Query Formulation for Systematic Review Literature Search. In *Proceedings of The Web Conference 2020*. 1071–1081.

[65] Harrisen Scells, Guido Zuccon, Bevan Koopman, and Justin Clark. 2020. A Computational Approach for Objectively Derived Systematic Review Search Strategies. In *European Conference on Information Retrieval*. Springer, 385–398.

[66] Harrisen Scells, Guido Zuccon, Mohamed A Sharaf, and Bevan Koopman. 2020. Sampling Query Variations for Learning to Rank to Improve Automatic Boolean Query Generation in Systematic Reviews. In *Proceedings of The Web Conference 2020*. 3041–3048.

[67] Michael Simon, Elke Hausner, Susan F Klaus, and Nancy E Dunton. 2010. Identifying Nurse Staffing Research in Medline: Development and Testing of Empirically Derived Search Strategies with the PubMed Interface. *BMC medical research methodology* 10, 1 (2010), 76.

[68] Gaurav Singh, Iain Marshall, James Thomas, and Byron Wallace. 2017. Identifying Diagnostic Test Accuracy Publications Using a Deep Model. In *CEUR Workshop Proceedings: Working Notes of CLEF 2017: Conference and Labs of the Evaluation Forum*, Vol. 1866.

[69] Jaspreet Singh and Lini Thomas. 2017. IIIT-H at CLEF eHealth 2017 Task 2: Technologically Assisted Reviews in Empirical Medicine.. In *CEUR Workshop Proceedings: Working Notes of CLEF 2017: Conference and Labs of the Evaluation Forum*.

[70] Claire Stansfield, James Thomas, and Josephine Kavanagh. 2013. 'Clustering' Documents Automatically to Support Scoping Reviews of Research: A Case Study. *Research synthesis methods* 4, 3 (2013), 230–241.

[71] Rodney Summerscales, Shlomo Argamon, Jordan Hupert, and Alan Schwartz. 2009. Identifying Treatments, Groups, and Outcomes in Medical Abstracts. In *Proceedings of the 6th Midwest Computational Linguistics Colloquium*.

[72] Rodney L Summerscales, Shlomo Argamon, Shangda Bai, Jordan Hupert, and Alan Schwartz. 2011. Automatic Summarization of Results from Clinical Trials. In *Proceedings of the 2011 IEEE International Conference on Bioinformatics and Biomedicine*. 372–377.

[73] James Thomas and Angela Harden. 2008. Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC medical research methodology* 8, 1 (2008), 45.

[74] Prem Timsina, Jun Liu, and Omar El-Gayar. 2016. Advanced Analytics for the Automation of Medical Systematic Reviews. *Information Systems Frontiers* 18, 2 (2016), 237–252.

[75] Mercedes Torres Torres and Clive E Adams. 2017. RevManHAL: towards automatic text generation in systematic reviews. *Systematic Reviews* 6, 1 (2017).

[76] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[77] Shuai Wang, Hang Li, Harrisen Scells, Daniel Locke, and Guido Zuccon. 2021. Mesh term suggestion for systematic review literature search. In *Proceedings of the 25th Australasian Document Computing Symposium*. 1–8.

[78] Shuai Wang, Hang Li, and Guido Zuccon. 2022. MeSH Suggester: A Library and System for MeSH Term Suggestion for Systematic Review Boolean Query Construction. *arXiv preprint arXiv:2212.09018* (2022).

[79] Shuai Wang, Harrisen Scells, Justin Clark, Bevan Koopman, and Guido Zuccon. 2022. From little things big things grow: A collection with seed studies for medical systematic review literature search. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3176–3186.

[80] Shuai Wang, Harrisen Scells, Bevan Koopman, and Guido Zuccon. 2022. Automated MeSH Term Suggestion for Effective Query Formulation in Systematic Reviews Literature Search. *Intelligent Systems with Applications* (2022), 200141.

[81] Huaying Wu, Tingting Wang, Jiayi Chen, Su Chen, Qinmin Hu, and Liang He. 2018. Ecnu at 2018 Ehealth Task 2: Technologically Assisted Reviews in Empirical Medicine. *Methods-a Companion to Methods in Enzymology* 4, 5 (2018), 7.

[82] Zhe Yu and Tim Menzies. 2017. Data Balancing for Technologically Assisted Reviews: Undersampling or Reweighting.. In *CEUR Workshop Proceedings: Working Notes of CLEF 2017: Conference and Labs of the Evaluation Forum*.