
How do clinicians become informed about how to treat their patients?


How do governments and institutions make health policy decisions?

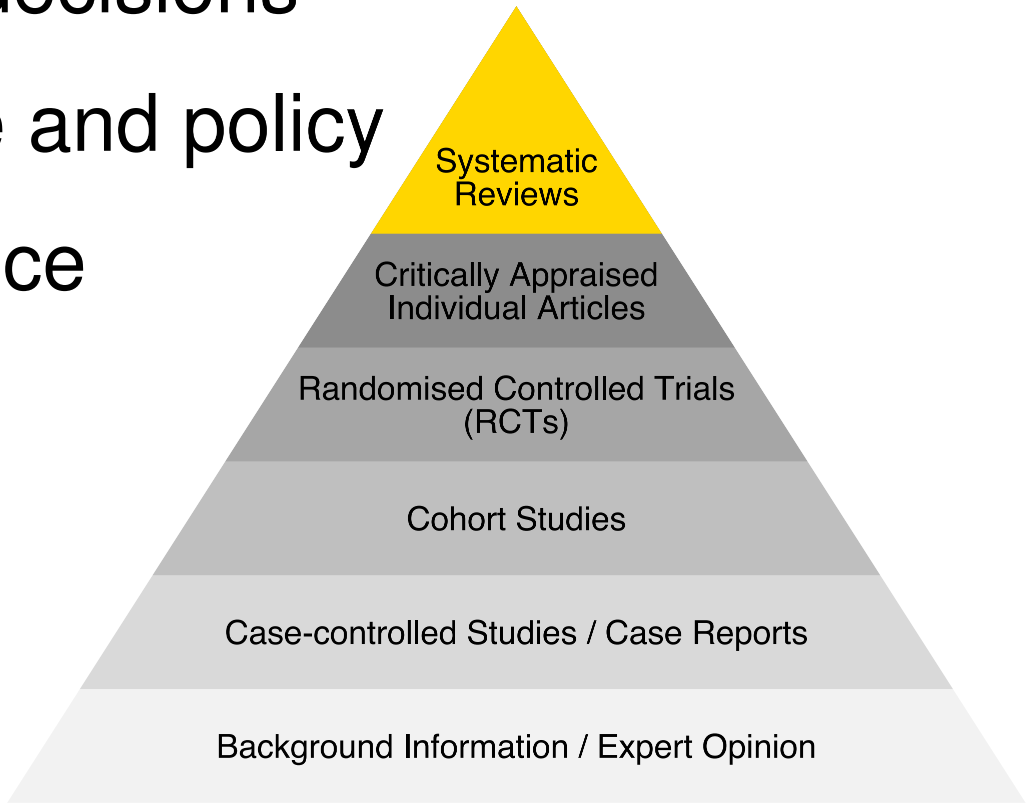
How do clinicians become informed about how to treat their patients?

Systematic Reviews

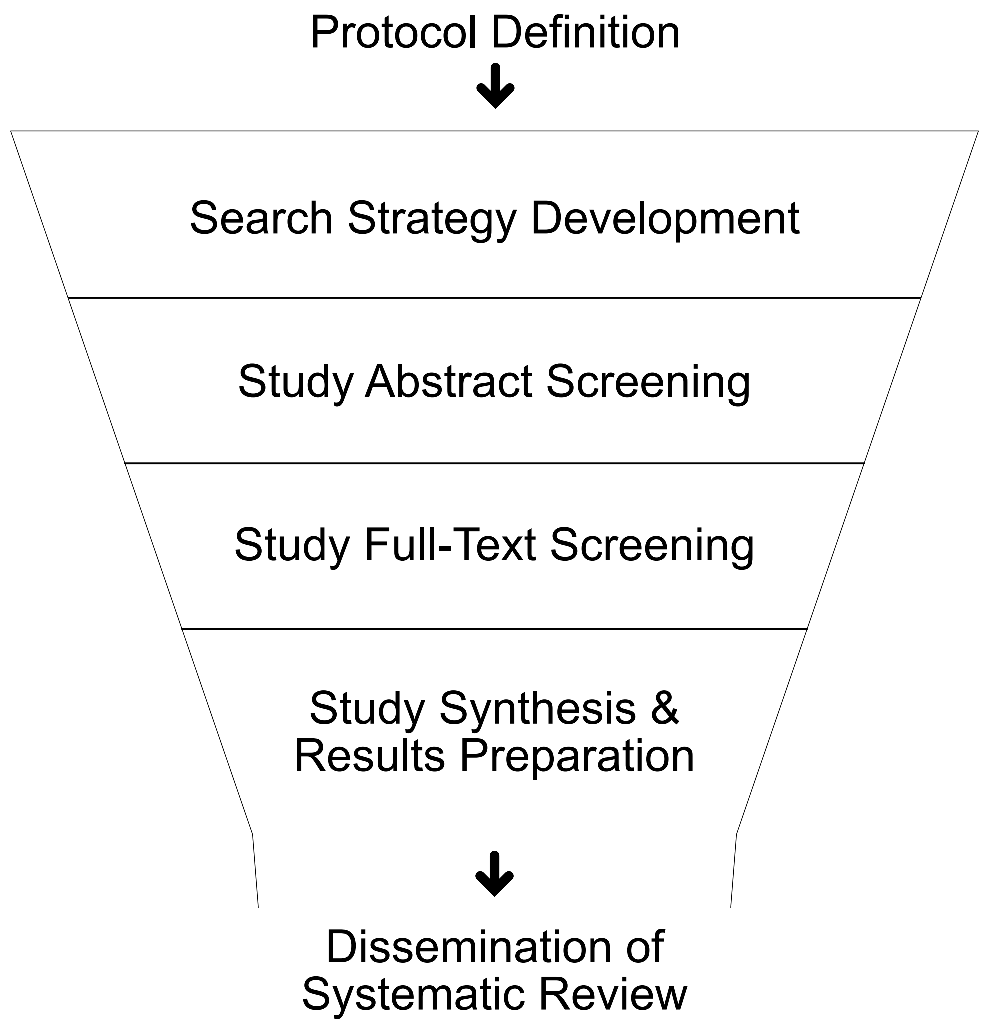
How do governments and institutions make health policy decisions?

Systematic Reviews
Overview

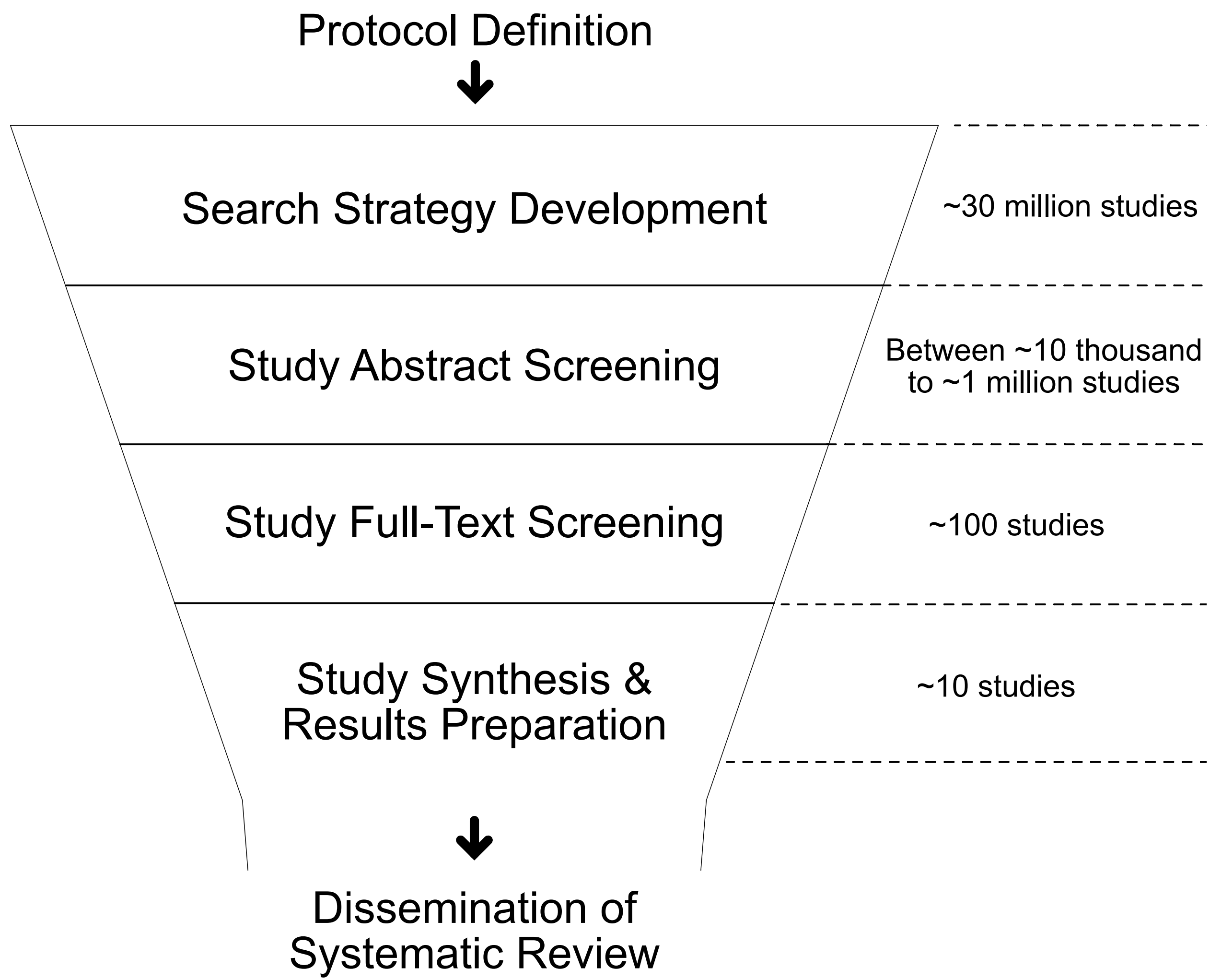
- **Guide** clinical decisions
- **Inform** practice and policy
- **Provide** evidence



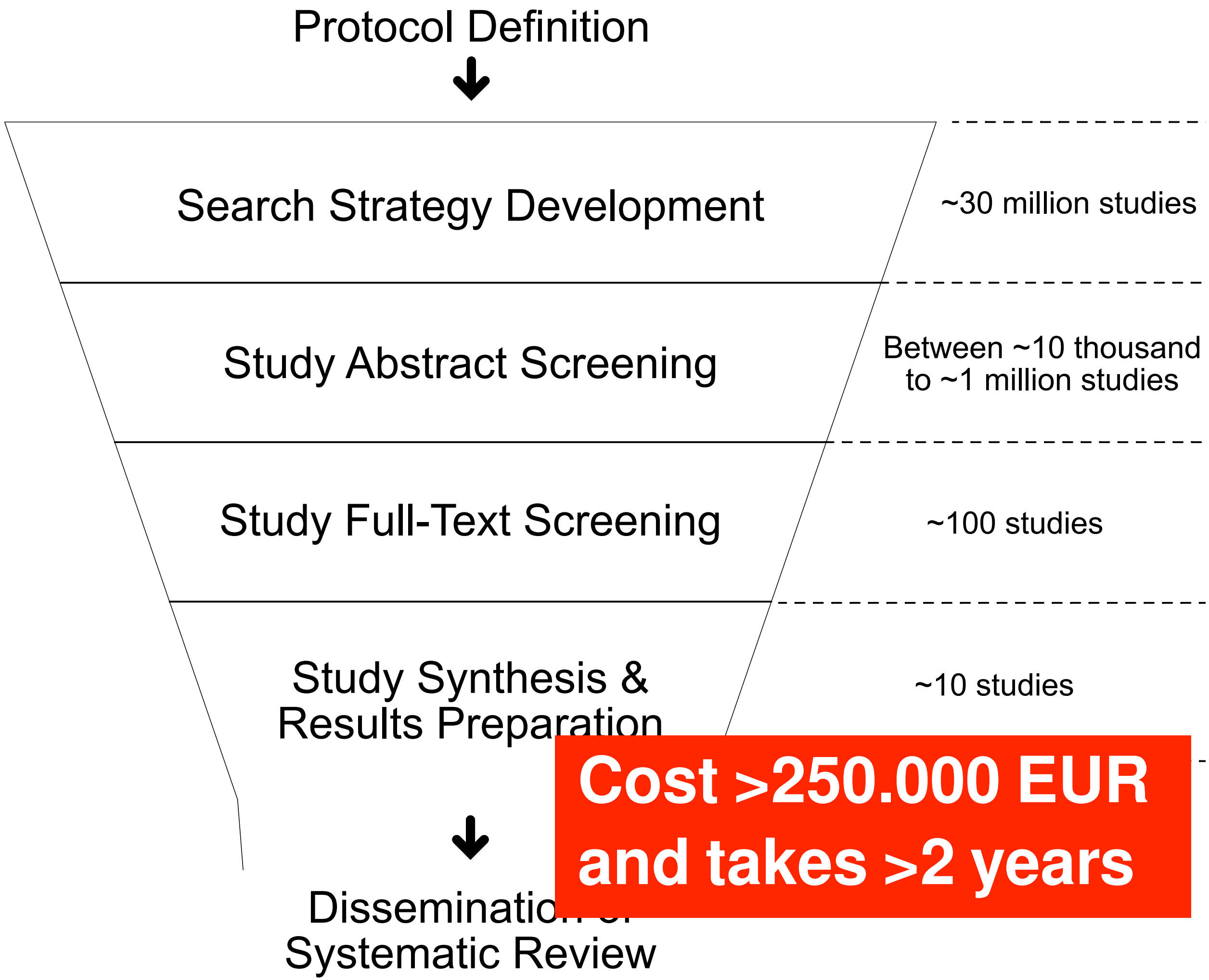
Systematic review creation is hard!



Why is systematic review creation hard?

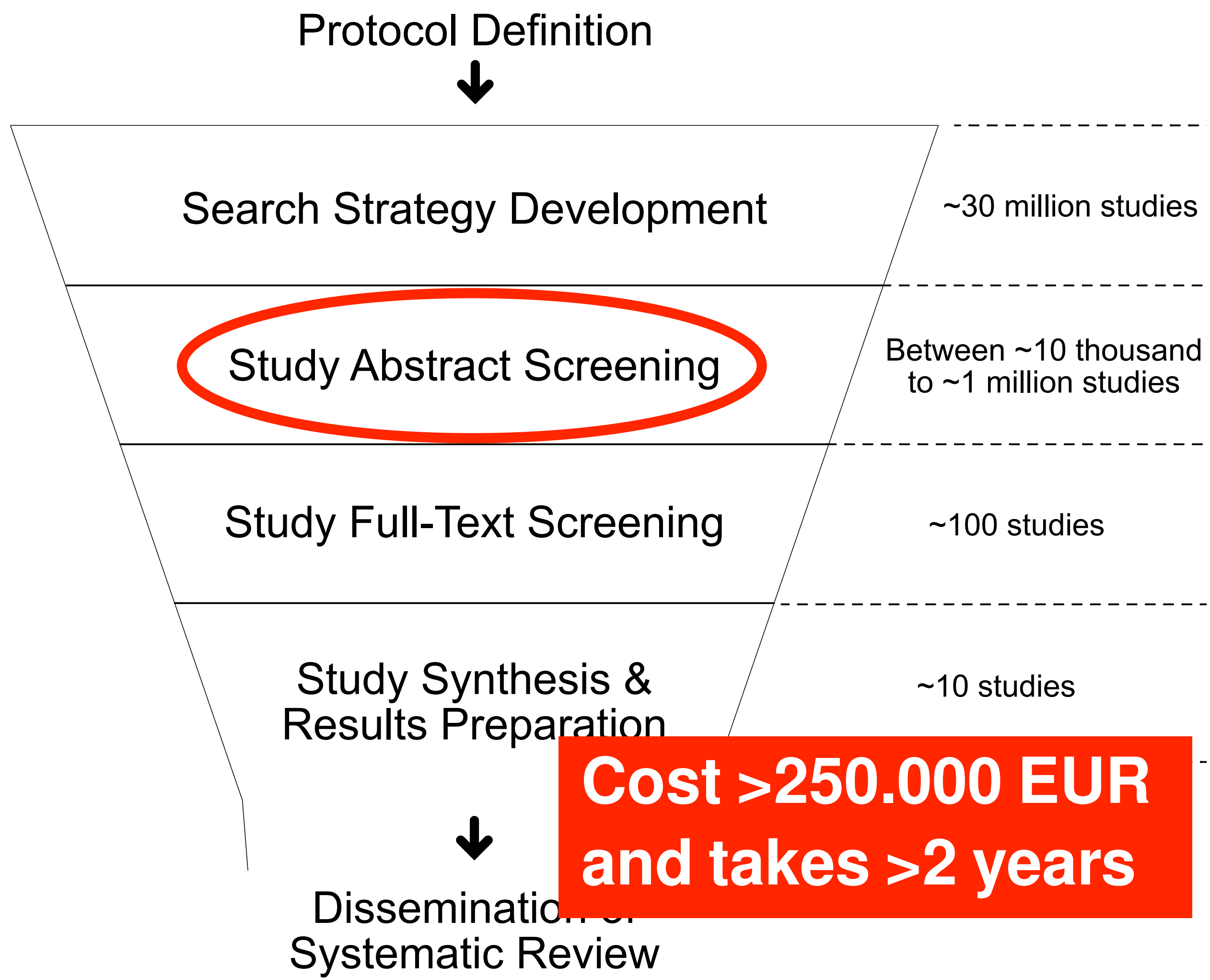


Why is systematic review creation hard?

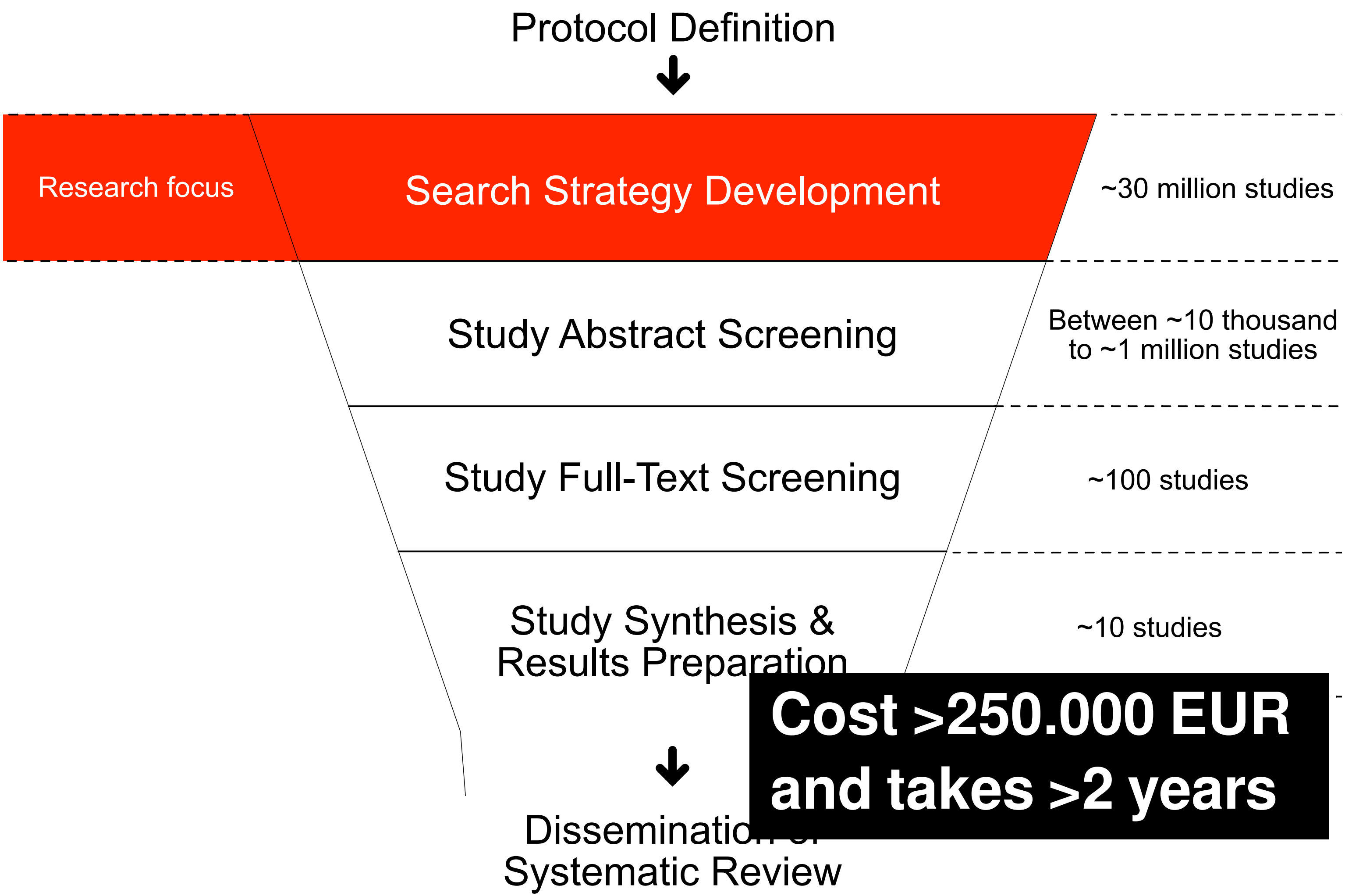


Cost >250.000 EUR and takes >2 years

Why is systematic review creation hard?



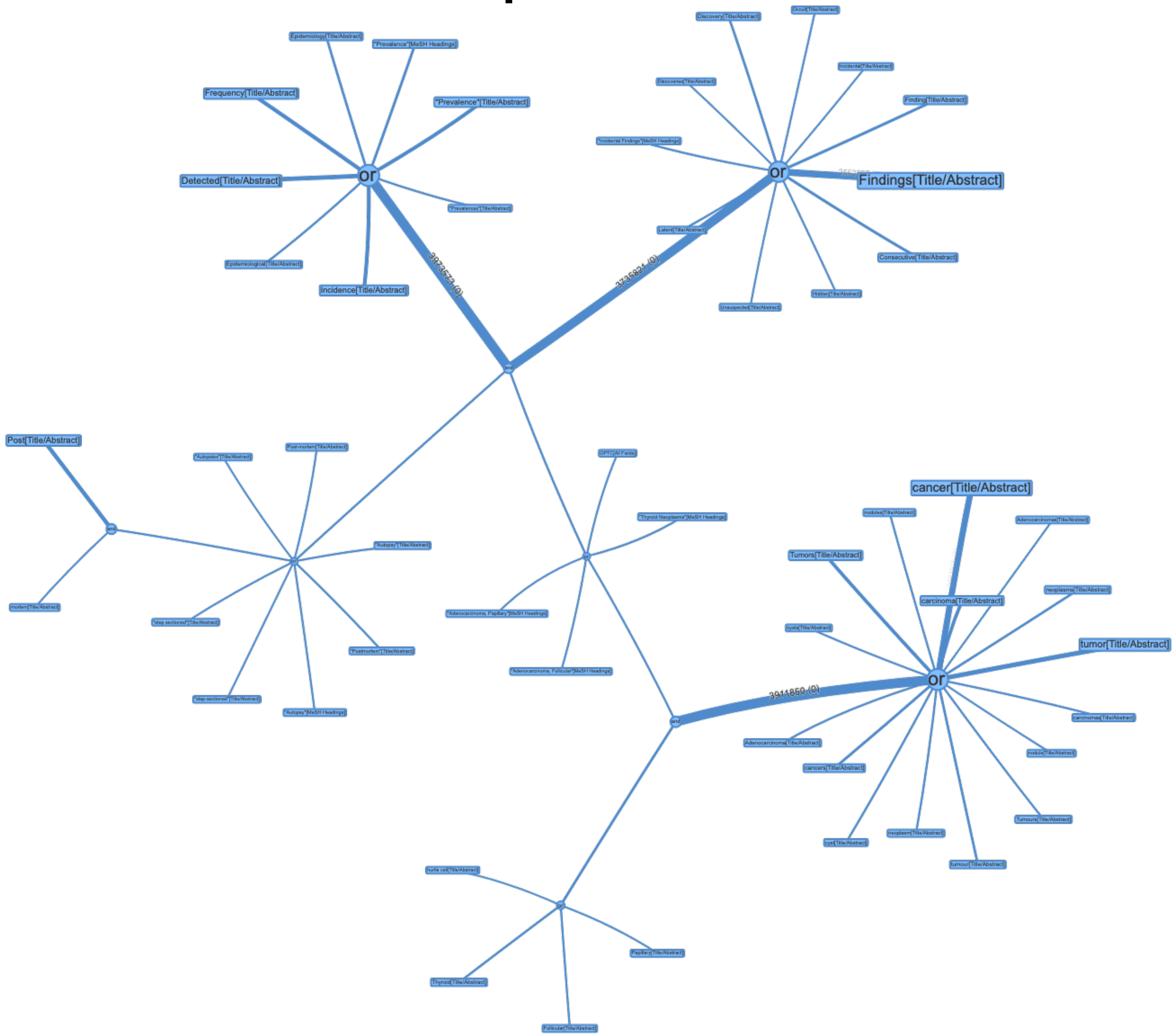
Why is systematic review creation hard?



Why such little research on queries?

```
 (("Thyroid Neoplasms"[MeSH] OR "Adenocarcinoma, Follicular"[MeSH]
OR "Adenocarcinoma, Papillary"[MeSH] OR OPTC OR ((Thyroid[tiab]
OR Follicular[tiab] OR Papillary[tiab] OR hurtle cell[tiab]) AND
(cancer[tiab] OR cancers[tiab] OR carcinoma[tiab] OR carcinomas[tiab]
OR Adenocarcinoma[tiab] OR Adenocarcinomas[tiab] OR neoplasm[tiab]
OR neoplasms[tiab] OR nodule[tiab] OR nodules[tiab] OR tumor[tiab]
OR tumour[tiab] OR Tumors[tiab] OR Tumours[tiab] OR cyst[tiab]
OR cysts[tiab]))) AND ("Autopsy"[MeSH] OR "Autopsy"[tiab] OR
"Autopsies"[tiab] OR "Postmortem"[tiab] OR Post-mortem[tiab] OR
"step-sectioned"[tiab] OR "step sectioned"[tiab] OR (Post[tiab]
AND mortem[tiab])) AND (Prevalence"[MeSH] OR Prevalence"[tiab] OR
Prevalences"[tiab] OR Incidence[tiab] OR Epidemiology[tiab] OR
Epidemiological[tiab] OR Frequency[tiab] OR Detected[tiab]) AND
("Incidental Findings"[MeSH] OR Incidental[tiab] OR Unsuspected[tiab] OR
Discovery[tiab] OR Discoveries[tiab] OR Findings[tiab] OR Finding[tiab]
OR Occult[tiab] OR Hidden[tiab] OR Latent[tiab] OR Consecutive[tiab]))
```

Why such little research on queries?



Why are Boolean queries used?

Reproducibility ➜ double check screening

Understandability ➜ control set size

Lecture Content
Overview

- Formulating Boolean queries
  - Use decoder model to automatically formulate Boolean queries
- Translating Boolean queries
  - Use decoder model to translate Boolean queries to alternative representations
- Automatically assessing documents
  - Use decoder model to automatically assess the relevance of documents

Formulating Boolean Queries

Overview

Use ChatGPT to formulate Boolean queries [Wang et al. 2023].

- Existing methods for Boolean query formulation use many heuristics.
- Use ChatGPT instead to automatically formulate Boolean queries.
- Compare effectiveness of all formulation methods to humans.

Formulating Boolean Queries

How Humans Formulate Queries

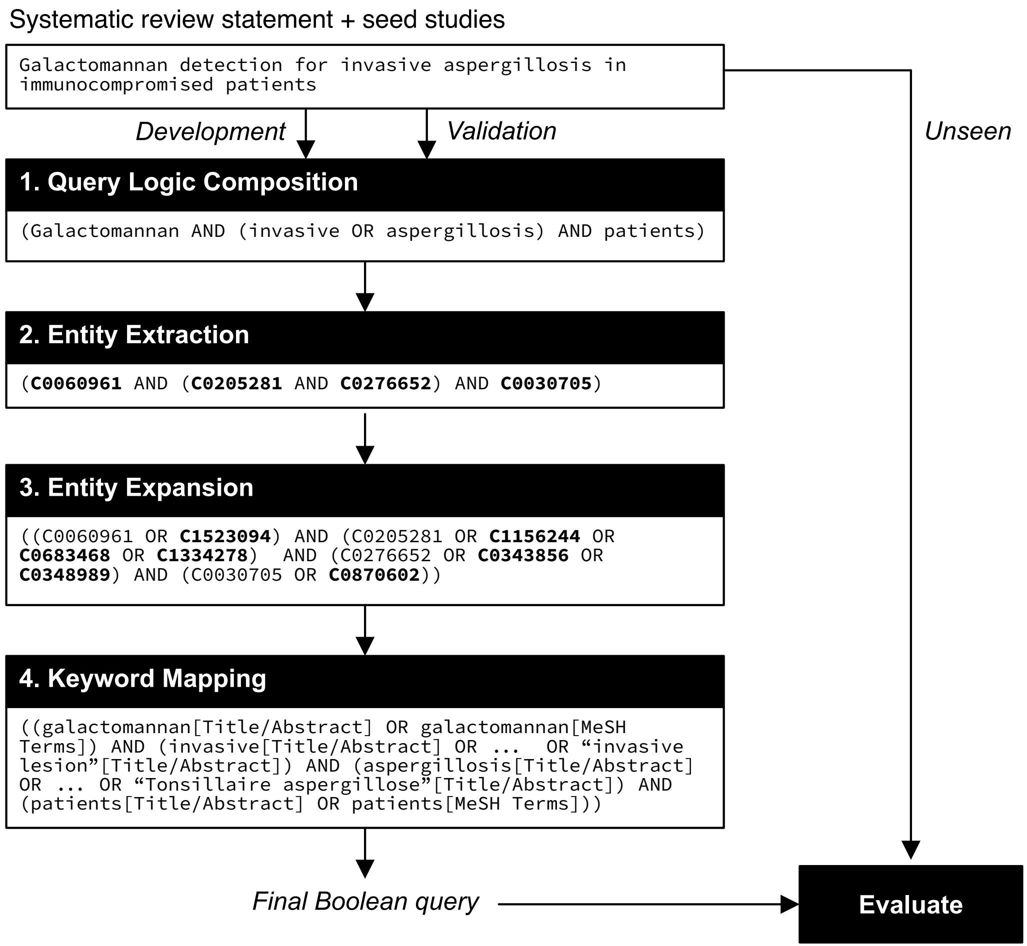
Conceptual method [Clark 2013] → Human expertise

Objective method [Hausner et al. 2012] → More algorithmic

Both methods → Seed studies

Formulating Boolean Queries

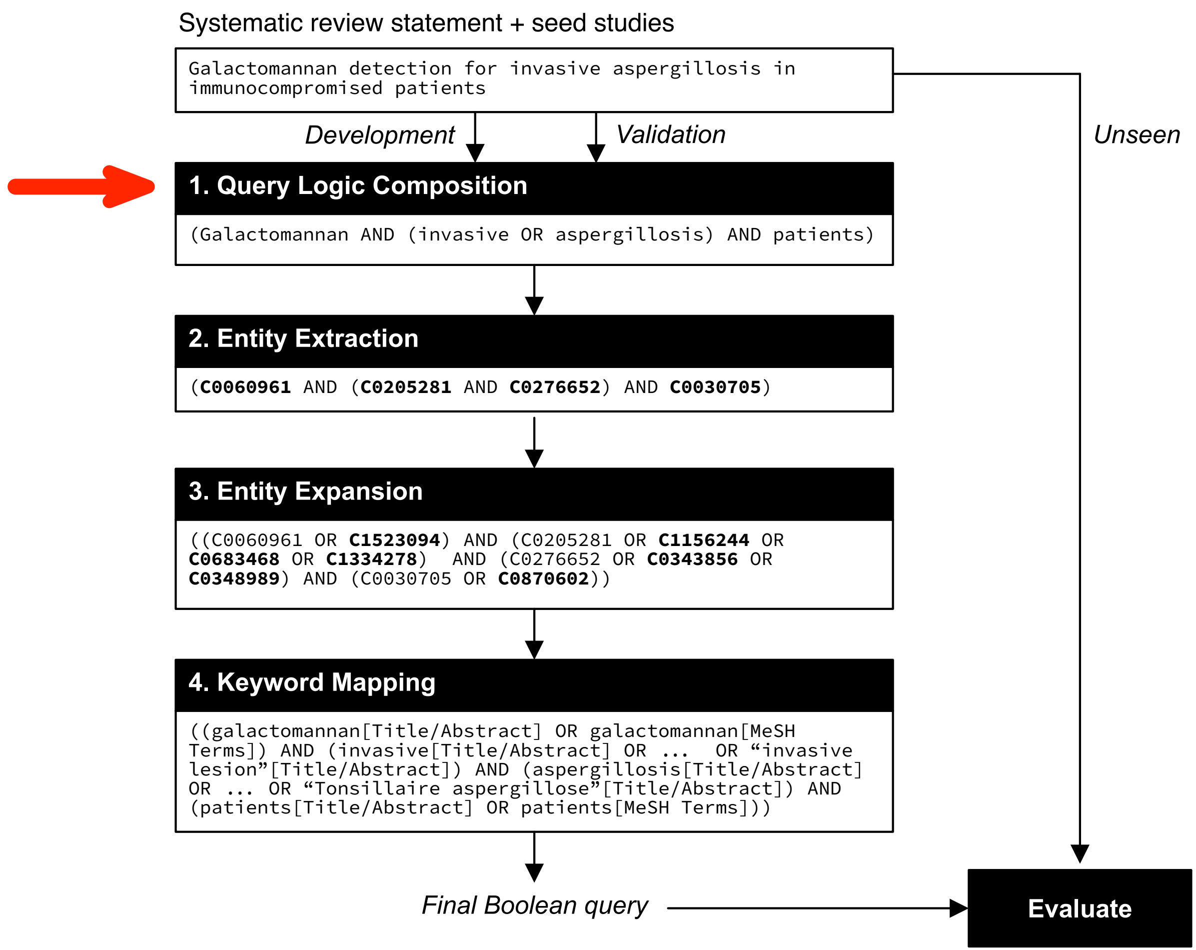
Automating the conceptual method



High level concepts ➜ broaden search ➜ iterate until satisfied

Formulating Boolean Queries

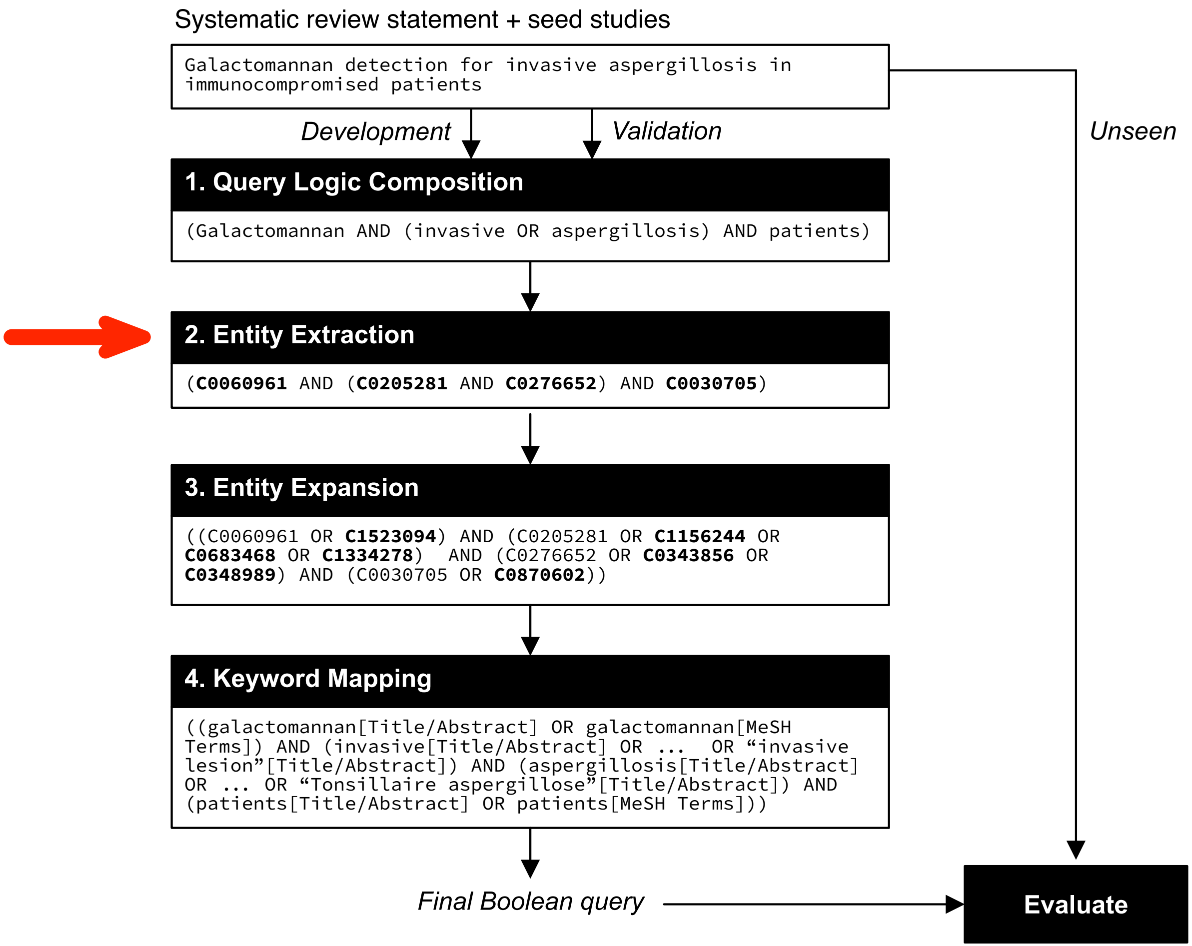
Automating the conceptual method



POS tagger ➜ parse grammar & segment words into noun phrases

Formulating Boolean Queries

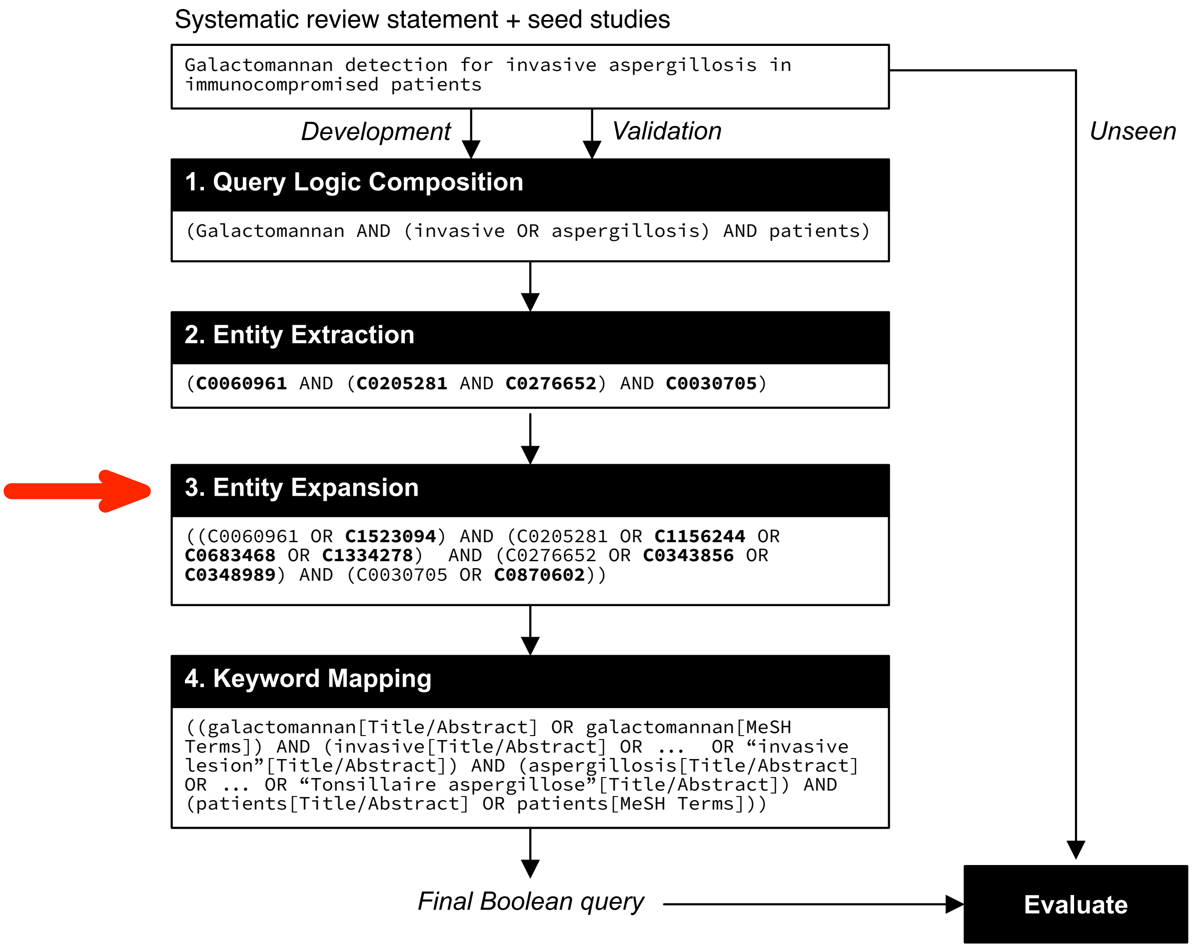
Automating the conceptual method



MetaMap ➜ extract CUIs from UMLS ontology

Formulating Boolean Queries

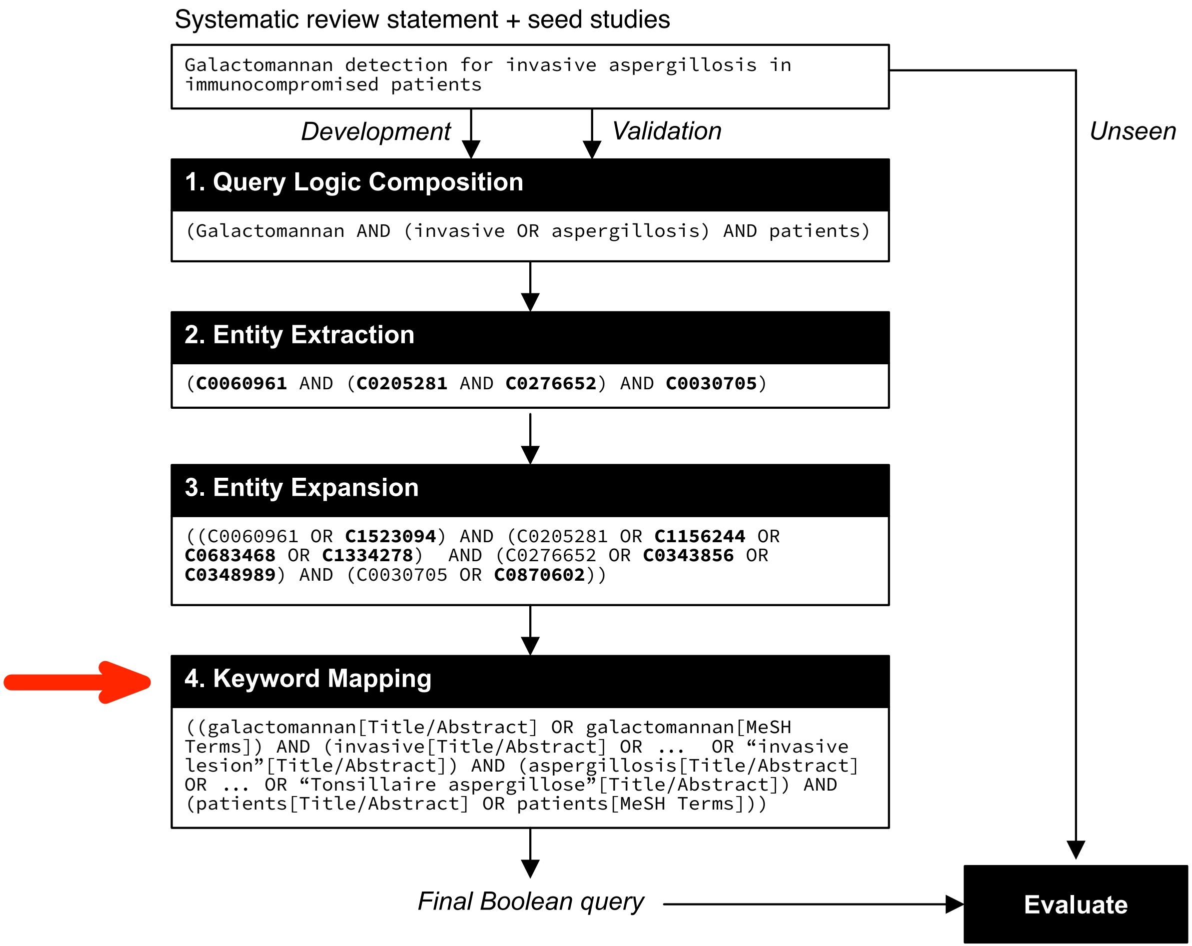
Automating the conceptual method



Skipgram model ➜ broaden scope

Formulating Boolean Queries

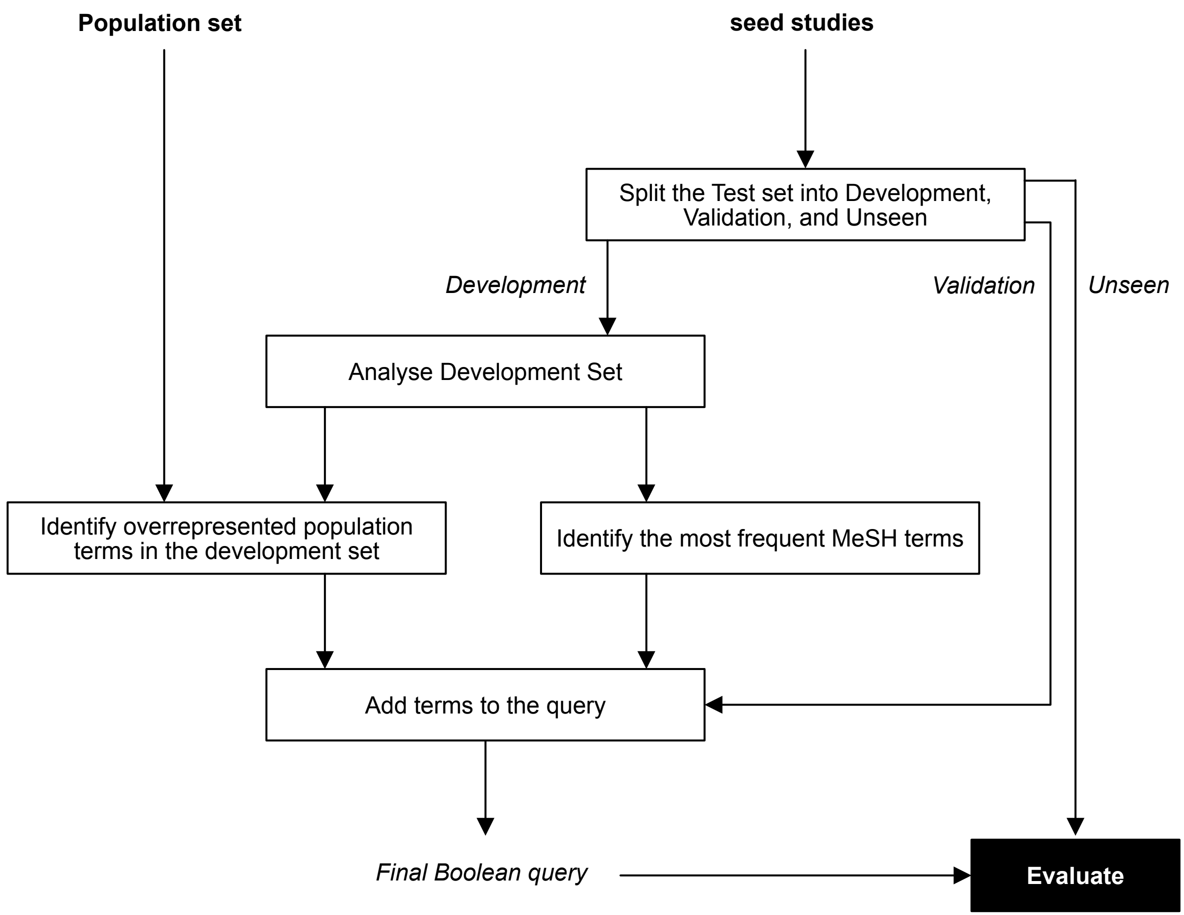
Automating the conceptual method



Map concepts (CUIs) to terms

Formulating Boolean Queries

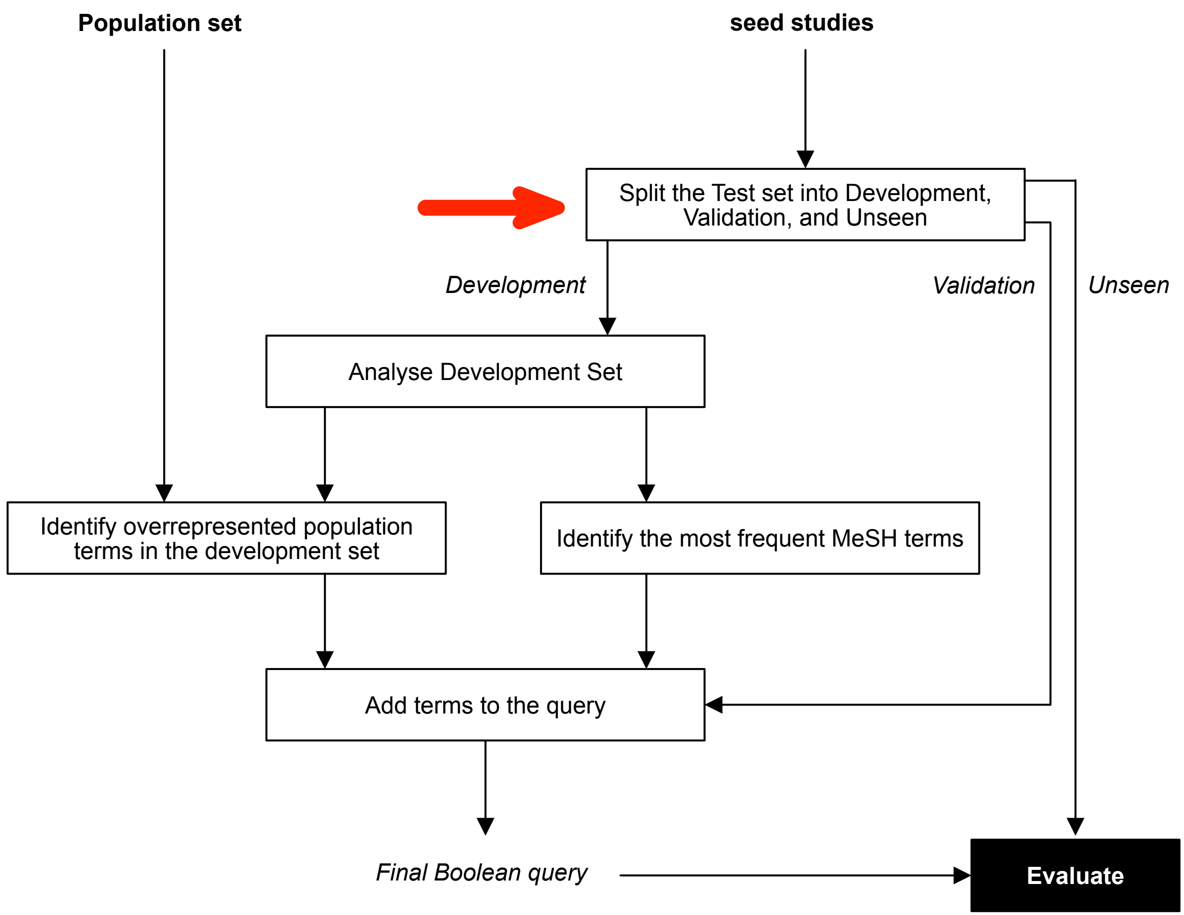
Automating the objective method



Find prominent terms from docs ➜ Add these terms to query

Formulating Boolean Queries

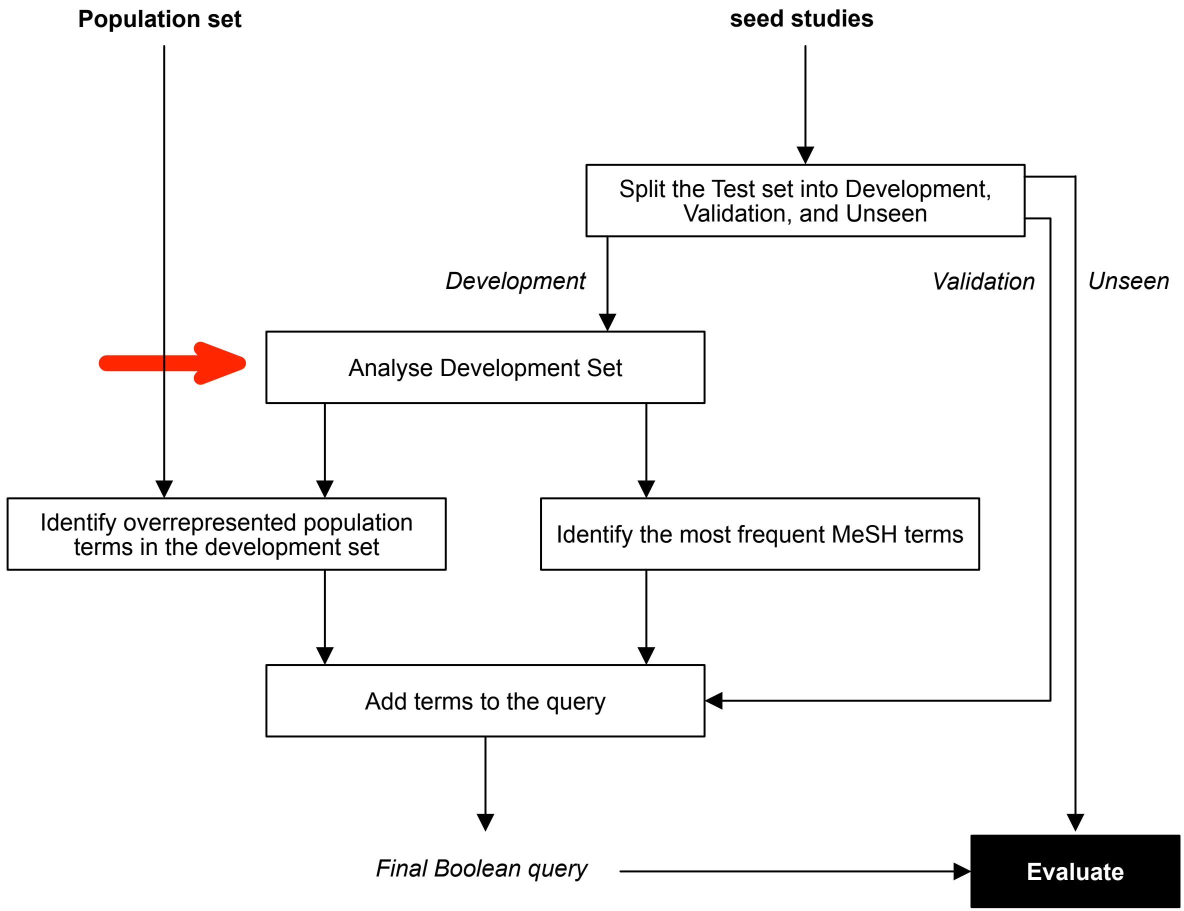
Automating the objective method



Extract list of keywords from seed studies

Formulating Boolean Queries

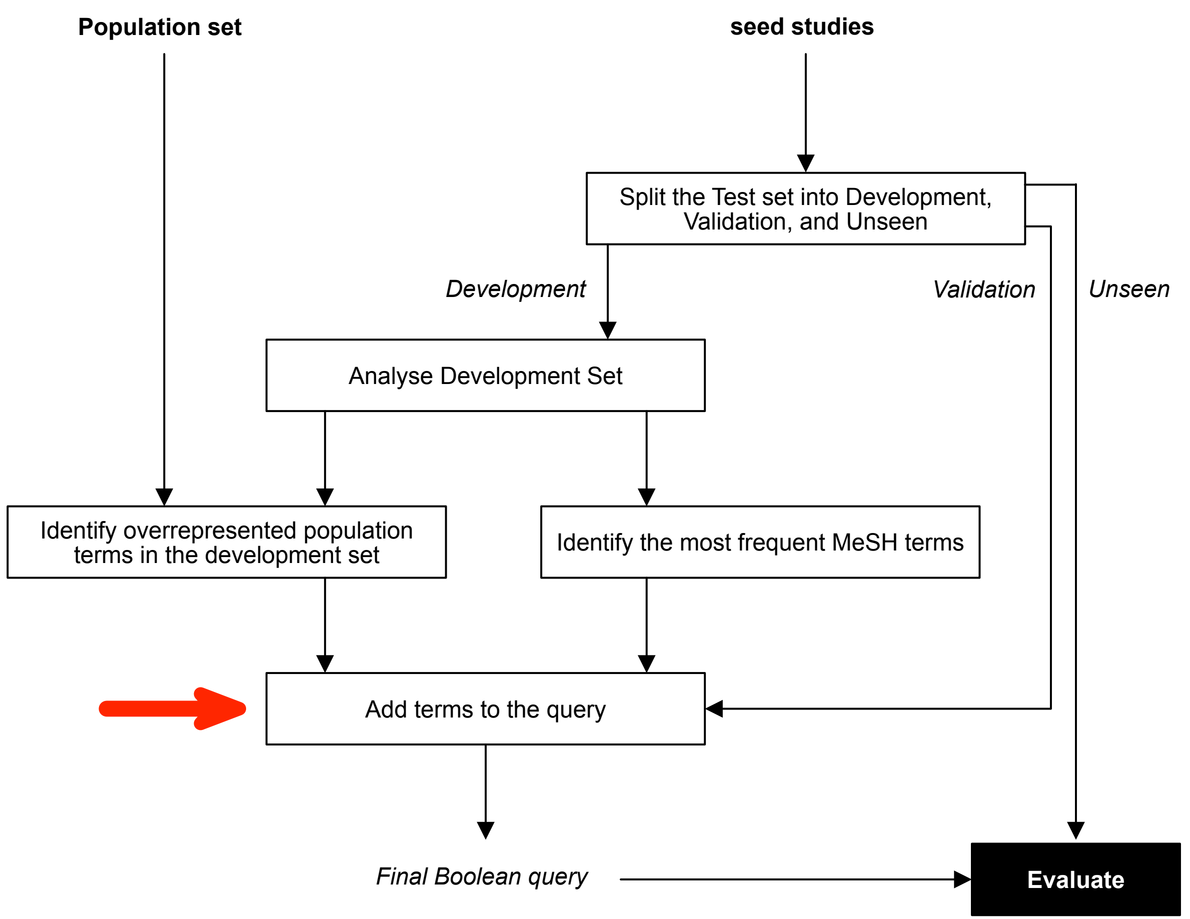
Automating the objective method



Rank documents using term frequency

Formulating Boolean Queries

Automating the objective method



Add keywords from documents to query

Formulating Boolean Queries

Conceptual versus objective results

CLEF TAR [Kanoulas et al. 2017, 2018]

	Precision	Recall
Conceptual	0.0014	0.6996
Objective	0.0002	0.9128

num. queries=72

Seed study collection [Wang et al. 2022]

Precision	Recall
0.0018	0.4138
0.0057	0.5192

num. queries=40

Formulating Boolean Queries

Conceptual versus objective results

CLEF TAR [Kanoulas et al. 2017, 2018]

	Precision	Recall
Human	0.0207	0.8317
Conceptual	0.0014	0.6996
Objective	0.0002	0.9128

num. queries=72

Seed study collection [Wang et al. 2022]

Precision	Recall
0.0367	0.7366
0.0018	0.4138
0.0057	0.5192

num. queries=40

Formulating Boolean Queries

Conceptual versus objective results

CLEF TAR [Kanoulas et al. 2017, 2018]

	Precision	Recall
Human	0.0207	0.8317
Conceptual	0.0014	0.6996
Objective	0.0002	0.9128

num. queries=72

Seed study collection [Wang et al. 2022]

Precision	Recall
0.0367	0.7366
0.0018	0.4138
0.0057	0.5192

num. queries=40

Objective provides highest recall, but larger trade-offs than conceptual

Formulating Boolean Queries

Conceptual versus objective results

CLEF TAR [Kanoulas et al. 2017, 2018]

	Precision	Recall
Human	0.0207	0.8317
Conceptual	0.0014	0.6996
Objective	0.0002	0.9128

num. queries=72

Seed study collection [Wang et al. 2022]

Precision	Recall
0.0367	0.7366
0.0018	0.4138
0.0057	0.5192

num. queries=40

Objective provides highest recall, but larger trade-offs than conceptual

Both methods are highly sensitive to initial seed studies

Formulating Boolean Queries

Conceptual versus objective results

CLEF TAR [Kanoulas et al. 2017, 2018]

	Precision	Recall
Human	0.0207	0.8317
Conceptual	0.0014	0.6996
Objective	0.0002	0.9128

num. queries=72

Seed study collection [Wang et al. 2022]

Precision	Recall
0.0367	0.7366
0.0018	0.4138
0.0057	0.5192

num. queries=40

Objective provides highest recall, but larger trade-offs than conceptual

Both methods are highly sensitive to initial seed studies

Conceptual method better starting point for precision, and objective for recall

Formulating Boolean Queries

Can ChatGPT formulate effective queries?

Hey ChatGPT, formulate a query for me!

Sure, no problem! ;-)

- Translation task: prompt ➜ query
- Two methods to formulate queries:
  - Unguided prompt *one shot to generate query*
  - Guided prompt *follows objective approach*
- Prevalence of Differentiated Thyroid Cancer in Autopsy Studies Over Six Decades: A Meta-Analysis

Formulating Boolean Queries

Unguided Prompt Examples – Simple

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?

(**"Prevalence of Differentiated Thyroid Cancer"** AND "Autopsy Studies"AND **"Six Decades"**) AND (systematic review OR meta-analysis)

Formulating Boolean Queries

Unguided Prompt Examples – Detailed

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.

(**"Differentiated Thyroid Cancer"** OR "Papillary Thyroid Cancer" OR "Follicular Thyroid Cancer") AND "Autopsy Studies" AND "Prevalence" AND **(("1960/01/01"[PDAT] : "2023/01/30"[PDAT]) OR ("1960"[PDAT]:"2023"[PDAT]))**

Formulating Boolean Queries

Unguided Prompt Examples – With Example

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.

(**"Differentiated Thyroid Cancer"[Mesh]** OR (differentiated[All Fields] AND thyroid[All Fields] AND cancer[All Fields]) OR Papillary Thyroid Cancer"[Mesh] OR (papillary[All Fields] AND thyroid[All Fields] AND cancer[All Fields]) OR Follicular Thyroid Cancer"[Mesh] OR (follicular[All Fields] AND thyroid[All Fields] AND cancer[All Fields])) AND ((prevalence[All Fields] OR incidence[All Fields]) AND (autopsy[All Fields] OR post-mortem[All Fields] OR mortuary[All Fields]) **AND (study[All Fields] OR studies[All Fields]) AND (six[All Fields] OR 60[All Fields] OR decades[All Fields])**)

Formulating Boolean Queries

Previous Query Formulation Results

CLEF TAR [Kanoulas et al. 2017, 2018]

	Precision	Recall
Human	0.0207	0.8317
Conceptual	0.0014	0.6996
Objective	0.0002	0.9128

Seed study collection [Wang et al. 2022]

Precision	Recall
0.0367	0.7366
0.0018	0.4138
0.0057	0.5192

Formulating Boolean Queries

Unguided Query Formulation Results

CLEF TAR [Kanoulas et al. 2017, 2018]

	Precision	Recall
Human	0.0207	0.8317
Conceptual	0.0014	0.6996
Objective	0.0002	0.9128
Simple	0.0543	0.1293
Detailed	0.1166	0.1310
Example	0.0752	0.5035

Seed study collection [Wang et al. 2022]

Precision	Recall
0.0367	0.7366
0.0018	0.4138
0.0057	0.5192
0.0501	0.0528
0.0983	0.0394
0.0283	0.1290

Formulating Boolean Queries

Unguided Query Formulation Results

CLEF TAR [Kanoulas et al. 2017, 2018]

	Precision	Recall
Human	0.0207	0.8317
Conceptual	0.0014	0.6996
Objective	0.0002	0.9128
Simple	0.0543	0.1293
Detailed	0.1166	0.1310
Example	0.0752	0.5035

Seed study collection [Wang et al. 2022]

Precision	Recall
0.0367	0.7366
0.0018	0.4138
0.0057	0.5192
0.0501	0.0528
0.0983	0.0394
0.0283	0.1290

Collection contains **seed studies**
➜ Use ChatGPT for objective method

Formulating Boolean Queries

Guided Prompt Examples – 1

Follow my instructions precisely to develop a highly effective Boolean query for a medical systematic review literature search. Do not explain or elaborate. First, Given the following statement and text from a relevant study, **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: `{statement}`. Text: `{text}`

1. **Differentiated thyroid cancer**
2. Prevalence
3. Autopsy studies
4. Meta-analysis
5. Occult carcinomas

...

50. Parameters.

Formulating Boolean Queries

Guided Prompt Examples – 2

...
50. Parameters.

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

...

50. (N/A) Parameters.

Formulating Boolean Queries

Guided Prompt Examples – 3

...
50. (N/A) Parameters.

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** OR **itemA2** or **itemA3**) AND (itemB1 OR itemB2 OR itemB3) AND (itemC1 OR itemC2 OR itemC3))

(((**differentiated**[Title/Abstract] OR **thyroid**[Title/Abstract] OR **carcinoma**[Title/Abstract] OR papillary[Title/Abstract] OR micro-carcinoma[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]))

Formulating Boolean Queries

Guided Prompt Examples – 4

...AND (autopsy[Title/Abstract] OR surgical[Title/Abstract] OR material[Title/Abstract] OR series[Title/Abstract] OR specimens[Title/Abstract] OR cases[Title/Abstract]))

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]))

Formulating Boolean Queries

Query Formulation Results on Seed Study Collection [Wang et al. 2022]

	Precision	Recall
Human	0.0367	0.7366
Conceptual	0.0018	0.4138
Objective	0.0057	0.5192
Simple	0.0501	0.0528
Detailed	0.0983	0.0394
Example	0.0283	0.1290

Formulating Boolean Queries

Query Formulation Results on Seed Study Collection [Wang et al. 2022]

	Precision	Recall
Human	0.0367	0.7366
Conceptual	0.0018	0.4138
Objective	0.0057	0.5192
Simple	0.0501	0.0528
Detailed	0.0983	0.0394
Example	0.0283	0.1290
Guided	0.0993	0.5171

Formulating Boolean Queries

Query Formulation Results on Seed Study Collection [Wang et al. 2022]

	Precision	Recall
Human	0.0367	0.7366
Conceptual	0.0018	0.4138
Objective	0.0057	0.5192
Simple	0.0501	0.0528
Detailed	0.0983	0.0394
Example	0.0283	0.1290
Guided	0.0993	0.5171

ChatGPT is more effective than automatic conceptual and objective methods

Formulating Boolean Queries

Query Formulation Results on Seed Study Collection [Wang et al. 2022]

	Precision	Recall
Human	0.0367	0.7366
Conceptual	0.0018	0.4138
Objective	0.0057	0.5192
Simple	0.0501	0.0528
Detailed	0.0983	0.0394
Example	0.0283	0.1290
Guided	0.0993	0.5171

ChatGPT is more effective than automatic conceptual and objective methods

ChatGPT is highly dependent on prompt and prone to hallucination

Translating Boolean Queries

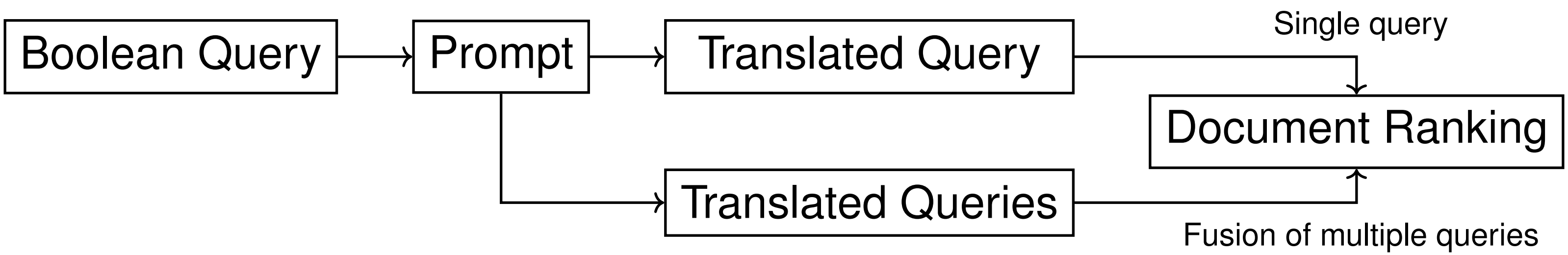
Overview

Use decoder models to translate Boolean queries into alternative representations [Wang et al. 2023].

- Translate the Boolean query into a natural language query.
- Rank documents retrieved by Boolean query using natural language query.
- Compare ranking effectiveness across baseline approaches from the literature.

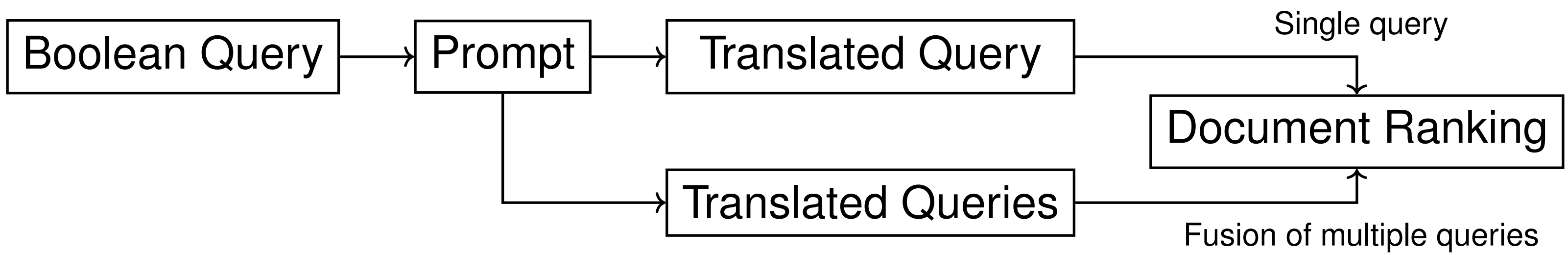
Translating Boolean Queries

Method Overview



Translating Boolean Queries

Method Overview



Boolean Query

(((“Thyroid Neoplasms”[MeSH] OR “Adenocarcinoma, Follicular”[MeSH] OR “Adenocarcinoma, Papillary”[MeSH] OR OPTC OR ((Thyroid[tiab] OR Follicular[tiab] OR Papillary[tiab] OR hurtle cell[tiab]) AND (cancer[tiab] OR cancers[tiab] OR carcinoma[tiab] OR carcinomas[tiab] OR Adenocarcinoma[tiab] OR Adenocarcinomas[tiab] OR neoplasm[tiab] OR neoplasms[tiab] OR nodule[tiab] OR nodules[tiab] OR tumor[tiab] OR tumour[tiab] OR Tumors[tiab] OR Tumours[tiab] OR cyst[tiab] OR cysts[tiab]))) AND (“Autopsy”[MeSH] OR “Autopsy”[tiab] OR “Autopsies”[tiab] OR “Postmortem”[tiab] OR Post-mortem[tiab] OR “step-sectioned”[tiab] OR “step sectioned”[tiab] OR (Post[tiab] AND mortem[tiab])) AND (“Prevalence”[MeSH] OR “Prevalence”[tiab] OR “Prevalences”[tiab] OR Incidence[tiab] OR Epidemiology[tiab] OR Epidemiological[tiab] OR Frequency[tiab] OR Detected[tiab]) AND (“Incidental Findings”[MeSH] OR Incidental[tiab] OR Unsuspected[tiab] OR Discovery[tiab] OR Discoveries[tiab] OR Findings[tiab] OR Finding[tiab] OR Occult[tiab] OR Hidden[tiab] OR Latent[tiab] OR Consecutive[tiab]))

Prompt

Construct a high-quality natural language query for the Boolean query of a systematic review:
{Boolean Query}
The effectiveness of the query will be determined by its capability to retrieve relevant documents when searching on a semantic-based search engine.

What is the prevalence or incidence of thyroid neoplasms, including adenocarcinoma, follicular, adenocarcinoma, papillary, OPTC, and other types of thyroid cancer, as detected through autopsy, post-mortem, or step-sectioned examinations?

Translated Query

Translating Boolean Queries

Single Query versus Multiple Queries

Decoder models have a temperature parameter that introduces randomness into token prediction.

- Single query generation sets temperature to 0.
  - Translation will produce the same query every time.
- Multiple query generation uses the default temperature parameter.
  - Translation will produce a different query every time.
  - Generate ten natural language queries per Boolean query.

Translating Boolean Queries

Single Query versus Multiple Queries

Decoder models have a temperature parameter that introduces randomness into token prediction.

- Single query generation sets temperature to 0.
  - Translation will produce the same query every time.
- Multiple query generation uses the default temperature parameter.
  - Translation will produce a different query every time.
  - Generate ten natural language queries per Boolean query.

What is the effectiveness of thyroid neoplasms, including adenocarcinoma, papillary neoplasms, and hurtle cell neoplasms, in the treatment of thyroid neoplasms? Specifically, I am interested in articles that discuss thyroid neoplasms in terms of cancer, neoplasms, carcinoma, or cysts, as well as articles that include thyroid neoplasms in step-sectioned or post-mortem studies.

What is the effectiveness of autopsy, postmortem, and step-sectioned neoplasms in the diagnosis of thyroid neoplasms, including Adenocarcinoma, Follicular Neoplasia, and Papillary Neoplasia? Specifically, I am interested in articles that discuss thyroid neoplasms, cancer, tumors, or cysts with respect to autopsy, postmortem, or step-sectioned.

What is the effectiveness of thyroid neoplasms, including adenocarcinoma, papillary thyroid neoplasm, and hurtle cell, in preventing or treating thyroid cancer, including subtypes of thyroid cancer such as papillomavirus, adenocarcinoma, and tumor?

What is the effectiveness of thyroid neoplasms, including adenocarcinoma, in cancer detection? Specifically, I am interested in studies that focus on the use of thyroid neoplasms in detecting hepatic fibrosis or cancer, as well as studies that utilize thyroid neoplasms to uncover latent or hidden thyroid conditions.

Translating Boolean Queries

Instruction fine-tuning

ChatGPT is a closed model, research with it has many limitations.

- Llama is a useable model trained on publicly available data [Touvron et al. 2023].
- However, on its own, Llama is 'just' a very large decoder-only transformer.
  - Adjusts aspects like the activation function and positional embeddings.
  - Several instances of the model, each with more parameters, 7B-65B.
  - For comparison, BERT has 110M parameters.
- Can fine-tune Llama in a similar way to make GPT-3 work like ChatGPT.
  - Self-instruct used to induce instruction following capabilities with minimal human-labelled data [Wang et al. 2023].
  - Bootstrap a model to generate instructions for itself and then learn to predict outputs for instructions.
- Alpaca is an instruction fine-tuned Llama model [Taori et al. 2023]
  - Here, Alpaca is further instruction fine-tuned for generating natural language queries.

Translating Boolean Queries

Single Query Results on Seed Collection [Wang et al. 2022]

Query	Method	MAP
Boolean terms	BM25	0.087
Boolean terms	QLM	0.085
Boolean terms	BioBERT	0.199

Translating Boolean Queries

Single Query Results on Seed Collection [Wang et al. 2022]

Query	Method	MAP
Boolean terms	BM25	0.087
Boolean terms	QLM	0.085
Boolean terms	BioBERT	0.199
ChatGPT	BioBERT	0.217
Alpaca	BioBERT	0.221

Translating Boolean Queries

Single Query Results on Seed Collection [Wang et al. 2022]

Query	Method	MAP
Boolean terms	BM25	0.087
Boolean terms	QLM	0.085
Boolean terms	BioBERT	0.199
ChatGPT	BioBERT	0.217
Alpaca	BioBERT	0.221
ChatGPT+Boolean terms	BioBERT	0.219
Alpaca+Boolean terms	BioBERT	0.230

Translating Boolean Queries

Single Query Results on Seed Collection [Wang et al. 2022]

Query	Method	MAP
Boolean terms	BM25	0.087
Boolean terms	QLM	0.085
Boolean terms	BioBERT	0.199
ChatGPT	BioBERT	0.217
Alpaca	BioBERT	0.221
ChatGPT+Boolean terms	BioBERT	0.219
Alpaca+Boolean terms	BioBERT	0.230

Instruction fine-tuned Alpaca produces most effective queries.

Translating Boolean Queries

Single Query Results on Seed Collection [Wang et al. 2022]

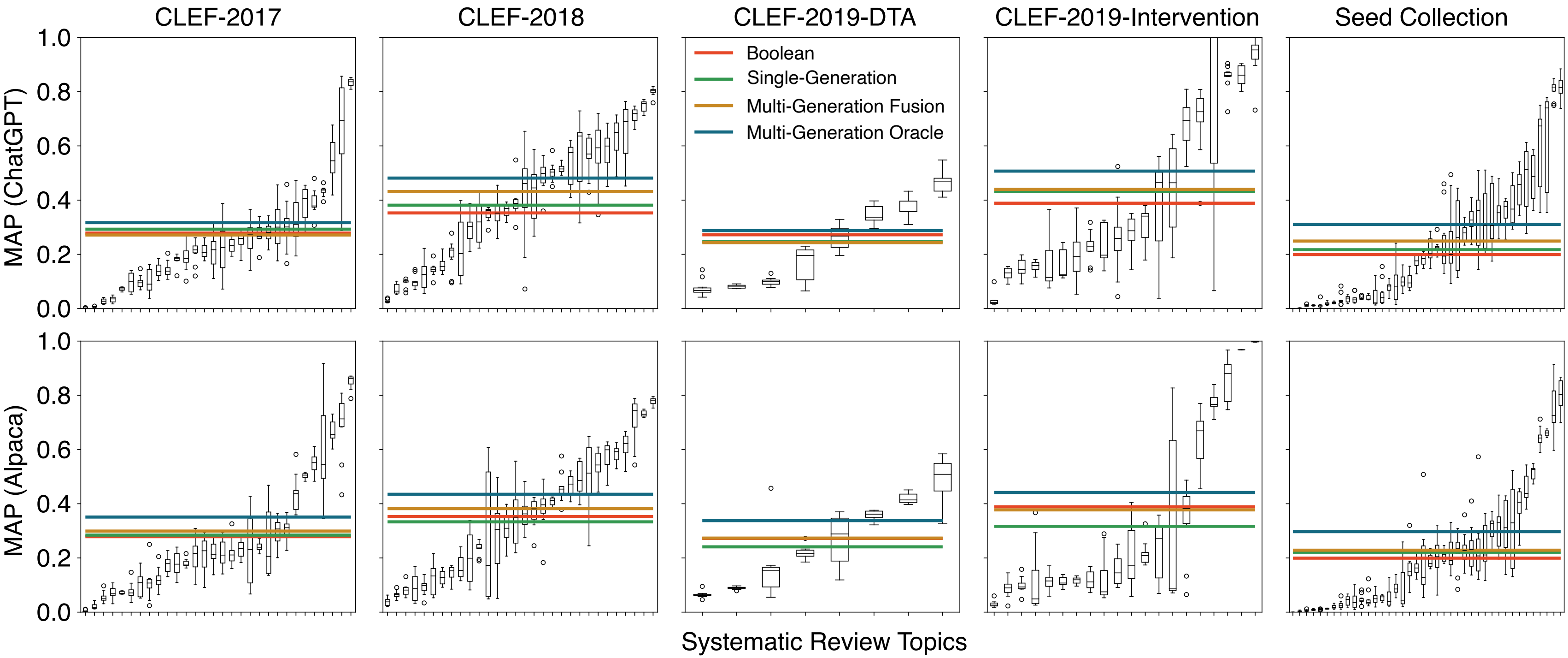
Query	Method	MAP
Boolean terms	BM25	0.087
Boolean terms	QLM	0.085
Boolean terms	BioBERT	0.199
ChatGPT	BioBERT	0.217
Alpaca	BioBERT	0.221
ChatGPT+Boolean terms	BioBERT	0.219
Alpaca+Boolean terms	BioBERT	0.230

Instruction fine-tuned Alpaca produces most effective queries.

Adding terms from the Boolean query improved ranking.

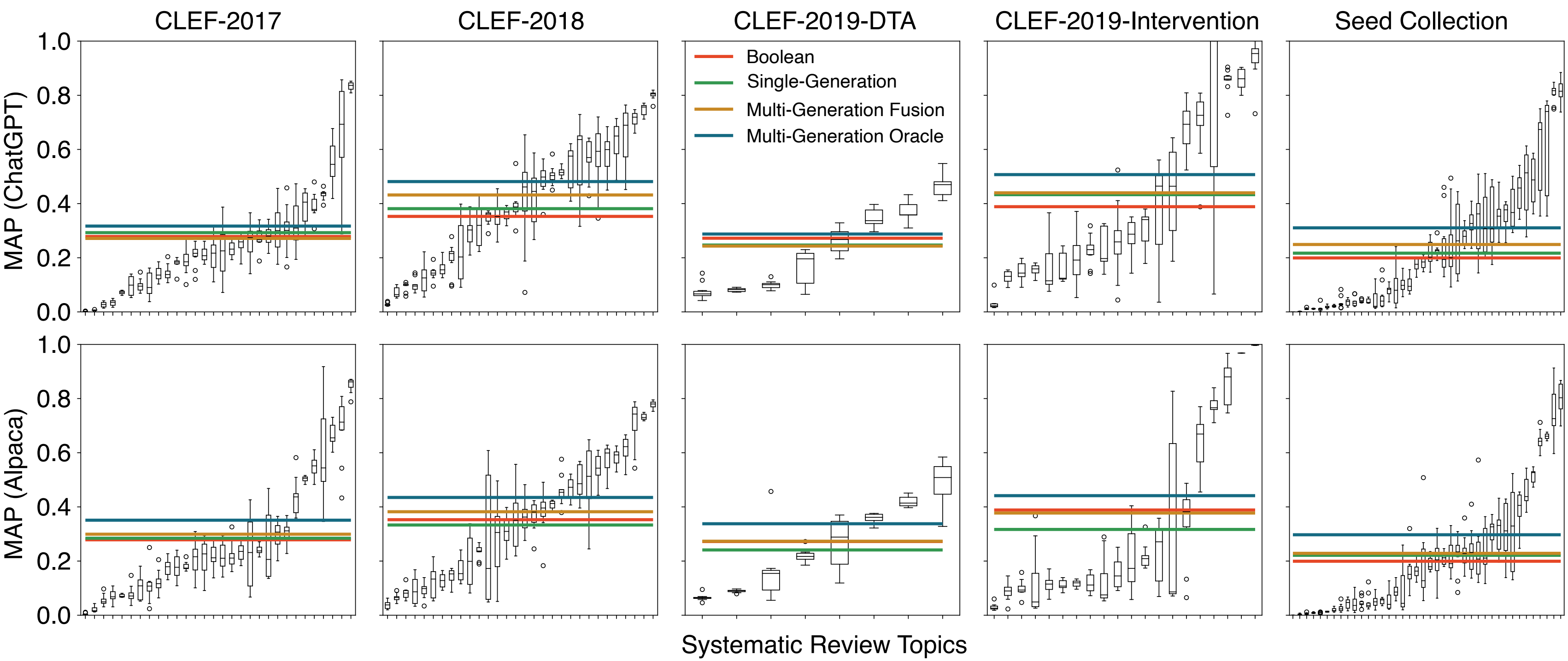
Translating Boolean Queries

Multiple Query Results on Seed Collection [Wang et al. 2022]



Translating Boolean Queries

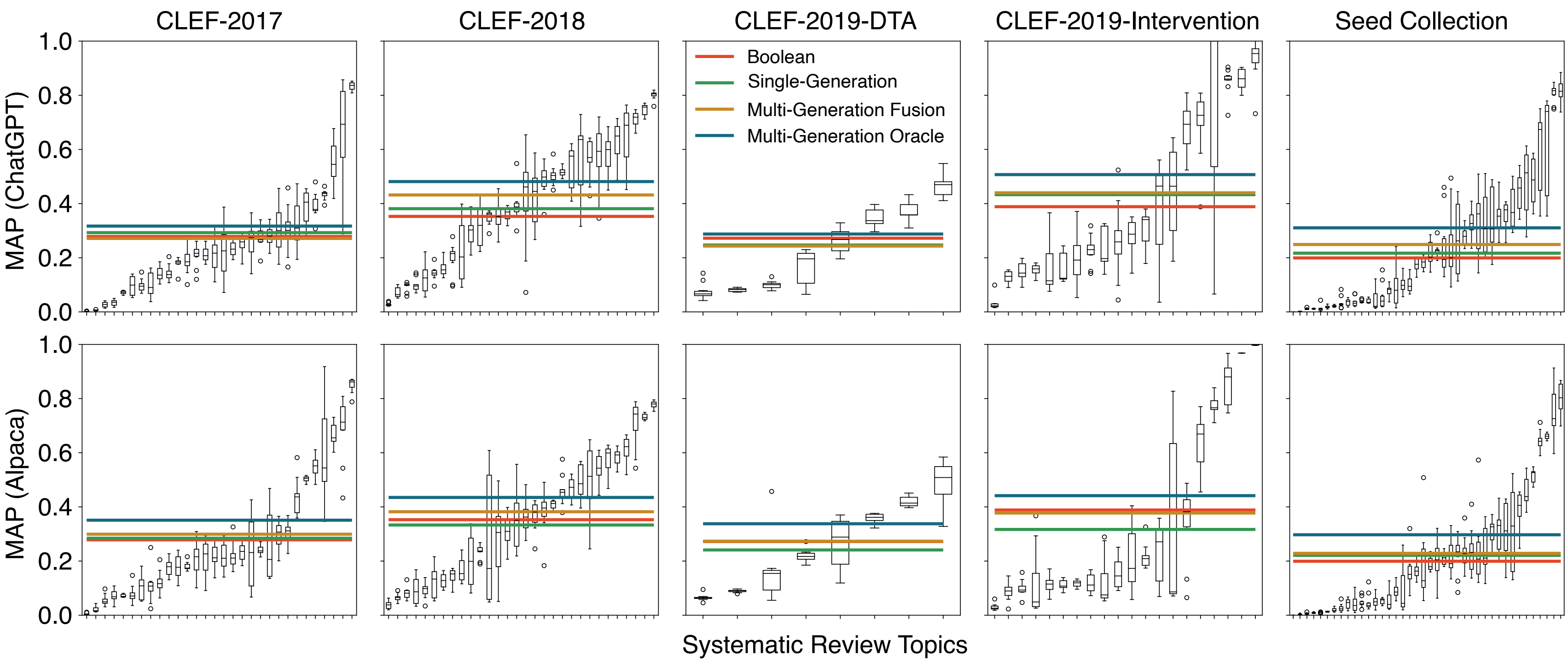
Multiple Query Results on Seed Collection [Wang et al. 2022]



Fusion of multiple generated queries typically results in higher ranking effectiveness.

Translating Boolean Queries

Multiple Query Results on Seed Collection [Wang et al. 2022]



Fusion of multiple generated queries typically results in higher ranking effectiveness.

Oracle suggests that individual queries exist that are highly effective.

Automatically Assessing Documents

Overview

Use decoder models to automatically classify documents retrieved by Boolean query as relevant or non-relevant [Wang et al. 2024].

- Prompts a decoder-only model with instructions to output whether a document is relevant to a given systematic review.
- Everything done in a ‘zero-shot’ fashion: Models are not fine-tuned.
- Use token probabilities to determine degree of relevance to review.
- Learn a threshold to tighten the bound on precision.

Automatically Assessing Documents

Overview

Answer the following yes/no question. Should the following retrieved document be included in the systematic review?
Review: `{review information}`
Document: `{document text}`
The answer is ‘

Prompt

Decoder Model

`[0.11,0.112,0.123,...,0.001]`

Next token probabilities

$P(\texttt{yes}|r,d)$
$P(\texttt{no}|r,d)$

Uncalibrated setting

$$I(d,t) = \begin{cases} 1, & \text{if } P(\text{yes}|d,t) \geq P(\text{no}|d,t) \\ 0, & \text{otherwise.} \end{cases}$$

Calibrated setting

$$S(d,t) = P(\texttt{yes}|r,d) - P(\texttt{no}|r,d)$$

$$I(d,t) = \begin{cases} 1, & \text{if } S_{\text{norm}}(d,t) \geq \theta \\ 0, & \text{otherwise.} \end{cases}$$

Automatically Assessing Documents

Decoder Models

Following large decoder models used in a zero-shot setting:

- LLaMa
- Alpaca
- LLaMa-2
  - Identical model architecture to LLaMa, but trained on more data.
- Guanaco
  - Memory-efficient fine-tuned LLaMA model.
- Falcon
  - First ‘open’ model of its kind, freely available and permissive license.
  - Decoder-only model with different yet architectures to LLaMA and GPT.

Every document retrieved by the query requires a forward pass through each model.

- Using GPT-3.5-turbo would cost USD$4,000 and GPT-4 USD$80,000.

Automatically Assessing Documents

Uncalibrated Results on Seed Collection [Wang et al. 2022]

Model	Precision	Recall	F3
BioBERT	0.04	0.93	0.24
Alpaca-7b	0.04	0.90	0.22
Falcon-7b	0.04	0.93	0.22
Guanaco-7b	0.04	1.00	0.23
LLaMA-2-7b	0.05	0.90	0.27
LLaMA-2-13b	0.13	0.48	0.28

Automatically Assessing Documents

Uncalibrated Results on Seed Collection [Wang et al. 2022]

Model	Precision	Recall	F3
BioBERT	0.04	0.93	0.24
Alpaca-7b	0.04	0.90	0.22
Falcon-7b	0.04	0.93	0.22
Guanaco-7b	0.04	1.00	0.23
LLaMA-2-7b	0.05	0.90	0.27
LLaMA-2-13b	0.13	0.48	0.28

LLaMa-2-7b gives best trade-off between precision and recall.

Automatically Assessing Documents

Uncalibrated Results on Seed Collection [Wang et al. 2022]

Model	Precision	Recall	F3
BioBERT	0.04	0.93	0.24
Alpaca-7b	0.04	0.90	0.22
Falcon-7b	0.04	0.93	0.22
Guanaco-7b	0.04	1.00	0.23
LLaMA-2-7b	0.05	0.90	0.27
LLaMA-2-13b	0.13	0.48	0.28

LLaMa-2-7b gives best trade-off between precision and recall.

LLaMa-2-13b can dramatically increase precision at the cost of recall.

Automatically Assessing Documents

Calibrated Results on Seed Collection [Wang et al. 2022]

Model	Precision	Recall	F3
BioBERT (unc)	0.04	0.93	0.24
BioBERT (cal)	0.04	0.83	0.22
LLaMA-2-7b (unc)	0.05	0.90	0.27
LLaMA-2-7b (cal)	0.05	0.97	0.26
LLaMA-2-13b (unc)	0.13	0.48	0.28
LLaMA-2-13b (cal)	0.06	0.87	0.25

Automatically Assessing Documents

Calibrated Results on Seed Collection [Wang et al. 2022]

Model	Precision	Recall	F3
BioBERT (unc)	0.04	0.93	0.24
BioBERT (cal)	0.04	0.83	0.22
LLaMA-2-7b (unc)	0.05	0.90	0.27
LLaMA-2-7b (cal)	0.05	0.97	0.26
LLaMA-2-13b (unc)	0.13	0.48	0.28
LLaMA-2-13b (cal)	0.06	0.87	0.25

Calibrated LLaMa-2-7b assesses documents better than calibrated LLaMa-2-13b.

Automatically Assessing Documents

Calibrated Results on Seed Collection [Wang et al. 2022]

Model	Precision	Recall	F3
BioBERT (unc)	0.04	0.93	0.24
BioBERT (cal)	0.04	0.83	0.22
LLaMA-2-7b (unc)	0.05	0.90	0.27
LLaMA-2-7b (cal)	0.05	0.97	0.26
LLaMA-2-13b (unc)	0.13	0.48	0.28
LLaMA-2-13b (cal)	0.06	0.87	0.25

Calibrated LLaMa-2-7b assesses documents better than calibrated LLaMa-2-13b.

Similar results can be observed on other test collections.