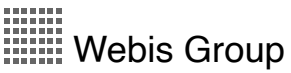
Generative Search

ChatGPT and Large Language Models for Medical Applications

Martin Potthast and Harry Scells

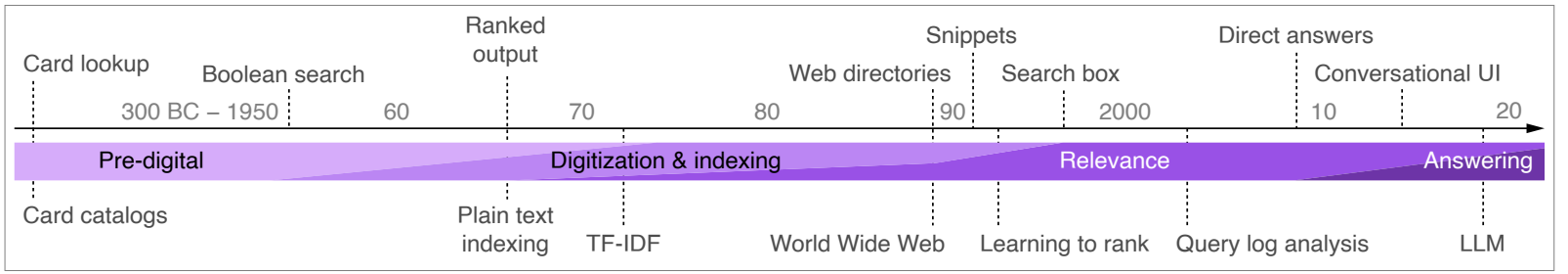


A Short History of Search Engines

A Short History of Search Engines

User

System

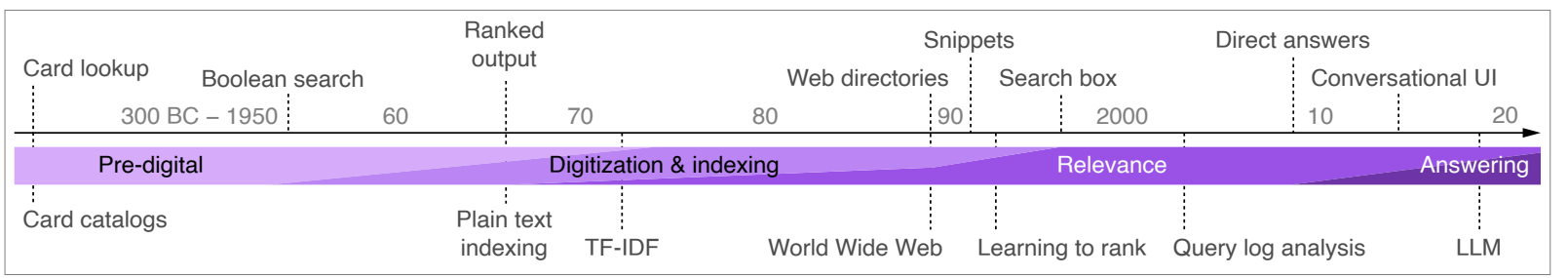


- M. Potthast, M. Hagen, B. Stein (2020). The dilemma of the direct answer.

A Short History of Search Engines

User

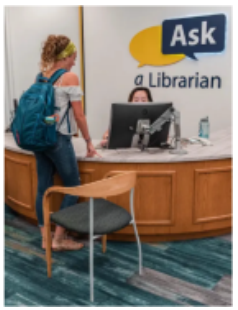
System



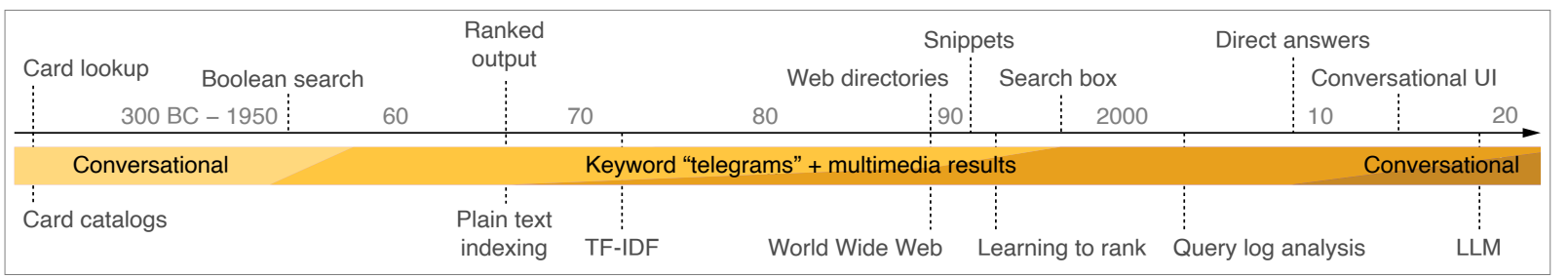
- M. Potthast, M. Hagen, B. Stein (2020). The dilemma of the direct answer.

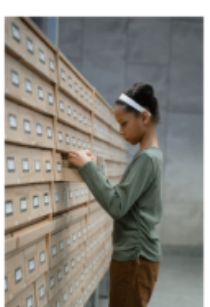
A Short History of Search Engines

User



System



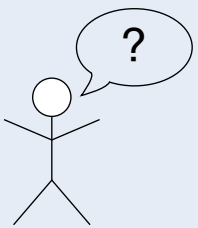


- M. Potthast, M. Hagen, B. Stein (2020). The dilemma of the direct answer.

Web Search Architecture

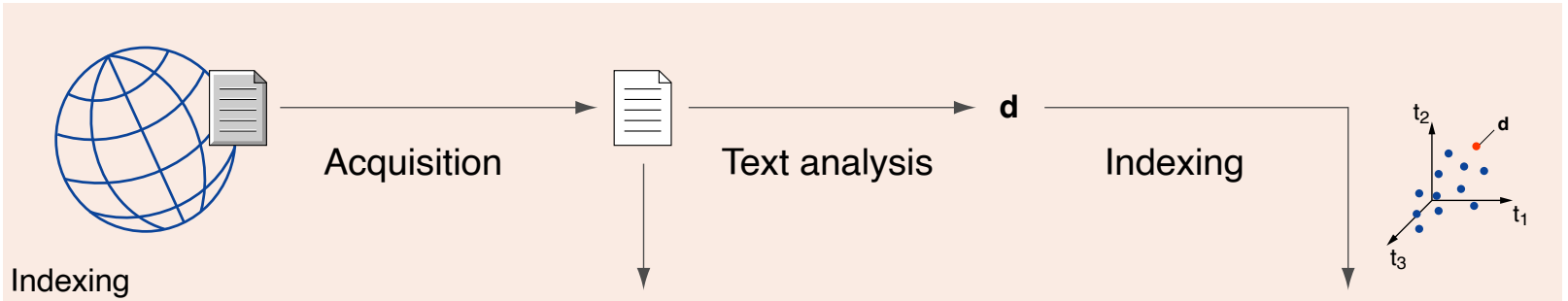
Indexing

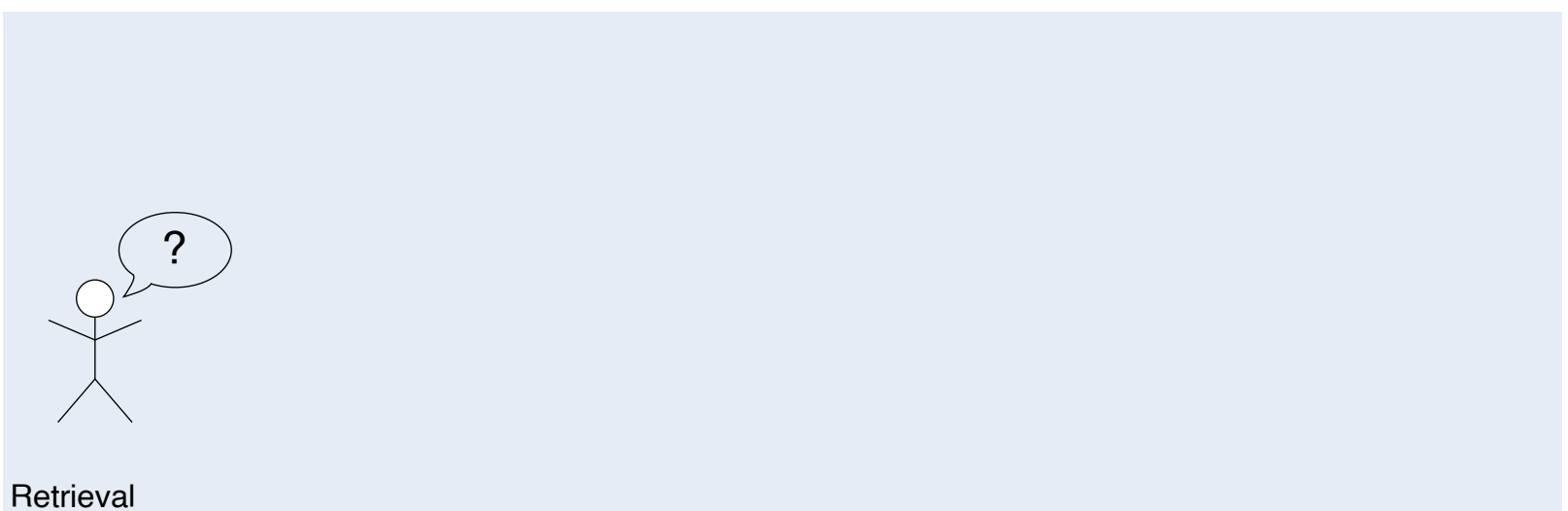
Storage



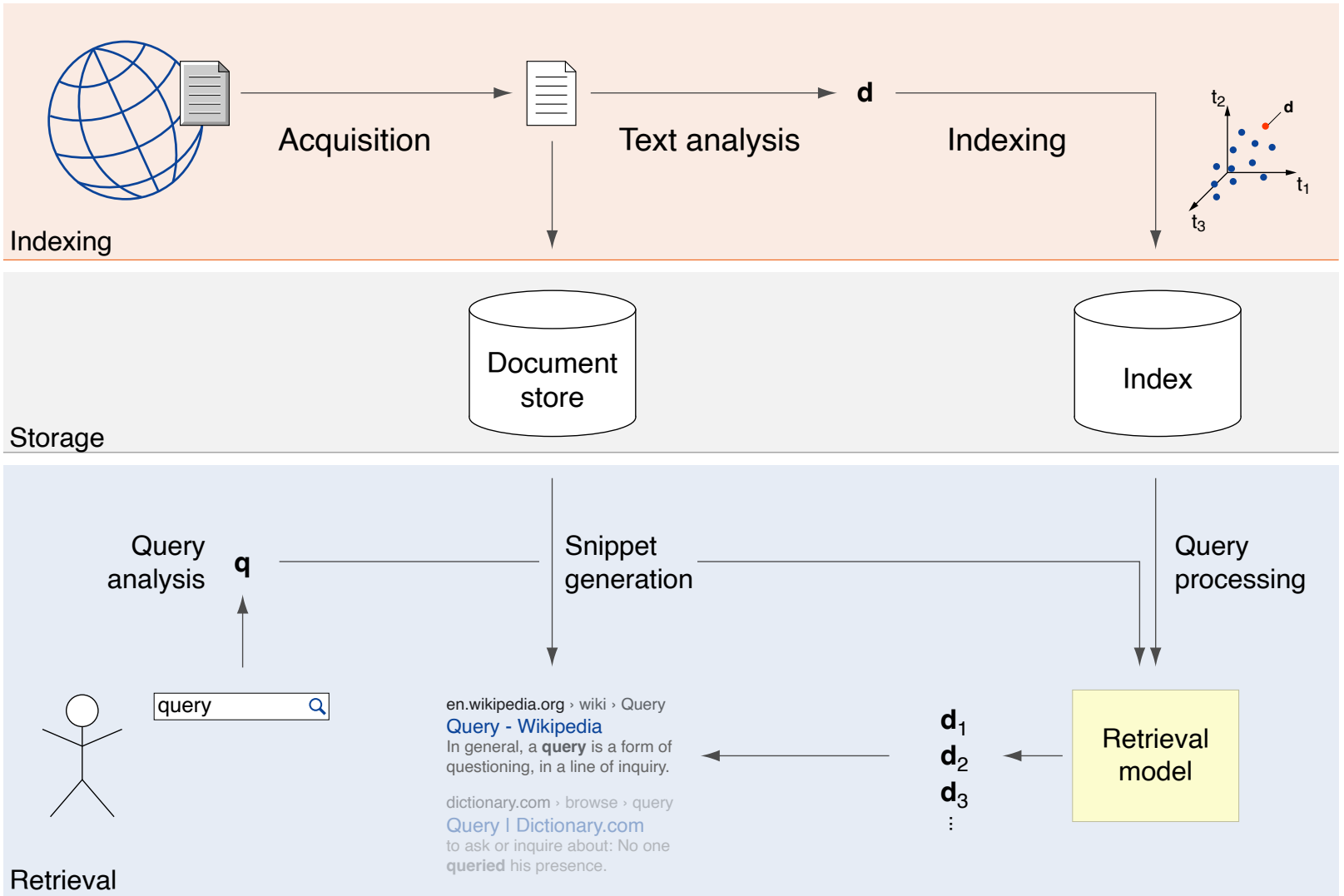
Retrieval

Web Search Architecture

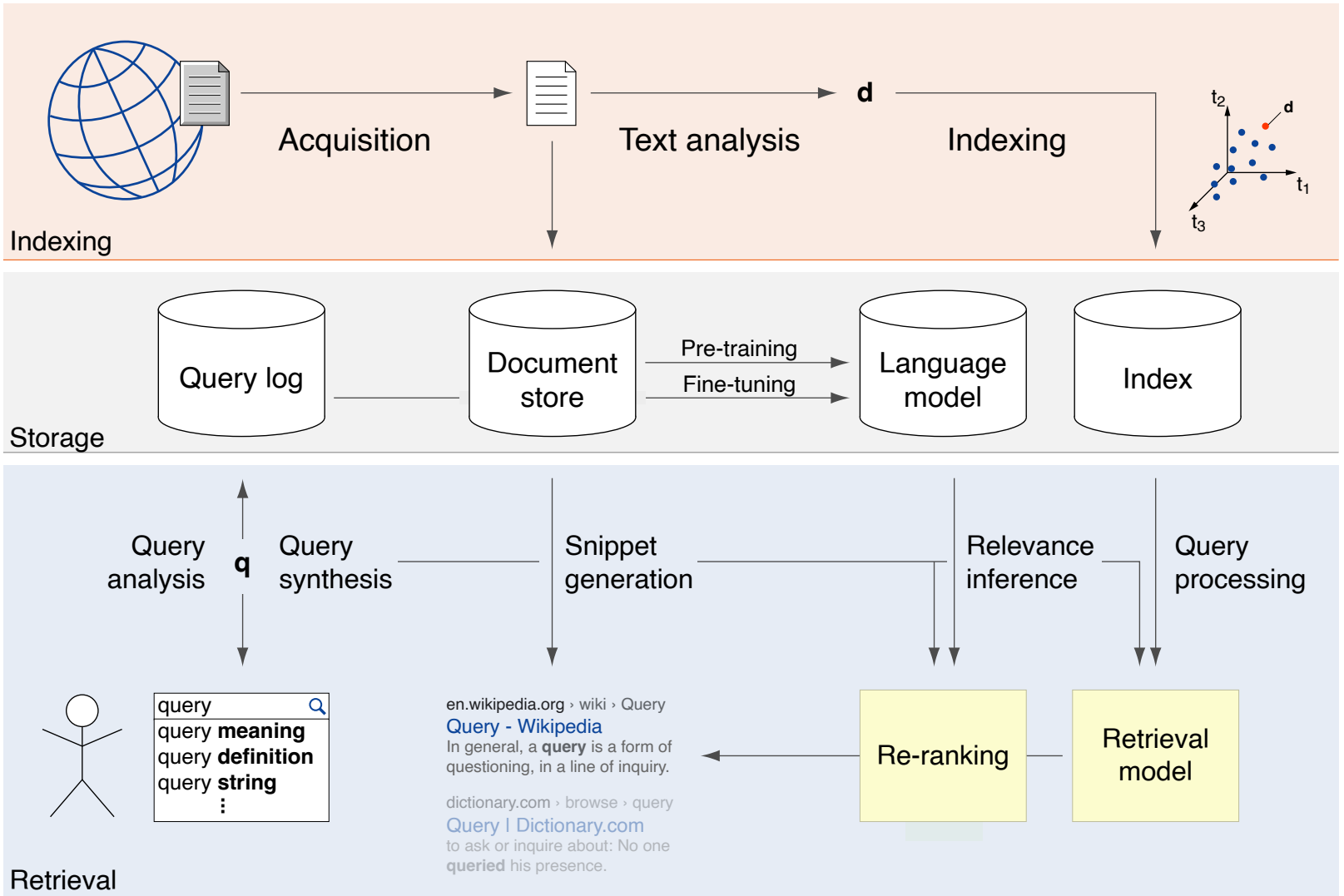




Web Search Architecture

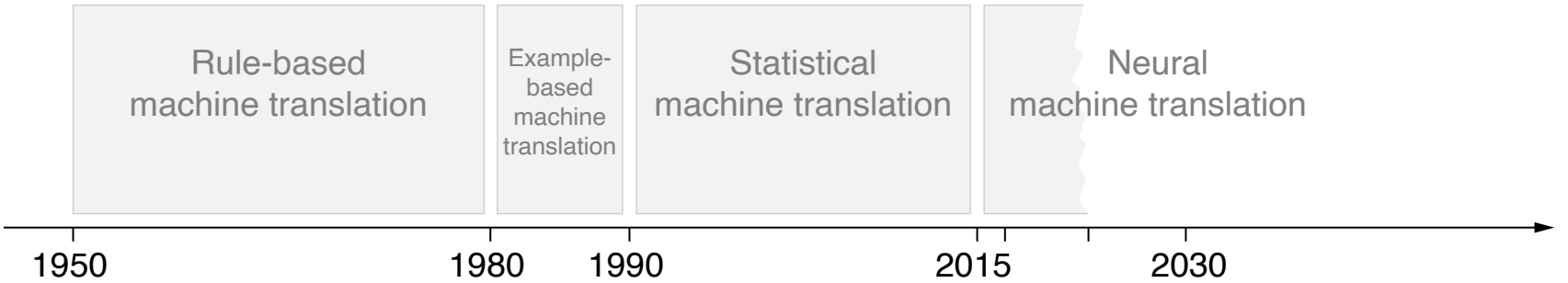


Web Search Architecture

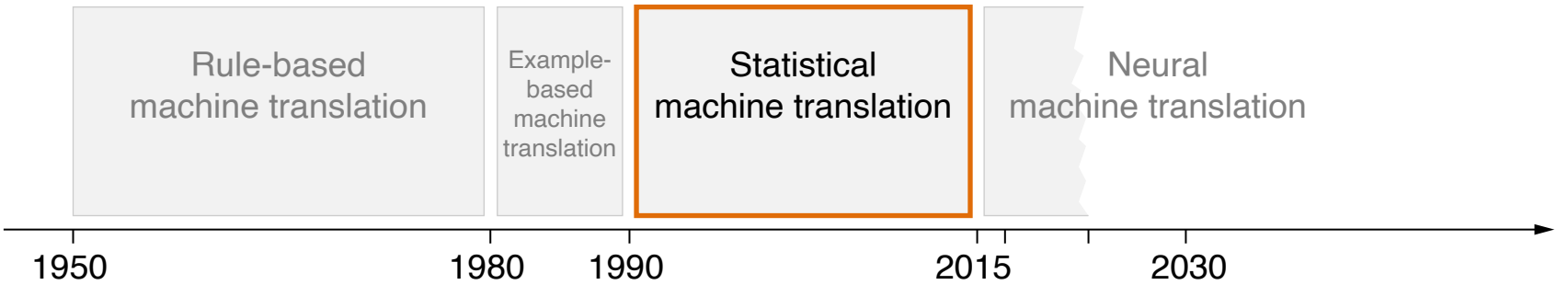


A Short History of Language Models

A Short History of Language Models

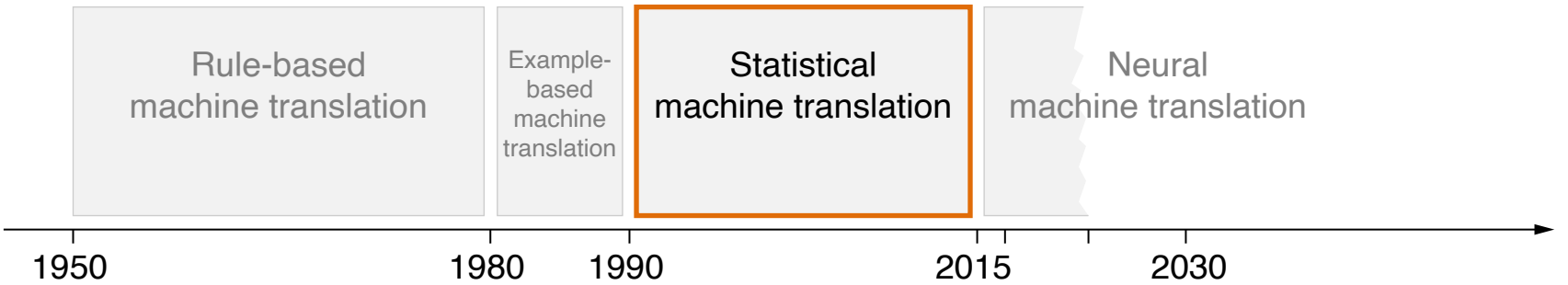


A Short History of Language Models



A statistical language model
is a probability distribution over all possible texts.

A Short History of Language Models



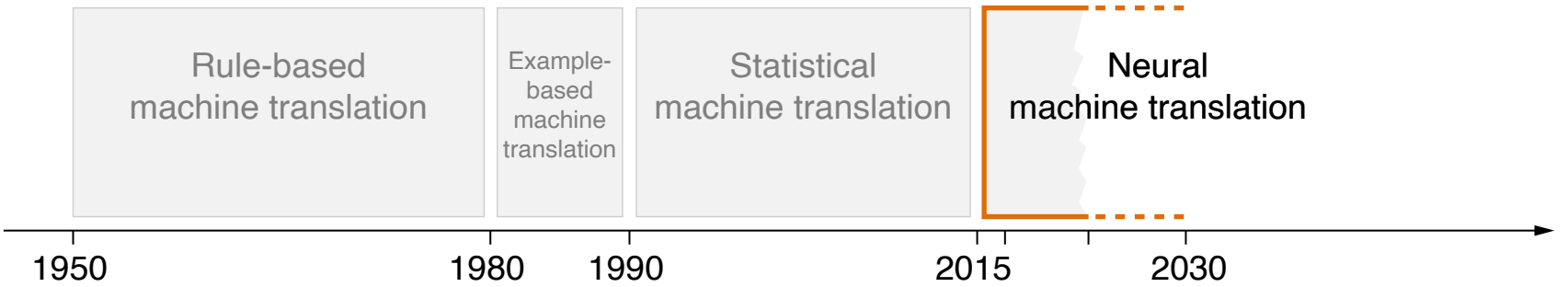
A statistical language model
is a probability distribution over all possible texts.

Illustration:

(1) `i love my ?`

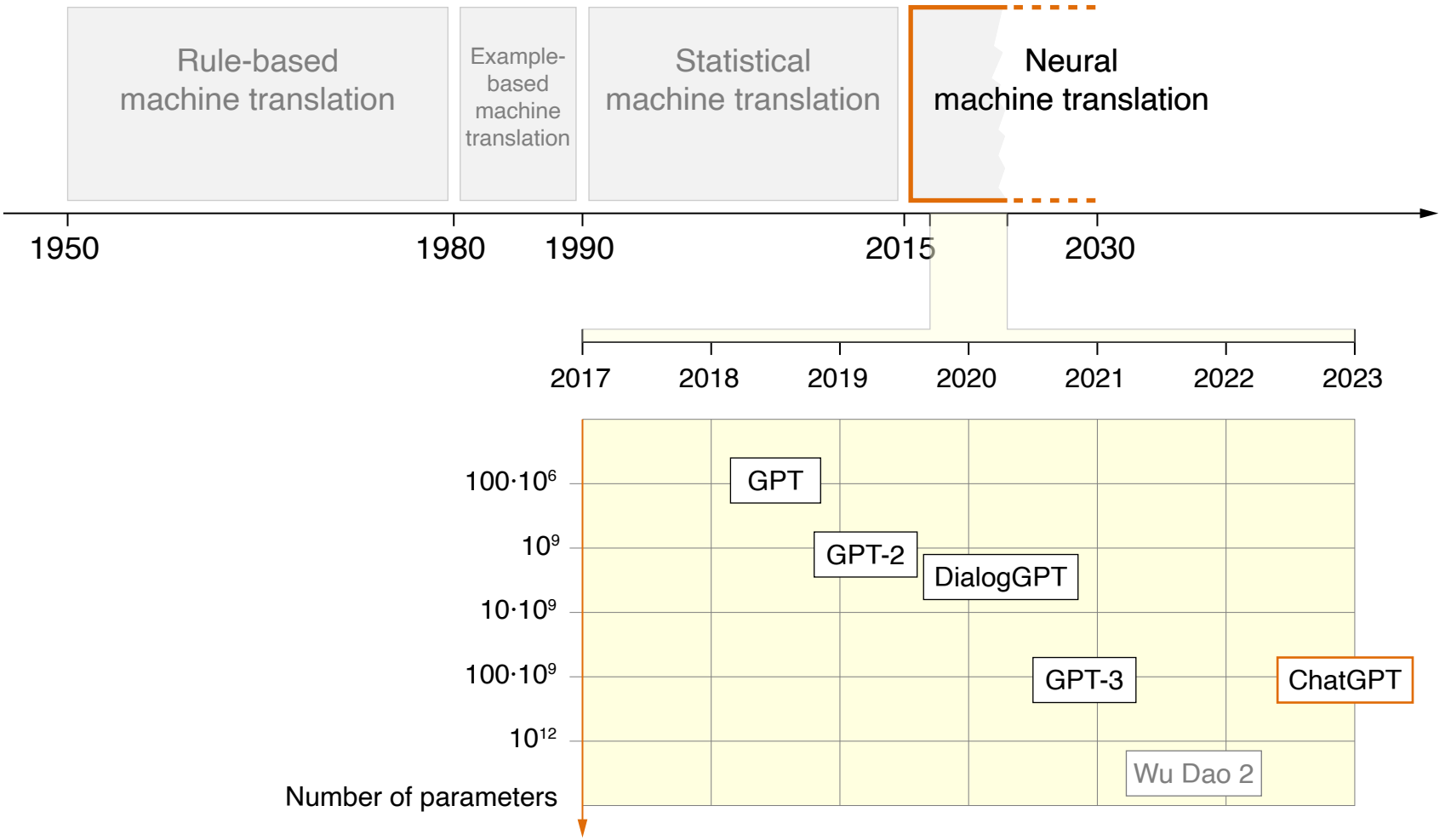
(2) `see ... works.`

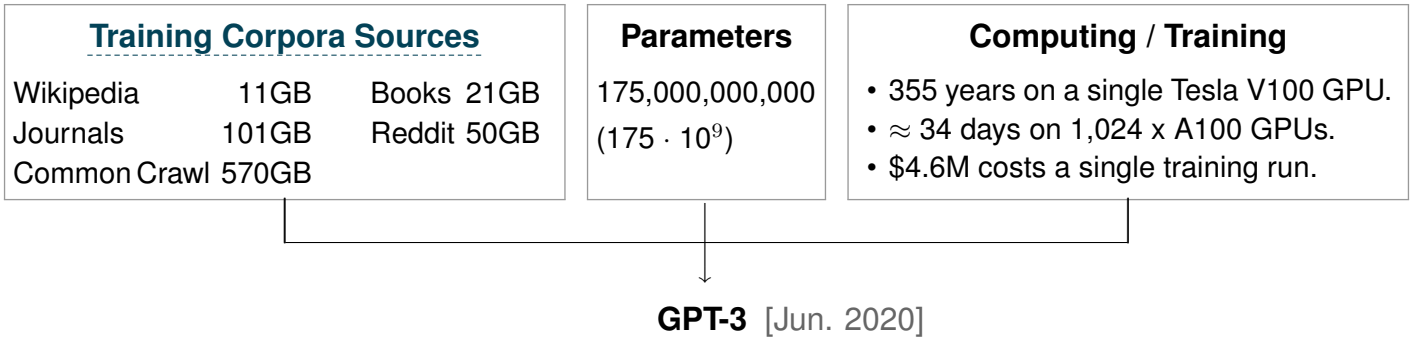
A Short History of Language Models

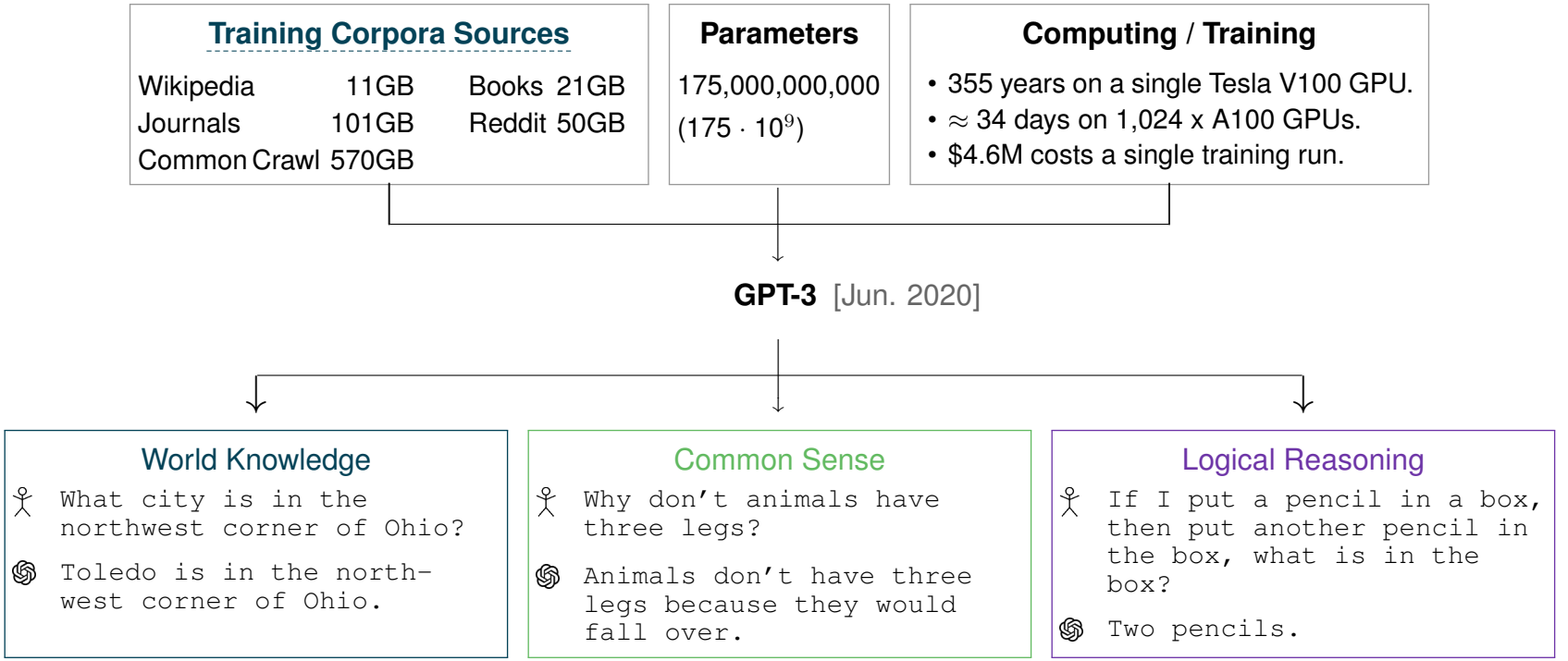


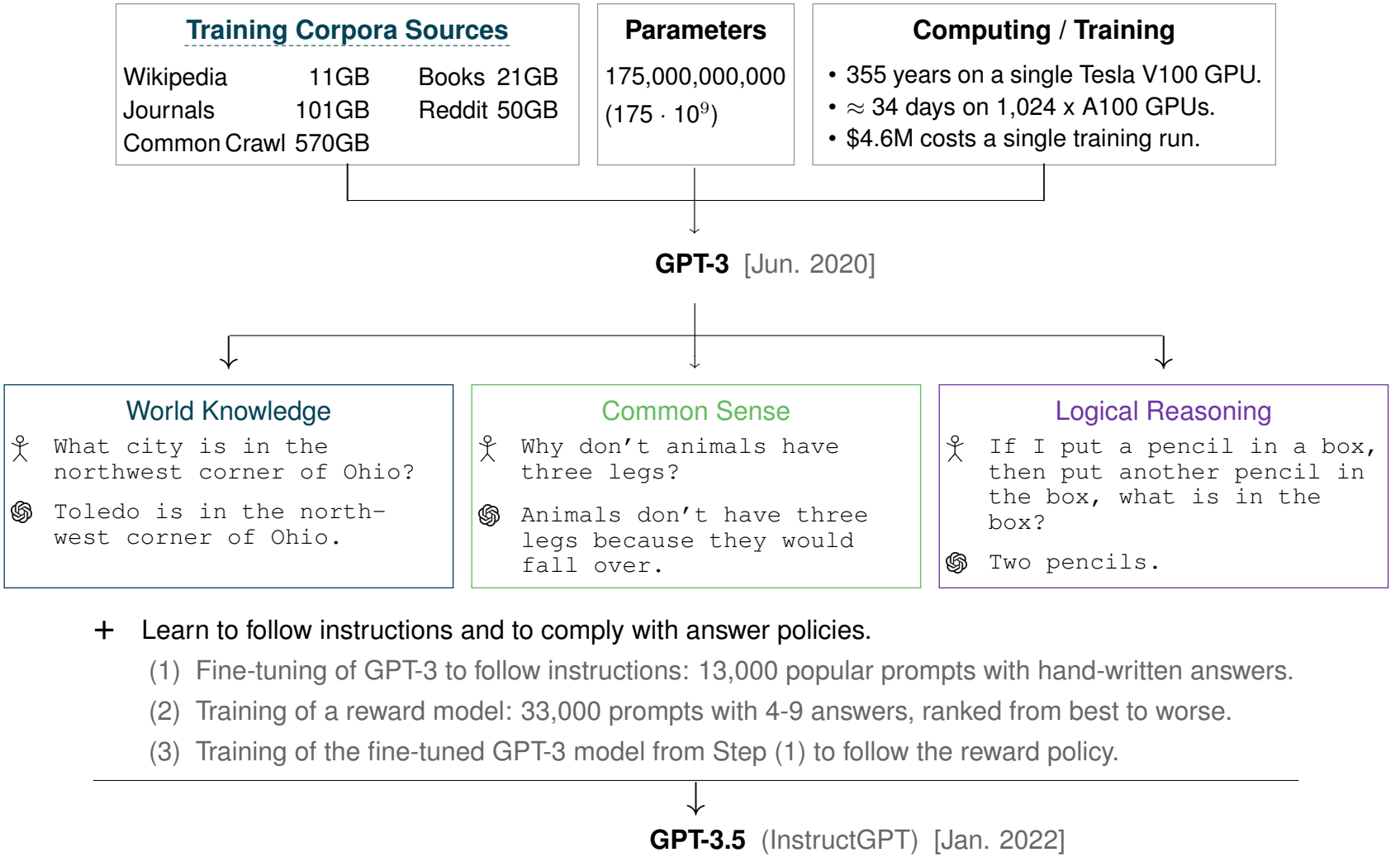
A neural language model
approximates a statistical language model.

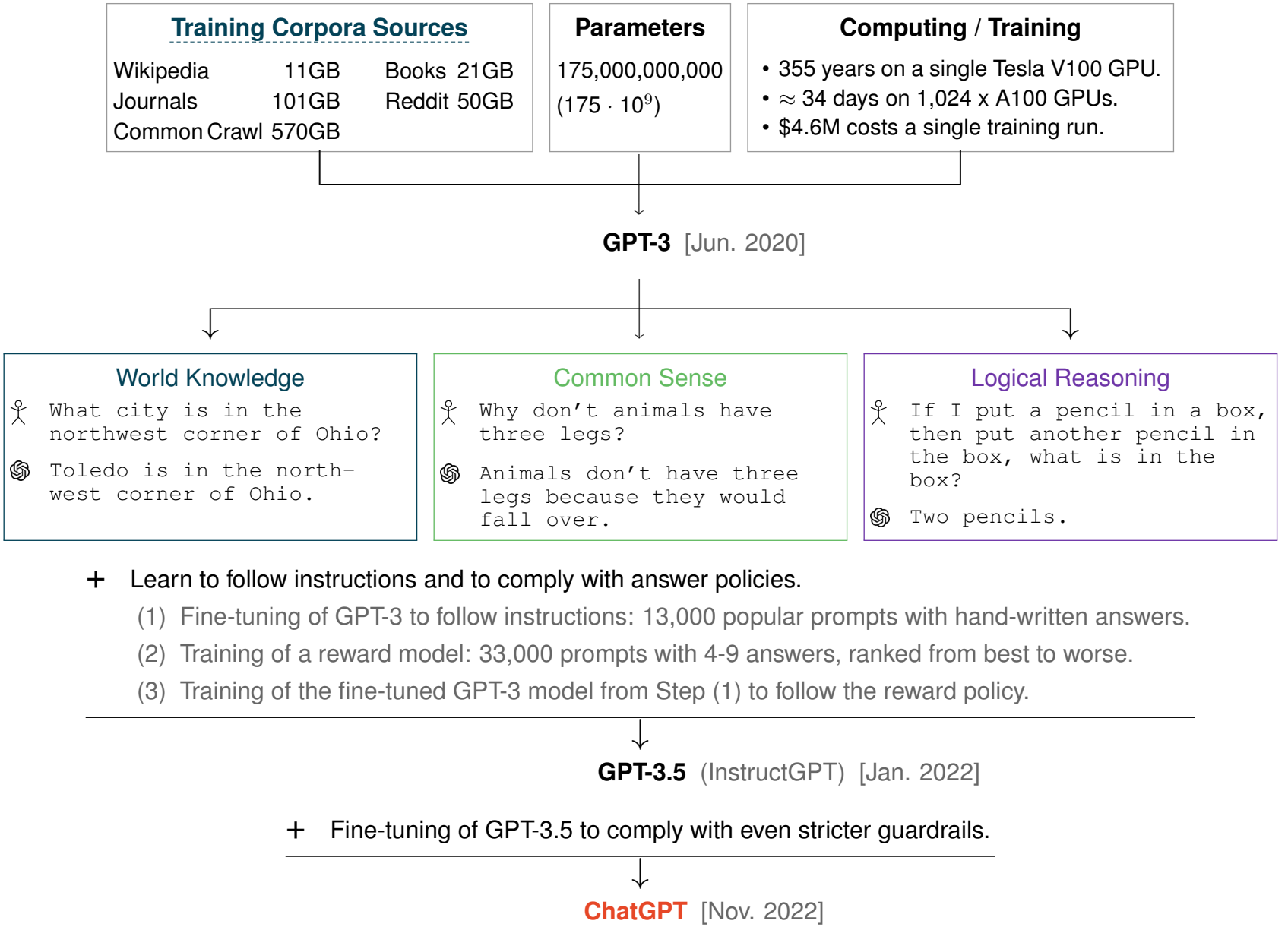
A Short History of Language Models




Training Corpora Sources
Wikipedia 11GB
Books 21GB
Journals 101GB
Reddit 50GB
Common Crawl 570GB
Parameters
175,000,000,000
$(175 \cdot 10^9)$
Computing / Training
• 355 years on a single Tesla V100 GPU.
• ≈ 34 days on 1,024 x A100 GPUs.
• $4.6M costs a single training run.
GPT-3 [Jun. 2020]


Training Corpora Sources
Wikipedia 11GB
Books 21GB
Journals 101GB
Reddit 50GB
Common Crawl 570GB
Parameters
175,000,000,000
(175 · 10^9)
Computing / Training
• 355 years on a single Tesla V100 GPU.
• ≈ 34 days on 1,024 x A100 GPUs.
• $4.6M costs a single training run.
GPT-3 [Jun. 2020]
World Knowledge
What city is in the northwest corner of Ohio?
Toledo is in the north-west corner of Ohio.
Common Sense
Why don't animals have three legs?
Animals don't have three legs because they would fall over.
Logical Reasoning
If I put a pencil in a box, then put another pencil in the box, what is in the box?
Two pencils.

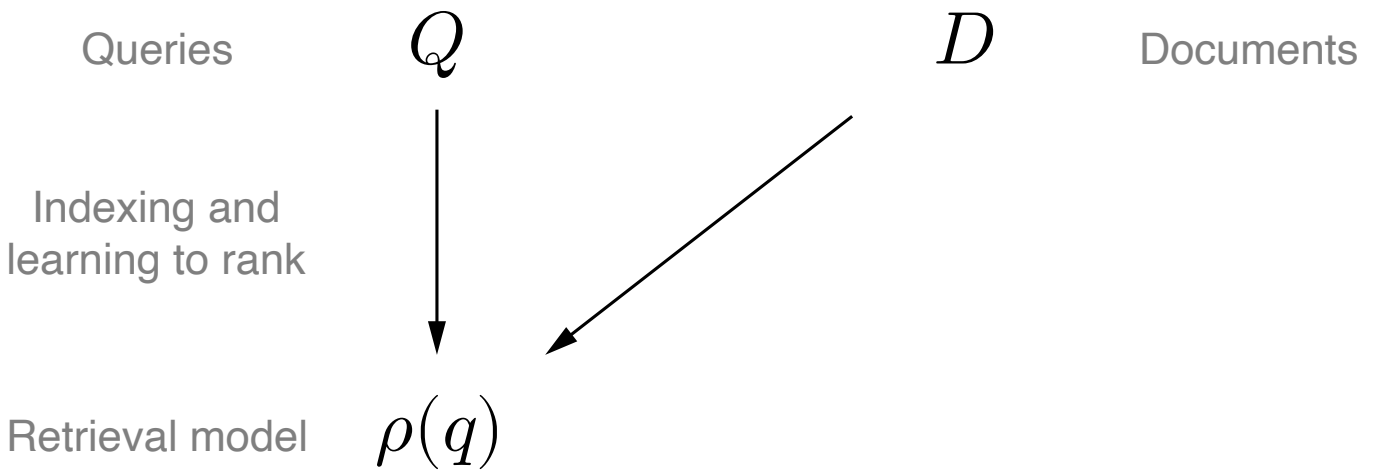

Training Corpora Sources
Wikipedia 11GB
Books 21GB
Journals 101GB
Reddit 50GB
Common Crawl 570GB
Parameters
175,000,000,000
(175 · 10^9)
Computing / Training
• 355 years on a single Tesla V100 GPU.
• ≈ 34 days on 1,024 x A100 GPUs.
• $4.6M costs a single training run.
GPT-3 [Jun. 2020]
World Knowledge
What city is in the northwest corner of Ohio?
Toledo is in the northwest corner of Ohio.
Common Sense
Why don't animals have three legs?
Animals don't have three legs because they would fall over.
Logical Reasoning
If I put a pencil in a box, then put another pencil in the box, what is in the box?
Two pencils.
+ Learn to follow instructions and to comply with answer policies.
(1) Fine-tuning of GPT-3 to follow instructions: 13,000 popular prompts with hand-written answers.
(2) Training of a reward model: 33,000 prompts with 4-9 answers, ranked from best to worse.
(3) Training of the fine-tuned GPT-3 model from Step (1) to follow the reward policy.
GPT-3.5 (InstructGPT) [Jan. 2022]


Training Corpora Sources
Wikipedia 11GB
Books 21GB
Journals 101GB
Reddit 50GB
Common Crawl 570GB
Parameters
175,000,000,000
(175 · 10^9)
Computing / Training
• 355 years on a single Tesla V100 GPU.
• ≈ 34 days on 1,024 x A100 GPUs.
• $4.6M costs a single training run.
GPT-3 [Jun. 2020]
World Knowledge
What city is in the northwest corner of Ohio?
Toledo is in the north-west corner of Ohio.
Common Sense
Why don't animals have three legs?
Animals don't have three legs because they would fall over.
Logical Reasoning
If I put a pencil in a box, then put another pencil in the box, what is in the box?
Two pencils.
+ Learn to follow instructions and to comply with answer policies.
(1) Fine-tuning of GPT-3 to follow instructions: 13,000 popular prompts with hand-written answers.
(2) Training of a reward model: 33,000 prompts with 4-9 answers, ranked from best to worse.
(3) Training of the fine-tuned GPT-3 model from Step (1) to follow the reward policy.
GPT-3.5 (InstructGPT) [Jan. 2022]
+ Fine-tuning of GPT-3.5 to comply with even stricter guardrails.
ChatGPT [Nov. 2022]

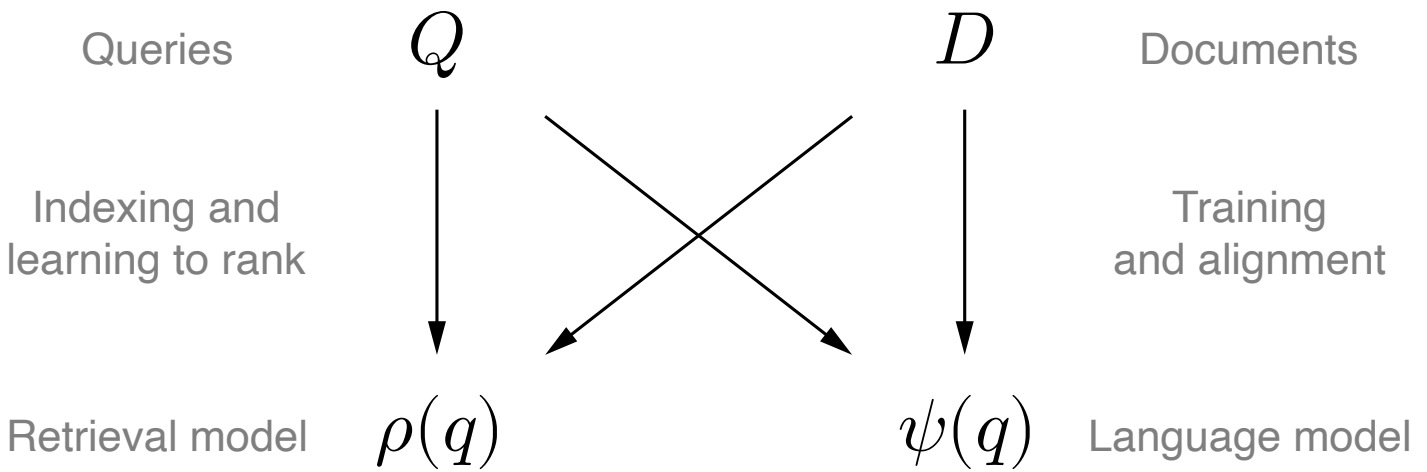
A Short History of Language Models

Queries Q D Documents

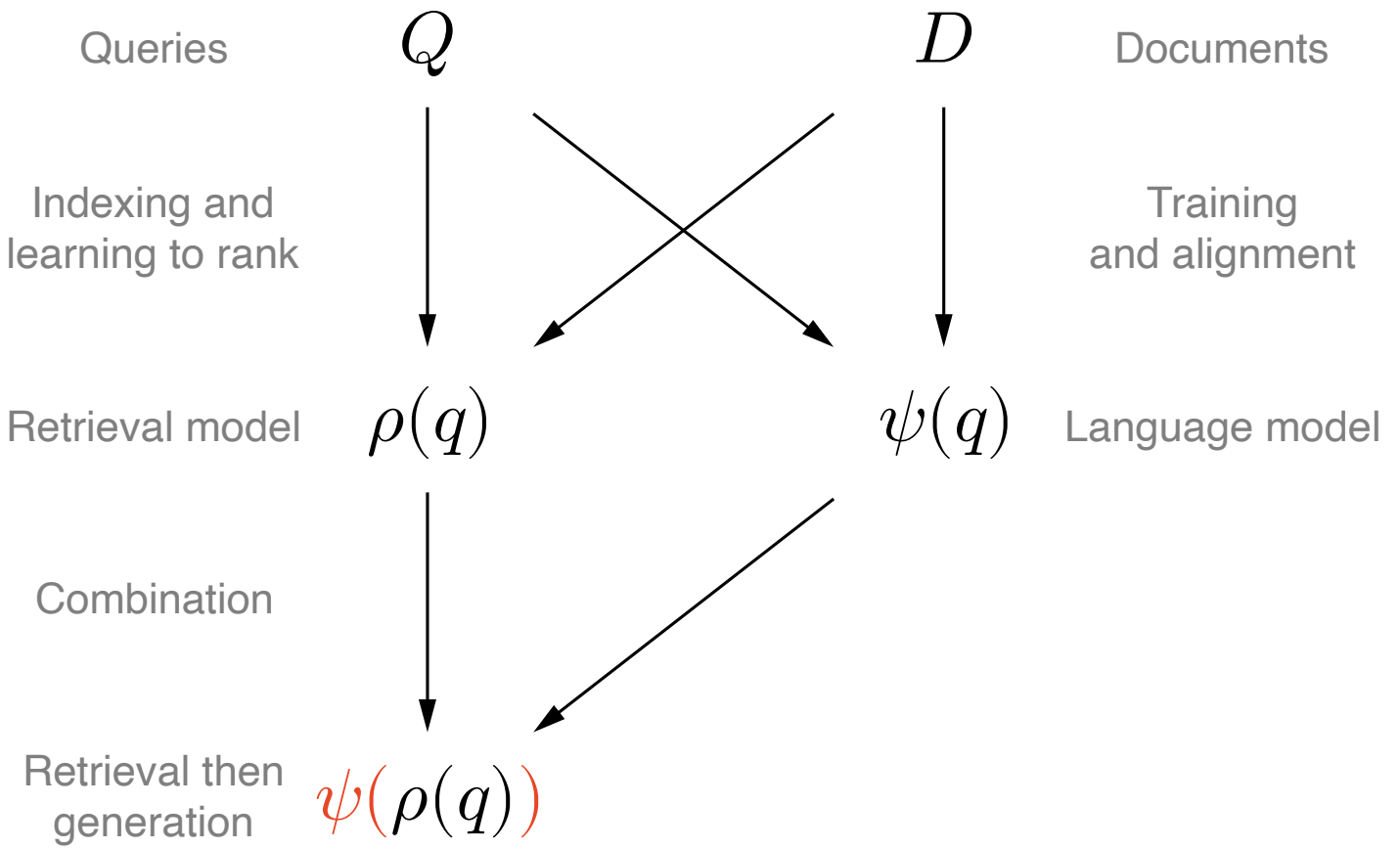
A Short History of Language Models



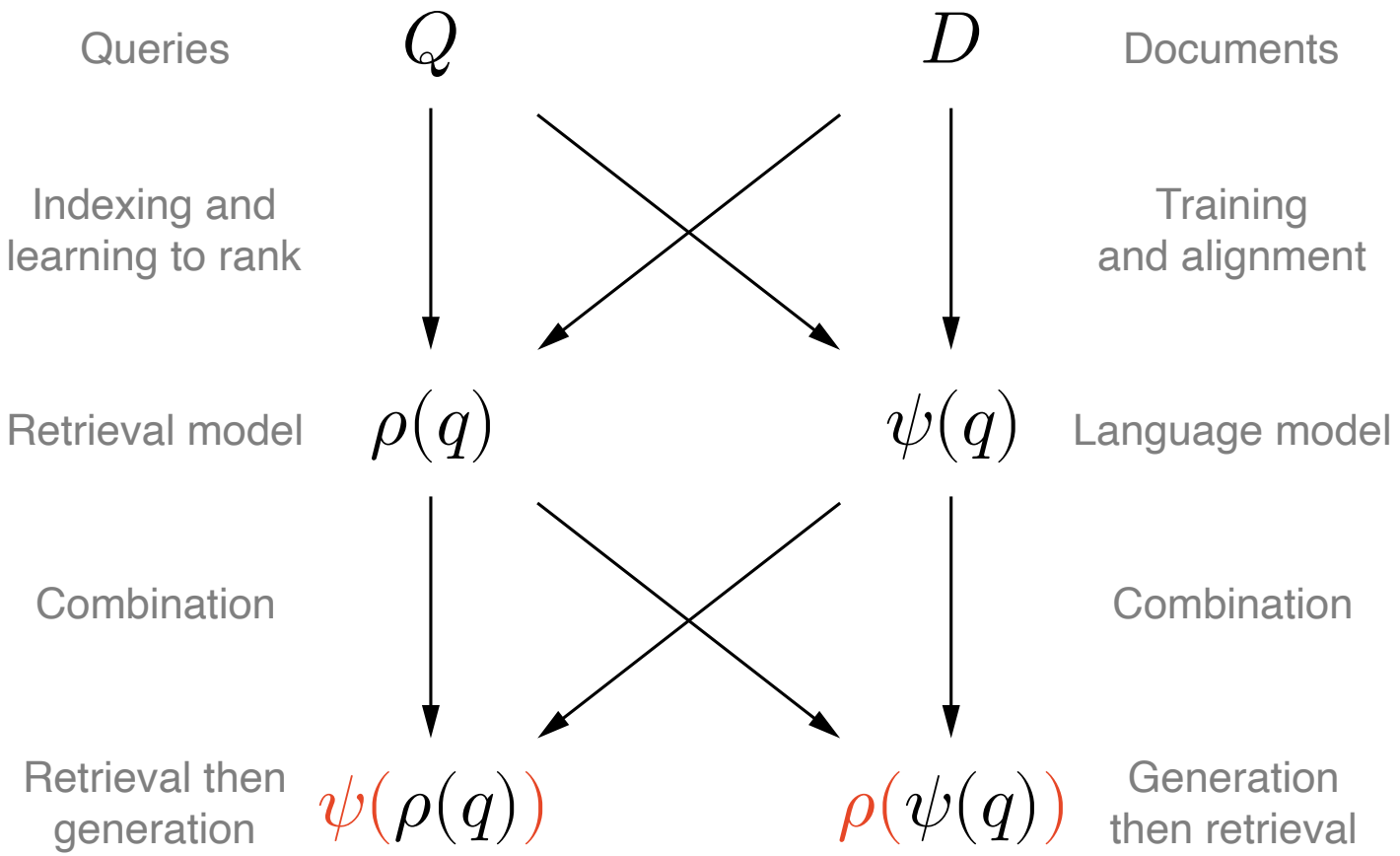
A Short History of Language Models



A Short History of Language Models



A Short History of Language Models

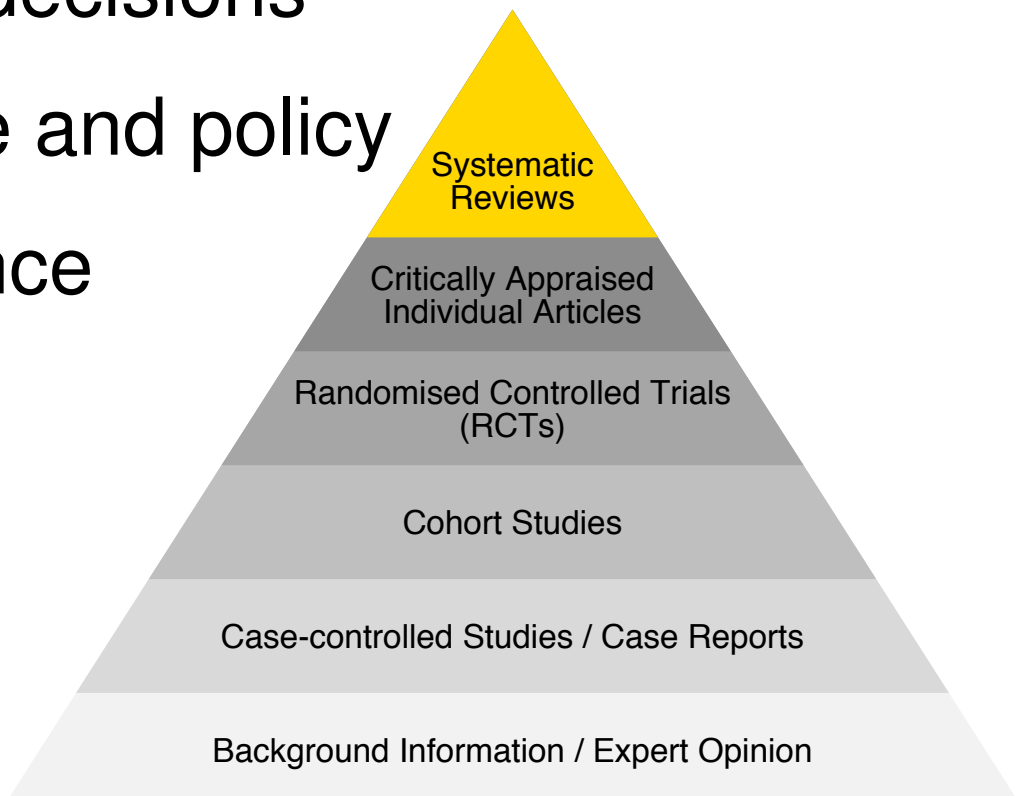


Medical Applications

Systematic Reviews

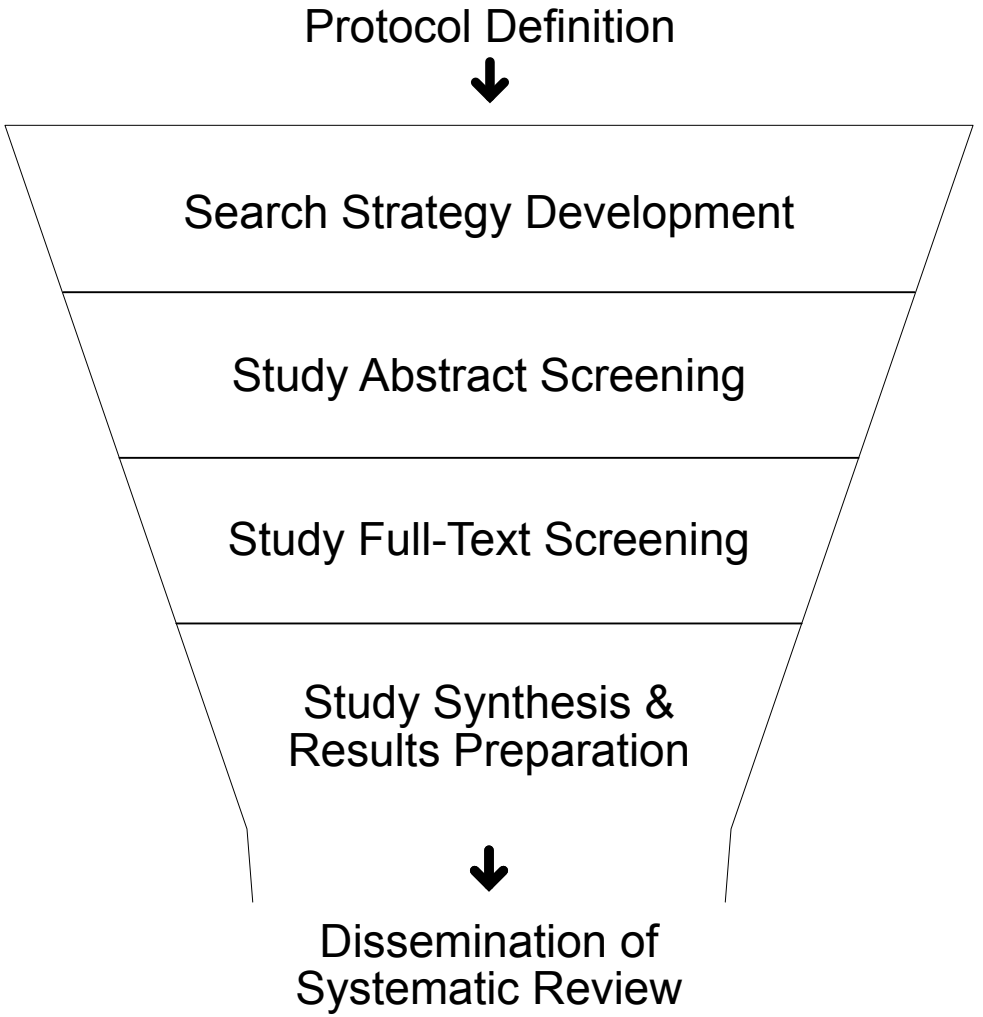
Overview

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



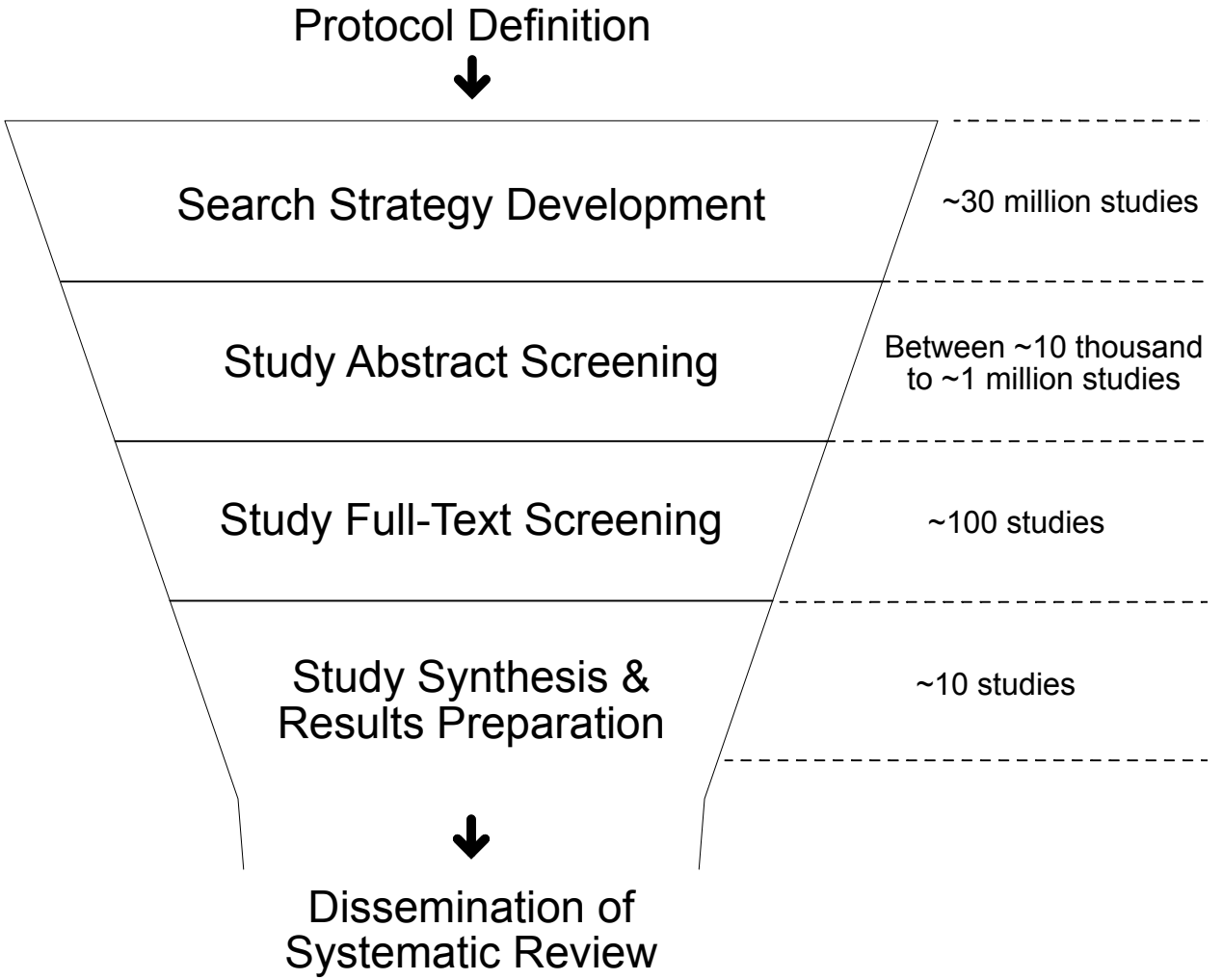
Systematic Reviews

Systematic review creation is hard!



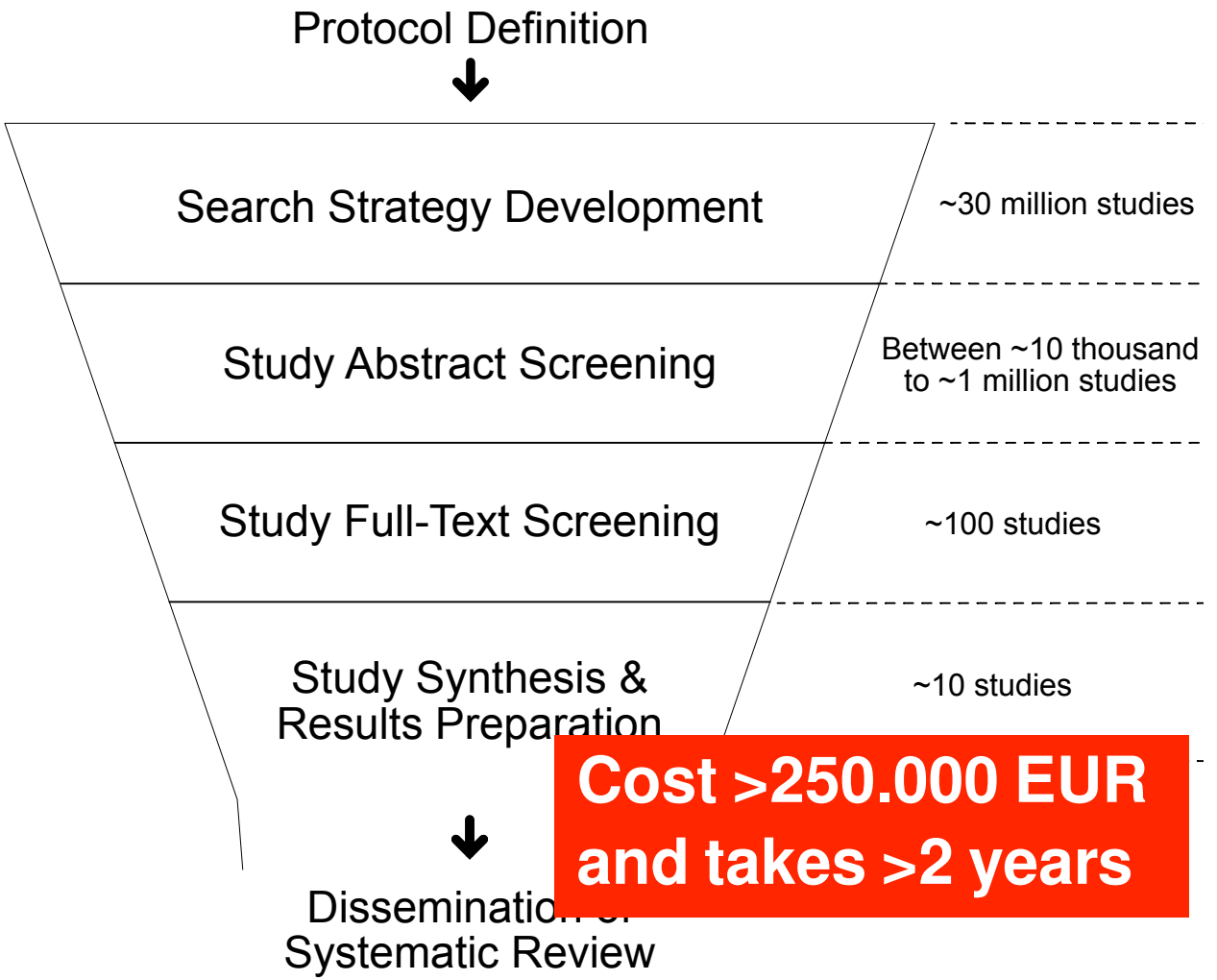
Systematic Reviews

Why is systematic review creation hard?



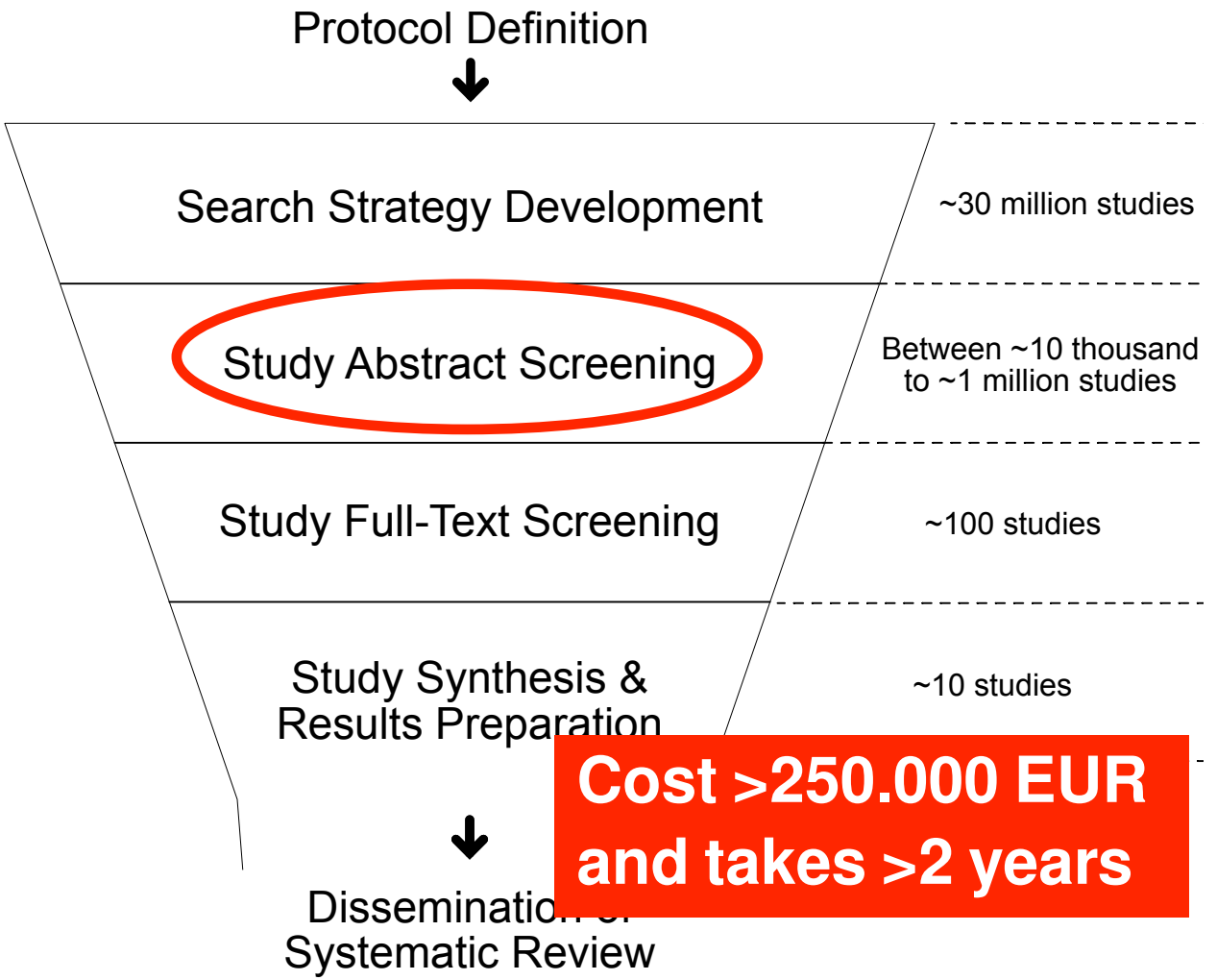
Systematic Reviews

Why is systematic review creation hard?



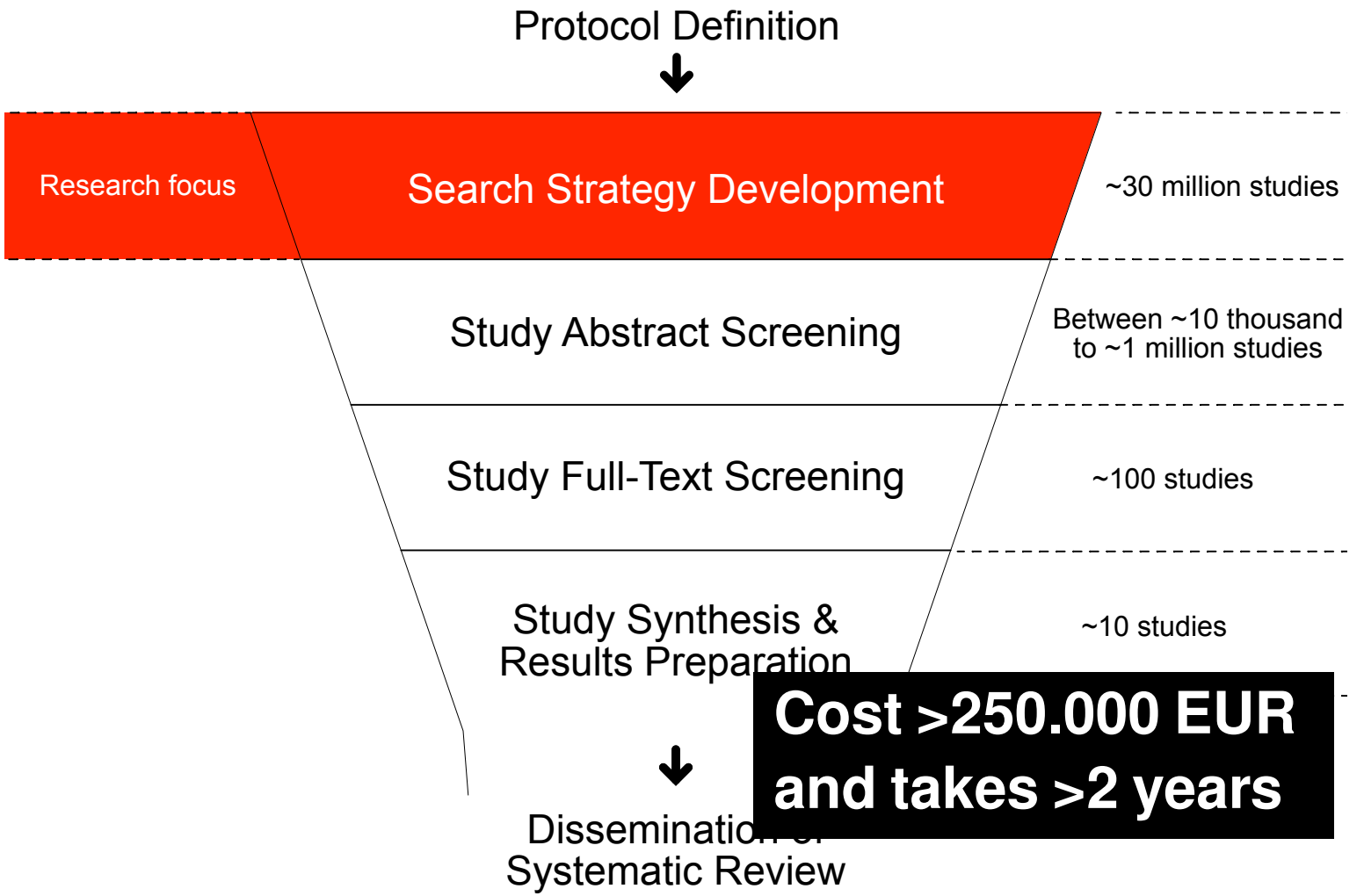
Systematic Reviews

Why is systematic review creation hard?



Systematic Reviews

Why is systematic review creation hard?



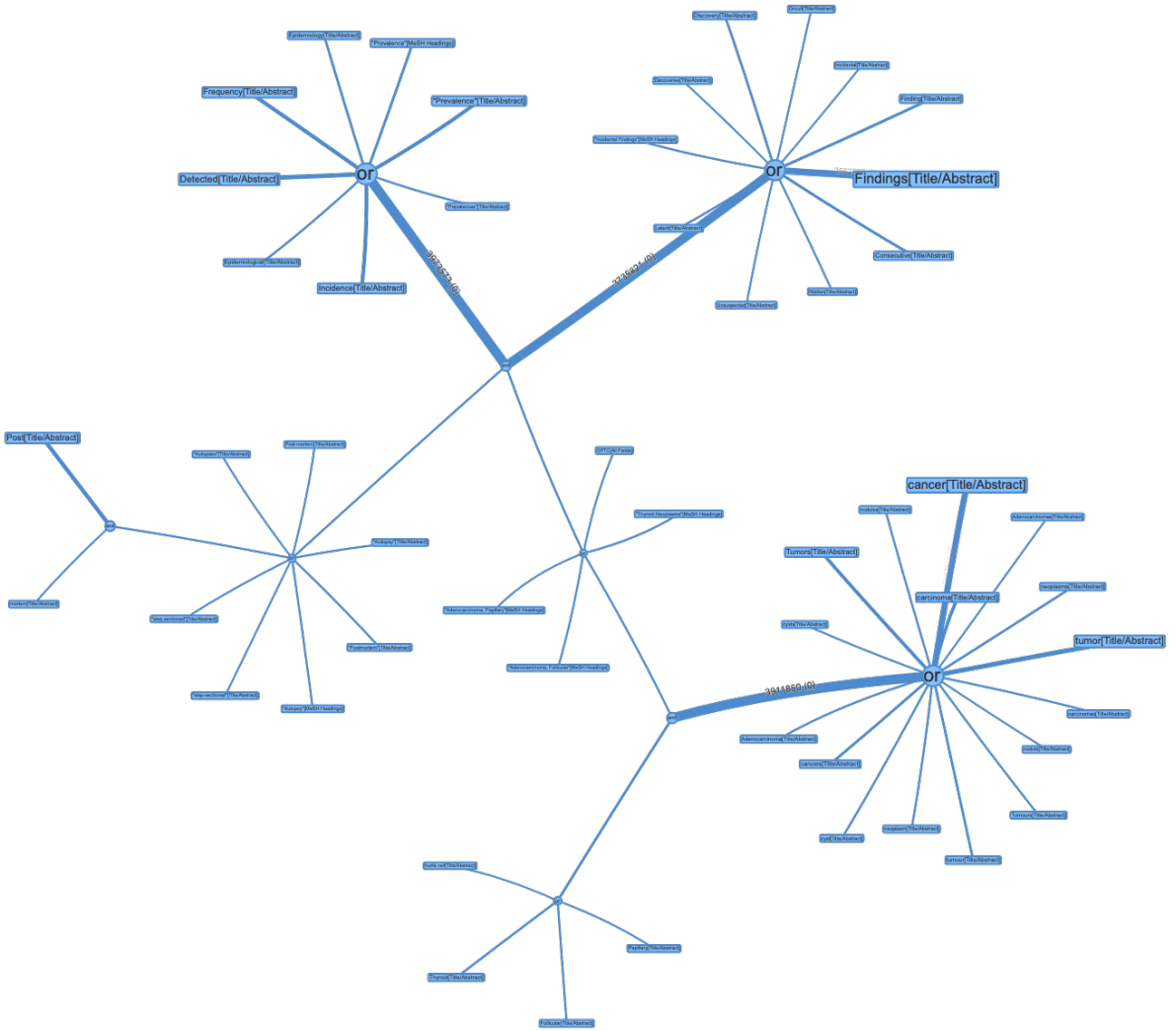
Systematic Reviews

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]))
```

Systematic Reviews

Why such little research on queries?



Systematic Reviews

Why are Boolean queries used?

Reproducibility ➜ double check screening

- Boolean query published alongside the review.
- Screening can be reproduced if outcome of review is in doubt.

Understandability ➜ control document set size

- Systematic reviews have a screening budget.
- Boolean query can be used to adjust number of studies to screen.

How humans formulate queries

Overview

Two approaches experts use to develop Boolean queries for systematic reviews:

- Conceptual method [Clark 2013] ➜ Human expertise
- Objective method [Hausner et al. 2012] ➜ More algorithmic

Both make use of seed studies.

- Studies known about before the review begins.
- Used to weakly validate effectiveness of Boolean query during development.

How humans formulate queries

Overview

Two approaches experts use to develop Boolean queries for systematic reviews:

- Conceptual method [Clark 2013] ➜ Human expertise
- Objective method [Hausner et al. 2012] ➜ More algorithmic

Both make use of seed studies.

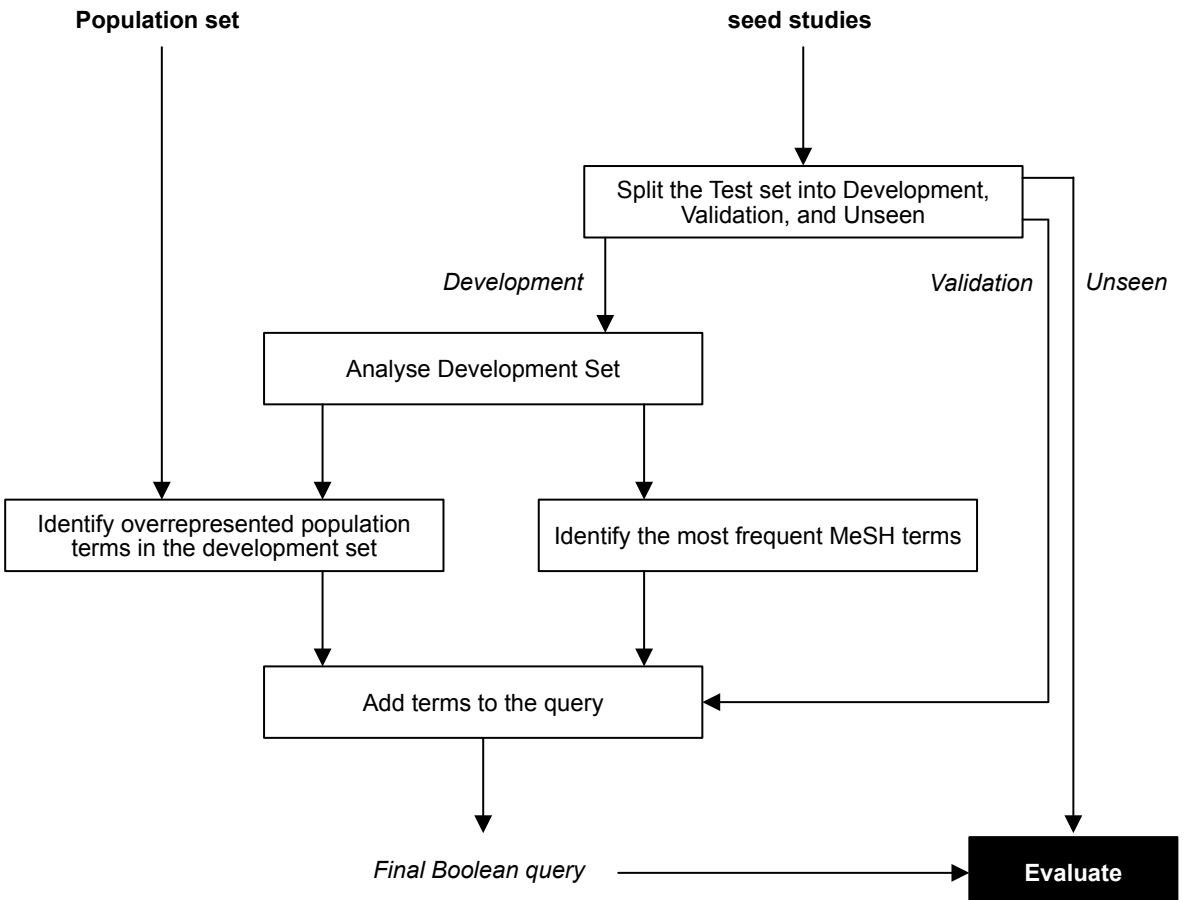
- Studies known about before the review begins.
- Used to weakly validate effectiveness of Boolean query during development.

In this talk:

1. Show a rule-based operationalisation of objective method.
2. Show how we adapted the objective method to be used in ChatGPT.
3. Compare the two approaches to human-formulated queries.

Automating the objective method

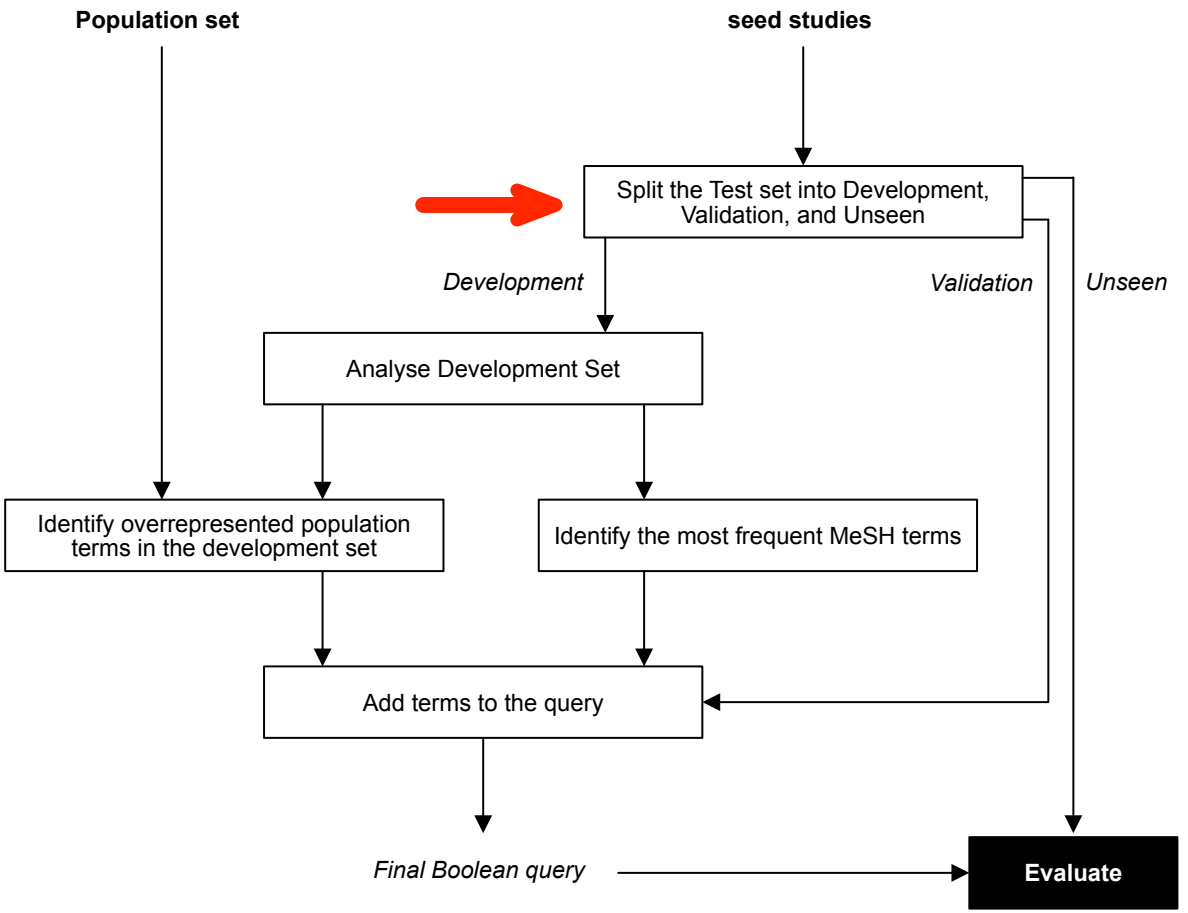
Overview



Find prominent terms from docs ➜ Add these terms to query

Automating the objective method

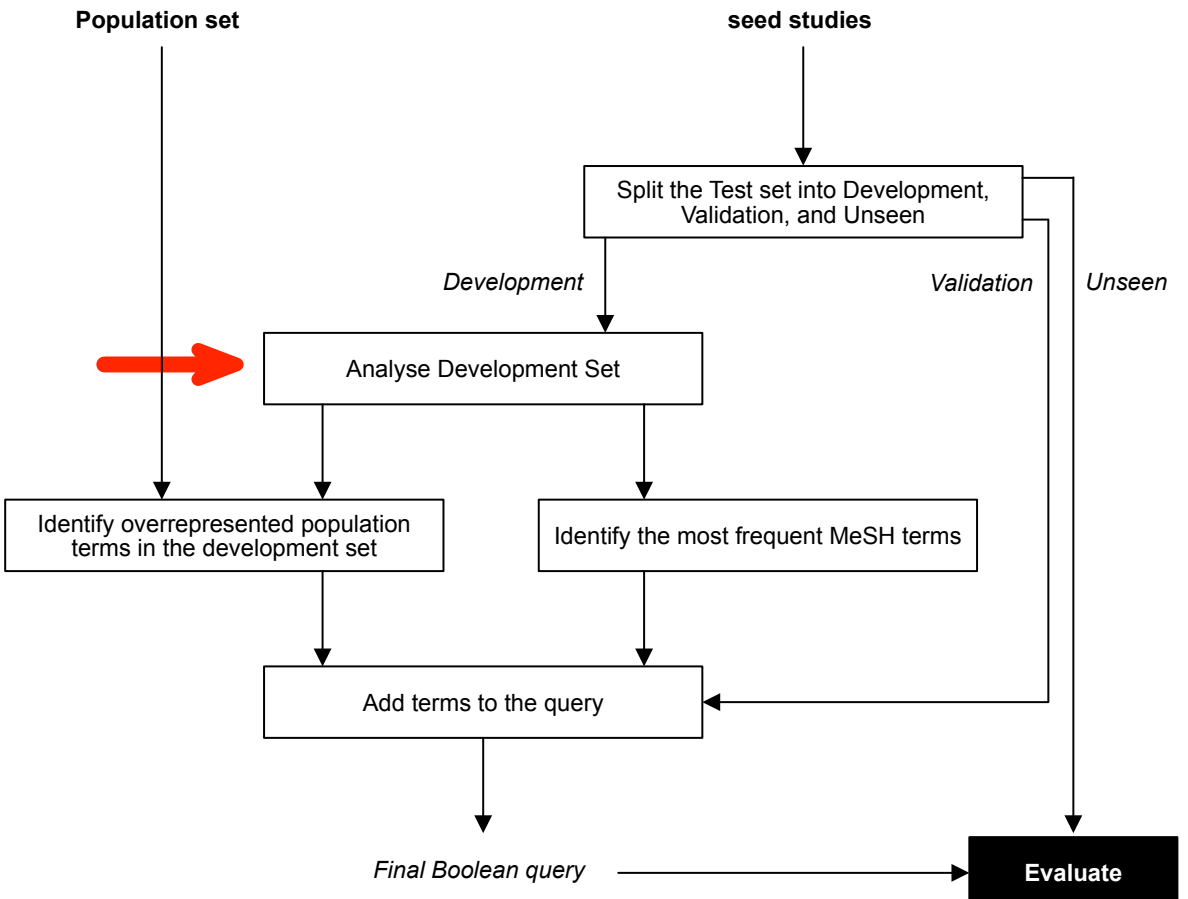
Step 1



Extract list of keywords from seed studies

Automating the objective method

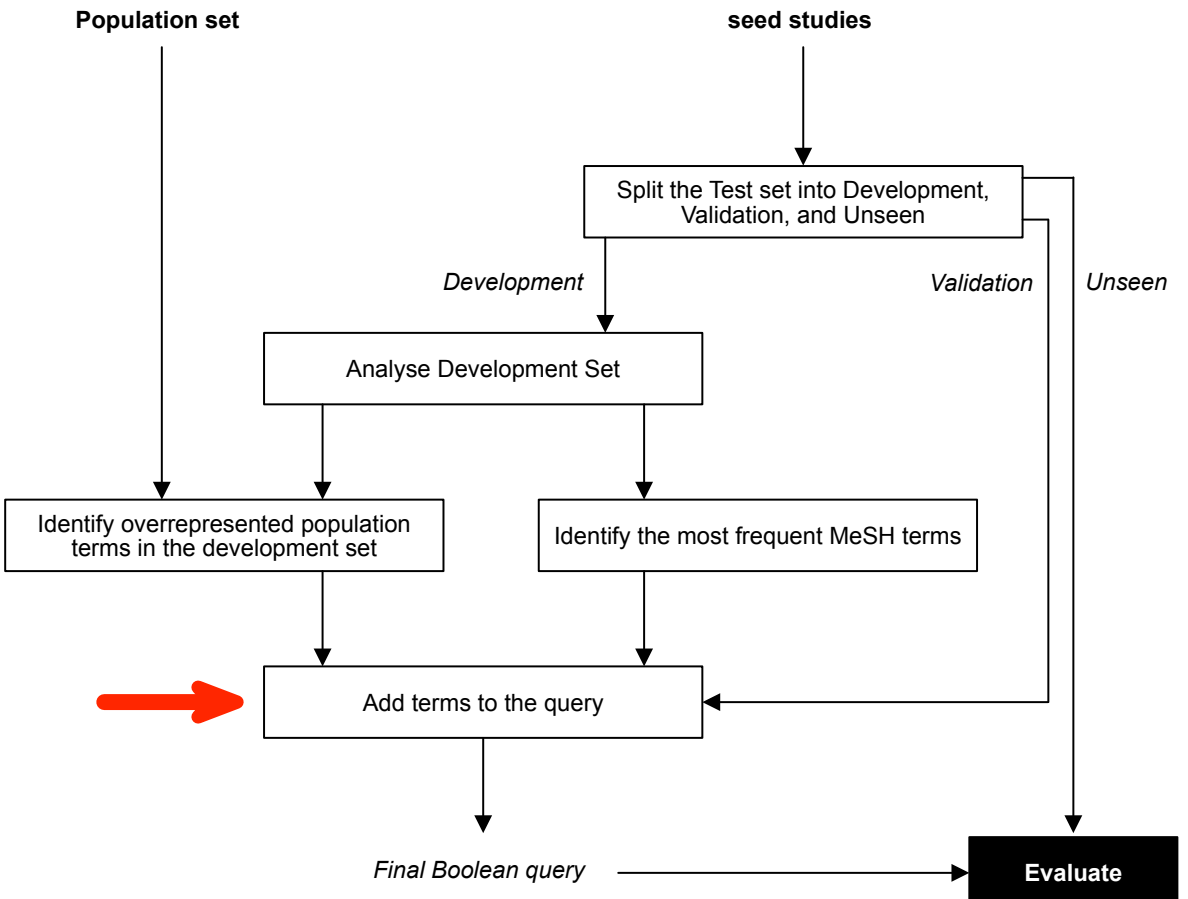
Step 2



Rank documents using term frequency

Automating the objective method

Step 3



Add keywords from documents to query

Can ChatGPT formulate effective queries?

Overview

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

Can ChatGPT formulate effective queries?

Step 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.

Can ChatGPT formulate effective queries?

Step 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.

Can ChatGPT formulate effective queries?

Step 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 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]))

Can ChatGPT formulate effective queries?

Step 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]))

Can ChatGPT formulate effective queries?

Results

	~No. Docs	Recall
Human	1000	83%
Objective (Rule-based)	500000	51%
Simple (ChatGPT)	1000	5%
Detailed (ChatGPT)	500	3%
Example (ChatGPT)	1000	12%
Objective (ChatGPT)	500	51%

Can ChatGPT formulate effective queries?

Results

	~No. Docs	Recall
Human	1000	83%
Objective (Rule-based)	500000	51%
Simple (ChatGPT)	1000	5%
Detailed (ChatGPT)	500	3%
Example (ChatGPT)	1000	12%
Objective (ChatGPT)	500	51%

ChatGPT is more effective than rule-based objective method

Can ChatGPT formulate effective queries?

Results

	~No. Docs	Recall
Human	1000	83%
Objective (Rule-based)	500000	51%
Simple (ChatGPT)	1000	5%
Detailed (ChatGPT)	500	3%
Example (ChatGPT)	1000	12%
Objective (ChatGPT)	500	51%

ChatGPT is more effective than rule-based objective method

ChatGPT is highly dependent on prompt and prone to hallucination

Current Concerns and Future Work

Problems and Guidance [Guimarães et al. 2024]

- Lack of terms synonymous with the main search terms.
  - ChatGPT adds related terms but not synonymous terms.
- Lack of clinical jargon.
  - ChatGPT adds relevant terms but few recognised technical terms.
- Lack of appropriate date restrictions.
  - ChatGPT struggles to choose appropriate dates for the search.
  - e.g., For COVID-19, the cut-off should be 2019 onwards.
- Lack of validated search filters.
  - ChatGPT does not automatically add, e.g. RCT filters, to the queries.

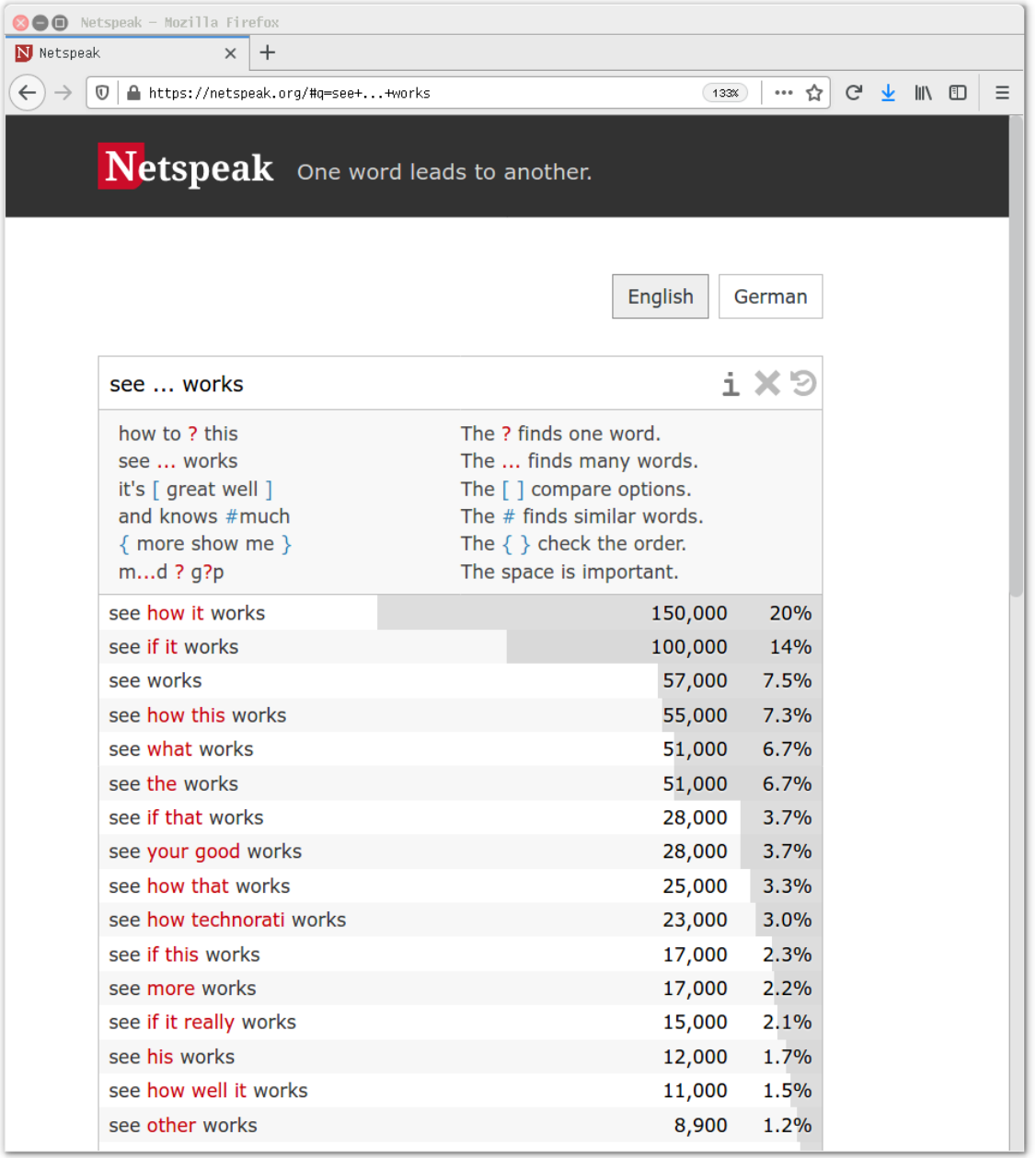
Current Concerns and Future Work

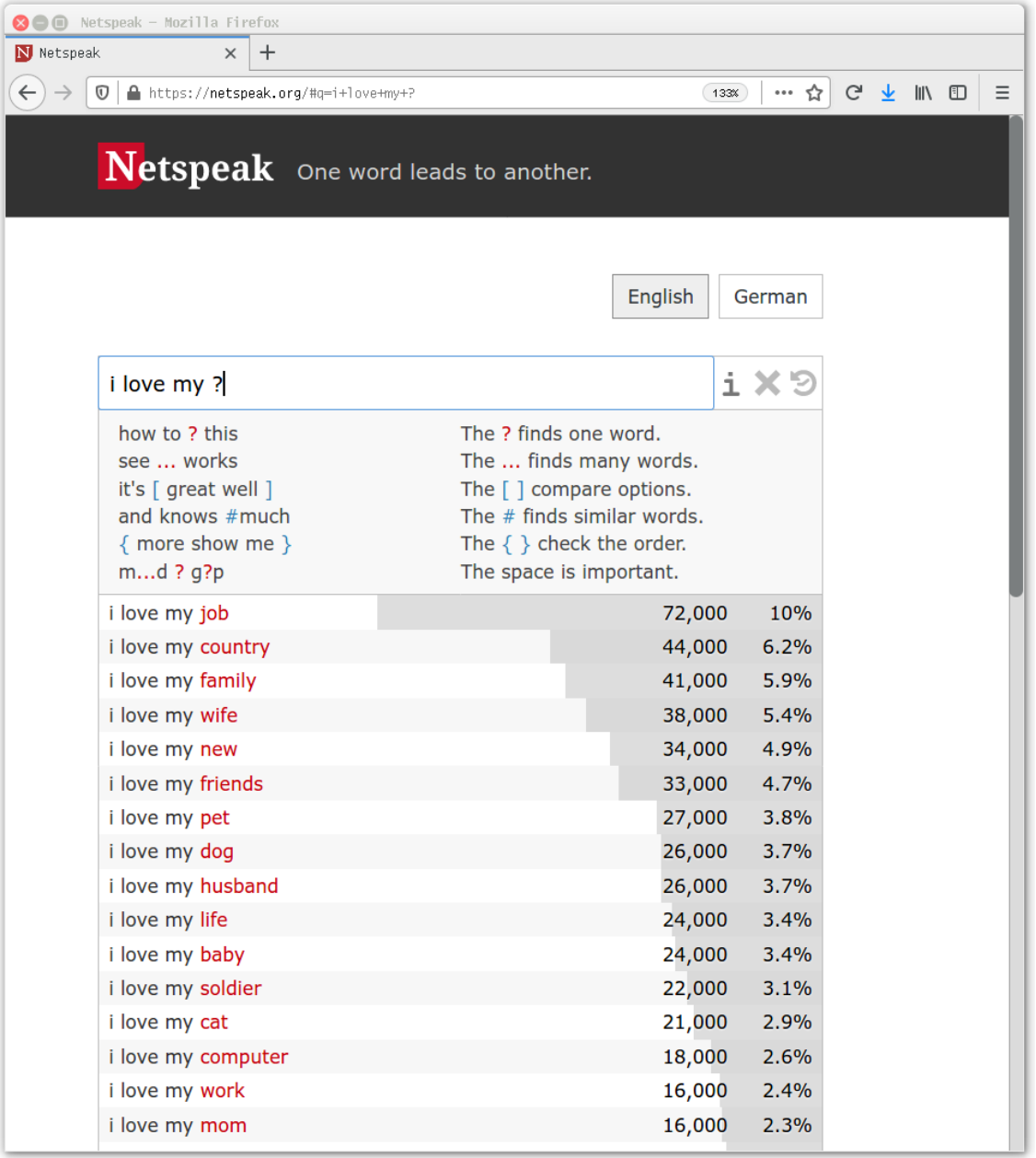
Problems and Guidance [Guimarães et al. 2024]

- Lack of terms synonymous with the main search terms.
  - ChatGPT adds related terms but not synonymous terms.
- Lack of clinical jargon.
  - ChatGPT adds relevant terms but few recognised technical terms.
- Lack of appropriate date restrictions.
  - ChatGPT struggles to choose appropriate dates for the search.
  - e.g., For COVID-19, the cut-off should be 2019 onwards.
- Lack of validated search filters.
  - ChatGPT does not automatically add, e.g. RCT filters, to the queries.

We are working on new methods that address these concerns.

- New models like ChatGPT for medical use-cases.
- Aligning these models to better perform tasks in the medical domain.


Netspeak – Mozilla Firefox
Netspeak
https://netspeak.org/#q=see+...+works
133%
Netspeak One word leads to another.
English
German
see ... works
how to ? this
see ... works
it's [great well]
and knows #much
{ more show me }
m...d ? g?p
The ? finds one word.
The ... finds many words.
The [ ] compare options.
The # finds similar words.
The { } check the order.
The space is important.
see how it works 150,000 20%
see if it works 100,000 14%
see works 57,000 7.5%
see how this works 55,000 7.3%
see what works 51,000 6.7%
see the works 51,000 6.7%
see if that works 28,000 3.7%
see your good works 28,000 3.7%
see how that works 25,000 3.3%
see how technorati works 23,000 3.0%
see if this works 17,000 2.3%
see more works 17,000 2.2%
see if it really works 15,000 2.1%
see his works 12,000 1.7%
see how well it works 11,000 1.5%
see other works 8,900 1.2%


Netspeak – Mozilla Firefox
Netspeak
https://netspeak.org/#q=i+love+my+?
133%
Netspeak One word leads to another.
English
German
i love my ?
how to ? this
see ... works
it's [great well]
and knows #much
{ more show me }
m...d ? g?p
The ? finds one word.
The ... finds many words.
The [ ] compare options.
The # finds similar words.
The { } check the order.
The space is important.
i love my job 72,000 10%
i love my country 44,000 6.2%
i love my family 41,000 5.9%
i love my wife 38,000 5.4%
i love my new 34,000 4.9%
i love my friends 33,000 4.7%
i love my pet 27,000 3.8%
i love my dog 26,000 3.7%
i love my husband 26,000 3.7%
i love my life 24,000 3.4%
i love my baby 24,000 3.4%
i love my soldier 22,000 3.1%
i love my cat 21,000 2.9%
i love my computer 18,000 2.6%
i love my work 16,000 2.4%
i love my mom 16,000 2.3%