

Webis at TREC BioGen 2024: Retrieval-Augmented Generation for Biomedical Questions

November 21, 2024



Jan Heinrich
Merker



Matthias
Hagen

Friedrich-Schiller-Universität Jena

<https://webis.de>

Webis at TREC BioGen 2024

Medical Q&A

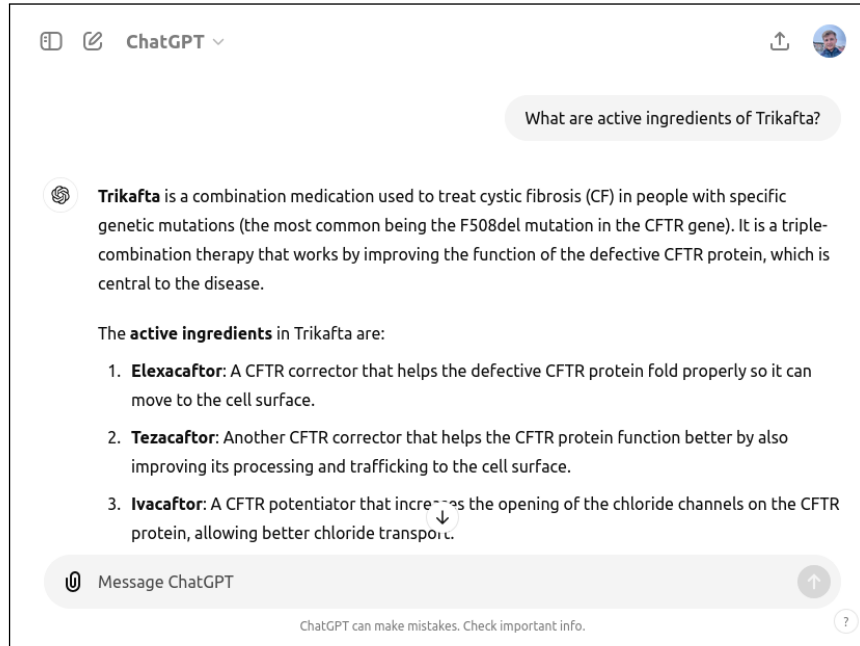
Example: What are active ingredients of Trikafta?



Webis at TREC BioGen 2024

Baselines

Example: What are active ingredients of Trikafta?



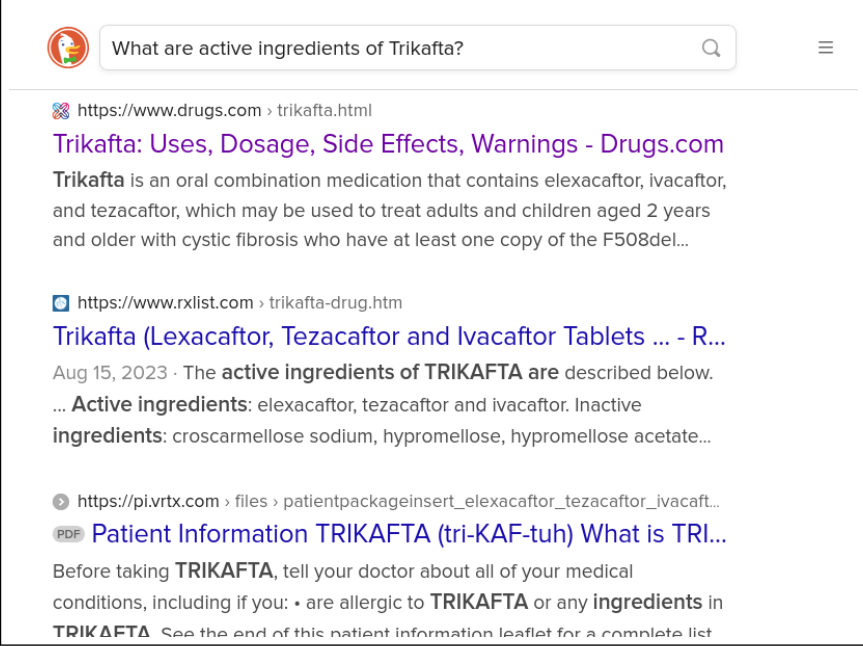
 Why not just use GPT ... ?

(correct ingredients, no dosage, no sources)

Webis at TREC BioGen 2024

Baselines

Example: What are active ingredients of Trikafta?



The screenshot shows a search engine interface with the query "What are active ingredients of Trikafta?". The search results are as follows:

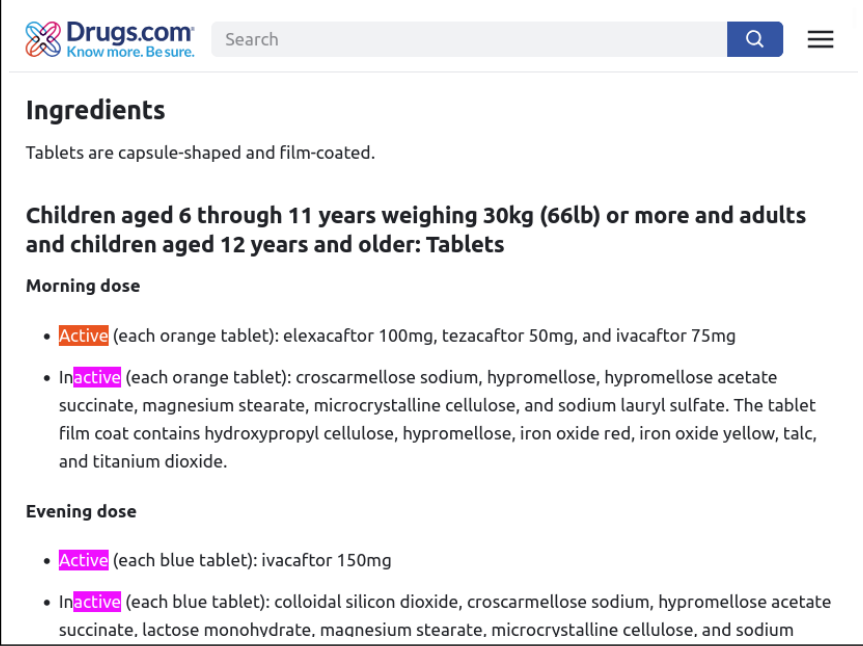
- [https://www.drugs.com > trikafta.html](https://www.drugs.com/trikafta.html)
Trikafta: Uses, Dosage, Side Effects, Warnings - Drugs.com
Trikafta is an oral combination medication that contains elexacaftor, ivacaftor, and tezacaftor, which may be used to treat adults and children aged 2 years and older with cystic fibrosis who have at least one copy of the F508del...
- [https://www.rxlist.com > trikafta-drug.htm](https://www.rxlist.com/trikafta-drug.htm)
Trikafta (Lexacaftor, Tezacaftor and Ivacaftor Tablets ... - R...
Aug 15, 2023 · The **active ingredients of TRIKAFTA** are described below.
... **Active ingredients:** elexacaftor, tezacaftor and ivacaftor. **Inactive ingredients:** croscarmellose sodium, hypromellose, hypromellose acetate...
- [https://pi.vrtx.com > files > patientpackageinsert_elexacaftor_tezacaftor_ivacaft...](https://pi.vrtx.com/files/patientpackageinsert_elexacaftor_tezacaftor_ivacaftor.pdf)
PDF Patient Information TRIKAFTA (tri-KAF-tuh) What is TRI...
Before taking **TRIKAFTA**, tell your doctor about all of your medical conditions, including if you: • are allergic to **TRIKAFTA** or any **ingredients** in **TRIKAFTA**. See the end of this patient information leaflet for a complete list

Q ... or a quick web search ... ?

Webis at TREC BioGen 2024

Baselines

Example: What are active ingredients of Trikafta?



The screenshot shows the Drugs.com website interface. At the top, there is a search bar with the Drugs.com logo and the tagline "Know more. Be sure." To the right of the search bar is a magnifying glass icon and a hamburger menu icon. Below the search bar, the page is titled "Ingredients". Underneath, it states "Tablets are capsule-shaped and film-coated." The next section is titled "Children aged 6 through 11 years weighing 30kg (66lb) or more and adults and children aged 12 years and older: Tablets". Below this, there are two sections: "Morning dose" and "Evening dose".

Ingredients

Tablets are capsule-shaped and film-coated.

Children aged 6 through 11 years weighing 30kg (66lb) or more and adults and children aged 12 years and older: Tablets

Morning dose

- **Active** (each orange tablet): elexacaftor 100mg, tezacaftor 50mg, and ivacaftor 75mg
- **Inactive** (each orange tablet): croscarmellose sodium, hypromellose, hypromellose acetate succinate, magnesium stearate, microcrystalline cellulose, and sodium lauryl sulfate. The tablet film coat contains hydroxypropyl cellulose, hypromellose, iron oxide red, iron oxide yellow, talc, and titanium dioxide.

Evening dose

- **Active** (each blue tablet): ivacaftor 150mg
- **Inactive** (each blue tablet): colloidal silicon dioxide, croscarmellose sodium, hypromellose acetate succinate, lactose monohydrate, magnesium stearate, microcrystalline cellulose, and sodium



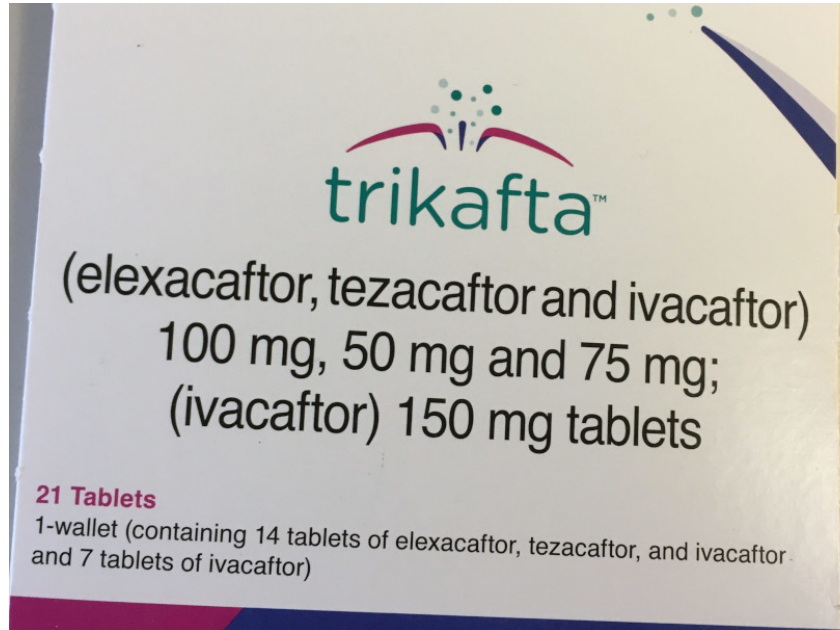
... with Ctrl+F on the first result?

(correct ingredients and dosage, good source, but takes longer)

Webis at TREC BioGen 2024

Baselines

Example: What are active ingredients of Trikafta?



☹️ ... and what about this?

(correct ingredients and dosage, good source, fastest?!)

Webis at TREC BioGen 2024

Starting Point: Medical RAG (∈ Medical Q&A?)

Can we learn from CLEF BioASQ?



- ❑ Similar QA-style evaluation setup (but no citations)
 - ❑ Training data available
 - 5046 questions until 2023
 - question, summary answer, “exact” answer, relevant docs./passages
 - ❑ Shifts towards RAG evaluation (Phase A+)
- Improve and re-evaluate our systems submitted there [Merker et al., CLEF 2024]

Webis at TREC BioGen 2024

RAG Pipeline for Medical Questions



Stages

- ❑ Document retrieval
 - Find relevant medical articles (from PubMed).
 - Extract passages, rank by relevance.
- ❑ Answer generation
 - Generate summary answer and "exact" answer.
 - Cite retrieved abstracts/passages.
- ❑ RAG / augmentation patterns
 - Combine retrieval- and generation-focused components.

Webis at TREC BioGen 2024

Approaches: Document Retrieval



Goal: Find relevant biomedical articles (from PubMed).

- ❑ Custom BM25 index with metadata + full text (Elasticsearch)
- ❑ Match abstract, title, MeSH terms, and/or previous answer (for GAR)
- ❑ Disallow articles with empty title/abstract (optional)
- ❑ Disallow non-peer-reviewed publication types (optional)

→ Is BM25 still any good after 30 years?

Webis at TREC BioGen 2024

Approaches: Passage Extraction and Re-Ranking



Goal: Extract concise passages from the article's abstract (or title).
Rank extracted passages by relevance to the question.

- ❑ Rule-based: Split abstract in sentences
- ❑ Passages: full title + sentence n -grams (1–3 sentences) from abstract
- ❑ Re-rank pointwise (monoT5, TAS-B, ANCE, or TCT-CoBERT; optional)
- ❑ Re-rank pairwise (duoT5; optional)

→ Which re-ranker to get “good” contexts?

Webis at TREC BioGen 2024

Approaches: Answer Generation with LLMs



Goal: Generate summary answer (and exact, e.g., yes-no, answer), and cite.

- ❑ Modular LLM “programming” with DSPy
- ❑ LLM: Mistral-7B-Instruct-v0.3 or GPT-4o mini
- ❑ Context: top- k passages (IDs re-mapped), previous answer (for refinement)
- ❑ Manual prompt “headers” for summary/exact answer generation
- ❑ Automatic prompts (direct/CoT) and few-shot examples via DSPy

→ Do we need manual prompting?

Webis at TREC BioGen 2024

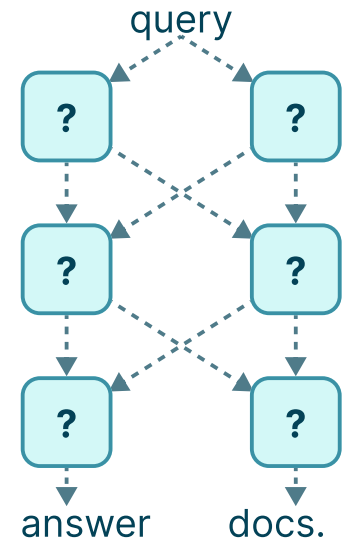
Approaches: Augmentation Patterns



Goal: Combine retrieval and generation components to augment each other.

- ❑ No augmentation: retrieve/generate separately
- ❑ RAG: Retrieve to “guide” answer generation
- ❑ GAR: Retrieve to “verify” generated answer
- ❑ Recurrent: “guide” + “verify”
- ❑ Either independent augment. or “cross-augment”

→ Does “recurrent” augmentation help?



Webis at TREC BioGen 2024

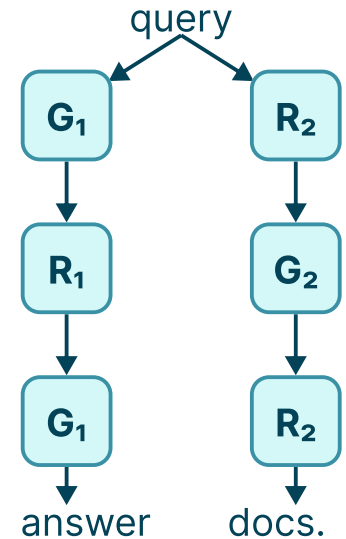
Approaches: Augmentation Patterns



Goal: Combine retrieval and generation components to augment each other.

- ❑ No augmentation: retrieve/generate separately
- ❑ RAG: Retrieve to “guide” answer generation
- ❑ GAR: Retrieve to “verify” generated answer
- ❑ Recurrent: “guide” + “verify”
- ❑ Either independent augment. or “cross-augment”

→ Does “recurrent” augmentation help?



Webis at TREC BioGen 2024

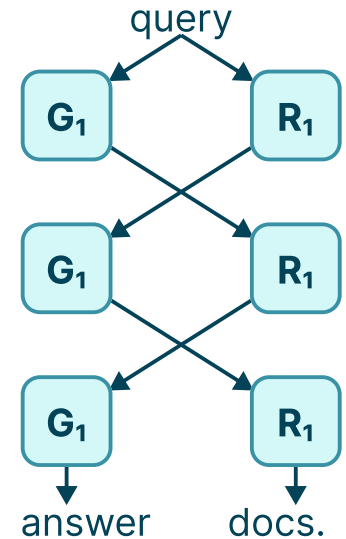
Approaches: Augmentation Patterns



Goal: Combine retrieval and generation components to augment each other.

- ❑ No augmentation: retrieve/generate separately
- ❑ RAG: Retrieve to “guide” answer generation
- ❑ GAR: Retrieve to “verify” generated answer
- ❑ Recurrent: “guide” + “verify”
- ❑ Either independent augment. or “cross-augment”

→ Does “recurrent” augmentation help?



Webis at TREC BioGen 2024

Submitted Runs

- Tune hyperparameters with Optuna and DSPy
 - Training data: 10 random Q&A pairs from BioASQ 2024
 - Objectives: Recall@1000, nDCG, ROUGE-1, ROUGE-L,
 1. 100 initial trials per LLM (GPT-4o mini and Mistral-7B)
 2. Top-10 trials per LLM re-evaluated with two additional objectives: faithfulness, answer relevance (RAGAS)
 3. Up to 5 best parameter choices per LLM used for TREC submission

- Semi-automatic post-processing to fix sentence splitting

- 7 submitted runs (4 Mistral-based, 3 GPT-based)

Webis at TREC BioGen 2024

Results

- ❑ LLMs: GPT-4o mini > Mistral-7B
- ❑ Context tends to “confuse” GPT-4o (highest accuracy w/o retrieval)
- ❑ Mistral often misses important aspects of answer
- ❑ No clear winner of the augmentation patterns
- ❑ Often bad recall (both answer recall and document recall)
 - Caveat: Only referenced documents considered

Webis at TREC BioGen 2024

Summary

- ❑ Mixed results due to relying on “random” parameter choices from Optuna
- ❑ Limitations:
 - No common baseline
 - Retrieval evaluation not isolated
- ❑ Future work: More comprehensive parameter tuning
- ❑ Question: Re-use extracted passages for nugget-based evaluation?

Code and Data

🔗 github.com/webis-de/TREC-24



Webis at TREC BioGen 2024

Summary

- ❑ Mixed results due to relying on “random” parameter choices from Optuna
- ❑ Limitations:
 - No common baseline
 - Retrieval evaluation not isolated
- ❑ Future work: More comprehensive parameter tuning
- ❑ Question: Re-use extracted passages for nugget-based evaluation?

Code and Data

🔗 github.com/webis-de/TREC-24



Thank you!

(... see you at CLEF BioASQ and TREC BioGen 2025)

Webis at TREC RAG 2024: Recall-Oriented Retrieval + Manual RAG

TREC 2024, 18–22 November 2024, Gaithersburg, USA

TODO: Author order, at the moment alphabetical/all equal contribution



**Maik
Fröbe**¹



**Lukas
Gienapp**²



**Harry
Scells**²



**Matti
Wiegmann**³



**Martin
Potthast**³



**Matthias
Hagen**¹



¹
FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA



²
UNIVERSITÄT
LEIPZIG

³
Bauhaus-
Universität
Weimar

webis.de

Webis at TREC RAG 2024

First-Stage Retrieval with ChatNoir

TODO: Shortly visit web page

Webis at TREC RAG 2024

First-Stage Retrieval with ChatNoir: A Short History

2012:

ChatNoir: A Search Engine for the ClueWeb09 Corpus

Martin Potthast

Matthias Hagen

Benno Stein

Jan Graßegger

Maximilian Michel

Martin Tippmann

Clement Welsch

- ❑ Needed and used for multiple PAN workshops (TODO: shortly visit web page)
- ❑ Developed from scratch, i.e., everything developed from scratch
- ❑ Difficult to scale, but achieved its goal to support paraphrasing research

Webis at TREC RAG 2024

First-Stage Retrieval with ChatNoir: A Short History

2018:

Elastic ChatNoir: Search Engine for the ClueWeb and the Common Crawl

Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast

- ❑ Replace all custom development with standard tools
- ❑ Elasticsearch (128 node cluster) for retrieval + snippets
- ❑ 3 Petabyte HDFS for random document access
- ❑ Mainly used for argument retrieval tasks at CLEF
- ❑ Better scalability, hosts all ClueWebs, and two CommonCrawls

Webis at TREC RAG 2024

First-Stage Retrieval with ChatNoir: A Short History

2025: (Demo for ECIR planned)



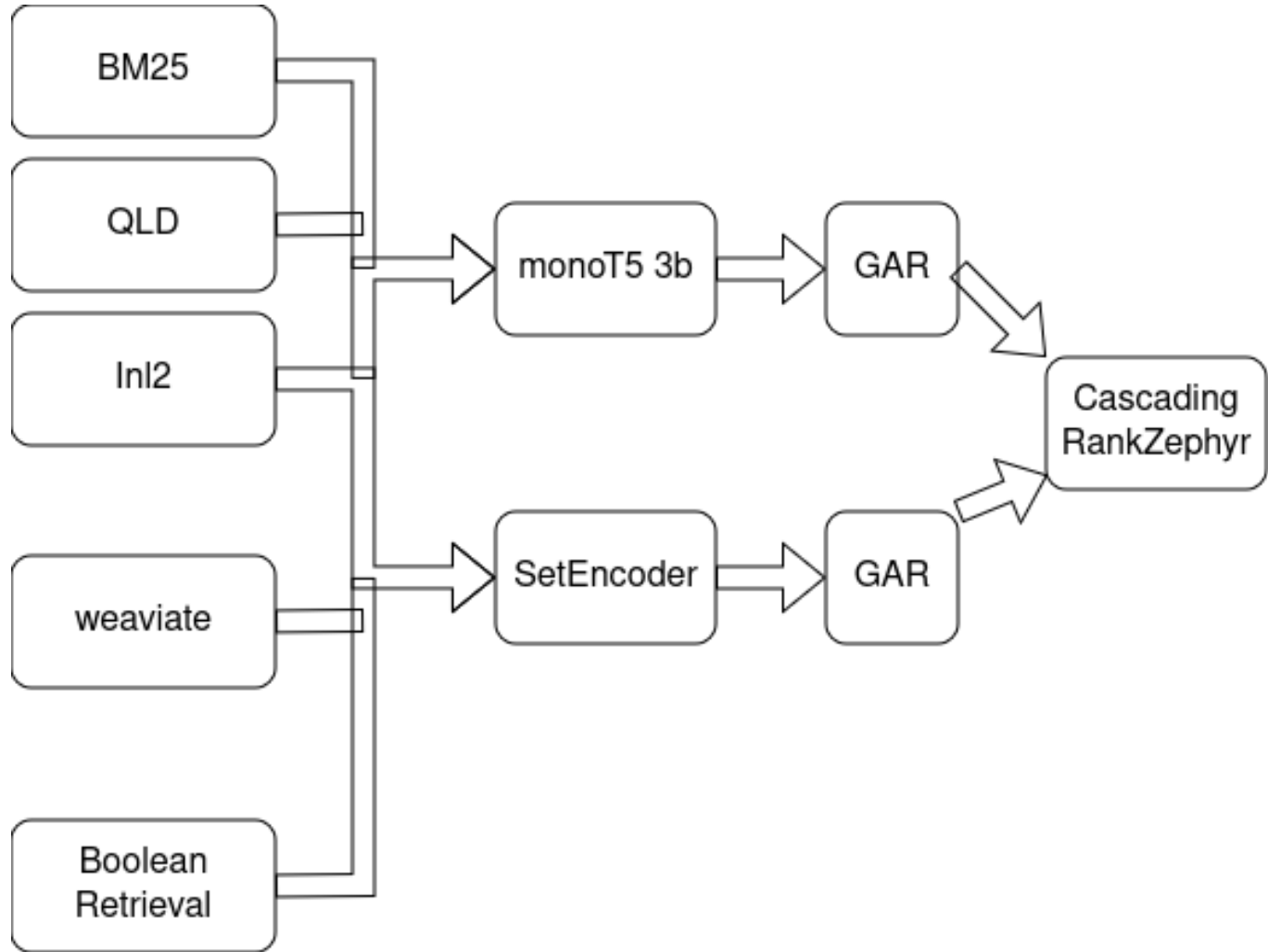
chatnoir-pyterrier

Use the ChatNoir REST-API in PyTerrier for retrieval/re-ranking against large corpora such as ClueWeb09, ClueWeb12, ClueWeb22, or MS MARCO.

- ❑ Goal: very fast entry level experimentation
- ❑ ir_datasets integration
- ❑ 12 Petabyte Ceph (100 nodes) for random document access
- ❑ Ceph much better scalable than HDFS

Webis at TREC RAG 2024

Overview Automatic Retrieval Pipeline



Webis at TREC RAG 2024

Overview Automatic Retrieval Pipeline: Boolean Retrieval

Motivation:

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

Harrison 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

Webis at TREC RAG 2024

Overview Automatic Retrieval Pipeline: Boolean Retrieval

Motivation:

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

Harrison 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

Implementation (with ChatGPT-4o + Llama3.1 70b + Mixtral):

- ❑ Step 1: Generate aspects of a query
- ❑ Step 2: Formulate boolean Queries for each aspect

Webis at TREC RAG 2024

Overview Automatic Retrieval Pipeline: Boolean Retrieval

Motivation:

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

Harrison 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

Implementation (with ChatGPT-4o + Llama3.1 70b + Mixtral):

- ❑ Step 1: Generate aspects of a query
- ❑ Step 2: Formulate boolean Queries for each aspect

Example: how does religion show in public school

- ❑ Aspects:
 - Religion might impact school holidays
 - Schools teach about various religions

Webis at TREC RAG 2024

Overview Automatic Retrieval Pipeline: Boolean Retrieval

Motivation:

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

Harrison 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

Implementation (with ChatGPT-4o + Llama3.1 70b + Mixtral):

- ❑ Step 1: Generate aspects of a query
- ❑ Step 2: Formulate boolean Queries for each aspect

Example: how does religion show in public school

- ❑ Aspects:
 - Religion might impact school holidays
 - Schools teach about various religions
- ❑ Derived boolean queries:
 - Religion AND ('school holiday' OR 'day off') AND (Christmas OR ...)
 - ...

Webis at TREC RAG 2024

Overview Automatic Retrieval Pipeline: Graph Adaptive Re-Ranking

- ❑ Take top-15 documents of monoT5/set-encoder
- ❑ Submit them as queries to ChatNoir to find similar documents
- ❑ Re-score them

Webis at TREC RAG 2024

Overview of Submitted Augmented Generations Task-RAG

- ❑ Divide-and-conquer RAG into three tasks (gpt-4o-mini, Llama 3.1)
- ❑ Task 1: Extract
- ❑ Task 2: Combine
- ❑ Task 3: Condense

Webis at TREC RAG 2024

Overview of Submitted Augmented Generations Task-RAG

- ❑ Divide-and-conquer RAG into three tasks (gpt-4o-mini, Llama 3.1)
- ❑ Task 1: Extract
- ❑ Task 2: Combine
- ❑ Task 3: Condense

Reuse-RAG

- ❑ Stitch together sentences with references from top-retrieved passages
- ❑ Offline sentence embeddings to remove redundancy and provide structure:
 - Intro
 - Arguments/details
 - Conclusions

Webis at TREC RAG 2024

Overview of Submitted Augmented Generations Task-RAG

- ❑ Divide-and-conquer RAG into three tasks (gpt-4o-mini, Llama 3.1)
- ❑ Task 1: Extract
- ❑ Task 2: Combine
- ❑ Task 3: Condense

Reuse-RAG

- ❑ Stitch together sentences with references from top-retrieved passages
- ❑ Offline sentence embeddings to remove redundancy and provide structure:
 - Intro
 - Arguments/details
 - Conclusions

Manual RAG

Let humans formulate RAG responses

Webis at TREC RAG 2024

Manual RAG: Long Term Goal

How do humans formulate RAG responses?

- ❑ We used TREC RAG for pilot experiments
- ❑ Similar to our previous paraphrasing work:
https://webis.de/publications.html?q=reuse#potthast_2013c
- ❑ Mid-term goal: Large scale study for SIGIR

Webis at TREC RAG 2024

Manual RAG: Experience from TREC 2024

Setup

- ❑ Custom Rag Response Collector
- ❑ User interactions monitored with BigBro
- ❑ Full versioning
- ❑ Starting point: Official baseline + Task RAG
- ❑ ToDo: Shortly show this

Webis at TREC RAG 2024

Manual RAG: Experience from TREC 2024

Setup

- ❑ Custom Rag Response Collector
- ❑ User interactions monitored with BigBro
- ❑ Full versioning
- ❑ Starting point: Official baseline + Task RAG
- ❑ ToDo: Shortly show this

TL;DR experience:

- ❑ Every topic takes between one and two hours
- ❑ All found this a very valuable experience
- ❑ The more time you spend with a topic, the more flaws you find in automatic responses
 - Few minutes: yeah, automatic response looks good
 - One hour: Automatic response looks really bad

Webis at TREC RAG 2024

Manual RAG: Some (unofficial) Testimonials



Maik Fröbe 08/23/2024 9:01 AM

It is super interesting, but also very time consuming.

For example, I am now rewriting the responses for "what are some way news stations are trying to limit bias" for around 30 minutes, and I am still only half through it. My interesting observation is that both baseline systems that I use to start do not really aggregate and create a "higher level" abstraction or structure.

For example, for the query `what are some way news stations are trying to limit bias`, the response is basically just an unorganized "shopping list" of biases.

But after reading through them, I came to the conclusion that there are two classes of biases: intentional and unintentional. Now I reorganize everything, which gives a much better high level structure I think, but really takes time.

I think I can maybe spend so much eye to the details for around 5 or 10 topics xD

Webis at TREC RAG 2024

Manual RAG: Some (unofficial) Testimonials



Martin Potthast 08/25/2024 8:00 AM

I did one topic yesterday evening. I agree with you that

- the generated answers have no real "structure" to their argumentation or to how they organize the information
- they distort and conflate sources

Webis at TREC RAG 2024

Manual RAG: Some (unofficial) Testimonials



Leaves 08/23/2024 1:45 PM

I agree with Maik in that this is a very insightful exercise.

My first topic `Do Reform UK's election claims on tax, immigration and environment add up` was hard because the SERP was bad. The models tried to include the references and then misclassified 'COVID' as an environment problem. Every 2nd sentence was "The references do not say if this is good or bad".

The second one is `how does our use of electronics cause natural disasters` and the models only write about how coronal mass ejection and solar storms are bad for our electronic devices.

Webis at TREC RAG 2024

Manual RAG: Doing some topics is fast

Topic:

how does this complication contribute to the central conflict of the play?

Webis at TREC RAG 2024

Manual RAG: Doing some topics is fast

Topic:

how does this complication contribute to the central conflict of the play?

GPT-4o Response (Baseline):

The complication of Romeos banishment significantly contributes to . . .

Webis at TREC RAG 2024

Manual RAG: Doing some topics is fast

Topic:

how does this complication contribute to the central conflict of the play?

GPT-4o Response (Baseline):

The complication of Romeos banishment significantly contributes to . . .

Manual response:

Search Request

how does this complication contribute to the central conflict of the play?

Response Editor

A play may have several conflicts where a complication is a situation that intensifies a conflict [1]. The question does not fully specify which complication of which play is referred to. Please provide the name of the play and which conflict you mean.

Webis at TREC RAG 2024

Manual RAG: Some topics should be fast, but still take above an hour Topic:

how did old riano residents protest relocation?

Caveat: No documents n old riano/riano/new riano exist in the corpus.

Webis at TREC RAG 2024

Manual RAG: Some topics should be fast, but still take above an hour Topic:

how did old riano residents protest relocation?

Caveat: No documents n old riano/riano/new riano exist in the corpus.

GPT-4o Response (Task-RAG):

Old Riano residents protested their relocation by organizing a 29-day ...

Webis at TREC RAG 2024

Manual RAG: Some topics should be fast, but still take above an hour Topic:

how did old riano residents protest relocation?

Caveat: No documents n old riano/riano/new riano exist in the corpus.

GPT-4o Response (Task-RAG):

Old Riano residents protested their relocation by organizing a 29-day ...

Manual response:

```
The corpus does not provide references that describe how residents of Old Riaño did protest against their relocation. However, there are other forms of documented protests against relocation, for example, residents of Hawai protested against evictions [2].
```

```
You can consider to submit your query to a large commercial web search engine like Google or Bing, as they might find relevant references on the web.
```

Webis at TREC RAG 2024

ToDo: potentially browse more topics in the RAG response collector

Webis at TREC RAG 2024

Caveats in the Corpora

Some quotes from the task page:

- ❑ A notable concern with the original corpus (MS MARCO V2) was the presence of duplicate passages, making evaluation harder (more holes, inconsistencies, etc.)

MS MARCO V2.1 document corpus as basis:

- ❑ Cleaned with deduping
- ❑ Removed 10% of corpus as redundant

Cool! I started my PhD with near-duplicates and their effects on evaluation. So I am happy the corpus was deduped.

Webis at TREC RAG 2024

However: Sliding Window includes Huge redundancy

Quote:

The resultant process leverages a sliding window size of 10 sentences and a stride of 5 sentences to create segments of text, roughly between 500-1000 characters, making it more manageable for users and baselines

Webis at TREC RAG 2024

However: Sliding Window includes Huge redundancy

Quote:

The resultant process leverages a sliding window size of 10 sentences and a stride of 5 sentences to create segments of text, roughly between 500-1000 characters, making it more manageable for users and baselines

Assume we have a document with 3 passages: A, B, and C whereas passage B is relevant

- ❑ Results in 2 segments: AB, BC
- ❑ Redundancy: 33%
- ❑ Which of the segments should be included?
- ❑ Which aggregation, etc.