OpenSource Connections

RAGmatazz at the LiveRAG Challenge

Matthias Krüger with David Fisher

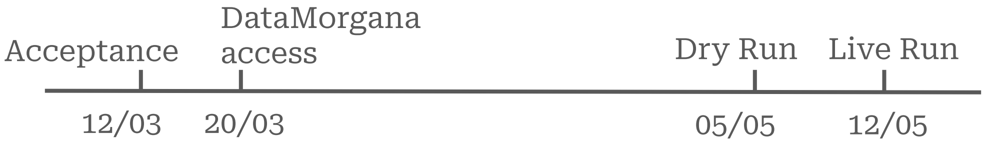
LiveRAG Challenge

- SIGIR 2025 Conference

- “RAGmatazz”

- Task: Generate answers to 500 questions grounded in a given corpus within 2 hours
  - Corpus: Fineweb-10BT
  - LLM: Falcon3-10B
  - Questions LLM generated (“DataMorgana”)
  - Judged on correctness and faithfulness

- Time



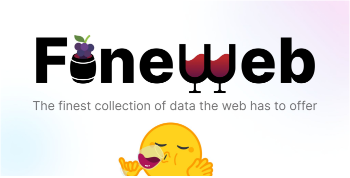
Acceptance 12/03

DataMorgana access 20/03

Dry Run 05/05

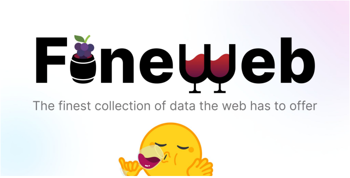
Live Run 12/05

Corpus: FineWeb-10BT (Hugging Face)



- Common Crawl data “refined”
- Web page content extracted using Trafilatura
- Heuristic documents filtering
- Random sample of 14.86M documents, almost all English
- It’s web data (examples)...

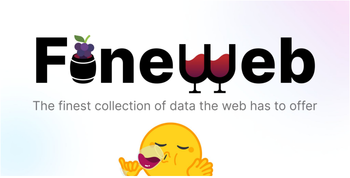
Corpus: FineWeb-10BT (Hugging Face)



The fishbowl is reimagined with this 2.5 gallon, hand-blown container with its own mountain range. This stunning conversation piece was designed by Johan Liden and Rinat Aruh of the New York concept and design group Aruliden. Not into fish? Use it as a terrarium. Or just fill it with a delicious punch and ice-shaped fish! Not surprisingly it took the Red Dot Award. Hand wash. Fish not included.

- 2.5 gallon handmade glass fishbowl
- Approximate size: 12.7" x 7.9

Corpus: FineWeb-10BT (Hugging Face)



Hey everyone! I'm here today to discuss a very important aspect of hockey off-season writing: lists.

To be more specific, this is about "best of" lists. You know the ones where writers rank the best _____ in franchise history. They are a staple this time of year. Whether it is by position, coaches, total lines or even building an all-time Mount Rushmore-style building of players, normally at least one large organization makes a month out of July and/or August creating a "best of" list for all 30 teams.

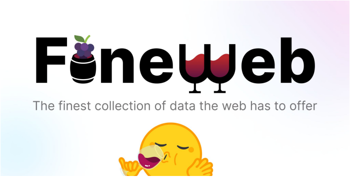
But Wild fans, can we all agree to skip out of these? Please? No one's feelings will be hurt.

Look, it is nice to feel included. Everyone likes seeing some news about their team and getting the ensuing attention. There would be some awkwardness mixed with cries of bias if only 29 teams were included for a "best of" series.

There is also awkwardness in having a "best of" list for a franchise still in its infancy. How can there be a Minnesota Wild "Mount Rushmore" if the Wild have existed for less time than it actually took to build Mount Rushmore? 13 years is enough time for George Washington to show up, but that doesn't mean John Adams and his Alien and Sedition Act deserve to be honored alongside Washington and Jefferson just because there are only three Presidents to choose.

What makes these "best of" lists worse is that every time a Wild one comes out the list gets preceded by an apology. One that's pretty half-ass yet true, like "coming up with

Corpus: FineWeb-10BT (Hugging Face)



Supreme Court Definition and Legal Meaning

On this page, you'll find the legal definition and meaning of Supreme Court, written in plain English, along with examples of how it is used.

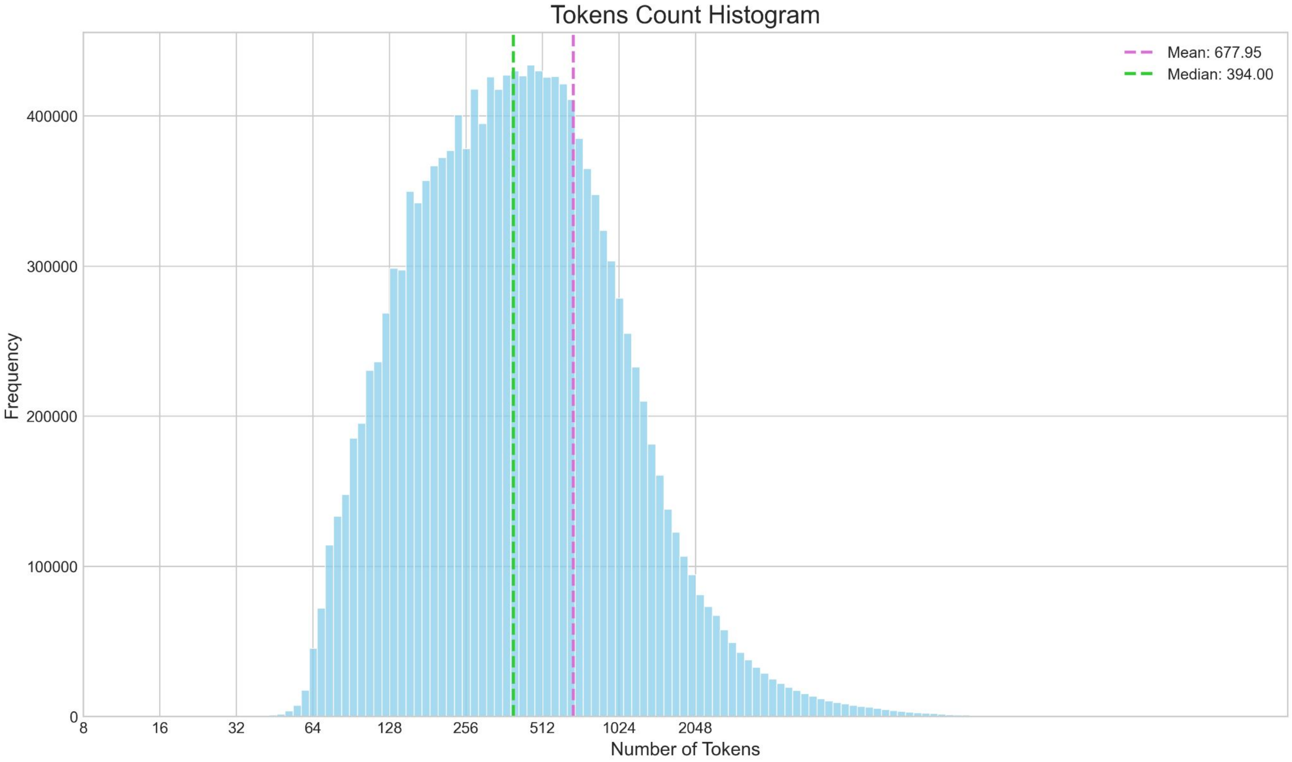
What is Supreme Court?

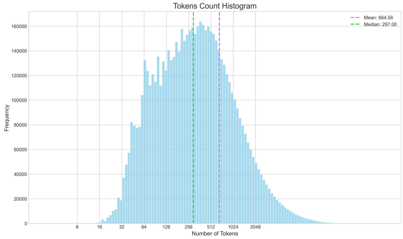
(n) Supreme court is the highest court of the land where such name is used to represent the court which has the ultimate power to determine the issues related to constitutional subjects, intra state disputes, cases in which federal or central government is a party etc are decided. Supreme court judges are appointed by the head of the state

Corpus: FineWeb-10BT (Hugging Face)

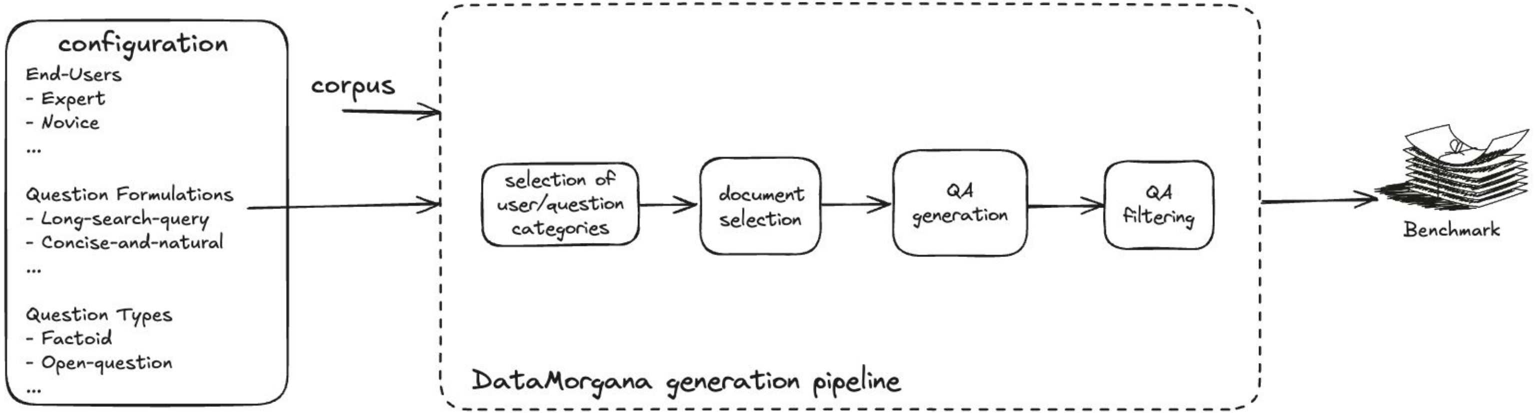
FineWeb

The finest collection of data the web has to offer





Questions: DataMorgana



- Output:
  - used configuration, document(s), question, answer

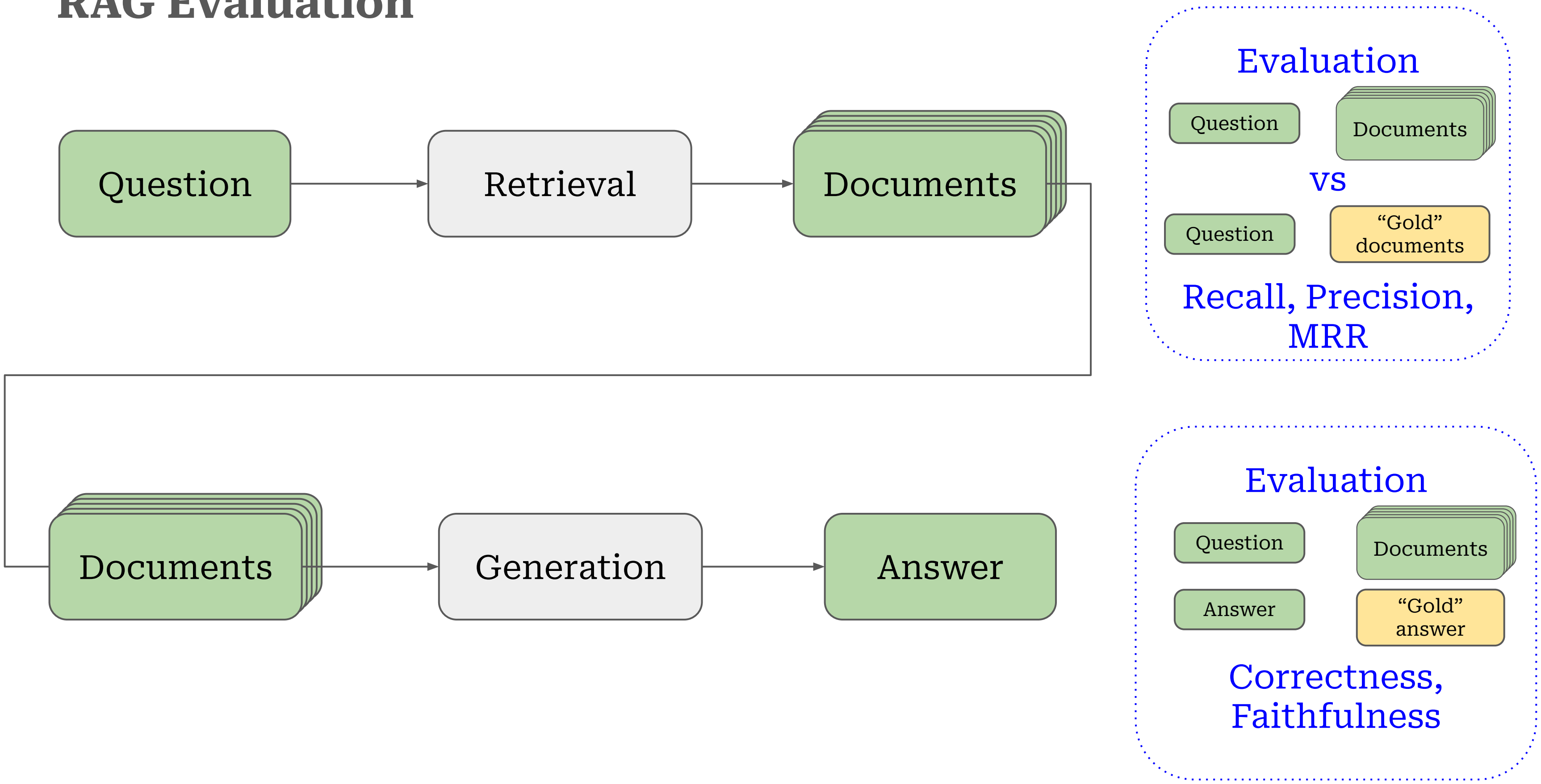
- Provided via API

Generation Model: Falcon3-10B

Technology Innovation Institute

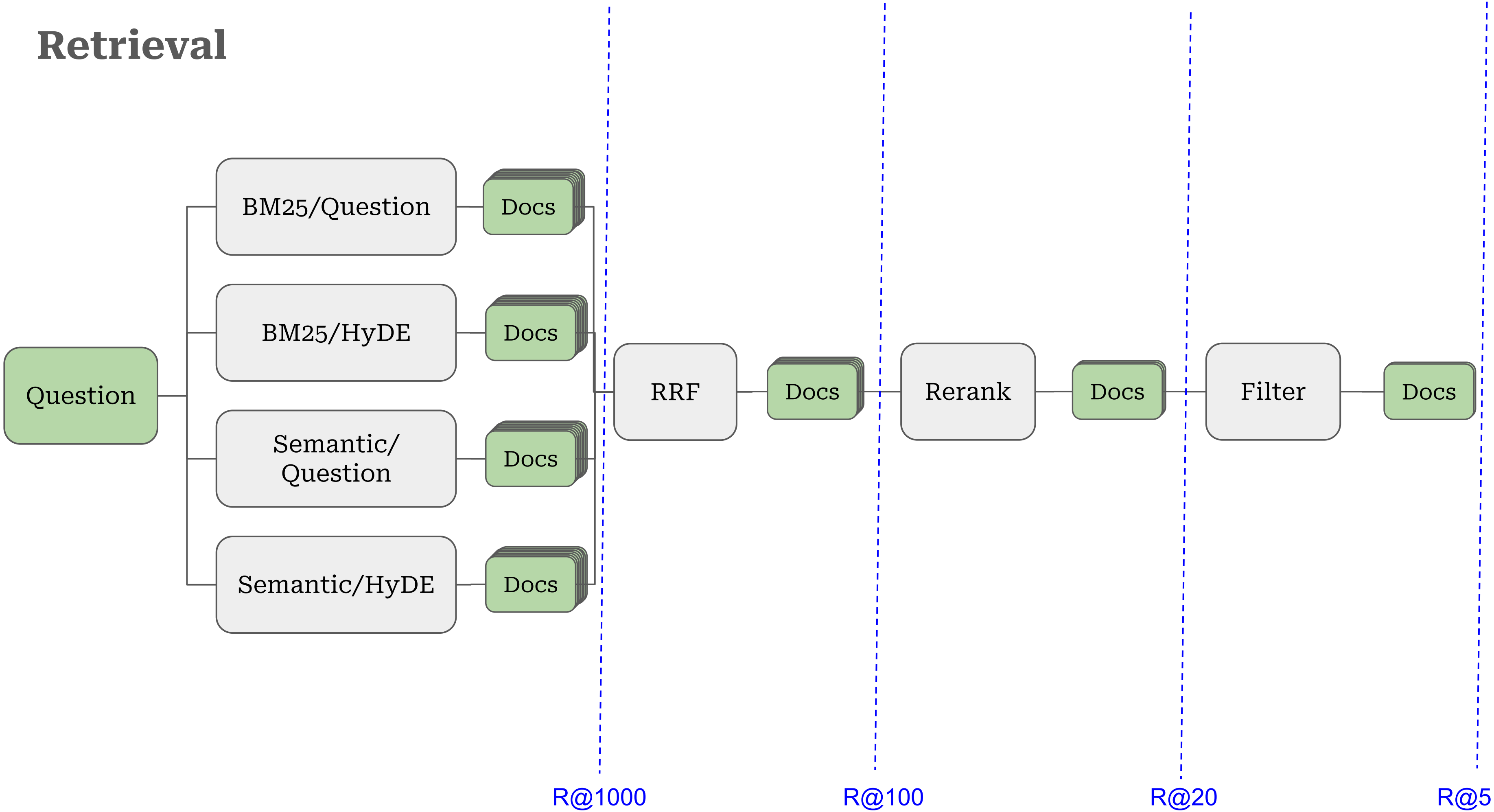
- One of the best models of its class (Dec 2024)
- Provided as API or download

RAG Evaluation



OpenSource Connections

Retrieval


Question
BM25/Question
BM25/HyDE
Semantic/
Question
Semantic/HyDE
Docs
Docs
Docs
Docs
RRF
Docs
Rerank
Docs
Filter
Docs
R@1000
R@100
R@20
R@5

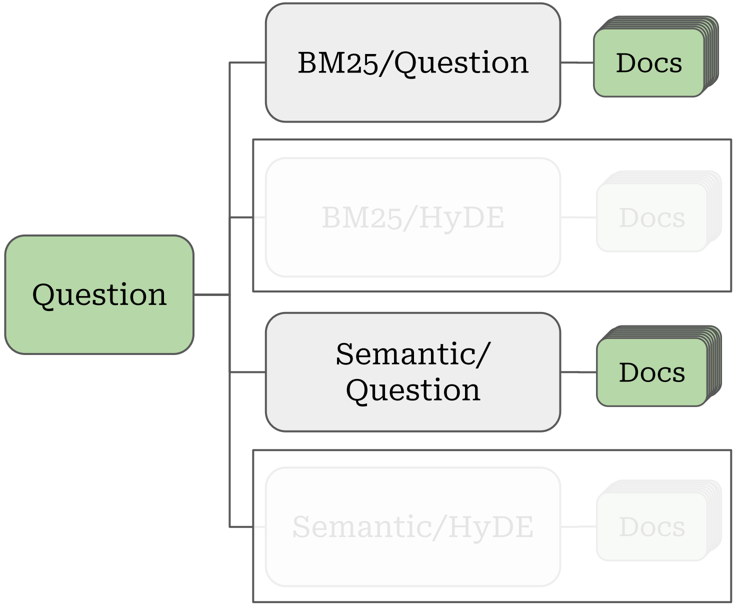
Initial Retrieval

BM25

- Index: bm25s
- Stemmer, stopword filter

Semantic

- Embedding Model Selection: MTEB Retrieval Task
- Challenge: Embeddings Creation
- Index: USearch



Initial Retrieval

Results

	R@1	R@5	R@20	R@40	R@100	R@1k
bm25	**0.216**	**0.396**	**0.546**	0.616	0.71	0.877
bge-base-en-v1.5	0.132	0.262	0.405	0.468	0.574	0.788
multilingual-e5-large-instruct	0.102	0.216	0.334	0.396	0.482	0.705
snowflake-arctic-embed-l-v2.0	0.151	0.341	0.523	**0.62**	**0.717**	**0.892**

Initial Retrieval

Embeddings Creation (pytorch / transformers)

	MacBook M1 (fp32)		TP RTXA5000 (bf16)	
	e/s	ttc	e/s	ttc
BAAI/bge-base-en-v1.5	40	103h	352	12h
intfloat/multilingual-e5-large-instruct	14	294h	115	36h
Snowflake/snowflake-arctic-embed-l-v2.0	2.8	171d	20	206h

Let’s use these AWS credits...

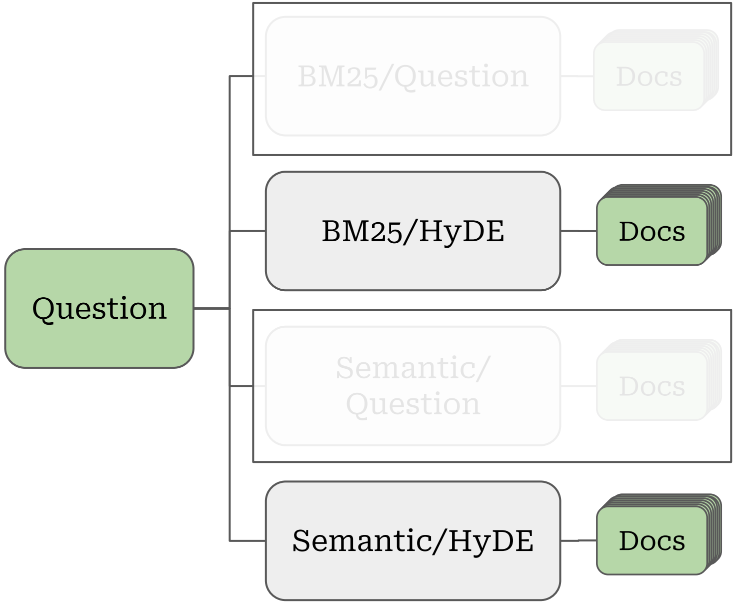
Initial Retrieval

HyDE (Hypothetical Document Embeddings)

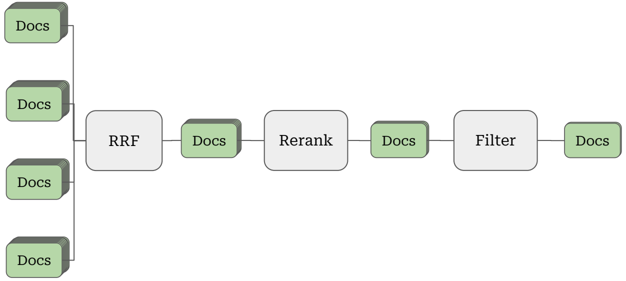
- Query for an LLM generated paragraph

~~~
You are given a topic. Produce a creative sample
paragraph from a website loosely related to that
topic. Imagine it to be a paragraph from an online
discussion forum or an online news source. Evaluate
your generated paragraph and include your reasoning
in your response. You will be rated for creativity.
Please respond with the following JSON format:
```
{
  "reasoning": "<your-reasoning>",
  "paragraph": "<your-paragraph>"
}
```

This is the topic:
{{question}}
~~~



Retrieval Refinement



Combining Results

- RRF with grid-search
  - any combination of retrievers
  - any value of k [1..100]
- R@100: 0.71 => 0.81

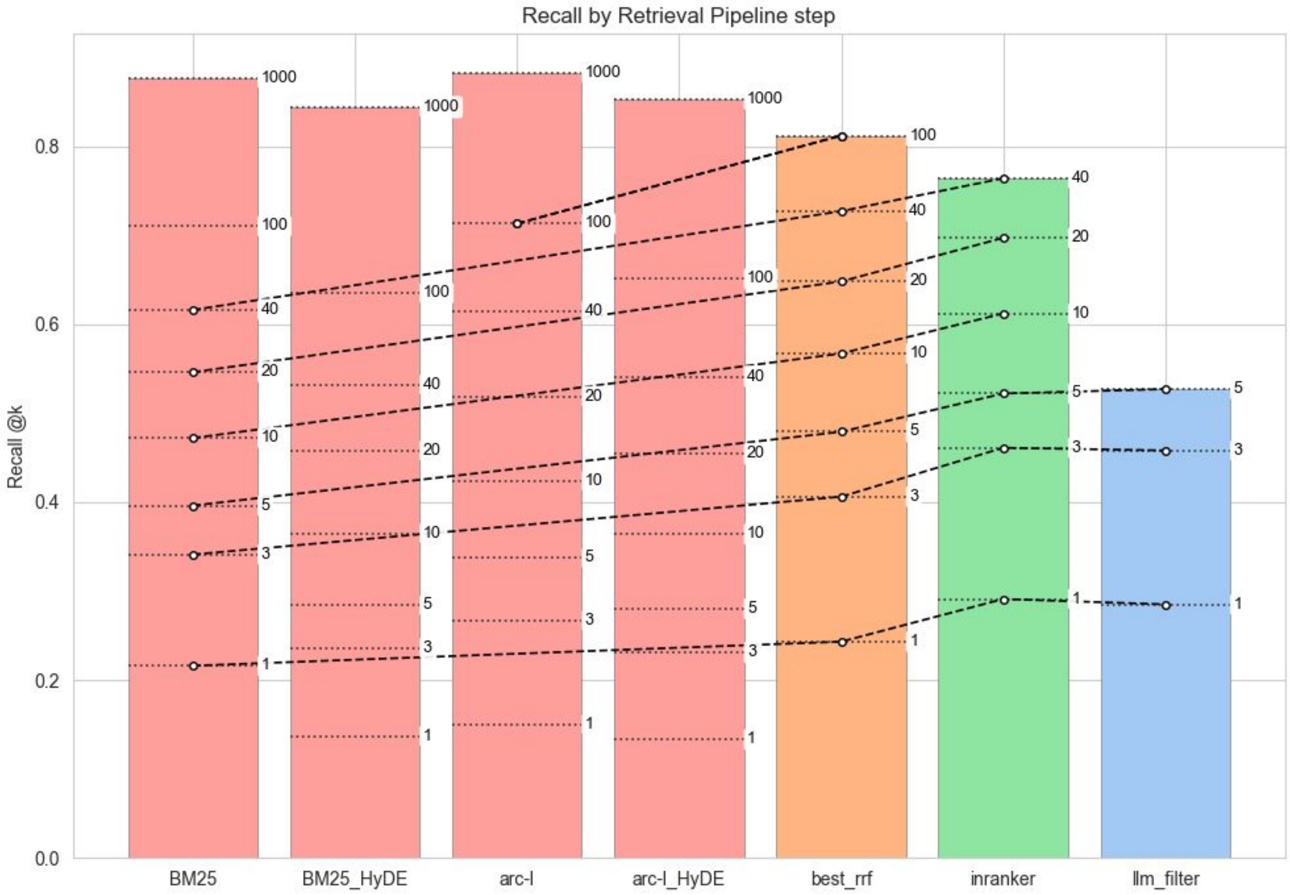
Reranker

- unicamp-dl/InRanker-base (220M)
- Simple character-level chunker with overlap
- R@20: 0.65 => 0.7

LLM based Relevance Filter

- Point-wise filtering of irrelevant documents
- No measurable positive impact

Retrieval Results



OpenSource Connections

Generation Step

- Basic prompt with documents truncated
- Request for document reference
- Fallback answer option
- 5 attempts with more docs but shorter text (5 x 8192 ... 30 x 1365)

Challenge Result

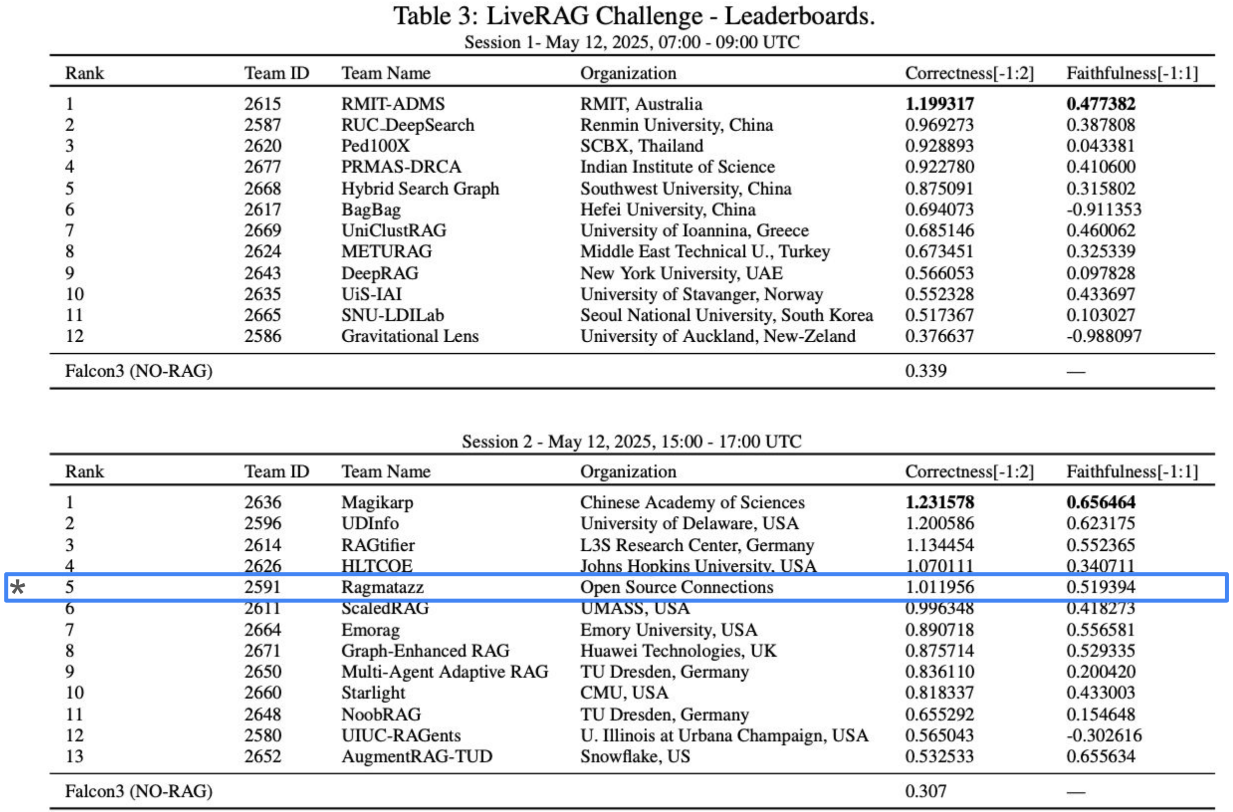


Table 3: LiveRAG Challenge - Leaderboards.

Session 1- May 12, 2025, 07:00 - 09:00 UTC

Rank	Team ID	Team Name	Organization	Correctness[-1:2]	Faithfulness[-1:1]
1	2615	RMIT-ADMS	RMIT, Australia	**1.199317**	**0.477382**
2	2587	RUC_DeepSearch	Renmin University, China	0.969273	0.387808
3	2620	Ped100X	SCBX, Thailand	0.928893	0.043381
4	2677	PRMAS-DRCA	Indian Institute of Science	0.922780	0.410600
5	2668	Hybrid Search Graph	Southwest University, China	0.875091	0.315802
6	2617	BagBag	Hefei University, China	0.694073	-0.911353
7	2669	UniClustRAG	University of Ioannina, Greece	0.685146	0.460062
8	2624	METURAG	Middle East Technical U., Turkey	0.673451	0.325339
9	2643	DeepRAG	New York University, UAE	0.566053	0.097828
10	2635	UiS-IAI	University of Stavanger, Norway	0.552328	0.433697
11	2665	SNU-LDILab	Seoul National University, South Korea	0.517367	0.103027
12	2586	Gravitational Lens	University of Auckland, New-Zeland	0.376637	-0.988097
Falcon3 (NO-RAG)				0.339	—

Session 2 - May 12, 2025, 15:00 - 17:00 UTC

Rank	Team ID	Team Name	Organization	Correctness[-1:2]	Faithfulness[-1:1]
1	2636	Magikarp	Chinese Academy of Sciences	**1.231578**	**0.656464**
2	2596	UDInfo	University of Delaware, USA	1.200586	0.623175
3	2614	RAGtifier	L3S Research Center, Germany	1.134454	0.552365
4	2626	HLTCOE	Johns Hopkins University, USA	1.070111	0.340711
* 5	2591	Ragmatazz	Open Source Connections	1.011956	0.519394
6	2611	ScaledRAG	UMASS, USA	0.996348	0.418273
7	2664	Emorag	Emory University, USA	0.890718	0.556581
8	2671	Graph-Enhanced RAG	Huawei Technologies, UK	0.875714	0.529335
9	2650	Multi-Agent Adaptive RAG	TU Dresden, Germany	0.836110	0.200420
10	2660	Starlight	CMU, USA	0.818337	0.433003
11	2648	NoobRAG	TU Dresden, Germany	0.655292	0.154648
12	2580	UIUC-RAGents	U. Illinois at Urbana Champaign, USA	0.565043	-0.302616
13	2652	AugmentRAG-TUD	Snowflake, US	0.532533	0.655634
Falcon3 (NO-RAG)				0.307	—

* 6th, 8th in human re-evaluation

Things we could have done

- Throw more compute at the problem
  - More powerful reranker
  - Rerank more documents

- Human evaluation
  - What are properties of queries where our R@1000 = 0?
  - Is our dataset a good proxy to measure retrieval quality?

- Detect query categories and create more specific pipelines

- Generation Prompt Improvements

- Chunking

- Content Cleansing

Takeaways

- Competitions are fun and a great way to learn
- Keeping it simple and getting the basics right can get you very far
- Everything beyond that requires a incrementally more effort

- Reports: https://huggingface.co/datasets/LiveRAG/Reports