BENCHMARKS, DATA, AND EVALUATION FOR ROBUST RETRIEVAL AND RETRIEVAL-AUGMENTED GENERATION ON HETEROGENEOUS DOMAINS AND LANGUAGES

23/3/2026

Nandan Thakur
David R. Cheriton School of Computer Science

Advisor: Prof. Jimmy Lin

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS

Overview

- Introduction: Retrieval & RAG
  - Benchmarks, Data and Evaluation Challenges
- Background
- Part I: Towards Retrieval Benchmarks
  - Revisiting Argument & Multilingual Retrieval Benchmarks
  - Building Realistic Benchmarks on Technical Documents
- Part II: Towards Data Quality
  - Large-scale Multilingual Generation
  - Supervised Data Curation and Relabeling
- Part III: Towards Retrieval-Augmented Generation (RAG) Evaluation
  - Framework for RAG evaluation and Automatic Support Evaluation with LLMs
  - Multilingual Relevance & Automatic RAG Arena Construction
- Conclusion
- Future Work

Introduction – Retrieval and RAG

- **Information Retrieval** (IR) is a task of finding relevant information from a large collection given a user query, involving first-stage retrieval and reranking stages.
- **Retrieval-Augmented Generation** (RAG) enhances the large language model (LLM) response by augmenting the relevant contextual information during generation.

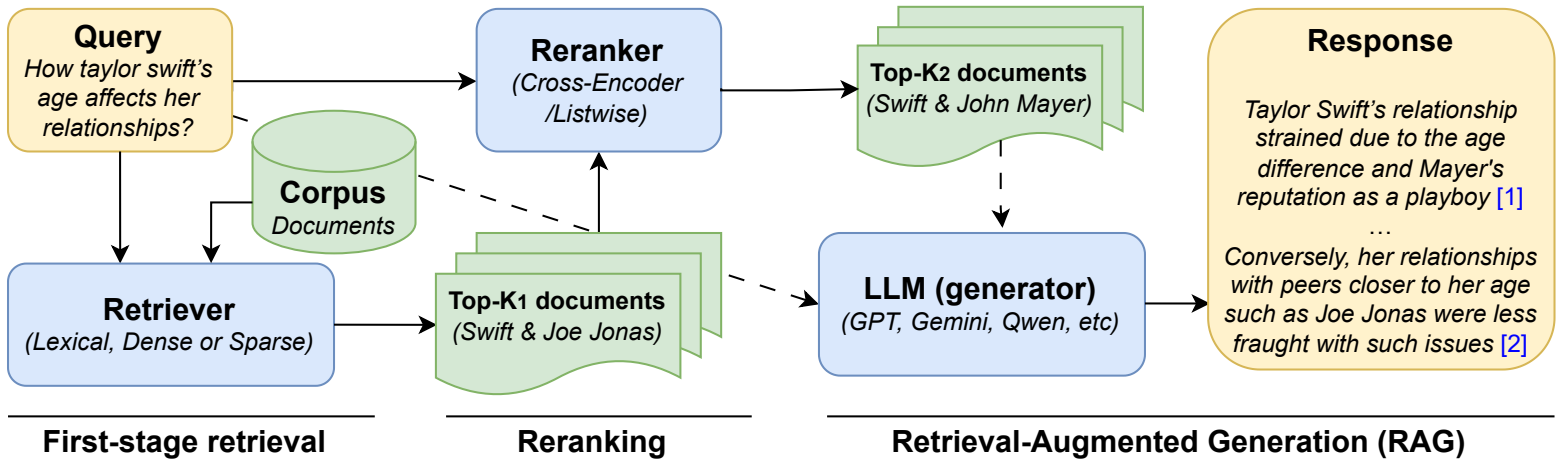


Figure: An illustration of a modern retrieval, reranker and RAG system.

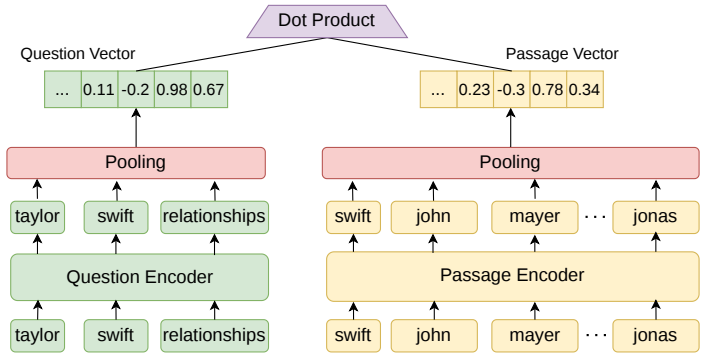
Introduction – Challenges

- **Retrieval Benchmarks**
  - Existing benchmarks are constructed over static and homogeneous corpora, often saturated.
  - Fails to capture out-of-distribution (OOD) generalization on domain-specific and multilingual settings.

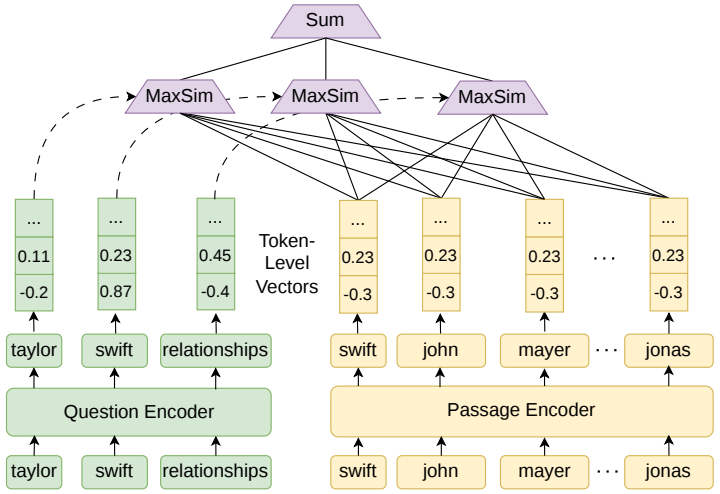
- **Data Quality**
  - Synthetic training data in multiple languages remains unevenly distributed and scarce.
  - Supervised datasets in English contain sparse human judgments leading to false negatives.

- **RAG Evaluation**
  - Automatic evaluation of features, such as attribution and faithfulness versus human alternatives.
  - Limitations in heuristic and arena-based evaluation of RAG systems.

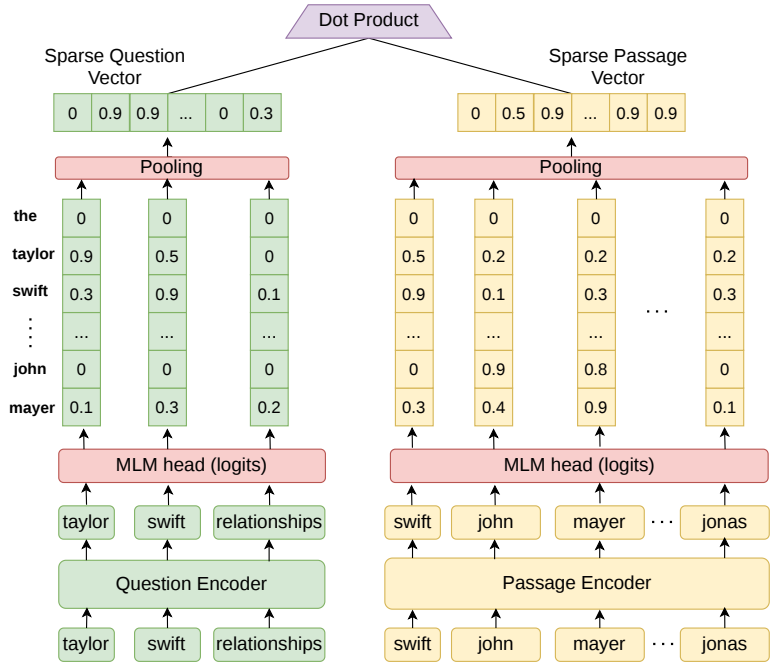
Background – Neural Retrieval



Dense Retrieval



Multi-Vector Retrieval



Sparse Retrieval

Background – BEIR Benchmark

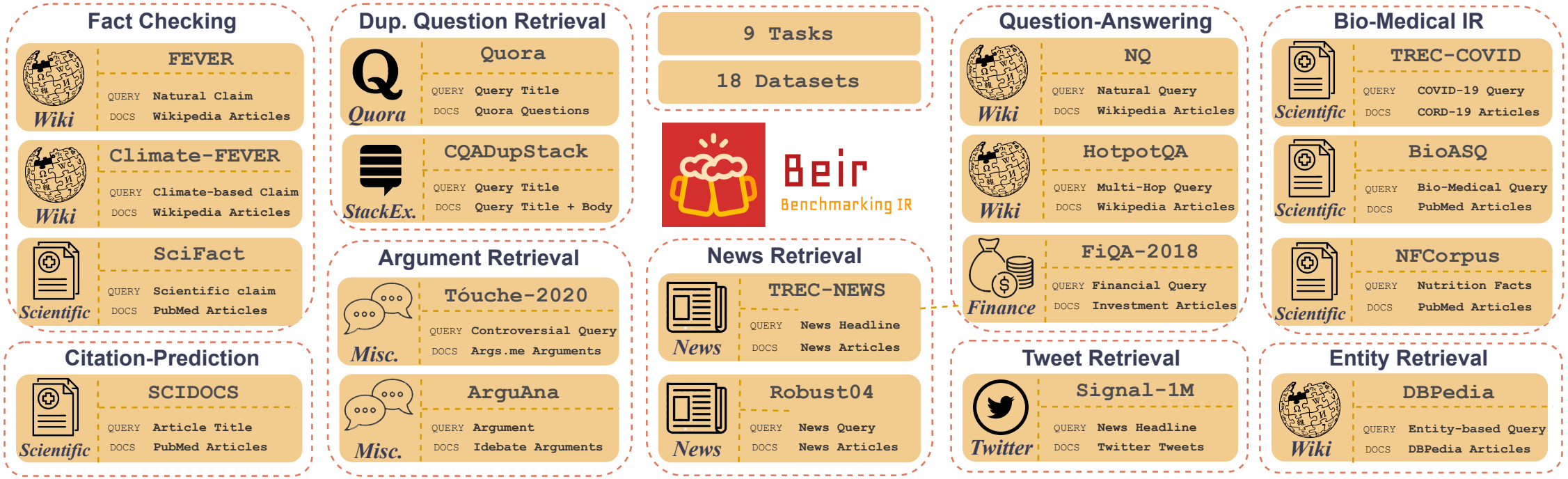


Figure: The diverse tasks and datasets used for zero-shot evaluation in the **BEIR benchmark** (Thakur et al. 2021).

PART I: TOWARDS RETRIEVAL BENCHMARKS

Revisiting Touché 2020: Argument Retrieval Subset of BEIR

- **Objective:** Rank top-k debate arguments based on topical relevance on debate titles.
- **Motivation:** Touché 2020 is one of the few BEIR datasets, where **BM25** continues to outperform ***all*** dense, sparse and multi-vector models.

Model limitations?

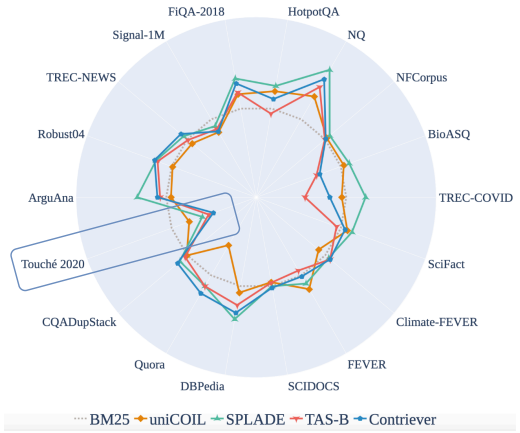


Figure: Comparison of neural retrieval models versus BM25 on Touché 2020, based on average nDCG@10.

Dataset limitations?

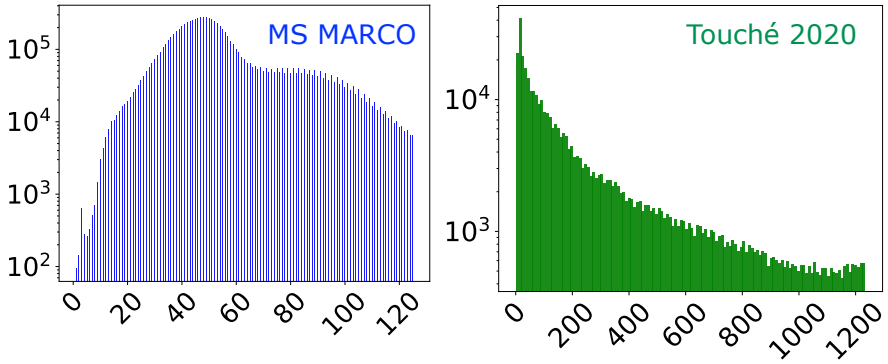


Table: Document length distribution & frequency in MS MARCO vs. Touché 2020.

Revisiting Touché 2020: Argument Retrieval Subset of BEIR

- **Idea** (*model-centric*): **Black-box evaluation** of top-10 retrieved documents by models.
- **Observation:** Neural retrievers retrieve **shorter** documents, with a high lexical overlap with argument conclusion (or *title*) but **noisy** argument premise (or *body*).

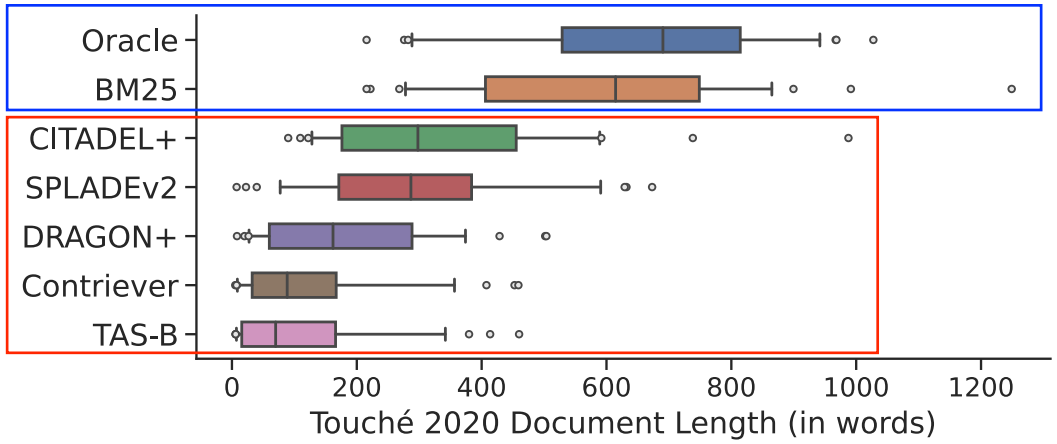


Figure: Distribution of top-10 retrieved document lengths by retrieval models.

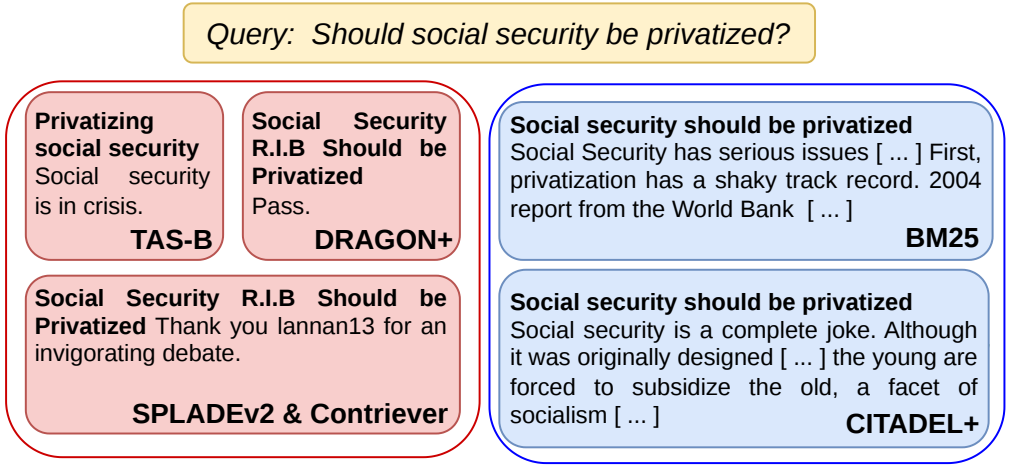


Figure: Example of top-1 document retrieved by retrieval models.

Revisiting Touché 2020: Argument Retrieval Subset of BEIR

- **Idea** (*data-centric*): **Denoising** short documents and conducting **post-hoc judgments**.
- **Observation:** Removed 78K short (<20 words) arguments & titles; added 827 (77% of 1064) **relevant** post-hoc judgments; BM25 continues to outperform **all** neural models!

	Original	+ Denoised	++ Post-hoc
# Documents	382,545	303,732	303,732
Avg. length	293.5	358.7	358.7
# Judgments	2,214	1,785	2,849
# Relevance = 2	636	620 (16 ↓)	1,136 (516 ↑)
# Relevance = 1	296	265 (31 ↓)	576 (311 ↑)
# Relevance = 0	1,282	900 (382 ↓)	1,137 (237 ↑)

Table: Touché 2020 dataset statistics; original / denoising / post-hoc judgments.

	Original		+ Denoised		++ Post-hoc	
Model	nDCG@10	hole@10	nDCG@10	hole@10	nDCG@10	δ inc.
BM25	0.367	61.6%	0.467	51.8%	0.785	△ 0.418
CITADEL+	0.339	60.2%	0.362	62.5%	0.703	△ 0.364
SPLADEv2	0.272	66.3%	0.326	63.3%	0.679	△ 0.407
DRAGON+	0.249	69.2%	0.340	63.9%	0.718	△ 0.469
Contriever	0.205	71.4%	0.303	65.9%	0.650	△ 0.445
TAS-B	0.162	77.8%	0.306	67.5%	0.682	△ 0.520

Table: nDCG@10 on Touché 2020; original / denoising / post-hoc judgments.

Expanding Multilingual Retrieval Benchmarks: MIRACL

- Multilingual retrieval benchmarks are **scarce** in languages or human judgments.
- **Motivation:** MIRACL is one of the **largest** human-labeled multilingual retrieval benchmarks; constructed in a **two-phase** framework with **31 native speakers**.
  - Phase I: *Query generation* by showing Wikipedia snippets as prompts.
  - Phase II: *Relevance assessment* after retrieving potential candidate passages.

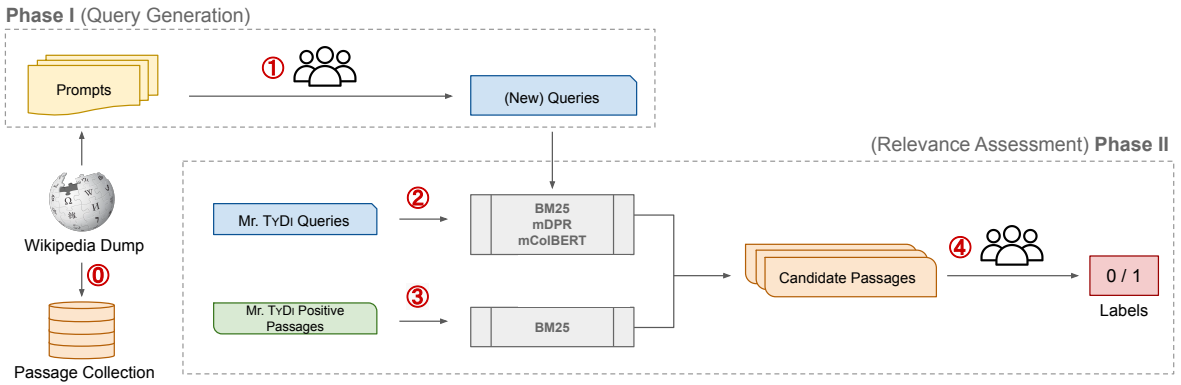


Figure: Two-phase human annotation framework to construct MIRACL.

Dataset	Natural Queries	Natural Passages	Human Labels	# Lang	Avg #Q	Avg #Labels/Q	Total #Labels
NeuCLIR	✓	✓	✓	3	160	32.74	5.2k
MKQA	×	✓	✓	26	10k	1.35	14k
mMARCO	×	×	✓	13	808k	0.66	533k
CLIRMatrix	×	✓	×	139	352k	693	34B
Mr. TyDi	✓	✓	✓	11	6.3k	1.02	71k
MIRACL	✓	✓	✓	18	4.3k	9.23	726k

Table: A Comparison of MIRACL versus existing retrieval benchmarks.

Realistic Benchmark Construction on Technical Documentation

- **Problem:** Existing benchmarks (e.g., NQ, HotPotQA) are currently not well-suited for evaluation due to **benchmark saturation** and **leaderboard overfitting**.
- **Solution:** FreshStack, a holistic framework to construct automatic realistic benchmarks in niche technical domains with real user-asked queries and curated answers.

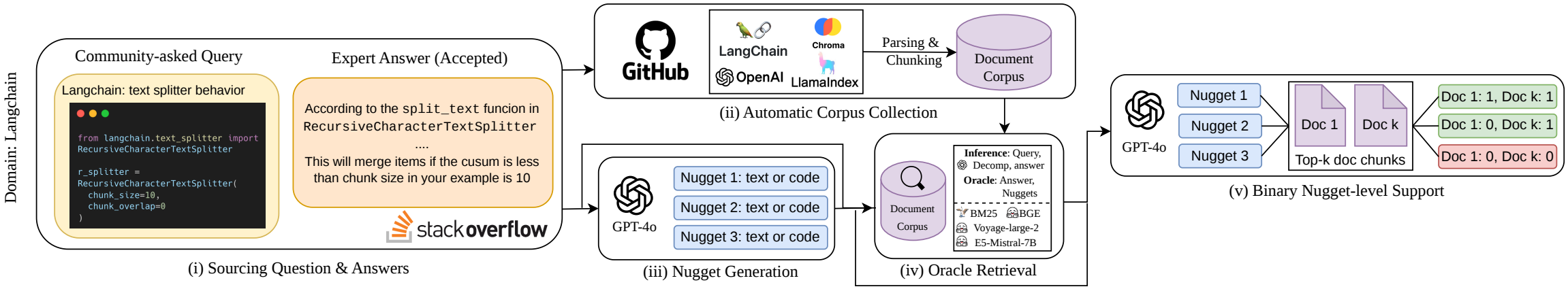


Figure: The FreshStack framework to automatically construct retrieval benchmarks on niche technical domains.

Realistic Benchmark Construction on Technical Documentation

Model	LangChain			Yolo v7 & v8			Laravel 10 & 11			Angular 16, 17 & 18			Godot4		
	αN@10	C@20	R@50	αN@10	C@20	R@50	αN@10	C@20	R@50	αN@10	C@20	R@50	αN@10	C@20	R@50
Inference Setting: Retrieving documents using only the Stack Overflow (SO) query.															
BM25	0.230	0.475	0.261	0.137	0.342	0.337	0.319	0.602	0.441	0.259	0.551	0.340	0.144	0.268	0.200
BM25 + Reranker	0.322	0.587	0.294	0.337	0.590	0.424	0.414	0.729	0.509	0.346	0.647	0.385	0.251	0.407	0.244
BGE (Gemma-2)	0.216	0.548	0.337	0.258	0.547	0.430	0.348	0.699	0.574	0.323	0.571	0.378	0.199	0.479	0.419
BGE (Gemma-2) + Reranker	0.349	0.662	0.387	0.388	0.666	0.459	0.306	0.646	0.571	0.296	0.595	0.387	0.324	0.576	0.471
E5 Mistral (7B)	0.304	0.654	0.393	0.243	0.552	0.394	0.250	0.565	0.470	0.262	0.548	0.368	0.217	0.444	0.359
E5 Mistral (7B) + Reranker	0.385	0.701	0.439	0.364	0.628	0.468	0.305	0.613	0.510	0.306	0.601	0.375	0.315	0.566	0.426
Voyage-large-2	0.246	0.528	0.309	0.270	0.570	0.453	0.345	0.701	0.543	0.304	0.625	0.427	0.282	0.522	0.458
Voyage-large-2 + Reranker	0.345	0.648	0.355	**0.418**	0.670	0.514	0.302	0.653	0.529	0.300	0.600	0.414	**0.342**	0.598	0.511
Fusion (4 models)	0.337	0.700	0.477	0.304	0.627	0.534	**0.426**	**0.748**	**0.646**	**0.385**	**0.719**	**0.532**	0.265	0.550	0.505
Fusion (4 models) + Reranker	**0.397**	**0.729**	**0.501**	0.416	**0.733**	**0.592**	0.319	0.671	0.614	0.318	0.641	0.488	0.340	**0.627**	**0.545**
Best Scores in the Oracle Setting: Upper Baselines on the FreshStack dataset															
SO Answer: Fusion (4 models)	0.484	0.821	0.619	0.546	0.854	0.788	0.564	0.892	0.820	0.470	0.805	0.695	0.449	0.741	0.683
SO Nuggets: Fusion (4 models)	0.519	0.881	0.655	0.601	0.876	0.825	0.566	0.888	0.818	0.544	0.881	0.756	0.476	0.814	0.719

Table: Retrieval results on FreshStack; best scores in the inference setting are highlighted in **bold**. The reranker is the *Voyage AI rerank-2* reranking the top 100 documents, if the reranker improves the retrieval score, it is highlighted in blue else red.

PART II: TOWARDS DATA QUALITY

Leveraging LLMs for Synthesizing Multilingual Training Data

- **Problem:** Multilingual retrievers require *a lot* of human-supervised training pairs, which are **uneven** and **scarce** across languages, or **machine-translated** from English.
- **Solution:** Large-scale dataset providing **28M training pairs** covering **33 languages** with **synthetic queries** generated using PaLM 2.

Model (mT5)	Pre-Train?	Finetune?	XOR-Retrieve (7L) (Avg. Success@5kt)	MIRACL (18L) (Avg. nDCG@10)	XTREME-UP (20L) (Avg. MRR@10)
Zero-shot baselines (English-only Supervision)					
mDPR-EN	-	MS MARCO	0.393	0.398	0.063
mContriever-EN	mC4	MS MARCO	0.440	0.378	0.079
Supervised baselines (English + Language-specific Supervision)					
mDPR-X	-	MS MARCO + FT	0.582	0.396	0.084
mContriever-X	mC4	MS MARCO + FT	0.596	0.554	0.124
Synthetic baselines (Our Work)					
SWIM-X (best)	mC4	SWIM-IR	0.667	0.464	0.261

Table: Retrieval results on three benchmarks: XOR-Retrieve (cross-lingual), MIRACL (monolingual) and XTREME-UP (cross-lingual).

Leveraging LLMs for Synthesizing Multilingual Training Data

How cheap is synthetic data generation?

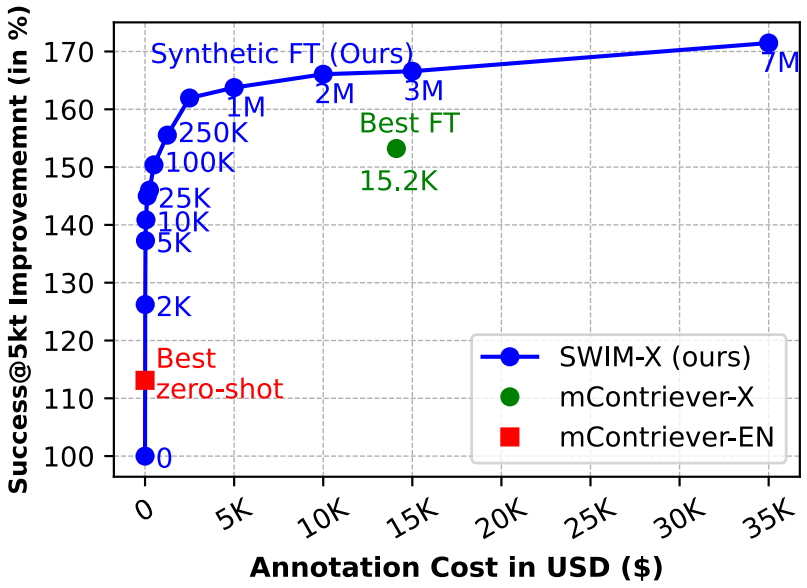


Figure: Success@5kt versus annotation cost on XOR-Retrieve.

How to extend to more languages?

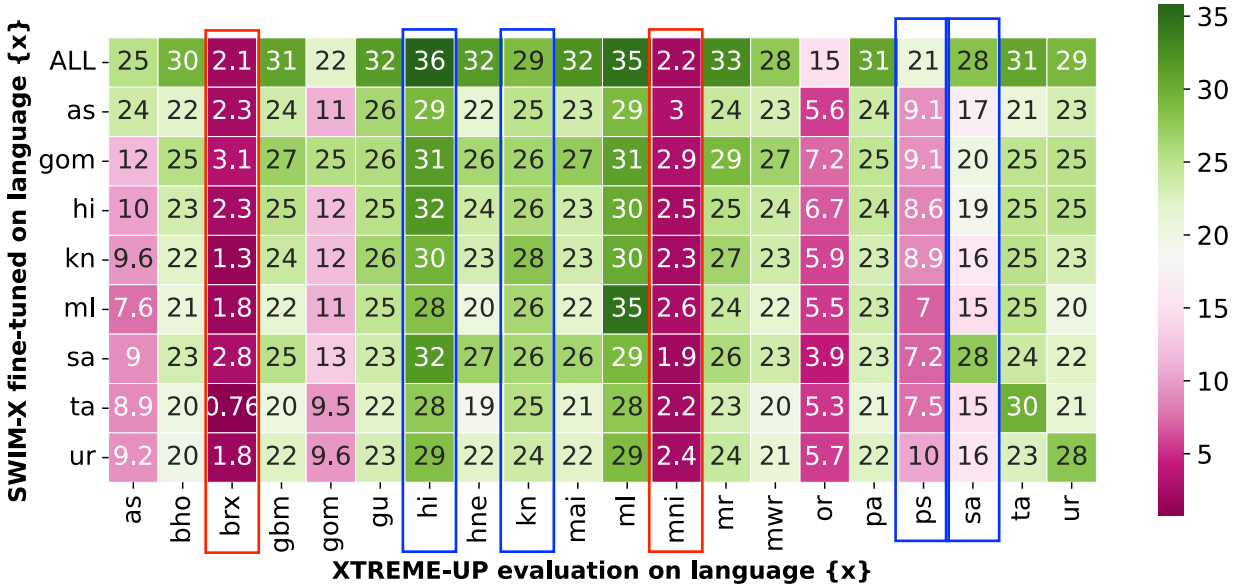
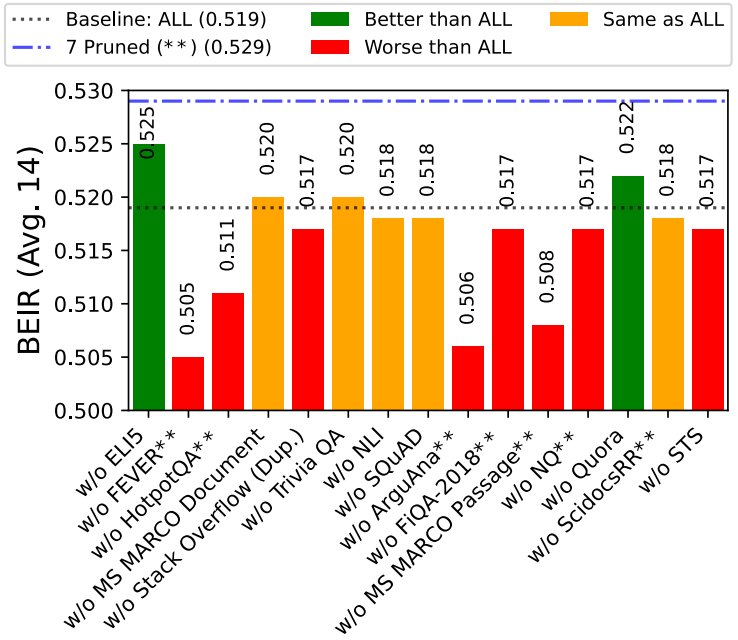


Figure: MRR@10 of SWIM-X fine-tuned evaluated on 20 Indic languages in XTREME-UP.

Supervised Dataset Pruning and Curation

- **Question:** Does every training dataset contribute towards model improvement?
- **Leave-one-out Dataset** (retrain dense retriever by removing one dataset at a time).



- **Green** (Substantially ***higher*** *accuracy*)
- **Orange** (***Similar*** *accuracy*)
- **Red** (Substantially **lower** accuracy)

Pruning **8/15** datasets (**57%** size reduction) improves avg. nDCG@10 (BEIR, 14 datasets) from 0.519 to 0.529.

Figure: Average nDCG@10 on BEIR; 15 datasets from BGE collection.

Supervised Dataset Pruning and Curation

- **Problem:** Sparse judgments introduce **false negatives** in training datasets, with prior work either avoids them during training, or adopts KL distillation.
- **Idea: Identify** and **relabel** false negatives with cascading LLM judges (*GPT-4o* & *-mini*).

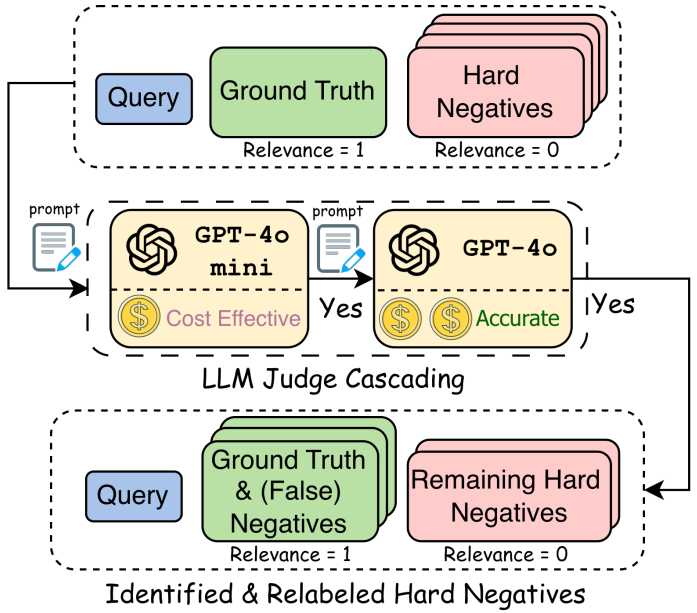


Figure: RLHN Methodology for identifying and relabeling false negatives.

Dataset	# Train Pairs	False Hard Negatives Stage 1	False Hard Negatives Stage 2
MS MARCO	485,823	391,965	326,301
HotpotQA	84,516	11,268	4,756
NQ	58,568	32,184	19,199
FEVER	29,096	7,764	3,577
ScidocsRR	12,655	2,068	351
FiQA-2018	5,500	3,632	1,833
ArguAna	4,065	0	0

Table: Training Pairs with false negatives identified with RLHN.

Supervised Dataset Pruning and Curation

- **Question:** False hard negatives — *remove* or r*elabel*? How does data curation compare against: Cross-Encoder (CE) distillation and HN mining?

BEIR Benchmark	*No Filtering*	*Baselines*		*RLHN (Stage 2)*		
	Default	HN Mining	CE Distill	Remove	Remove HN	RLHN
TREC-COVID†	0.783	0.789	0.793	0.794	0.785	0.809
NFCorpus†	0.378	0.377	0.363	0.380	0.382	0.390
NQ	0.595	0.601	0.624	0.573	0.598	0.591
HotpotQA	0.737	0.734	0.741	0.741	0.736	0.735
FiQA-2018	0.439	0.434	0.417	0.441	0.445	0.448
ArguAna	0.701	0.697	0.725	0.700	0.700	0.692
Touché-2020†	0.256	0.286	0.305	0.218	0.265	0.266
DBPedia	0.438	0.444	0.446	0.433	0.441	0.447
SCIDOCS	0.242	0.243	0.216	0.245	0.243	0.242
FEVER	0.878	0.878	0.889	0.881	0.876	0.871
Climate-FEVER	0.391	0.386	0.377	0.382	0.384	0.367
SciFact	0.735	0.735	0.727	0.744	0.735	0.740
TREC-NEWS†	0.465	0.466	0.458	0.464	0.473	0.484
Robust04†	0.442	0.451	0.452	0.447	0.458	0.497
Signal-1M (RT)†	0.275	0.272	0.271	0.274	0.270	0.274
BioASQ†	0.378	0.375	0.413	0.384	0.384	0.394
Avg. 16 (All)	0.508	0.511	0.514	0.506	0.511	**0.515**
Avg. 7 (OOD)	0.425	0.431	0.436	0.423	0.431	**0.445**

Table: nDCG@10 on the BEIR benchmark with E5-base model; best scores are highlighted in **bold**; seven OOD datasets are highlighted in **blue**.

Impact of curated dataset size?

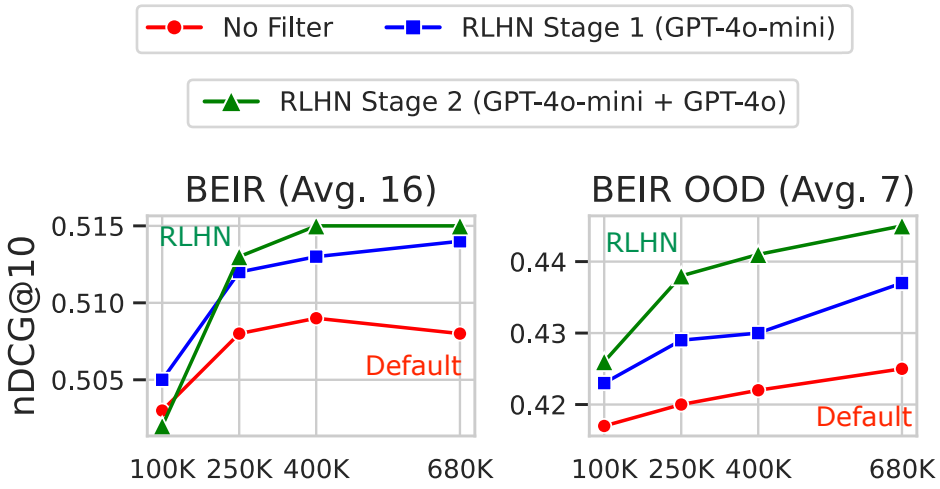


Figure: Avg. nDCG@10 scores on BEIR (all) and (OOD, 7 datasets) with variation in the number of curated training pairs.

PART III: TOWARDS RAG EVALUATION

Ragnarök: A framework for Reusable Baselines at TREC 2024 RAG

- Designed a **reusable framework** for RAG evaluation, constructed the **MS MARCO V2.1** corpus, **topics** with **long-form answers**, and used for providing **RAG baselines**.

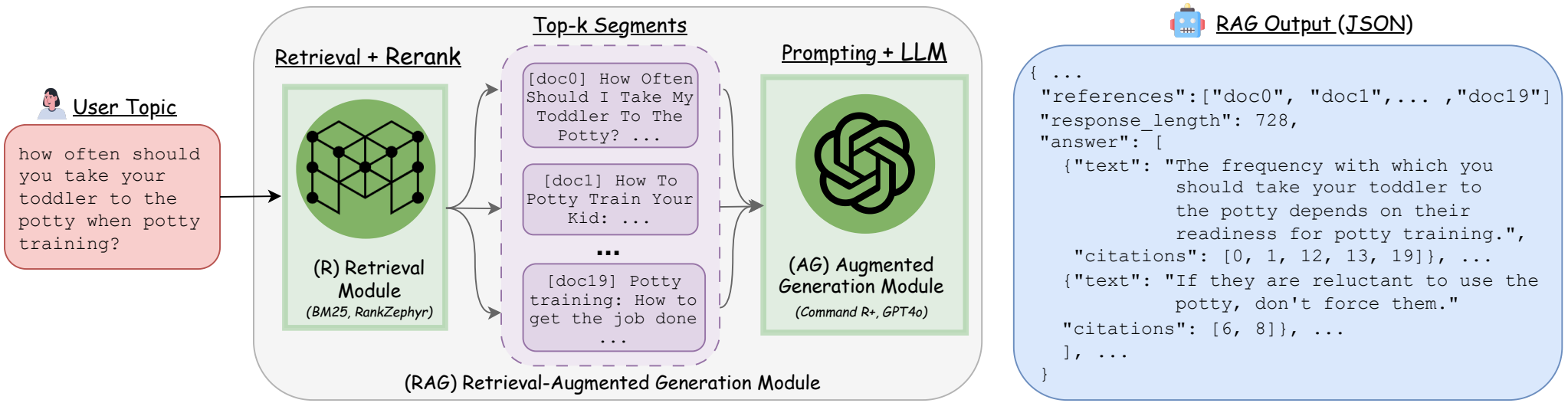


Figure: A Schematic diagram of **Ragnarök**, used for providing reusable baseline responses at TREC 2024 RAG.

Support Assessment via LLM judges at TREC 2024 RAG

- **Problem:** Evaluating **support** (*whether information in an answer is factually supported by its cited documents*) via human judges is cumbersome and expensive.
- **Idea:** Pairwise assessment with GPT-4o judge; validated against human judgments.

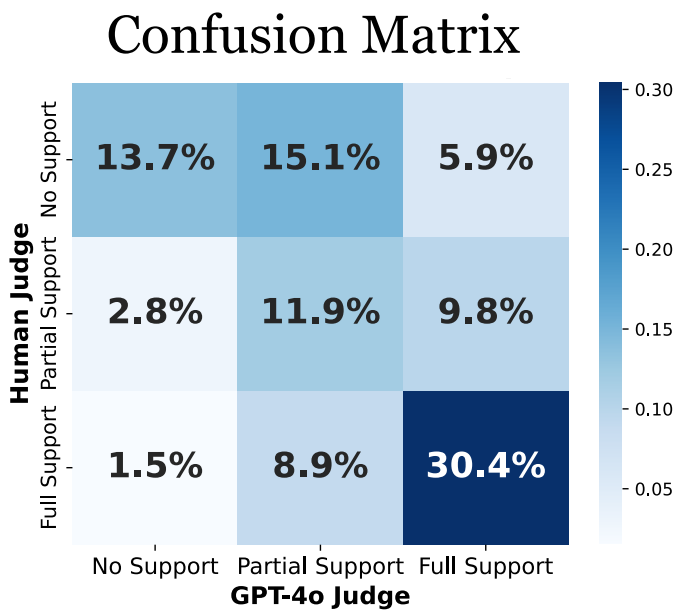


Figure: Confusion matrix comparing support judgments.

	Cohen's Kappa	Manual from Scratch	
		GPT-4o	Human (NIST)
(1)	Expert (human)	0.29	– 0.03
(2)	LLAMA-3.1 (405B)	0.60	– 0.20

Table: Inter-assessor agreement scores in the disagreement task.

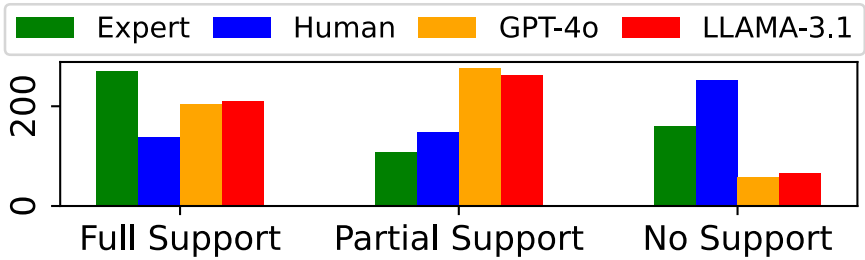
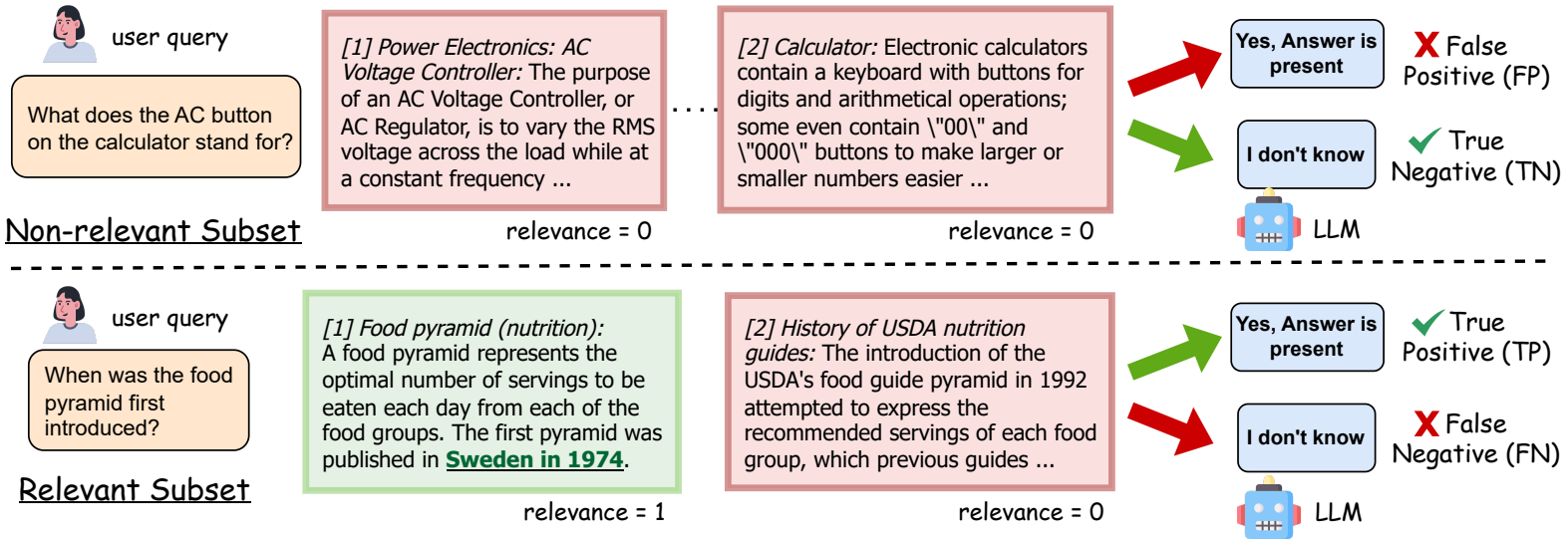


Figure: Label distribution by assessors in the disagreement task.

Multilingual LLM Relevance Assessment on NoMIRACL

- **Idea:** Convert the **LLM relevance assessment** into a **binary classification task**.
- Non-relevant subset measures **LLM abstention** when relevant information is not present, whereas the relevant subset measures **identification** of relevant information.



- **Non-relevant subset:** *unanswerable* or *invalid* queries in MIRACL.
- **Relevant subset:** MIRACL queries.

Figure: LLM relevance evaluation as a binary tree classification in NoMIRACL.

Multilingual LLM Relevance Assessment on NoMIRACL

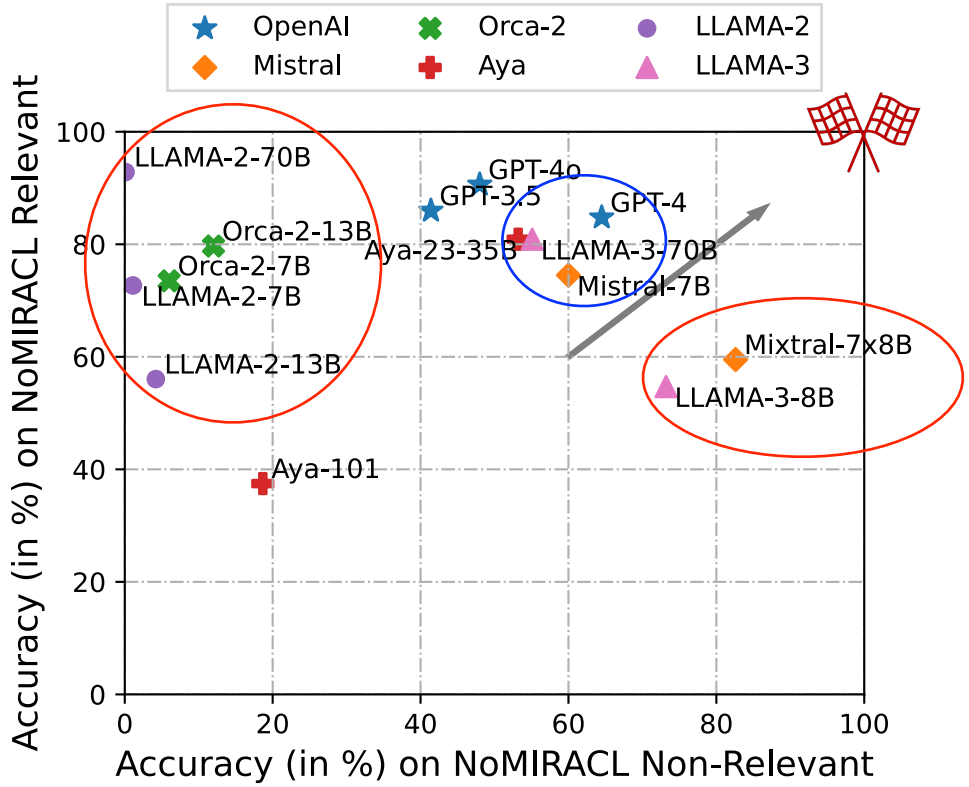


Figure: Average accuracy (in %) on 18 languages in NoMIRACL on both non-relevant and relevant subsets.

Prompt Optimization

	Prompt Optimization	NoMIRACL (Accuracy)	
		Non-relevant	Relevant
	Original	60.0	74.5
(1)	+ Role (You are an evaluator...)	51.3	80.8
(2)	+ Repeat (repeat instruction)	44.1	89.7
(3)	+ Explanation (CoT reasoning)	69.7	66.2

Table: Prompt optimization ablation on average accuracy on 18 languages in NoMIRACL with Mistral 7B model.

Expanding Automatic Multilingual RAG benchmarks

- **Problem: Heuristic features** are difficult to combine & requires gold truth; **arena setups** are expensive requiring multiple LLM API calls.
- **Idea:** Train a surrogate learning-to-rank (L2R) judge to generate a **synthetic RAG arena**, by training solely on heuristic features.

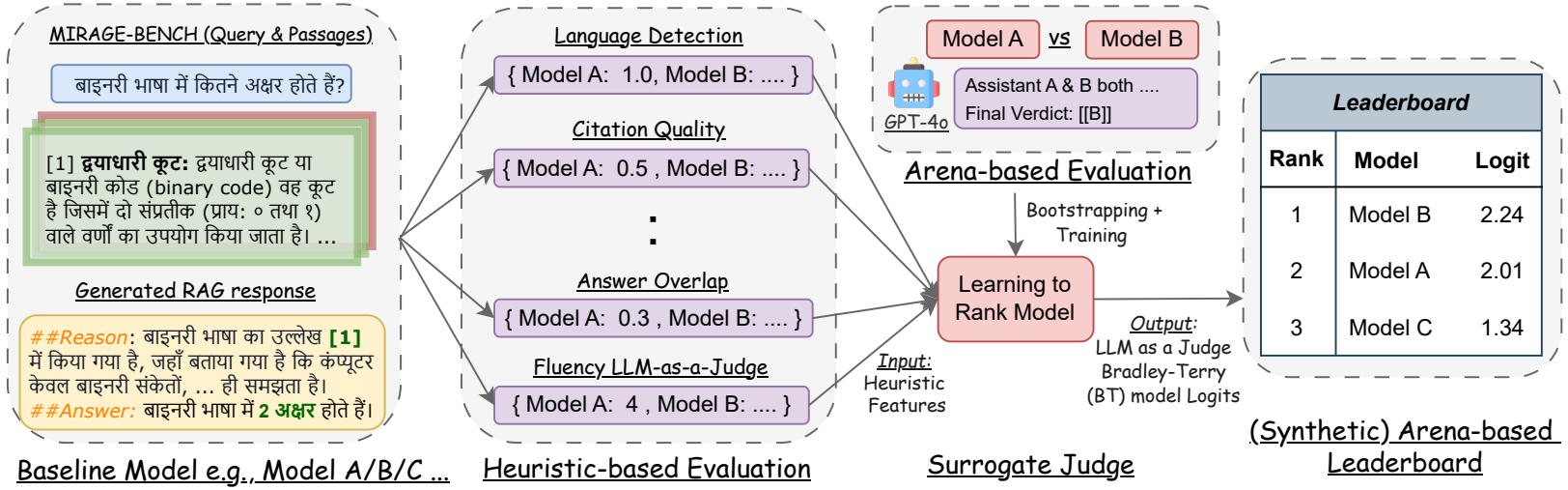
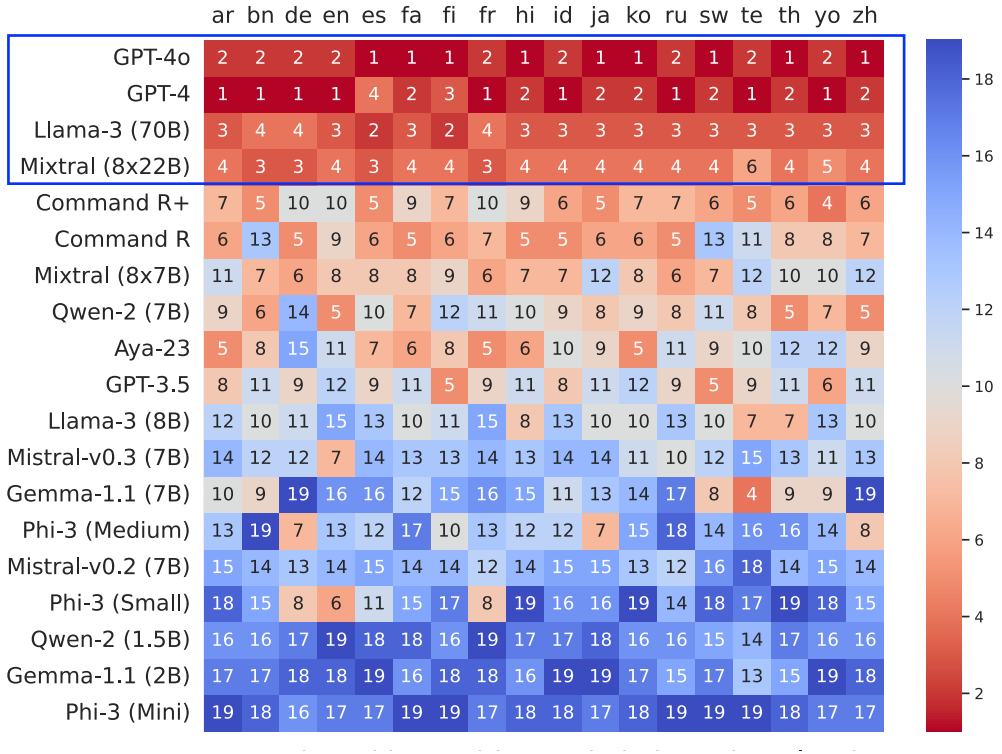


Figure: Flowchart used to train a surrogate judge, and at inference generate a synthetic arena leaderboard.

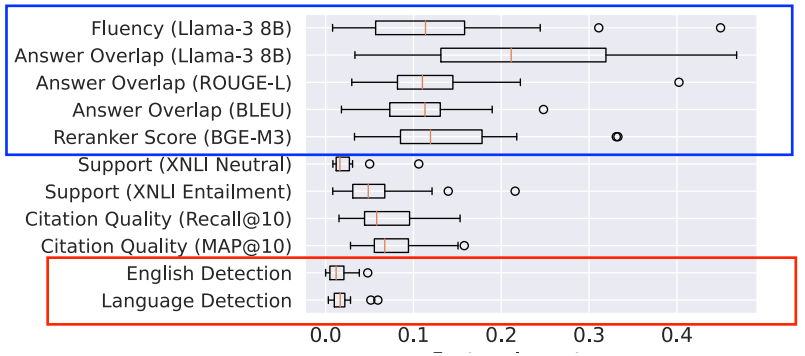
Expanding Automatic Multilingual RAG benchmarks



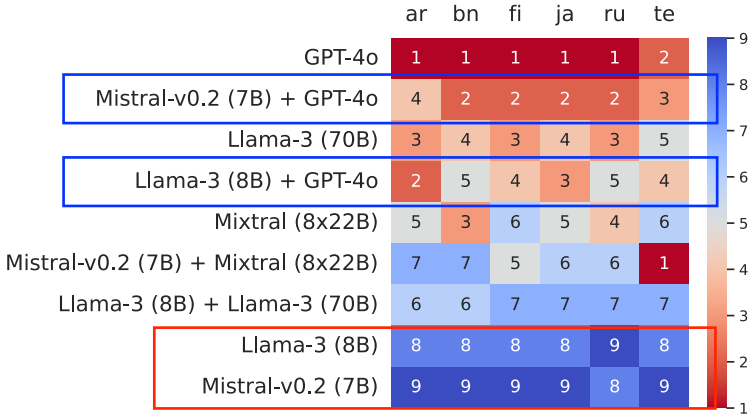
Avg. Kendall Tau

(L2R & GPT-4o judge) = 0.909

Heuristic Feature Importance



Fine-tuning on Smaller LLMs



Conclusion

- **Part I: Towards Retrieval Benchmarks**
  - Revisiting Touché 2020: Argument Retrieval Subset of BEIR.
  - Expanding Multilingual Retrieval Benchmarks: MIRACL.
  - Realistic Benchmark Construction on Technical Documentation.
- **Part II: Towards Data Quality**
  - Leveraging LLMs for Synthesizing Multilingual Training Data.
  - Supervised Dataset Pruning and Data Curation.
- **Part III: Towards RAG Evaluation**
  - Ragnarök: A framework for Reusable Baselines at TREC 2024 RAG.
  - Support Assessment via LLM judges at TREC 2024 RAG.
  - Multilingual LLM Relevance Assessment on NoMIRACL.
  - Expanding Automatic Multilingual RAG benchmarks.

Future Work

- **Constructing Effective Deep Research Benchmarks**
  - TREC 2025 RAG track incorporated narrative-style queries, requiring multi-step reasoning.
  - Construct private retrieval and RAG benchmarks with a “refresh” option.

- **Generating Reasoning-Intensive Training Datasets**
  - Construct multi-hop reasoning training datasets for search agents trained with Reinforcement Learning (RL), such as GRPO.

- **Expanding RAG Evaluation at TREC 2026 RAG**
  - Incorporate newer evaluation features such as safety and counterfactual robustness.

Thank you!

Other Selected Works

- ***Injecting Domain Adaptation with Learning-to-hash for Effective and Efficient Zero-shot Dense Retrieval***
  Nandan Thakur, Nils Reimers, Jimmy Lin
  ReNeuIR @ SIGIR 2023.
- ***SPRINT: A Unified Toolkit for Evaluating and Demystifying Zero-shot Neural Sparse Retrieval***
  Nandan Thakur, Kexin Wang, Iryna Gurevych, Jimmy Lin
  SIGIR 2023 (Resource Track).
- ***Resources for Brewing BEIR: Reproducible Reference Models and Statistical Analyses***
  Ehsan Kamalloo, **Nandan Thakur**, Carlos Lassance, Xueguang Ma, Jheng-Hong Yang, Jimmy Lin
  SIGIR 2024 (Resource Track).
- ***The Great Nugget Recall: Automating Fact Extraction and RAG Evaluation with Large Language Models***
  Ronak Pradeep, **Nandan Thakur**, Shivani Upadhyay, Daniel Campos, Nick Craswell, Jimmy Lin
  SIGIR 2025 (Main Track: Long Paper).
- ***Chatbot Arena Meets Nuggets: Towards Explanations & Diagnostics in the Evaluation of LLM Responses***
  Sahel Sharifymoghaddam*, Shivani Upadhyay*, **Nandan Thakur***, Ronak Pradeep, Jimmy Lin
  Preprint 2025. abs/2504.20006.
- ***Overview of the TREC 2025 Retrieval Augmented Generation (RAG) Track***
  Shivani Upadhyay, **Nandan Thakur**, Ronak Pradeep, Nick Craswell, Daniel Campos, Jimmy Lin
  Preprint 2026. abs/2603.09891.