SimpleText Best of Labs in CLEF-2024: Application of Large Language Models for Scientific Text Simplification

Nicholas Largey, Reihaneh Maarefdoust, Shea Durgin & Behrooz Mansouri

Department of Computer Science, University of Southern Maine

CLEF 2025, Madrid, Spain

Problem Statement: Science is Hard to Read

Challenge: Scientific literature is dense, full of jargon and buzzwords

- Assumes a large amount of existing knowledge
- Creates a barrier for students, the public, and researchers

Lab Goal: Make scientific information more accessible to a broader audience

Our Approach: Utilize the Large Language Models (LLMs) and explore construction of prompts as a skill to refine

Ermakova et al., Overview of the CLEF 2024 SimpleText Track: Improving Access to Scientific Texts for Everyone. ECIR 2024

SimpleText 2024 Tasks

Task 1: Content Selection

- Find the right background documents to help explain a complex paper
- Corpus of scientific abstracts
- Query: ad-hoc query + article from which query was generated

Task 2: Complexity Spotting

- Automatically identify and explain difficult terms in a scientific text

Task 3: Text Simplification

- Rewrite complex scientific text into simpler language

Ermakova et al., Overview of the CLEF 2024 SimpleText Track: Improving Access to Scientific Texts for Everyone. ECIR 2024

Task 1: Retrieving Passages to Include in a Simplified Summary

LLM for Query Expansion

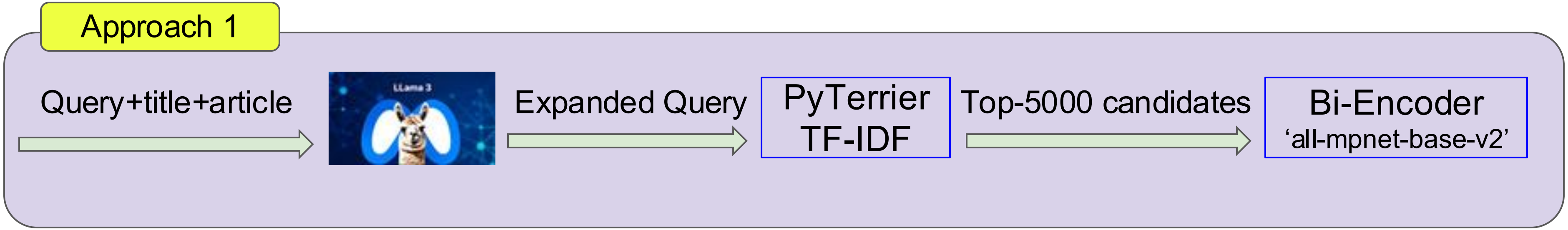
Pass the query, the related article title, and context to LLaMA3 to expand the initial query

LLM as a Pair-wise Re-ranker

Passing query + article title + (context)

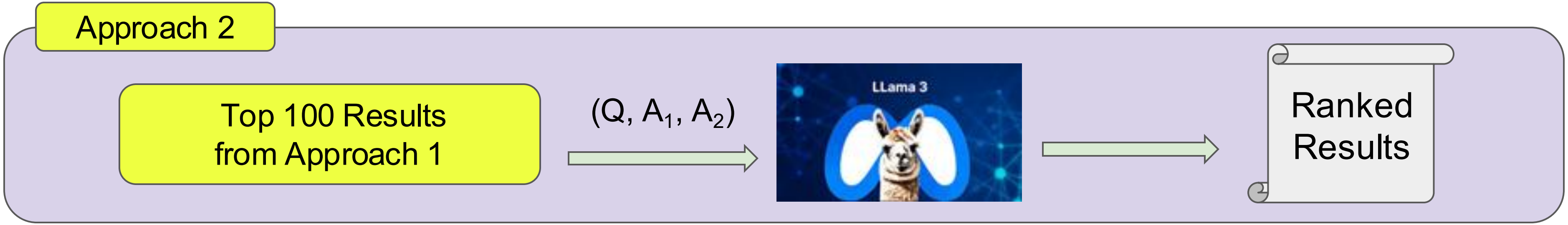
Task	System Message
Query Expansion	Being a ranking model your first Task is to do query expansion. For an information need, you will add more context to it. Contextualize the query as best as you can in one or two short sentences, for a given information need and context.
Re-ranking	You are a ranking model for information retrieval. Given a query and two documents, you will say which one is more relevant. If Document 1 is more relevant say yes, otherwise say no.

Task 1: Overview of our Top Approaches



Approach 1

Query+title+article → Llama 3 → Expanded Query → PyTerrier TF-IDF → Top-5000 candidates → Bi-Encoder ‘all-mpnet-base-v2’



Approach 2

Top 100 Results from Approach 1 → (Q, A_1, A_2) → Llama 3 → Ranked Results

Task 1: Results and Analysis

Model	MRR	P@10	P@20	NDCG@10	NDCG@20	Bpref	MAP
LLaMABiEncoder	**0.94**	**0.82**	**0.55**	**0.62**	**0.52**	**0.36**	**0.23**
LLaMAReranker2	0.93	0.79	0.54	0.60	0.50	0.35	0.22

LLaMABiEncoder results, for 90% of topics, the MRR value is 1

Topic Id	Original Query	Expanded Query	MRR	P@10
T11.1	character relationship	character relationship network map The Witcher	1	1
G02.C1	concerns related to the handling of sensitive information by voice assistants	voice assistants handling sensitive information concerns Apple Siri recordings	0.33	0.7

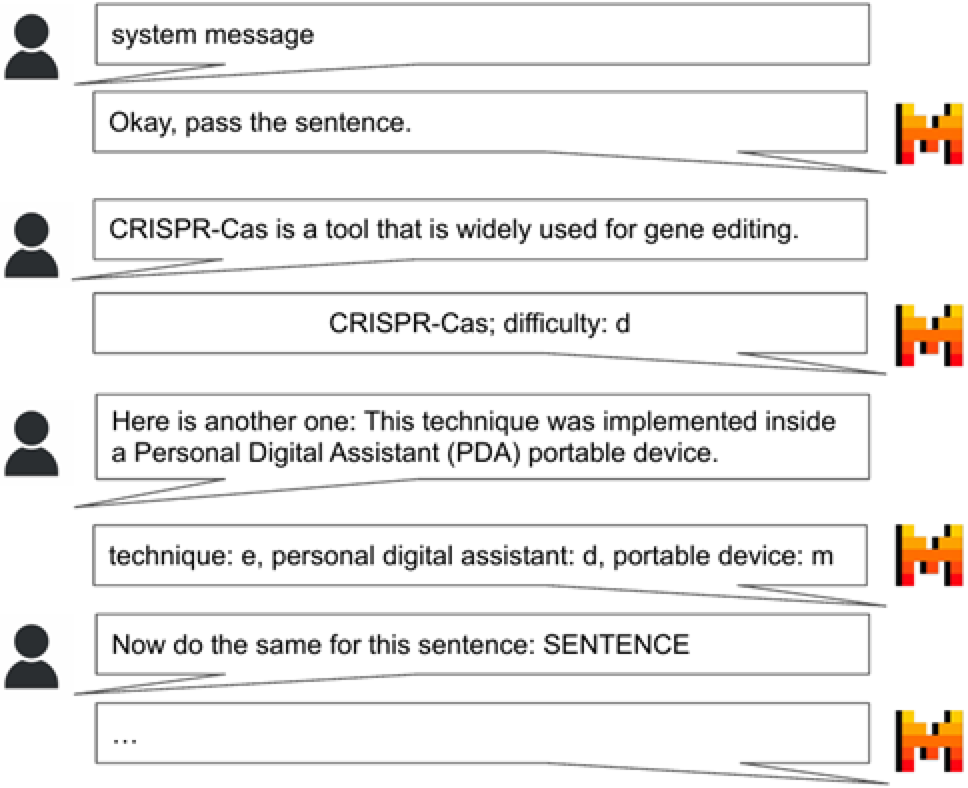
Task 2: Identifying and Explaining Difficult Concepts

- LLaMA-3 and Mistral with few-shot prompting
- Different system prompts for each subtask

Mistral

System Message
Task 2.1 (Detecting Difficult Terms)

Role: You are a helpful assistant.
Task: Given a sentence, you will just output the unclear technical term or terms (up to 5 terms). You choose the terms that should be defined in order to understand the sentence. Each sentence can have up to 5 phrases.
Answer Format: You will decide the difficulty of unclear terms with scales easy (e), medium (m), hard (d).
Negative Sample: Note that easy does not include terms such as shown, pronouns, or numbers



system message

Okay, pass the sentence.

CRISPR-Cas is a tool that is widely used for gene editing.

CRISPR-Cas; difficulty: d

Here is another one: This technique was implemented inside a Personal Digital Assistant (PDA) portable device.

technique: e, personal digital assistant: d, portable device: m

Now do the same for this sentence: SENTENCE

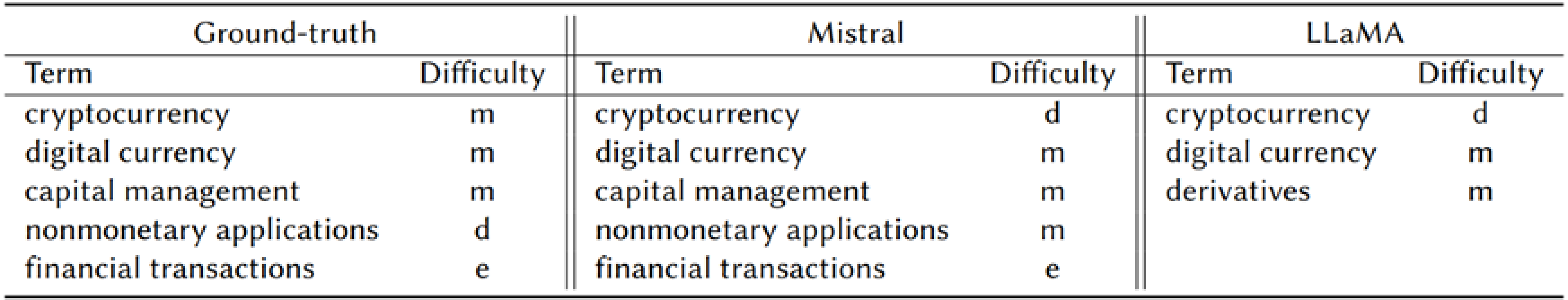
...

Task 2: Results and Analysis

Model	Recall	Precision	Recall of Difficult Terms	Precision of Difficult Terms
Mistral	**0.41**	**0.69**	0.19	0.49
LLaMA	0.28	0.65	**0.26**	**0.67**

LLaMA extracts fewer terms compared to Mistral

- More precise on detecting difficulty levels
- LLaMA has less consistency compared to Mistral for the same sentences



Ground-truth		Mistral		LLaMA	
Term	Difficulty	Term	Difficulty	Term	Difficulty
cryptocurrency	m	cryptocurrency	d	cryptocurrency	d
digital currency	m	digital currency	m	digital currency	m
capital management	m	capital management	m	derivatives	m
nonmonetary applications	d	nonmonetary applications	m		
financial transactions	e	financial transactions	e		

Task 3: Simplify Scientific Text

Instruction-Tuned LLaMA3:

- Encode the provided training corpus with QLoRA to Instruction-Tune a LLaMA3 model
- Each run consists of new system message which is used for both sub-Tasks

Instruction Tuning

System Message (Run 1)

Task: Simplify this text for science students in college.

Output Format: Maximize the use of simple words and short sentences but include key words from the original text.

Optimize the output FKGL, SARI and BLEU scores.

Negative Sample: Don't give an explanation, just output the simplified text.

Instruction Tuning

LLaMA3 Input Prompt

System Message +
Source Sentence/Abstract +
Target Sentence/Abstract

Task 3: Simplify Scientific Text

Training

Initial Training Corpus

QLoRA Encoded Training Corpus

Fine-Tuned Model

Approach

Test Corpus

Simplified Sentences
or
Abstracts Results

Task 3: Results and Analysis

Task 3.1

Run	FKGL	SARI	BLEU
1	8.39	**40.58**	**7.53**
2	**9.47**	40.36	6.26

Task 3.2

Run	FKGL	SARI	BLEU
1	**9.07**	**43.44**	**11.73**
3	10.17	43.21	11.03

- Instructing LLaMA3 during training to target specific metrics lead to more desirable output
- In the output for run 1, the instructed “science student in college” scored lower on the FKGL scale than expected

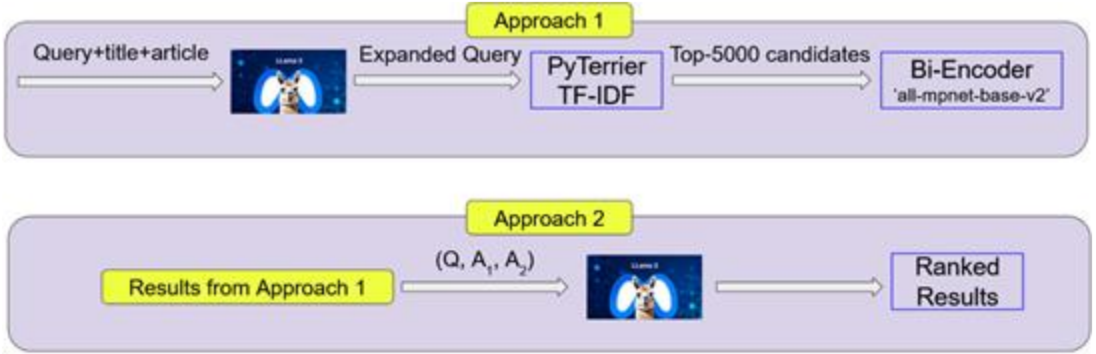
Reference Sentence	Submitted Sentence
Many Personal Digital Assistants (PDAs) are not able to understand decimal numbers	Most PDA CPUs are integer-only CPUs, meaning they can't perform calculations with decimal numbers

Conclusion

AIIR Lab approaches in SimpleText’24 involved using LLMs for all tasks

- Used for query expansion and re-ranking for Task 1 with promising results
- Used for extracting difficult terms and explanation for Task 2
  - Less competitive results
  - Future work will involve using retrieval augmented generation (RAG) for definition generation
- Fine-tuned for Task 3 with QLoRA, also, with promising results
  - We plan to explore Chain-of-Thoughts to teach the models how to reason about simplification

SimpleText Best of Labs in CLEF 2024



Nicholas Largey, Reihaneh Maarefdoust, Shea Durgin & Behrooz Mansouri
Contact: Nicholas.Largey@Maine.edu

AIR
Artificial Intelligence & Information Retrieval

UNIVERSITY OF
SOUTHERN MAINE

MADRID
CLEF 2025
Conference and Labs of the Evaluation Forum