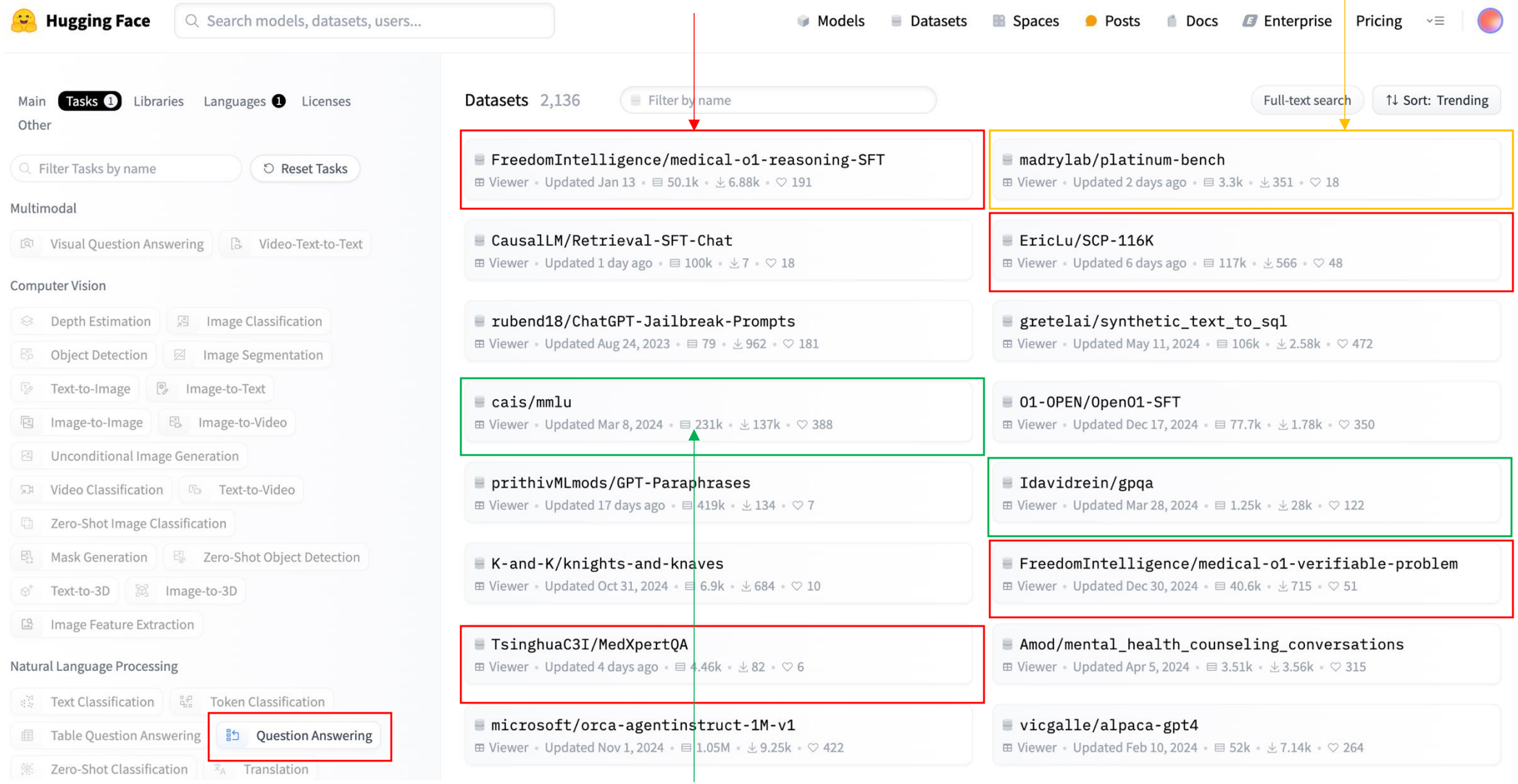
ADVANCED PRACTICAL: DIVERSIFIED ANSWER GENERATION QA SYSTEM

Mara-Eliana Popescu
Supervised by Ashish Chouhan
Data Science Group
WS24/25

Motivation

Scientific questions expecting very precise answers

Mostly one correct answer, no "diversification"



Multiple choice

https://huggingface.co/datasets?modality=modality:text&task_categories=task_categories:question-answering&language=language:en&sort=trending

Diversified Question Answering

Instead of focusing on narrow questions, expecting very precise answers, we are interested in broader questions allowing answers from multiple (expert) perspectives and covering several topics. This is what we mean in this setting by diversity or diversification.

This leads to the following questions:

1. *Are there Question Answering datasets designed for diversification?*
2. *Can current LLMs generate a diversified answer given an appropriate question?*
3. *How can we evaluate diversification in answer generation?*

Objectives

- Create a custom diversified QA dataset.
- Create a knowledge base providing the necessary information to answer the questions.
- Construct a RAG pipeline for (ideally) diversified answer generation.
- Evaluate the pipeline focusing on diversification.

QA Dataset

About the data:

- The ground truth data consists of blog posts by Ask EP.
- Those blog posts answer questions asked by citizens about various topics (e.g. mental health, climate politics, elections etc.).
- The blog posts cover several viewpoints or topics for the given question.

Protection of personal data and privacy

Your right to protection of your personal data is enshrined in the EU charter of fundamental rights. In 2016, the EU adopted the General Data Protection Regulation – often referred to as 'GDPR rules'. The regulation applies to all companies that process their users' data within the EU. Under the GDPR, social media companies must obtain explicit consent from their users to access and process their data. It establishes a series of rights for citizens including the rights to:

The new EU digital rulebook

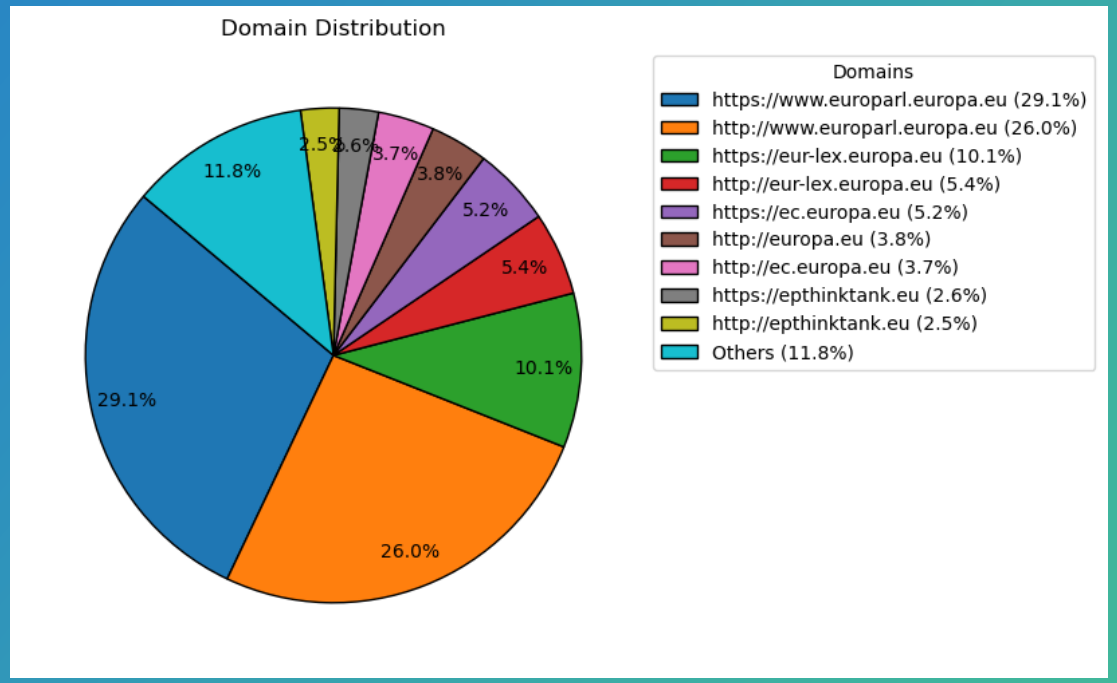
One of the priorities of the von der Leyen Commission was to make Europe 'fit for the digital age'. In 2022, Parliament and EU governments brought in two major new laws to create a fairer and safer online world: the Digital Markets Act and the Digital Services Act. Broadly, the idea is that 'what is illegal offline should be illegal online'.

The Digital Markets Act – limiting the power of big digital companies

https://epthinktank.eu/2024/06/28/regulating-social-media-what-is-the-european-union-doing-to-protect-social-media-users/

QA Dataset: data collection and storage

- Scraping blog post content using BeautifulSoup (also tried Crawl4AI).
- We extract from each blog post:
  - html, url, date of publication, title, multiple languages flag, links, tags and tag urls,
  - question,
  - answer
- The question is sometimes directly formulated in the title of the blog post and sometimes only in the first paragraph (no consistency here). We used GPT-4 to decide in each case which of the two contains the question.
- The answer is stored as a dictionary, where the keys are section titles, and the values are the corresponding section paragraphs.
- All contents are then stored in an OpenSearch index.
- There are 192 documents (all blog posts by Ask EP until today).
- Continuous process of scraping, processing and indexing.



Most contents are from the legal sector judging by the domains.

Knowledge Base

- We scrape the text content from all the external links of the blog posts to construct the knowledge base.
- After removing duplicates, invalid urls and urls causing various http errors, we gathered a list of **1668 non-pdf urls** and **135 pdf urls**.
- Scraping the text from non-pdf urls with BeautifulSoup with additional cleaning gives good results.
- For the pdf urls, we use MarkItDown, which is a tool developed by Microsoft to convert various formats to Markdown. We decided to use it because it can handle tabular data quite well and 87 out of the 135 pdfs contain tables.

Knowledge Base: Embedding and Storing

- Selected a good embedding model for the legal domain using MMTEB from HuggingFace: text-embedding-3-small (OpenAI embedding model).
- Benefits: remote use through API for a low cost, max. input of 8191 tokens and embedding dimension 1536.
- We also considered working with an open-source embedding model (e.g. mxbai-embed-large) and running it locally with Ollama to embed all documents, however the resource consumption is quite high.
- We used the RecursiveCharacterTextSplitter from LangChain to chunk the documents.
- We chunked, embedded and indexed the documents in OpenSearch with the following information: url, chunk id, text and embedding.
- Knowledge base contains 14914 documents.
- Challenges: circuit_breaking_exception from OpenSearch after indexing 99% of the data.

RAG Pipelines

- We experimented with two retrieval strategies:
  - Simple Retriever
  - Retriever using Maximum Marginal Relevance
- The simple retriever uses the default retrieval algorithm from OpenSearch, i.e HNSW for k-NN.
- The retriever with the maximum marginal relevance strategy balances relevance and **diversity**:

$$\boldsymbol{MMR(D_i) = argmax_{D_i \in R \setminus S}[\lambda Sim(D_i, Q) - (1-\lambda) max_{D_j \in S} Sim(D_j, D_i)]}$$

- It selects documents which are similar to the query Q, but disimilar to the already selected documents in S.
- We then use gpt-4o-mini to generate a response to a query given the retrieved documents.

RAG Pipelines

- We also make sure to prompt the generative model to diversify its answer:

Question: {question}
The following documents provide relevant information: **{context}**
Please answer the question only by using the provided information. Make sure to provide a diversified response that covers different perspectives and details from the provided documents. Your answer should include multiple viewpoints and insights from the context, not just a single perspective. If necessary, highlight different interpretations, opinions, or additional context that is relevant to the question. Answer the question comprehensively, using the information from the documents provided.

- We use LangChain to construct the pipeline combining retrieval and generation.

Evaluation

- Unfortunately, current evaluation frameworks like RAGAS, DeepEval, Arize etc. do not have metrics which evaluate diversity in answer generation or in retrieval.
- One way to evaluate diversity at retrieval level is by using the *alpha*-nDCG metric. This metric considers several intents of the given question and then reflects the value of a document based on the number of intents it covers.
- Exact formula in Clarke, Charles LA, et al. "Novelty and diversity in information retrieval evaluation.“.
- There are few implementations of the metric and it is often unclear how to use them.
- We have implemented our own version of the metric, however it is incomplete because it requires to have a so-called ideal cumulative gain and computing it is an NP-complete problem.

Evaluation

- Therefore, we only evaluate the RAG pipeline using classic metrics such as: Context Precision, Context Recall, Context Relevancy, Answer Relevancy and Faithfulness.
- RAGAS raised several problems for evaluation: numerous timeout errors, rate limit errors (both for open and closed models) and high time consumption.
- Therefore, we switched to DeepEval, an open-source evaluation framework.
- On a sample of 50 questions from the ground truth dataset we obtain the following results (with simple retriever here):
  - Contextual Precision: 0.803333
  - Contextual Recall: 0.643244
  - Contextual Relevancy: 0.677705
  - Answer Relevancy: 0.891988
  - Faithfulness: 0.961027

Conclusions and Future Work

- Evaluating diversity at retrieval or generation level remains difficult due to the lack of metrics or available implementations.
- Current ways to encourage diversification are in the retrieval strategy and in prompt engineering.
- Another way of diversifying answer generation would be to cluster the retrieved documents according to covered topics and create an answer for each such cluster.
- Methods from search result diversification could potentially be employed at retrieval level in our setting.

QUESTIONS?