



# Better Retrieval for Generation

Ankit Saha - AI21BTECH11004

Pranav Balasubramanian - AI21BTECH11023



# Problem Statement

## **What is Retrieval-Augmented Generation (RAG)?**

To address the issue of Large Language Model (LLM) hallucinations and irrelevant outputs, we provide relevant information from a set of retrieved passages (related to the query) to provide the LLM with the necessary context to keep the answer grounded to the query being asked.

This project aims to develop better retrieval strategies that out-perform a set baseline in terms of retrieving and generation quality.



## Datasets

1. [MS MARCO](#) - Retrieval-based question-answer dataset in English
2. [Indic MARCO](#) - Retrieval dataset in 11 Indian languages



## MS MARCO

- [Dataset homepage](#)
- [Dataset description and download](#)
- Retrieval-based question answering dataset maintained by Microsoft
- Task is to generate a sound answer given a query (question) and a list of 10 related passages (retrieval)
- Dataset contains tuples of
  - answers
  - list of 10 relevant passages
  - index of best passage
  - query
  - type of query (e.g., description, numeric, entity, person)

# MS MARCO Examples



**query:** price to install tile in shower

**query\_type:** numeric

**passages:**


"In regards to tile installation costs, consumers can expect to pay an average of \$25 per square foot, depending on the grade of material that is used. For a medium-sized shower, the price can cost about \$2,000. Tile installation materials

.....

More luxurious showers can cost in excess of \$10,000 to install.", "1 Install ceramic tile floor to match shower-Average prices for installation are between \$11 to \$22 per square foot; 2 A light/fan combination-Averages at \$180 and one hour of installation; 3 Insulate and re-finish ceilings and walls-Fiberglass wall insulation with R-30 value will cost \$2.25 per square foot."

**answer:** \$11 to \$22 per square foot

# MS MARCO Examples



**query:** what is rappelling

**query\_type:** description

**passages:**

"Rappelling is the process of coming down from a mountain that is usually done with two pieces of rope. Use a natural anchor or a set of bolts to rappel from with help from an experienced rock climber in this free video on rappelling techniques. Part of the Video Series: Rappelling & Rock Climbing.",

.....

a descent of a vertical cliff or wall made by using a doubled rope that is fixed to a higher point and wrapped around the body. abseil. mountain climbing, mountaineering-the activity of climbing a mountain. descent-the act of changing your location in a downward direction."

**answer:** Rappelling is the process of coming down from a mountain that is usually done with two pieces of rope.



## Literature Survey

- [1] [Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks](#)
- [2] [Dense Passage Retrieval for Open-Domain Question Answering](#)
- [3] [Mixtral of Experts](#)
- [4] [MS MARCO: A Human Generated MACHine Reading COMprehension Dataset](#)
- [5] [Searching for Best Practices in Retrieval-Augmented Generation](#)
- [6] [IndicIRSuite: Multilingual Dataset and Neural Information Models for Indian Languages](#)



## Approaches in [5]

- Query classification
- Chunking
- Passage reranking
- Passage repacking
- Passage summarization
- Vector databases
- Retriever-generator fine-tuning





## Proposed Novel Ideas

1. Likelihood ratio test for retrieved passage relevance
2. Mixture of experts for retrieval
3. Create prompt templates for generating responses based on retrieved documents



## Likelihood ratio test for retrieved passage relevance

- The paper [\[5\]](#) involves *query classification*, which takes a query as input and outputs whether or not retrieval is required for generating the answer
- We propose to add another layer of filtering after the query passes through the above classifier and relevant passages are retrieved
- This filtering is based on hypothesis testing, which tests whether the retrieved passages are really relevant to the query or not
- Inspired from [GMM-UBM models](#) used in speaker identification and verification



## Likelihood ratio test for retrieved passage relevance

$H_0$ : The query  $q$  is relevant to the retrieved passage  $c$

$H_A$ : The query  $q$  is irrelevant to the retrieved passage  $c$

$$\frac{p(q|c)}{p(q|\bar{c})} \begin{cases} \geq \theta & \text{accept } H_0 \\ < \theta & \text{reject } H_0 \end{cases}$$

- $\theta$  is the decision threshold for accepting or rejecting the null hypothesis
- $p(q|c)$  is the probability that query  $q$  comes from passage  $c$
- $p(q|\bar{c})$  is the probability that query  $q$  comes from the set of all passages except  $c$



## Mixture of experts for retrieval

- Inspired from the paper [\[3\]](#)
- We propose to use a set of retrievers to fetch the context related to query
- Each retriever is an expert for retrieving documents based on domain of query, for example, Retriever-1 is good at retrieving context for healthcare related queries, Retriever-2 is good at good at retrieving context for legal domain related queries
- Two approaches to get the retrieved context:
  - A router is initially used to forward the query to the retriever based on the domain which can best retrieve the context for it
  - Query is passed to the top-k or all the retrievers and a combination of the outputs can be used as the context



## Create prompt templates for generating responses

- After retrieving the context, our next step is to generate an answer faithful to the query and the context
- We propose to design prompts which include the query, the context and the expected answer (from the dataset) to instruction fine-tune the generator
- The aim here is to create better and effective prompts based on the input query to improve generation by making it more relevant, accurate and less ambiguous



## Baselines

- For retrieval performance
  - Basic RAG pipeline without applying the proposed ideas
- For generation performance
  - Basic RAG pipeline like above
  - Baselines in [\[1\]](#)
  - Baselines in [\[4\]](#)



## Plan of Action

- Decide on models for retriever and generator
- Deploy this pipeline and evaluate using appropriate metrics
- Apply the ideas proposed and compute new evaluation metrics
- Compare the results against the initially deployed pipeline and previous works in literature



## Evaluation Metrics

- <https://docs.confident-ai.com/docs/guides-rag-evaluation>
- Evaluation metrics for retrieval
  - Contextual Precision
  - Contextual Recall
  - Contextual Relevancy
- Evaluation metrics for generation
  - Answer Relevancy
  - Faithfulness
  - ROUGE-L (for comparison with baseline)
  - BLEU-1 (for comparison with baseline)



# Retriever evaluation metrics

- Contextual Precision:

$$\text{Contextual Precision} = \frac{1}{\text{Number of Relevant Nodes}} \sum_{k=1}^n \left( \frac{\text{Number of Relevant Nodes Up to Position } k}{k} \times r_k \right)$$

- Contextual Recall:

$$\text{Contextual Recall} = \frac{\text{Number of Attributable Statements}}{\text{Total Number of Statements}}$$

- Contextual Relevancy:

$$\text{Contextual Relevancy} = \frac{\text{Number of Relevant Statements}}{\text{Total Number of Statements}}$$

# Generator Evaluation metrics



- Answer Relevancy

$$\text{Answer Relevancy} = \frac{\text{Number of Relevant Statements}}{\text{Total Number of Statements}}$$

- Contextual Relevancy

$$\text{Contextual Relevancy} = \frac{\text{Number of Relevant Statements}}{\text{Total Number of Statements}}$$

- ROUGE-L

- BLEU-1