NLP M3 REPORT
2 QA Models with Retrieval

The objective of that milestone was to experiment with 2 different models and compare the accuracies between them, utilizing the concept of retrieval question answering where the model goes to find the proper context to use it in answering the question instead of directly giving it the context, both experiments were conducted on the ***squad_v2*** QA dataset a subset of [5000] entries.

(question) → [Retriever] → top-k relevant contexts → [Reader Model] → answer

Experiment 1:

Pretrained model: *deepset/roberta-base-squad2*
Already trained on SQuAD.
We use it as-is with a QA pipeline.
Steps were as follows:

Wrap Contexts as LangChain Documents: Wraps each unique context string as a Document, tagging it with a source ID.

Embed and Store with FAISS: Creates embeddings for all documents and stores them in a FAISS index for fast similarity search. (FAISS = docs embedded)

Load QA Pipeline Model: Loads *deepset/roberta-base-squad2* into a HuggingFace QA pipeline, and wraps it for LangChain use. And adds it to the pipeline right away.

Create a LangChain QA Chain with Retrieval: Connects the retriever (FAISS + context chunks) with the QA model to form a Retrieval-Augmented Generation (RAG) setup. STARTS TO CONNECT THE RETRIEVAL WITH THE MODEL, through the pipeline that we just created in the previous step (LLM).

Prediction and Evaluation: We employed the evaluate library to utilize the official SQuAD evaluation metrics for assessing the QA model's performance. The main metrics used are:

Exact Match (EM): Measures the percentage of predictions that exactly match the ground truth answers. It is a strict metric, requiring a perfect match.

F1 Score: Computes the harmonic mean of precision and recall at the token level between the prediction and ground truth, allowing for partial matches. This metric captures the degree of overlap and is more forgiving than EM.

These two metrics — **EM** and **F1** — are widely regarded as the standard for evaluating extractive Question Answering tasks on datasets like SQuAD due to their balance between strict correctness and partial credit.

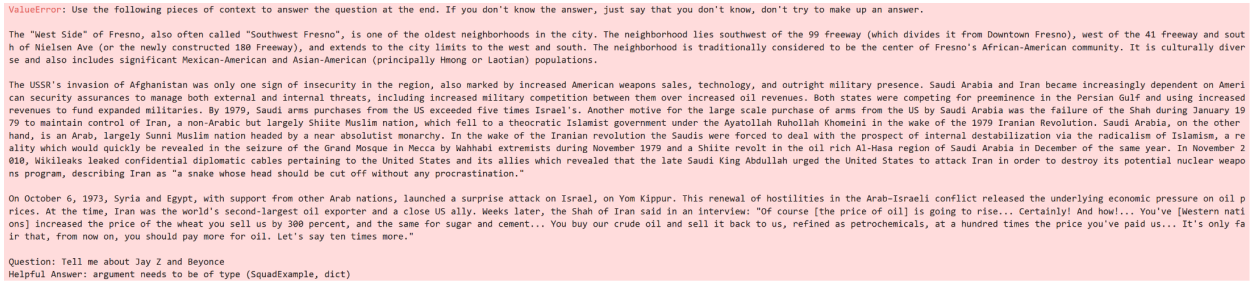
Additionally, we computed **ROUGE** metrics (rouge1 and rougeL) to analyze the quality of predicted answers in terms of lexical overlap and sequence similarity. ROUGE complements EM and F1 by providing insights into the similarity of answer wording and phrasing, especially useful for assessing partial or near-miss predictions.

Results:

Exact Match (EM): 20%
F1 Score: 21.42%
ROUGE-1: 17.5%
ROUGE-L: 17.9%

The **EM** and **F1** scores indicate the model's ability to predict precise and partially correct answers, while **ROUGE** scores give an additional perspective on textual similarity. These results highlight areas for improvement in both retrieval accuracy and answer extraction quality.

Use qa_chain



```
ValueError: Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.

The "West Side" of Fresno, also often called "Southwest Fresno", is one of the oldest neighborhoods in the city. The neighborhood lies southwest of the 99 freeway (which divides it from Downtown Fresno), west of the 41 freeway and sout
h of Nielsen Ave (or the newly constructed 180 Freeway), and extends to the city limits to the west and south. The neighborhood is traditionally considered to be the center of Fresno's African-American community. It is culturally diver
se and also includes significant Mexican-American and Asian-American (principally Hmong or Laotian) populations.

The USSR's invasion of Afghanistan was only one sign of insecurity in the region, also marked by increased American weapons sales, technology, and outright military presence. Saudi Arabia and Iran became increasingly dependent on Ameri
can security assurances to manage both external and internal threats, including increased military competition between them over increased oil revenues. Both states were competing for preeminence in the Persian Gulf and using increased
revenues to fund expanded militaries. By 1979, Saudi arms purchases from the US exceeded five times Israel's. Another motive for the large scale purchase of arms from the US by Saudi Arabia was the failure of the Shah during January 19
79 to maintain control of Iran, a non-Arabic but largely Shiite Muslim nation, which fell to a theocratic Islamist government under the Ayatollah Ruhollah Khomeini in the wake of the 1979 Iranian Revolution. Saudi Arabia, on the other
hand, is an Arab, largely Sunni Muslim nation headed by a near absolutist monarchy. In the wake of the Iranian revolution the Saudis were forced to deal with the prospect of internal destabilization via the radicalism of Islamism, a re
ality which would quickly be revealed in the seizure of the Grand Mosque in Mecca by Wahhabi extremists during November 1979 and a Shiite revolt in the oil rich Al-Hasa region of Saudi Arabia in December of the same year. In November 2
010, Wikileaks leaked confidential diplomatic cables pertaining to the United States and its allies which revealed that the late Saudi King Abdullah urged the United States to attack Iran in order to destroy its potential nuclear weapo
ns program, describing Iran as "a snake whose head should be cut off without any procrastination."

On October 6, 1973, Syria and Egypt, with support from other Arab nations, launched a surprise attack on Israel, on Yom Kippur. This renewal of hostilities in the Arab-Israeli conflict released the underlying economic pressure on oil p
rices. At the time, Iran was the world's second-largest oil exporter and a close US ally. Weeks later, the Shah of Iran said in an interview: "Of course [the price of oil] is going to rise... Certainly! And how!... You've [Western nati
ons] increased the price of the wheat you sell us by 300 percent, and the same for sugar and cement... You buy our crude oil and sell it back to us, refined as petrochemicals, at a hundred times the price you've paid us... It's only fa
ir that, from now on, you should pay more for oil. Let's say ten times more."

Question: Tell me about Jay Z and Beyonce
Helpful Answer: argument needs to be of type (SquadExample, dict)
```

Experiment 2:

Model to be fine-tuned: ***t5-small***
Not yet fine-tuned for QA.
We will fine-tune it using the SQuAD training set.

Steps are as follows:

Load and Prepare SQuAD v2 Dataset: Import all necessary libraries for data loading, model building, retrieval, and evaluation.

Select a Subset of SQuAD v2 for Training and Validation: Load the SQuAD v2 dataset and limit the size to 5,000 samples for both training and validation to reduce training time during experimentation.

Create a Dense Retriever with FAISS and Sentence Transformers: Build a FAISS vector store to retrieve relevant context paragraphs for each question using dense embeddings.

Prepare Input-Target Pairs for the Model Using Retrieved Contexts: Use the retriever to fetch relevant contexts and prepare inputs in the format expected by the T5 model: "question: ... context: ..." → "answer"

Load T5 Tokenizer: Load the tokenizer corresponding to the t5-small model to convert text to token IDs.
Tokenize Input and Target Texts: Tokenize the inputs and targets for training and evaluation. Ensures consistent sequence length for batching.

Compute Evaluation Metric (F1 & Exact Match): Compute SQuAD-style F1 and Exact Match metrics by comparing model predictions to ground truth answers.

Set Training Arguments: Define training configurations such as batch size, number of epochs, and whether to use mixed-precision (fp16) if a GPU is available.

Load T5 Model and Create Trainer: Initialize the *T5ForConditionalGeneration* model and wrap it with *Seq2SeqTrainer* for training and evaluation.

Train the Model: Start fine-tuning the T5 model on the retrieved-question-answer dataset.

Evaluate: We use the same *evaluate* library to load the official SQuAD metric script as we did in **Experiment 1**

Results:

Exact Match (EM): *12.84%*
F1 Score: *13.99%*
ROUGE-1: *14.38%*
ROUGE-L: *14.30%*