Advanced NLP Assignment 2: End-to-end NLP System Building

Ansh Khandelwal **Shrey Madeyanda**
{anshk, smadeyan}@andrew.cmu.edu

1 Data Creation

1.1 Knowledge Base Compilation

Our knowledge resources were compiled with a focus on Carnegie Mellon University and its Language Technologies Institute (LTI), its faculty, research output, teaching materials, and other university-related information. To keep out knowledge base relevance and within a manageable scope, we primarily included documents and data sources recommended for their direct relation to LTI and CMU's broader academic and social landscape. The sources included faculty listings from the LTI faculty directory, research papers published by them (going back to 2023), detailed course offerings across CMU departments, academic calendars, program-specific details from LTI, and significant events and historical information about CMU.

1.2 Data Extraction Process

The raw data was extracted using different tools based on the format of the original source. BeautifulSoup, a Python library for extracting data from HTML files, was used to parse the webpages. PDF documents, such as research papers and course handbooks, were parsed using PyPDF2. Research papers, and their corresponding metadata, for the LTI faculty were downloaded through the Semantic Scholar API.

1.3 Data Organization

The extracted raw data was further processed and organized as follows:

Research Papers The research paper PDFs were parsed into text files, with each file then split into page-wise chunks. Metadata (title, authors, year and tldr) was appended to the start and end of each page to provide a richer context for retrieval.

Course Schedules Course schedule webpages were parsed into text files, in a tabular format. They were then segmented into chunks of 60 rows each, with columns names added to each row for contextual clarity.

Program Handbooks The LTI course handbooks were divided into page-wise chunks. Unlike the research papers, no metadata was directly added to each chunk. Instead, relevant headers were added to help guide retrieval.

1.4 Data Annotation

We have opted not to finetune our RAG question answering system, and hence did not annotate any training data. For the purpose of testing our system we annotated a dataset comprising question-answer pairs across various knowledge categories. The following are some details about the test set:

Test Set Size A total of 60 question-answer pairs covering the following categories: academic schedules, course information, LTI program information, LTI as a whole, research publications, and general facts about CMU. See Figure 1 for the distribution among the categories.

Annotation Sources The questions were derived from the knowledge sources mentioned earlier, to ensure a broad representation.

Annotation Strategy The decision on the kind and amount of data to annotate was guided by the objective to cover a comprehensive range of topics and ensure a good representation of the diverse information sources. Question answer pairs were included from each knowledge sources in each category mentioned above.

Annotation Interface Our test dataset underwent manual annotation. Each annotator made multiple question-answer pairs, and the final test set was curated by taking a subset of these.

Since we did not opt to go for finetuning, we did not annotate any training data.

Category	Annotator A	Annotator B
Relevant	21	24
Partially Relevant	9	6
Irrelevant	0	0

Table 1: Results of paired t-test for selected model and baseline

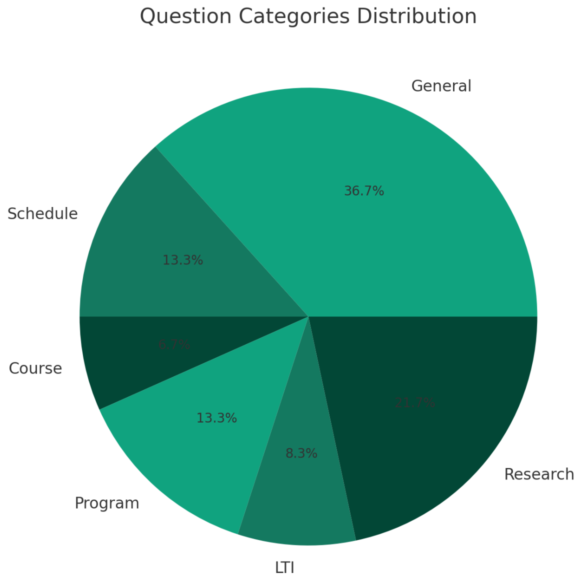

Figure 1: Rouge-1 Scores for Categories across Models

1.5 Estimating Annotation Quality

To estimate the quality of our annotations, we implemented an inter-annotator agreement (IAA) assessment involving ourselves as the two annotators (we are a team of two). Each annotator was tasked with evaluating 30 question-answer pairs selected from our test set. The pairs were selected in a manner that maintained the distribution across categories seen in the overall test set. This evaluation process was structured around categorizing each question-answer pair into one of three categories: Relevant, Partially Relevant, and Irrelevant. This categorization framework allowed us to measure the degree of relevance and accuracy of the annotated data in relation to the source material and the precision of the query.

Cohen's Kappa was employed as the metric for quantifying the level of agreement between the annotators, as it accounts for the possibility of the agreement occurring by chance.

In our evaluation rubric, we prioritized the conciseness of the answers and their relevance to the posed questions. We acknowledged that the optimal length of an answer could vary depending on the nature of the question; for instance, queries related to research papers might necessitate more detailed responses.

Table 1 shows the categorization of the question-answer pairs by the annotators. This gives a Cohen's Kappa value of 0.734, which shows substantial agreement among the annotators in their categorization.

2 Model Details

We experimented with various combinations of embeddings (for document retrieval) and reader models (for context-based question-answering) to optimize performance. Our aim was to identify the most effective pairing that could accurately handle a wide range of queries, including those necessitating detailed technical responses. Below, we detail the methodologies employed, our rationale for selecting these approaches, and the configurations that yielded the most promising results.

2.1 Embedding Techniques

For the vector storage of our embeddings, and hence for our retriever, we utilized the Facebook AI Similarity Search (FAISS) library, for its efficiency in clustering and searching large scale vectors. It significantly reduces search time for nearest neighbors in high-dimensional spaces, an advantage crucial for quickly retrieving relevant information in a QA system. We evaluated two different embeddings.

2.1.1 all-MiniLM-L6-v2

This model, available at all-MiniLM-L6-v2 from Hugging Face, is characterized by its embeddings of 384 dimensions. It supports a maximum of 256 tokens. For FAISS storage, this setup took approximately 69 MB.

2.1.2 UAE-Large-V1 (UAE)

The UAE embeddings were generated with the model from UAE-Large-V1 on Hugging Face, with a model size of 1.34 GB, generate embeddings in 1024 dimensions and accommodate up to 512 to-

Score	UAE + LLama Q5	MiniLM + FlanT5	p-value
ROUGE 1	0.2973	0.1842	0.0060
ROUGE 2	0.1817	0.0950	0.0172
ROUGE L	0.2802	0.1800	0.0115

Table 2: Results of paired t-test for selected model and baseline

kens. The FAISS storage footprint for UAE was around 130 MB.

Our experiments spanned various chunk sizes (512 to 2048 tokens) and overlap sizes, with the optimal configuration identified as 1024 tokens for chunk size and 100 tokens for overlap. The larger embedding dimensions and token capacity of the UAE model presented clear advantages, notably the ability to encapsulate more detailed information within a single embedding and to process larger portions of text at once, leading to our selection of UAE for further experiments.

2.2 Reader Models

For the reader component, which needs to answer the given questions using the documents provided by the retriever as context, we focused on evaluating models capable of extractive summarization. We experimented with:

2.2.1 flan-t5

We utilized the sjrhuschlee/flan-t5-large-squad2 variant from Hugging Face, which is a fine-tuned version of flan-t5-large on the SQuAD2.0 dataset. It is adept at handling both answerable and unanswerable questions, which is useful for our diverse query set.

2.2.2 llama-7b (5-bit quantized)

Due to the substantial size of llama-7b, we opted for a quantized version hosted at TheBloke/Llama-2-7B-Chat-GGUF on Hugging Face, llama-2-7b-chat.Q5KM.gguf, which strikes a balance between maintaining quality and minimizing memory footprint. It was quantized using llama.cpp.

2.3 Baseline and Final Selection

As a baseline, we employed the MiniLM embeddings in conjunction with the flan-t5 reader model. Despite its relatively low performance, this combination served as a valuable reference point, highlighting areas for improvement and guiding our subsequent experiments.

After testing with different combinations of embeddings and readers, we ultimately selected the UAE embeddings paired with the llama-2-7b reader. This configuration was better for generating detailed, technical responses required for complex queries. The verbose nature of the answers provided by llama-2-7b was deemed beneficial for questions demanding longer explanations (such as questions on explaining the focus of research papers), a capability where flan-t5 was found lacking.

3 Results

In this report, we evaluate the performance of a Retrieval-Augmented Generation (RAG) model by comparing its generated answers against a set of reference answers manually curated (test set). We assess the similarity between the generated answers and reference answers using various evaluation metrics. The primary goal is to determine the effectiveness of the RAG model in generating answers that align closely with the provided references.

3.1 Evaluation Metrics

F1 Score The F1 score, a common evaluation metric in natural language processing tasks, provides a balanced measure of a model's performance by considering both precision and recall. To compute the F1 score between a predicted answer and a ground truth answer, the text is first normalized and tokenized. The number of common tokens between the prediction and ground truth is then determined, from which precision and recall are calculated. Finally, the F1 score is computed as the harmonic mean of precision and recall, providing a single numerical value that reflects the model's ability to generate accurate and relevant answers. A higher F1 score indicates better alignment between the generated answers and the reference answers.

Rouge Scores (Rouge-1, Rouge-2, Rouge-L) Rouge scores evaluate the overlap between the model's predictions and the reference answers at the unigram, bigram, and longest common subsequence levels. Rouge-1 measures the overlap of unigrams, Rouge-2 measures the overlap of bigrams, and Rouge-L measures the overlap of the longest

Model	F1 Score	ROUGE-1	ROUGE-2	ROUGE-L	Recall Score
(M1) UAE + LLama Q5	**0.31**	**0.30**	**0.18**	**0.28**	**0.53**
(M2) UAE + LLama Q3	0.28	0.28	0.17	0.27	0.52
(M3) Only LLama Q5	0.09	0.11	0.02	0.10	0.37
(M4) Only LLama Q3	0.08	0.10	0.02	0.10	0.35
(M5) UAE + FlanT5	0.30	0.26	0.18	0.26	0.26
(M6) MiniLM + LLama Q5	0.26	0.26	0.15	0.25	0.43
(M7) MiniLM + FlanT5	0.19	0.18	0.10	0.18	0.17

Table 3: Performance of Various RAG systems on the curated test set

common subsequences. Higher Rouge scores indicate better agreement between the generated answers and the reference answers.

Recall Recall measures the ability of the model to find all relevant information in the reference answers. It is the ratio of the number of correct positive results to the number of all relevant samples. A higher recall score indicates that the model is effectively capturing relevant information from the reference answers.

3.2 Methodology

We conducted the evaluation by comparing reference answers to the generated answers by different models. Each file contains reference answers for the same set of questions, with each line corresponding to an answer. We used the following steps to compute the evaluation metrics:

1. Loaded the reference answers from both files.
2. Tokenized the answers and performed necessary preprocessing.
3. Calculated F1 score, Rouge scores (Rouge-1, Rouge-2, Rouge-L), and recall for each pair of corresponding answers from the two files.
4. Computed the average scores for each metric to provide an overall assessment of the model's performance. The average scores for each model are tabulated in Table 1.

3.3 Comparison of Results

To assess the performance improvements of our selected model configuration (UAE + llama-2-7b) over the baseline (MiniLM + flanT5), we employed pairwise t-testing on ROUGE-1, ROUGE-2, and ROUGE-L scores. The statistical analysis, conducted at a significance level of 0.05, is detailed in Table 2 of our report. The results show that the enhancements in performance metrics for our chosen model are statistically significant compared to the baseline. This analysis substantiates our decision to proceed with the UAE embeddings and llama-2-7b reader model, reinforcing their effectiveness in delivering high-quality question-answering capabilities.

4 Analysis

4.1 Overall Average Analysis

The results presented in Table 3 showcase the performance of each model across various metrics. Notably, the combination of UAE embedding with LLama Q5 achieved the highest F1 score of 0.31 and the highest recall score of 0.53, indicating a relatively strong performance in terms of both precision and the ability to capture relevant information.The ROUGE scores for this combination were moderate, with ROUGE-1, ROUGE-2, and ROUGE-L scores of 0.30, 0.18, and 0.28, respectively. These scores were still the highest in comparison to other RAG systems.

On the other hand, employing only the LLama reader without embeddings (closed-book without retrieve-and-augment strategy) yielded lower scores across all metrics compared to the UAE + LLama combination. Specifically, the F1 score dropped to 0.09 for LLama Q5 and 0.08 for LLama Q3, indicating a significant decrease in the model's ability to accurately generate answers. Similarly, the recall scores for these configurations decreased to 0.37 and 0.35, respectively.

The performance disparity between LLama Q5 and LLama Q3 can be attributed to the difference in quantization levels. LLama Q5, with its higher number of quantization levels, provides a finer-grained representation of embeddings compared to LLama Q3. This finer granularity allows LLama Q5 to retain more information during compression, resulting in a more faithful representation of the original embeddings. Consequently, LLama Q5 exhibits superior performance in question answering tasks by better preserving nuances and subtleties in the data, thus highlighting the importance of quantization level selection in optimizing model performance.

Incorporating FlanT5 as the reader model alongside UAE embeddings resulted in a slight improvement in the F1 score compared to only LLama, with a score of 0.30. However, the ROUGE scores remained relatively low for this configuration, suggesting potential limitations in capturing the nuances of the reference answers. Additionally, using MiniLM embeddings with either LLama or FlanT5 led to further reductions in performance across all metrics, indicating the importance of selecting appropriate embeddings for question answering tasks.

4.2 Categorical Analysis

For a more fine-grained evaluation of our models, it becomes imperative to discern the models' effectiveness in handling varied data domains. The test set comprises six distinct categories: general, LTI, research, program, schedule, and course. Each category encapsulates unique challenges and complexities, thereby necessitating a nuanced assessment of model performance. For instance, domains like courses demand meticulous data processing of tables to ensure accurate retrieval by the model. Conversely, categories such as general or research might entail relatively straightforward retrieval tasks, provided appropriate hashing mechanisms are applied to the data. By plotting each model's performance, particularly in terms of ROUGE-1 scores, across these diverse categories, we aim to elucidate the strengths and weaknesses of each model. Such analyses are instrumental in discerning the models' efficacy in different contexts, thereby facilitating informed decisions regarding model selection and refinement strategies.

Upon reviewing the plotted ROUGE-1 scores, we can observe that the UAE + LLamaQ5 model performs best in the LTI category but struggles the most in the Course category. This difficulty in the Course category is likely due to the model's trouble in extracting information from tabular data commonly found in course-related questions. Interestingly, a similar poor performance is seen in the Schedule category, indicating a consistent challenge with tabular data across these two categories.

Moreover, when comparing different models across categories, it's notable that all models perform poorly in the Course category. However, the UAE + FlanT5 model (M5) stands out with a better score in the Schedule category. This could be because FlanT5 tends to provide concise outputs, which may be beneficial in scenarios like schedules where streamlined information retrieval is crucial.

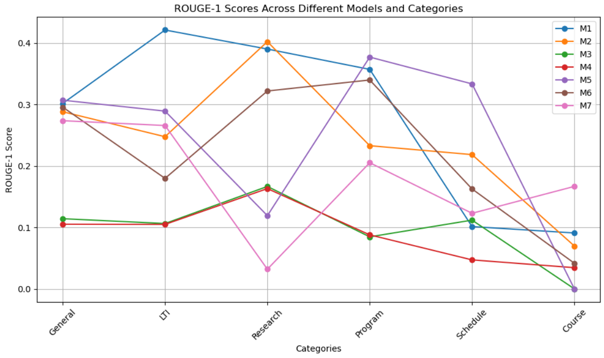

Figure 2: Rouge-1 Scores for Categories across Models

This analysis highlights the importance of understanding how different models interact with specific types of data. It also emphasizes the need for careful consideration when selecting models for different tasks, ensuring they align well with the nature of the data they'll be handling.

4.3 Retrieve-and-augment strategy vs Closed-book Models

From the observations gleaned from Table 3, it becomes evident that there is a stark decline in performance when queries are solely directed to the readers without the incorporation of a knowledge base via retrievers. This decline is reflected in the average performance across all metrics. A notable comparison between the ROUGE-1 scores across various categories for models M3 and M4, in contrast to models M1 and M2, reveals a significant deterioration in performance. Particularly striking are the green and red lines representing M3 and M4, respectively, which consistently exhibit the lowest scores across all categories in the plotted analysis.

This decline in performance can be attributed to the absence of a knowledge base, which deprives the LLama reader of contextual information necessary for accurate query response. Consequently, there is a clear imperative for implementing a RAG strategy to furnish the LLama reader with a robust context for answering queries effectively, especially when those queries get very domain-specific. By integrating retrievers to provide relevant knowledge, the RAG approach promises to enhance the performance of LLama-based question answering

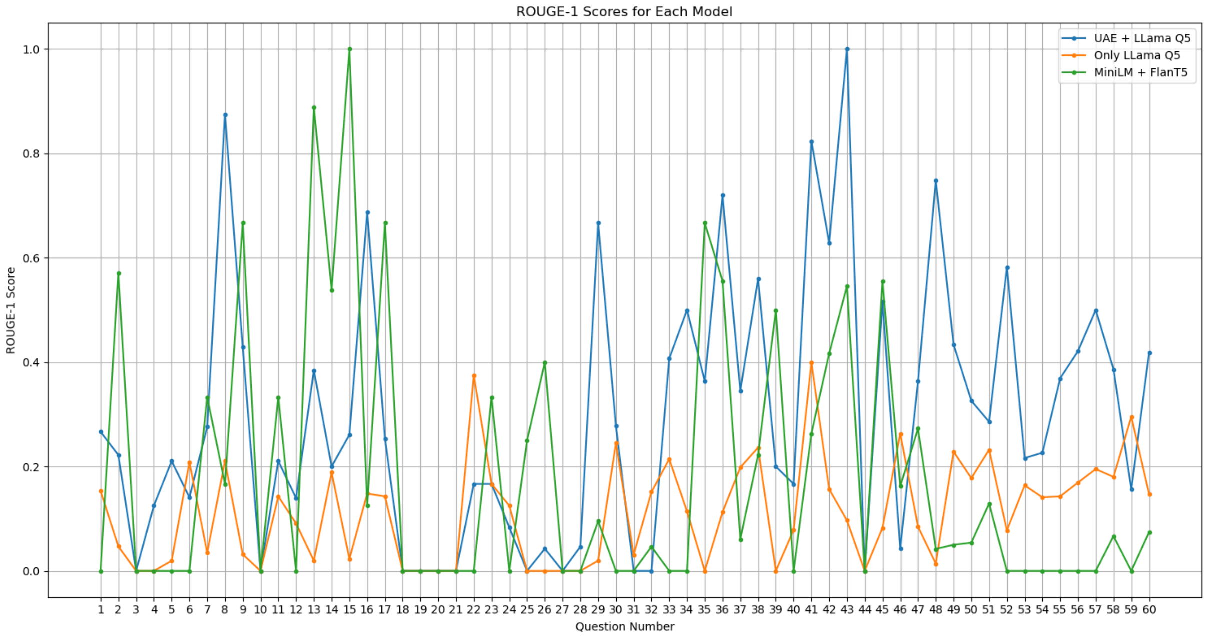

Figure 3: Rouge-1 Scores for all questions-answer pairs of curated test set across 3 models

systems and mitigate the limitations observed in the standalone reader-only approach.

4.4 Qualitative Analysis

For qualitative analysis, a selection of three models has been made: LLamaQ5 + UAE Embedding (Model 1), Only LLamaQ5 (closed book) (Model 2), and Flant5 + MiniLM Embedding (Model 3). These choices were deliberate, aimed at facilitating comparisons across various readers and embeddings, while also considering the performance of a closed book model without retrieval or augmentation strategies.

The plot in Figure 3 illustrates the ROUGE-1 scores for these models across all question-answer pairs in the test set. By examining the scores for each question, we gain insight into the individual performance of these models, thereby facilitating qualitative analysis.

For further investigation, two specific questions have been chosen from the plots. Evaluating the model outputs on these selected questions enables us to delve deeper into their performance variations. This analysis seeks to ascertain whether the employment of retrieval and augmentation strategies indeed influences model performance.

Example 1 For this analysis, we have selected Question 5.
Question: When will CMU's main commencement ceremony take place
Reference Answer: Sunday, May 12
Output Model 1: The main commencement ceremony for Carnegie Mellon University will take place on Sunday, May 12, 2024.
Output Model 2: Einzeln, Inc. CMU's main commencement ceremony is typically held in May of each year. The exact date and time of the ceremony can vary from year to year, but it usually takes place on a Saturday in the middle of the month. CMU's 2023 commencement ceremony is scheduled to take place on Saturday, May 14, 2023, at Heinz Field. The ceremony will begin at 9:30 a.m. and will be livestreamed for those who are unable to attend in person.
Output Model 3: Ceremony10–11:30
Upon comparing these outputs, it becomes evident that Model 1 has performed the best, providing a precise answer aligned with the reference. Conversely, Model 2 has generated irrelevant output, likely due to the absence of retrieval and augmentation strategies. Model 3's output is entirely incorrect, likely because it uses smaller embeddings which provide with lesser context and has a reader with smaller number of parameters (FlanT5). This discrepancy in output quality is reflected in the differences observed in the ROUGE-1 score for Question 5 in the plot. It's noteworthy that even Model 1's ROUGE-1 score isn't significantly high, indicating the reason due to verbosity.

Example 2 For this analysis, we have selected Question 35. **Question:** How many units is the course: Law of Computer Technology?
Reference Answer: 12 units.
Output Model 1: The Law of Computer Technology course is 12 units.
Output Model 2: Law of Computer Technology is a course that covers various legal and ethical issues related to the use of computer technology. The course is designed to provide students with a comprehensive understanding of the legal and ethical implications of computing and information technology.
Output Model 3: 12
In this scenario, Model 3 outperforms the other models, including Model 1, despite providing the correct answer. Model 2, however, fails to provide any accurate answer due to its reliance solely on the external knowledge of its trained data without access to the retrieval knowledge base. Model 3's superior performance over Model 1 can be attributed to its succinct response, which closely aligns with the reference answer. This precision in response likely contributes to a higher level of relevance, resulting in a better match with the reference answer. Consequently, Model 3 achieves a higher ROUGE-1 score for Question 35, despite Model 1 providing the correct answer. This observation underscores the importance of not only providing correct answers but also ensuring the relevance and precision of responses. While verbosity may sometimes yield correct answers, a more concise and precise response, as demonstrated by Model 3, can lead to better quantitative measures, such as ROUGE scores, reflecting the model's effectiveness in capturing the essence of the reference answer.