ANLP Homework 2

Team members: Vinayak Khare, Medha Sinha, Jai Dalvi

Data creation (10 points):

- How did you compile your knowledge resource, and how did you decide which documents to include?

  Ans: We scraped the data from (General Info and History of Pittsburgh/CMU, Events in Pittsburgh and CMU, Music and Culture and Sports) primarily using 3 libraries: BeautifulSoup, Selenium and Pdfplumber.

  We decided to choose documents based on the following:

  - **Diversity of Information** – The links cover a broad range of topics, including history, events, museums, food festivals, and sports, ensuring a well-rounded knowledge base.
  - **Credibility and Authority** – We included reliable sources such as Wikipedia, official city websites, sports, music webpages and reputable encyclopedias (e.g., Encyclopedia Britannica) to ensure accuracy.
  - **Relevance to the Topic** – The selected links focus on general information, history, events, culture, and sports in Pittsburgh and CMU, aligning with my objective
  - **Usefulness for Scraping and Research** – The resources provide structured data that can be easily navigated or extracted, such as event calendars and city regulations.

- How did you extract raw data? What tools did you use?

- **Ans:** We extracted raw data using web-scraping techniques by fetching relevant information from structured web pages. This involved parsing HTML content, identifying key data points such as event listings, historical information, and cultural highlights, and then cleaning the extracted data for usability.
- We used **BeautifulSoup** for parsing and extracting data from HTML pages, **pdfplumber** for scraping data from pdfs and **Selenium** for handling dynamic web pages that require JavaScript execution
- What data was annotated for testing and training (what kind and how much)?

  Ans: We annotated event-related data, historical records, and cultural information extracted from the web pages. The dataset included structured event listings, venue details, and descriptions. The total annotated dataset consisted of approximately 300 records, ensuring a balanced representation across different event categories and topics

- How did you decide what kind and how much data to annotate?

  Ans: We focused on annotating data that was essential for training and evaluation, prioritizing high-frequency event types, key historical facts, sports,music, food programs, cities and cultural references. The amount of data was determined based on the coverage needed for accurate predictions while maintaining a feasible annotation workload.

- What sort of annotation interface did you use?
- How did you estimate the quality of your annotations? (IAA)

Annotator Ratings Table

Query	Ground Truth	Annotator 1 Rating	Annotator 2 Rating
What are the two largest healthcare providers in Pittsburgh?	UPMC and Allegheny Health Network.	Full	Full
What was the first military hospital in U.S. history?	General Edward Hand Hospital, serving from 1777 to 1845.	Partial	Full
Where does the Pittsburgh Symphony Orchestra perform?	Heinz Hall.	Partial	Full
What were Andrew Carnegie's major contributions to the steel industry and philanthropy?	Revolutionized steel industry, sold Carnegie Steel to J.P. Morgan, and funded libraries and education.	Partial	Partial

There are two categories: **Full** and **Partial**.

- **Full/Full:** 1 instance (Query 1)
- **Partial/Partial:** 1 instance (Query 4)
- **Partial/Full:** 2 instances (Queries 2 and 3)
- **Full/Partial:** 0 instances

The observed agreement is the proportion of queries where both annotators agreed:

Observed Agreement=0.5

First, we compute the proportions for each category for each annotator:

- **Annotator 1:**
  - Full: 14=0.2541=0.25
  - Partial: 34=0.7543=0.75
- **Annotator 2:**
  - Full: 34=0.7543=0.75
  - Partial: 14=0.2541=0.25

The expected agreement for each category is the product of the probabilities for that category:

- For **Full:** 0.25×0.75=0.18750.25×0.75=0.1875
- For **Partial:** 0.75×0.25=0.18750.75×0.25=0.1875

Thus, the total expected agreement is:

Expected Agreement=0.1875+0.1875=0.375

Cohen's Kappa is defined as:

κ=Observed Agreement−Expected Agreement1−Expected Agreementκ=1−Expected AgreementObserved Agreement−Expected Agreement

Substitute the values:

κ=0.2

The Cohen's Kappa for the annotator ratings is **0.2**, indicating a slight agreement between the annotators.

Model details (10 points): clearly describe your model(s). Please include the following details,

What kind of methods (including baselines) did you try? Explain at least two variations (more is welcome). This can include variations of models, which data it was trained on, training strategy, embedding models, retrievers, re-rankers, etc.

We used 2 AI models: Mistral ("mistralai/Mistral-7B-Instruct-v0.3") and Llama ("Llama 3.1-8B-Instruct") in our experiment.

Mistral-7B-Instruct-v0.3: This is a 7-billion parameter model fine-tuned for instruction-following tasks. It is optimized for generating structured responses, making it well-suited for answering factual questions, summarizing documents, and precisely following user prompts. Mistral performs strongly in retrieval-augmented generation (RAG) setups, effectively incorporating external knowledge sources before formulating answers.

llama3-8b-8192 is a large language model with 8 billion parameters and an extended 8192-token context window, designed to capture long-range dependencies for improved coherence in tasks such as text generation, summarization, translation, and conversational AI. Its advanced architecture and training strategies strike a balance between computational efficiency and robust performance, making it suitable for both research and practical NLP applications.

Embedding Models:

We experimented with three embedding models for retrieval: all-MiniLM-L6-v2, BAAI/bge-large-en, and mpnet-base-v2. The all-MiniLM-L6-v2 model offers lightweight, efficient sentence embeddings ideal for rapid retrieval tasks with minimal computational overhead. In contrast, the BAAI/bge-large-en model produces higher-quality embeddings that significantly enhance retrieval accuracy for complex queries. Additionally, mpnet-base-v2 leverages the advanced MPNet transformer architecture by combining masked and permuted language modeling, capturing nuanced contextual relationships to support effective semantic search, text clustering, and natural language understanding. This diversified approach enables us to balance efficiency and accuracy based on the specific needs of our application.

Retrieval Approach:

We used cosine similarity to compare the embeddings generated by the models with the query embedding. The retrieval process involved computing the cosine similarity between the query and precomputed document embeddings, followed by retrieving the top-k (e.g., k=3, 5, or 10) most relevant chunks for further processing. Additionally, we experimented with different response modes provided by frameworks such as LlamaIndex. In tree-summarize mode, the retrieved document chunks are recursively summarized in a hierarchical manner, aggregating intermediate summaries to produce a coherent final response. In refine mode, an initial answer is generated from the most relevant chunk and is then iteratively refined by incorporating additional context from subsequent chunks. Furthermore, we integrated LlamaIndex and LangChain into our pipeline to streamline document indexing and query processing, while few-shot prompting methods were employed to further enhance the contextual relevance and specificity of generated responses. These varied response strategies allow us to tailor the output to the complexity of the query, enhancing the overall quality and relevance of the final response.

What was your justification for trying these methods?

Ans: We selected these methods to balance efficiency and accuracy in our retrieval system. Cosine similarity is computationally efficient and works effectively with dense vector embeddings, making it ideal for comparing question embeddings with document chunks. Utilizing top-k retrieval ensures that only the most relevant information is selected, while our choice of embedding models—from the fast all-MiniLM-L6-v2 to the higher-quality BAAI/bge-large-en—allows us to optimize for both speed and precision depending on the task at hand. Moreover, incorporating different response modes like tree-summarize and refine enables the system to generate outputs that are appropriately detailed and coherent, adapting the final response to the complexity and nuance of the user query. The use of LlamaIndex, LangChain, and few-shot prompting methods further enhanced our system's capability to effectively index documents, process queries, and generate context-aware responses.

3. Results (10 points): report raw numbers from your experiments. Please include the following details,

We evaluated two models on our testing data: Mistral-7B-Instruct-v0.3 and Llama 3.1 8B. The evaluation focused on both F1 score and Exact Match (EM) metrics.

- **Mistral-7B-Instruct-v0.3:**
  - **F1 Score:** 0.22
  - The model often captured partial correctness but frequently omitted critical details, resulting in a lower F1 score.
  - It struggled with generating fully structured responses and often produced fragmented or incomplete answers.
  - Additionally, the model showed difficulty in distinguishing between closely related entities, leading to minor factual inaccuracies.
- **Llama 3.1 8B:**
  - **F1 Score:** 0.44
  - **Exact Match Score:** 0.13
  - Although the Llama model achieved a higher F1 score—indicating it generally captured more of the relevant information—the Exact Match score was lower. This suggests that while Llama can generate answers with many correct components, it often does not perfectly replicate the ground truth answer.

Statistical Significance

We conducted statistical tests (using paired t-tests/bootstrapping) to compare the performance of the two models. The difference in F1 scores between Mistral (0.22) and Llama (0.44) was found to be statistically significant ($p < 0.05$). This indicates that the Llama model's superior F1 performance is not due to random variation, even though its

Exact Match performance shows that there remains room for improvement in generating fully accurate, complete responses.

4. Analysis

- **Perform a comparison of the outputs on a more fine-grained level than just holistic accuracy numbers, and report the results. For instance, how did your models perform across various types of questions?**

1. Quantitative Overview

Llama 3.1 8B (First Set)

- **Strengths:**
  - Achieved perfect scores (Exact Match and F1 = 1.0) on many templated or straightforward questions (e.g., ice cream flavors, event titles, museum names).
  - Many responses were close to the ground truth, capturing key facts when the evidence was explicit.
- **Challenges:**
  - Several questions—particularly those requiring additional context or less prominent details—resulted in partial matches or "I don't know" answers.
  - F1 scores for nuanced queries ranged from around 0.0 up to 1.0, showing high variability.

Mistral-7B-Instruct-v0.3 (Second Set)

- **Strengths:**
  - Demonstrated moderate performance on some structured queries (e.g., identifying healthcare providers or listing the two airports), with the highest individual F1 reaching about 0.57.
- **Challenges:**
  - The average F1 score was approximately 0.225, indicating that many responses were missing key details.
  - Many questions—especially in specialized domains like historical facts, regulatory details, and nuanced cultural or geological topics—received very low F1 scores (often well below 0.1).

---

2. Qualitative Analysis by Question Type

A. Factual/Historical Queries

- **Llama 3.1 8B:**
  - **Strengths:** Generally good at retrieving direct factual details such as invention dates or event-specific data when evidence is clearly present.
  - **Opportunities:** Some answers omitted contextual qualifiers (e.g., location details like "at a drugstore"), leading to lower scores on nuanced queries.
- **Mistral-7B-Instruct-v0.3:**
  - **Strengths:** Capable of extracting lists (e.g., multiple healthcare providers) when the facts are explicit.

- **Opportunities:** Often struggled with less-prominent details, with several historical queries receiving very low scores. This suggests that retrieval or synthesis of nuanced historical context may be less robust.

B. Event Details and Organizational Information

- **Llama 3.1 8B:**
  - **Strengths:** Frequently delivered nearly exact matches on templated questions like event titles, venues, and organizational names.
  - **Opportunities:** Some responses missed minor but important details (e.g., full addresses or complete time ranges) that prevented perfect matches.
- **Mistral-7B-Instruct-v0.3:**
  - **Strengths:** Performed adequately on questions with clearly defined structural data (such as listing interstates or airports).
  - **Opportunities:** Often produced answers that were incomplete or vague for questions requiring precise event details (e.g., ticket deadlines or exact regulatory references).

C. Specialized Domains (Healthcare, Sports, Geology, Regulatory Information)

- **Llama 3.1 8B:**
  - **Strengths:** Successfully identified key entities in specialized domains when the evidence was available, although performance varied.
  - **Opportunities:** Some specialized queries (e.g., professional wrestling careers or detailed policy questions) received low scores due to missing critical contextual elements.
- **Mistral-7B-Instruct-v0.3:**
  - **Strengths:** Showed moderate success in extracting lists and numeric facts when clearly provided.
  - **Opportunities:** Tended to struggle with nuanced or multi-part queries in these domains—many answers were overly abbreviated, leading to F1 scores below 0.2 in several cases.

3. Comparative Insights

- **Overall Accuracy and Consistency:**
  - **Llama 3.1 8B** generally demonstrated a higher ceiling—achieving several perfect matches on templated queries—while also exhibiting variability on more context-dependent questions.
  - **Mistral-7B-Instruct-v0.3** produced lower average performance (with an average F1 of about 0.225), indicating more frequent omissions or incomplete answers, particularly in complex or specialized topics.
- **Templated vs. Open-Ended Questions:**
  - Both models excel on templated, fact-based queries. However, Llama 3.1 8B more consistently delivered high-quality answers for these questions.
  - For open-ended or synthesis questions that require integrating multiple details, both models showed challenges, but the drop in performance was more pronounced with Mistral-7B-Instruct-v0.3.
- **Domain-Specific Performance:**
  - In historical and event-detail questions, Llama 3.1 8B's outputs were closer to the ground truth, though some context was occasionally missed.
  - In specialized domains like healthcare details, geological processes, or regulatory requirements, both models showed room for improvement—but Mistral-7B-Instruct-v0.3, in particular, struggled to capture the necessary nuance.

Perform an analysis that evaluates the effectiveness of retrieve-and-augment strategy vs closed-book use of your models

Both strategies were evaluated using an F1 metric that measures overlap between the generated answer and the provided ground truth. In our comparison:

- **Closed-Book LLM:** Answers were generated without any external retrieval. These responses were based solely on the model's internal knowledge.
- **RAG Approach:** Answers were produced with the help of retrieved documents or evidence, allowing the model to ground its responses in external data.

Key Observations

1. **Accuracy on Direct Facts:**
    - **Closed-Book:** For early queries, the closed‑book responses were nearly perfect (F1 scores of 1.0 in several cases). However, for later queries, many answers were either incorrect or did not match the ground truth (many F1 scores of 0).
    - **RAG:** In contrast, while RAG sometimes produced "I don't know" on queries where evidence was lacking , it also provided correct or near‑correct details on many queries . This led to several high F1 scores even when an exact string match was not achieved.
2. **Fallback Behavior:**
    - **Closed-Book:** The model tended to provide an answer even if it was partially incorrect, which sometimes resulted in a misleading but confident response.
    - **RAG:** The RAG system sometimes returned "I don't know" for queries where the retrieved evidence was insufficient, avoiding the risk of generating factually incorrect details. While this yields a 0 F1 for that query, it may be preferable from a trust and safety perspective.
3. **Overall Average F1:**
    - **Closed-Book Model:** The average F1 score across 55 queries was calculated to be approximately **0.22**.
    - **RAG Model:** The average F1 score was approximately **0.67**—more than double that of the closed-book model. This suggests that incorporating retrieval helps the model provide answers that are closer to the ground truth overall.
4. **Query-Specific Variations:**
    - In some cases , the RAG approach provided answers that were correct in essence but differed in phrasing (yielding F1 scores of ~0.57 and ~0.33, respectively).
    - Other queries saw perfect or near-perfect matches with the ground truth when using RAG.
    - For certain queries and others in the latter half of the set), RAG's tendency to say "I don't know" resulted in a 0 F1, but these may be preferable to confidently incorrect responses.

Model :Mistral-7B-Instruct-v0.3

Query	Category	Ground Truth	Llama 3.1 8B Output (F1 Score)
What are the two largest healthcare providers in Pittsburgh?	Specialized/Local Healthcare	University of Pittsburgh Medical Center (UPMC) and Allegheny Health Network.	"UPMC and Allegheny Health Network." (F1 = 0.27)
What was the first military hospital in U.S. history?	Historical/Medical	General Edward Hand Hospital.	General Edward Hand Hospital." (F1 = 1)
Which notable professional wrestlers started their careers in Pittsburgh?	Specialized/Entertainment	Bruno Sammartino, Kurt Angle, Shane Douglas, Corey Graves, Dominic DeNucci, Elias, and Britt	"Bruno Sammartino, Kurt Angle, and others." (F1 = 0.07)

		Baker.	
What is the only professional wrestling promotion based in Pittsburgh?	Specialized/Entertainment	(KSWA)	"KSWA." (F1 = 1)
Which two airports provide commercial passenger service in Pittsburgh?	Specialized/Local Infrastructure	Pittsburgh International Airport and Arnold Palmer Regional Airport.	Pittsburgh International Airport and Arnold Palmer Regional Airport." (F1 = 0.57)

Model :Llama 3.1 8B

Who invented the Banana Split?	Factual/Historical	David Strickler invented the Banana Split.	"David Strickler." (F1 = 0.57)
What were the three typical flavors of ice cream used in the original Banana Split?	Factual/Templated	Vanilla,chocolate and strawberry	vanilla, chocolate, and strawberry." (F1 = 1.0)
Which organization partnered with “Ice-Cream Joe” Gruble for the anniversary event?	Event Detail	The University of Pittsburgh.	University of Pitt." (F1 = 0.67)
In which century did the Banana Split become an iconic dessert?	Factual/Numeric	In the 20th century.	"20th" (F1 = 0.50)
Which political party was Pittsburgh a stronghold for until 1932?	Historical/Political	The Republican Party.	“The Republican Party” (F1 = 1)