AMRITA
VISHWA VIDYAPEETHAM
अद्धावान् लभते ज्ञानम्
DEEMED TO BE UNIVERSITY

22AIE314
Natural Language Processing
Term Project Abstract

VelociRAPTOR

Ivin Joel Abraham - AM.EN.U4AIE22123
Satvik Mishra - AM.EN.U4AIE22148
Shrisharanyan Vasu -AM.EN.U4AIE22150

Functionalities

1. Contextual Information Retrieval
   - Retrieves relevant documents or data from an external knowledge base (e.g., ChromaDB, Pinecone, FAISS) to provide contextually rich responses
2. Enhanced Answer Accuracy
   - Improves factual correctness by grounding responses in retrieved external sources, reducing hallucinations common in LLMs.
3. Dynamic Knowledge Updating
   - Enables real-time updates by pulling from external sources, allowing models to remain up-to-date without requiring retraining.
4. Efficient Long-Term Memory Handling
   - Supports efficient knowledge storage and retrieval, overcoming the token limit constraints of LLMs by storing embeddings of large datasets.
5. Domain-Specific Adaptation
   - Tailors responses based on a specific domain by integrating custom knowledge bases, improving LLM performance in specialized areas like law, medicine, or finance.

What Domain Knowledge do you need to do this project

- RAPTOR RAG for retrieval-augmented generation
- Gaussian Mixture Models (GMM) for clustering
- Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction
- Expectation-Maximization (EM) for iterative parameter estimation (used to implement GMM at its core)
- Bayesian Information Criterion (BIC) to evaluate model complexity (used in GMM to get optimal number of clusters to be used by the algorithm)
- NumPy
- In this project, we will employ llama-3.2-3b-instruct (via LMStudio) for LLM-based text generation and bart-large-cnn (by Facebook) for summarization tasks.

Do you have the below modules ? Give (Yes/No)

- Frontend (Web based interface or Mobile App) - No
- Select a suitable pretrained LLM with proper justification why the model - Yes (llama-3.2-3b-instruct)
- Appropriate Fine tuning technique - No
- Appropriate Prompt Engineering Strategy - Yes
- RAG - Yes
- RLHF - No

Answer the questions below

Have you seen research papers which has done similar work?

Ans: Yes. LangChain RAG.

Do you have some interesting functionality which they missed out?
List them.

Ans: Uses GMM & UMAP for clustering results and also has enhanced query expansion & document indexing

Novelty components

List out what can be new/novel in your project (In functionality/methodology)-List them

1. Custom RAPTOR RAG pipeline
2. Automated Target Prioritization Using GMM
3. Improved document clustering via Bayesian Information Criterion
4. Structured report generation using BART-Large-CNN

Answer the questions below

- Do you have any open dataset for fine tuning?

  Ans: No.

- If above answer is NO, give the strategy to generate the dataset – through web scrap/from authentic source.

  Ans: Provide documents for knowledge base.

- What is the usage of your App. Who are your end users? How much useful is your App, for an enduser in the current scenario

  Ans: The end users will be researchers, scientists, journalists and other people who need to understand the nuances of complex documents. In the current scenario, it is extremely necessary as there is more content being uploaded into the internet than ever and it becomes hard to read and understand all of it manually.

Select suitable Pretrained LLM/LLMs to do all the 5 functionalities

Which pretrained model you think will best suit your Application? Compare with existing Pretrained models. Justify why you chose yours.

1. **Your chosen Pretrained model:**
   I. llama-3.2-3b-instruct, 3 billion parameters - Chatbot
   II. bart-large-cnn, 400 million parameters - Summarization of reconnaissance reports
2. **Comparison of other LLMs:**
   I. GPT-4o: Cost ineffective
   II. T5 (Text-to-Text Transfer Transformer): Good for summarising.
   III. Falcon-40B: Too large
3. **Justification for your Model Choice: Why this model?**
   I. Llama-3.2-3b: Lightweight, optimized for security-related text processing
   II. BART-Large-CNN: Fine-tuned for summarizing long-form text reports

Instruction :

In the following slides explain each functionality as per the sample project template we have put in sharepoint page

Document Retrieval

- **Input:** User query
- **Output:** Relevant text chunks
- **Processing:**
  - Uses **RAPTOR RAG pipeline**
  - **Indexes documents using ChromaDB**
  - **Retrieves top-k relevant passages**
- **LLM Need?** Yes (Improving retrieval quality)

Query Expansion

- **Input:** User-entered query
- **Output:** Expanded query with synonyms & variations
- **Processing:**
  - Uses **llama-3.2-3b-instruct** to rewrite queries
  - Helps **retrieve better document matches**
- **LLM Need?** Yes

Clustering & Indexing

- **Input:** Documents to be indexed
- **Output:** Grouped & structured text database
- **Processing:**
  - Uses **GMM + UMAP** for clustering
  - Optimized using **Bayesian Information Criterion (BIC)**
- **LLM Need?** No (Unsupervised ML handles this)

Summarization

- **Input:** Retrieved text passage
- **Output:** Concise summary
- **Processing:**
  - Uses **BART-Large-CNN** for text compression
  - Generates **structured bullet points**
- **LLM Need?** Yes

Report Generation

- **Input:** Retrieved & summarized data
- **Output:** JSON, CSV, or HTML reports
- **Processing:**
  - Converts text into **structured formats**
  - Integrates with **external analytics tools**
- **LLM Need?** Yes

Namah
Shivaya