Question Answering with RAG

Samaa Soliman

May 18, 2025

Milestone 3: Question Answering with RAG

Project Overview

This milestone focuses on implementing and evaluating a Retrieval-Augmented Generation (RAG) system for question answering using the SQuAD v2 dataset. The system leverages FAISS vector store for document retrieval and Llama 3 as the language model. The system is implemented in two parts: 1. Zero-shot prompting: Direct question answering based on retrieved context 2. Chain of Thought (CoT) prompting: Step-by-step reasoning before providing the answer

Implementation Steps

1. **Dataset Preparation**
   - Loaded the SQuAD v2 validation dataset
   - Extracted context passages as documents
   - Removed duplicate documents to create a clean corpus
2. **Vector Store Creation**
   - Used `intfloat/e5-small` embeddings from Hugging Face
   - Created a FAISS in-memory vector store for efficient similarity search
   - Configured the retriever to fetch top 5 most relevant documents
3. **LLM Integration**
   - Utilized Ollama to run Llama 3 locally
   - Set temperature to 0 for deterministic outputs
4. **Prompting Strategies**
   - Implemented two distinct prompting approaches:
     – Zero-shot prompting: Direct question answering based on retrieved context
     – Chain of Thought (CoT) prompting: Step-by-step reasoning before providing the answer
5. **Conversational Memory**
   - Integrated ConversationBufferMemory from LangChain to enable chat history retention
   - Modified the prompt template to include previous conversation context
   - Implemented dedicated testing for follow-up questions to demonstrate memory capabilities
   - Added debug features to visualize how chat history is incorporated into prompts
6. **Evaluation Framework**
   - Created a comprehensive evaluation system
   - Used ROUGE metrics (ROUGE-1, ROUGE-2, ROUGE-L) for answer quality assessment
   - Evaluated on 1000 samples from the validation set

Results Comparison

Metric	Zero-Shot Prompting	Chain of Thought (CoT)
ROUGE-1	0.5330	0.5202
ROUGE-2	0.3284	0.2876
ROUGE-L	0.5307	0.5193
Average time per question	0.50 seconds	1.71 seconds
Total evaluation time	503.33 seconds	1714.38 seconds

Key Findings

1. **Performance Analysis**
   - Zero-shot prompting consistently outperformed CoT across all ROUGE metrics
   - The simpler approach yielded better results for this particular task
   - CoT took approximately 3.4x longer to process the same number of questions
2. **Efficiency Considerations**
   - Zero-shot prompting was significantly more efficient
   - The additional computational overhead of CoT reasoning did not translate to better performance
3. **Conversational Capabilities**
   - Adding memory enabled the system to handle follow-up questions effectively
   - The model could reference information from previous exchanges
   - Debug output confirmed proper inclusion of chat history in prompts
   - This transformed the system from single-query QA to a conversational assistant
4. **Conclusion**
   - For straightforward QA tasks on SQuAD v2, the zero-shot approach is superior
   - The task's simplicity doesn't benefit from the additional complexity introduced by CoT
   - When retrieval quality is good, direct answering works better than verbose reasoning
   - Memory capabilities significantly enhance user experience for multi-turn interactions

Future Work

- Experiment with different embedding models for potential retrieval improvements
- Test hybrid prompting strategies that adapt based on question complexity
- Explore different context window sizes to optimize the retrieval process
- Try other LLM models to compare performance across different architectures
- Implement more sophisticated memory mechanisms (e.g., summary memory, entity memory) for extended conversations