DSL605/DS605: Deep Learning for Low Resource NLP (2024–25–W)

ClinIQLink 2025 - LLM Lie Detector Test

Report
Team: Solo Leveling

Name	Student ID
Srikar Mukkamala	12141590

1 Problem Statement

Problem Overview

This project aims to develop a machine learning model for the ClinIQLink 2025 - LLM Lie Detector Test, which evaluates the ability of generative models to produce factually accurate medical information. The model will focus on improving knowledge retrieval and minimizing hallucinations in medical question-answering (QA) tasks.

2 Motivation

Motivation for the Project

Current generative models for medical QA face two primary challenges:

- **Knowledge Retrieval Issues:** Many models struggle to retrieve precise and factual medical information, leading to incorrect or misleading answers.
- **Hallucinations in Generative Models:** Models often generate responses that are not grounded in factual medical knowledge, posing risks when applied in real-world clinical settings.

3 Objectives

Project Objectives

1. Develop a knowledge-enhanced generative model that retrieves and generates factually accurate medical information.
2. Implement a retrieval-augmented generation (RAG) framework to enhance factual grounding.
3. Introduce hallucination detection and mitigation strategies by incorporating confidence estimation and external fact-checking.
4. Evaluate the model's performance using the ClinIQLink dataset, focusing on precision, recall, F1 score, BLEU, ROUGE, METEOR, and semantic similarity metrics.

4 Relevant Study

Relevant Findings

Key studies and findings influencing the model design:

- **Retrieval-Augmented Generation (RAG):** Combining generative models with external knowledge retrieval improves factual correctness (Lewis et al., 2020).
- **Medical QA Systems:** Fine-tuning LLMs on medical datasets (e.g., MedQA, PubMed) enhances their knowledge base (Haoran et al., 2024).
- **Hallucination Detection:** Methods like confidence calibration, contrastive loss, and self-consistency reduce hallucinations in LLMs (Duy et al., 2024).

5 Proposed Solution

Solution Proposal: ClinIQ-RAG Model

The proposed model consists of three core components:

1. **Knowledge Retrieval Module:** Uses a hybrid approach with a dense retriever (e.g., BM25 + FAISS) to fetch relevant medical documents.
2. **Fact-Aware Response Generator:** Implements a fine-tuned LLM (e.g., LLaMA, GPT-4, or Med-PaLM) with knowledge-augmented input.
3. **Hallucination Mitigation Framework:** Includes confidence scoring, external fact-checking, and contrastive learning to minimize misinformation.

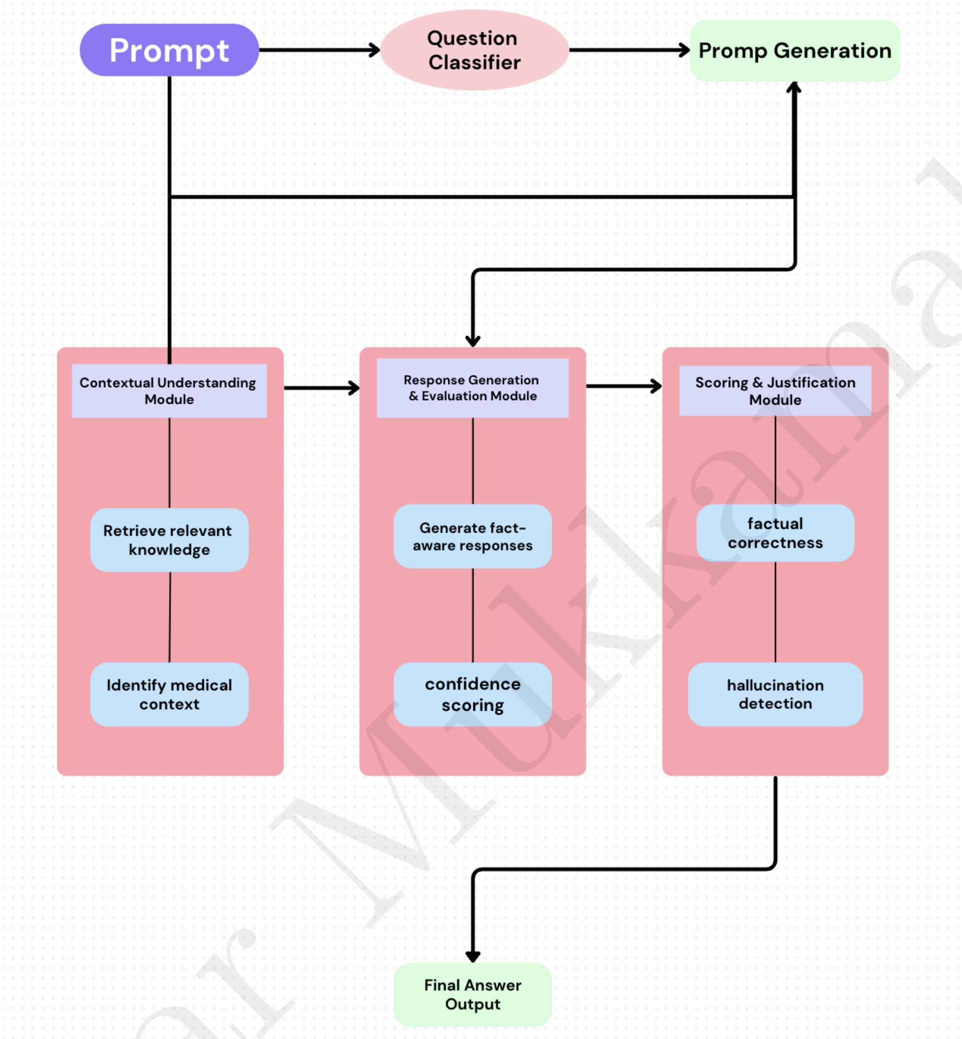
Prompt
Question Classifier
Promp Generation
Contextual Understanding Module
Retrieve relevant knowledge
Identify medical context
Response Generation & Evaluation Module
Generate fact-aware responses
confidence scoring
Scoring & Justification Module
factual correctness
hallucination detection
Final Answer Output

Figure 1: The architecture of the proposed model consists of a Question Type Classifier, a Contextual Understanding Module, a Response Generation & Evaluation Module, and a Scoring & Justification Module.

6 Methodology

This section outlines the end-to-end methodology for building a Retrieval-Augmented Generation (RAG)-based Large Language Model (LLM) system for medical question answering. The approach combines dense and sparse retrieval techniques with supervised fine-tuning of a domain-specific LLM, supported by prompt engineering and hallucination analysis frameworks. Figure 2 shows the process of the proposed methodology.

1. Data Acquisition and Knowledge Distillation
I used the MedRAG dataset, which includes three primary knowledge sources:

- Textbooks
- PubMed
- Wikipedia

To make the corpus computationally manageable, I performed knowledge distillation, reducing the combined dataset to approximately 30 GB of high-relevance content. This distillation involved filtering out noisy or low-relevance entries and retaining semantically rich, well-structured documents.

2. Document Encoding and Vector Store Construction
Data encoded using the all-MiniLM-L6-v2 SentenceTransformer model to generate dense vector representations. These embeddings capture the semantic structure of each document chunk.

I used FAISS (Facebook AI Similarity Search) to construct a dense vector index. Additionally, I stored a compressed representation in a 3GB pickle file for efficient memory usage and retrieval latency optimization.

3. Hybrid Document Retrieval Module
I implemented a hybrid retriever, combining:

- Dense retrieval via FAISS-based nearest-neighbor search on MiniLM embeddings.
- Sparse retrieval via BM25 keyword-based scoring.

To rank retrieved documents, I applied Reciprocal Rank Fusion (RRF), which merges rankings from both retrievers. The top-5 documents were selected as contextual evidence, maximizing relevance and lexical diversity to support downstream question answering.

4. Prompt Engineering and Classification
Categorized medical questions into seven subtypes based on structure and required reasoning:

- True/False
- Multiple Choice (MCQ)
- List-type
- Short Answer
- Multi-hop Reasoning
- Inverse Short Answer
- Inverse Multi-hop

For each category, I designed specialized prompt templates using prompt engineering principles. These prompts encapsulate both the retrieved evidence and question structure, formatted to instruct the LLM explicitly about the expected answer type.

5. LLM Supervised Fine-Tuning
I used BioMistral-7B, a 14.5GB, non-quantized transformer model pre-trained on biomedical corpora. This model uses Rotary Positional Embeddings (RoPE) for enhanced token ordering and generalization on long context sequences.

I fine-tuned the model using supervised learning across all seven task types using cross-entropy loss. Each input consisted of a question-specific prompt concatenated with the top-5 retrieved context documents, and the output was the ground-truth answer.

Hyperparameters:

- Batch size: 8
- Learning rate: 2e-5 with linear decay
- Optimizer: AdamW
- Number of epochs: 20 (early stopping based on validation F1)

To rank retrieved documents, I applied Reciprocal Rank Fusion (RRF), which merges rankings from both retrievers. The top-5 documents were selected as contextual evidence, maximizing relevance and lexical diversity to support downstream question answering.

6. Inference and Generation Control
At inference time, first classify the question type, generate the corresponding prompt, retrieve top-5 context documents, and format the complete input for the LLM. To ensure factuality, I tuned the temperature parameter between 0.1 and 0.3, minimizing generation randomness and hallucinations

- Dense retrieval via FAISS-based nearest-neighbor search on MiniLM embeddings.
- Sparse retrieval via BM25 keyword-based scoring.

To rank retrieved documents, I applied Reciprocal Rank Fusion (RRF), which merges rankings from both retrievers. The top-5 documents were selected as contextual evidence, maximizing relevance and lexical diversity to support downstream question answering.

7. Evaluation Metrics
I assessed the performance of the system using both lexical and semantic metrics:

- F1 Score (Exact Match & Token Overlap)
- BLEU (N-gram precision)
- ROUGE-L (Longest common subsequence recall)
- METEOR (Semantic similarity including synonyms and paraphrases)

Evaluation was done on a held-out medical QA test set stratified across all question types.

8. Post-Hoc Hallucination and Factuality Analysis
To analyze and mitigate hallucinations in the generated answers:

- I used RAGAS and LYNX, two hallucination detection frameworks that compare generated text with source context to score factual consistency.
- Additionally, I performed external fact-checking via biomedical databases and human evaluation, identifying overclaims, fabrications, and unsupported inferences.

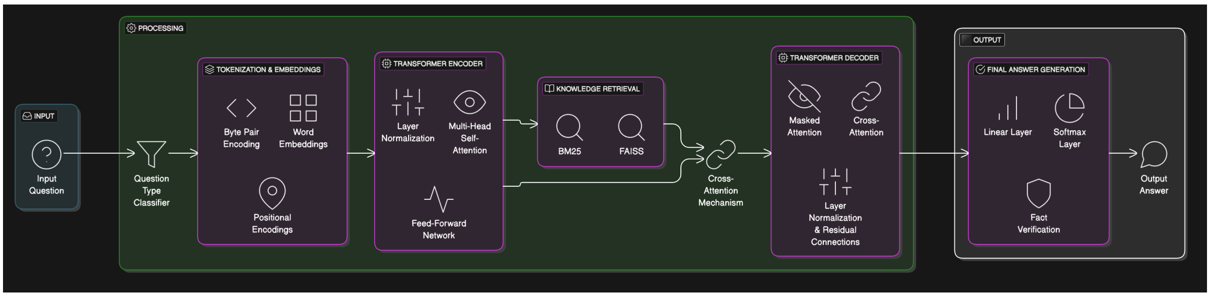

Figure 2: A more technically detailed architecture with the respective subcategories based on module wise classification.

References

Links

- ClinIQLink 2025 - LLM Lie Detector Test
- Lewis et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.
- Haoran et al. (2024). Enhancing Healthcare through Large Language Models: A Study on Medical Question Answering.
- Duy et al. (2024). Towards Reliable Medical Question Answering: Techniques and Challenges in Mitigating Hallucinations in Language Models.