**CVA Current System Logic and Workflow**

**1.1 System Overview**

The Medical Record Analysis System processes clinical questions by:

* Extracting relevant context from medical records using atoms (key terms)
* Leveraging Large Language Models (LLMs) to generate structured responses
* Supporting Elasticsearch

**1.2 Workflow Steps**

**Step 1: Data Ingestion**

* **Input**: Medical records in CSV format containing patient data, clinical notes, and metadata
* **Processing**:
  + Data is indexed into Elasticsearch for efficient retrieval
  + Each record contains: file\_name, page, content, audit\_id, line coordinates, and mrindex\_results

**Step 2: Question Processing**

* **Input**: Clinical questions defined in Excel format with:
  + Atoms (pipe-separated key terms for retrieval)
  + Target classes (document sections to search)
  + Prompt questions for LLM
  + Expected output format (Opinion or Table)
* **Processing**:
  + Parse atoms and target classes for each question
  + Prepare search parameters based on question type

**Step 3: Context Retrieval**

* **Retrieval Process**:
  + **Elasticsearch**: Query using atoms with fuzzy matching and phrase matching
  + **CSV-based**: Simple string matching in content column
* **Context Expansion**:
  + For each matched record, expand context by including 2 lines above and 2 lines below
  + Sort records by line coordinates to maintain document order
  + Extract expanded content window around matched atoms

**Step 4: Prompt Construction**

* **Context Preparation**:
  + Deduplicate identical context hits
  + Build context string by joining unique expanded contents
* **Prompt Template**:
  + System prompt: Strict instructions for factual clinical data extraction
  + User prompt: Includes context, question, and format instructions

**Step 5: LLM Inference**

* **Model**: Llama-3.1-8b-instruct (or similar)
* **Token Management**:
  + Current method: Simple character count estimation (len(text)//4)
  + If context exceeds token limit, truncate by removing entire context parts

**Step 6: Response Generation**

* **Output Types**:
  + **Opinion**: Yes/No with explanation following strict template
  + **Table**: Markdown table with extracted data
* **Output Formats**:
  + Excel reports (detailed with SME feedback columns)
  + PDF summaries (simplified for presentation)

**Step 7: Report Generation**

* **Excel**: Detailed report with all processing steps and context
* **PDF**: Clean, formatted summary for stakeholders

**1.3 System Architecture**

**Medical Records (CSV) → Elasticsearch Indexing → Context Retrieval**

**↑ ↓**

**Questions (Excel) → Question Processing → LLM Inference**

**↓ ↑**

**Report Generation ← Response Generation ← Token Management**

**A screenshot of a computer

AI-generated content may be incorrect.**

**2. Edge Cases and Challenges**

|  |  |  |  |
| --- | --- | --- | --- |
| **Edge Case** | **Description** | **Current System Impact** | **Production Risk** |
| **Token Limit Exceeded** | Context from all hits exceeds LLM's token limit (e.g., 7000 tokens) | Simple truncation by removing entire context parts, potentially losing critical information | **High** - May lead to incomplete or incorrect answers |
| **Late-Arriving Records** | New medical records arrive after initial processing for an audit ID | System processes only available records; late records are missed until next run | **High** - Answers may be incomplete if critical records arrive late |
| **High Atom Count** | Questions with many atoms (e.g., 25+) result in numerous context hits | Context becomes very large, increasing truncation risk | **High** - More likely to exceed token limits |
| **Duplicate Context** | Same content appears in multiple hits from different atoms | Deduplication occurs, but scoring doesn't prioritize relevance | **Medium** - Efficiency issue, but correctness maintained |
| **Varying Document Sections** | Different sections (discharge summary, progress notes) have different importance | All sections treated equally; no prioritization | **Medium** - May miss critical information in important sections |
| **Positional Relevance** | Earlier parts of documents may be more relevant (e.g., discharge summary) | No position-based scoring in context selection | **Medium** - May include less relevant later content |
| **Window Size Rigidity** | Fixed 2-line window may be too small or too large for different content | Cannot adapt to different content densities | **Medium** - May miss context or include too much noise |
| **Concurrent Processing** | Multiple audit IDs need processing simultaneously | No built-in parallelization | **Low** - Processing time may be long for large batches |
| **Error Handling** | Failures in indexing, retrieval, or LLM calls | Basic error handling with logging | **Medium** - May not recover gracefully from transient failures |