BRD Analysis - 2025-08-08

Okay, here's a summary of the key requirements extracted from the provided BRD content, categorized for clarity: **I. Core Functionality & API Endpoints:** * **Text Analysis:** The core function is document analysis via a POST request to `http://localhost:9080/file`. This includes: * **Content Analysis:** Processing a text query via the `content` parameter. * **RAG (Retrieval-Augmented Generation):** Analyzing documents (PDFs, Excel files, images, videos) to answer questions. * **Document Classification:** Identifying document types (e.g., Commercial Invoice, Bill of Lading). * **Excel Analysis:** Analyzing financial data within spreadsheets. * **Image Analysis:** Extracting information from images (e.g., invoices). * **Video Analysis:** Analyzing banking procedures or verification videos. **II. Model & Input Requirements:** * **Supported AI Models:** The API supports multiple models: * `simplyfillm`: The default, designed for high-quality, complex queries. * `gpt-4.1-mini`: Advanced reasoning and complex tasks. * `gemma3:4b`: Quick queries and responses. * `gemma3:12b`: General purpose, default option. * `llama3.2 - vision:latest`: Advanced visual document analysis * `mistral:latest`: Specialized financial expertise * **Input Formats:** The API accepts various file types: PDFs, Excel (XLS, XLSX), JPG, PNG, JPEG, MP4. * **File Size/Duration Limitations:** PDF (no size limit), Excel (unspecified), Images (300 DPI recommended), Videos (MP4, max 5 minutes). **III. Data Processing & Response:** * **RAG Process:** Documents are chunked and relevant segments are identified based on the query. * **Response Format:** The API returns a JSON response including: * `description`: Text content extracted from the analyzed document. * `audit_trail`: Timestamped events, request ID, metrics (total requests, errors, guardrail blocks, response time), and success rate. * `guardrail_info`: Indicates whether guardrail checks were performed and the verdict (safe/unsafe). **IV. Guardrail System & Error Handling:** * **Content Filtering:** A guardrail system actively filters potentially harmful content (profanity, fraud). * **Error Handling:** * `400 No prompt provided`: Requires a `content` parameter. * `403 Input flagged by guardrail`: Query revised due to flagged content. * `500 Error processing request`: Internal server error, needs to try again. **V. Best Practices & Usage Guidelines** * Use `simplyfillm` for critical queries. * Be specific in your questions. * Use RAG for detailed document Q&A.; * Classify documents first to understand file types. Do you want me to: * Drill down on a specific aspect of the requirements? * Create a visual diagram or flowchart of the process? * Generate example cURL commands for a specific scenario?