Project Name: AI-DocumentIntelligence-POC

Date: 22nd April 2025

[Purpose 1](#_Toc1747457271)

[Scope of Project 2](#_Toc1403455462)

[1. Q&A & Information Extraction 2](#_Toc1797477358)

[2. Contract Search and Query 2](#_Toc548668977)

[3. Model Training (TBD) 3](#_Toc1172818440)

[Deliverables 3](#_Toc1487186875)

[Timeline 4](#_Toc49147023)

[Resources Needed 4](#_Toc1611069556)

[Technology Stack 5](#_Toc1799809612)

[Out of Scope 5](#_Toc1374652268)

### **Purpose**

1. Show case AI capabilities of Radix.
2. Expand Radix AI portfolio by adding 4 AI developers and 1 AI manager.
3. Facilitate testing various AI models for accuracy.

### **Scope of Project**

#### **1. Q&A & Information Extraction**

* **Objective**: Develop a system to extract key information from contracts. Along with pre-defined extracted information, facilitate user to Q&A for specific contract. Selected contract should be viewable in browser.
* **Key Information**:
  + Agreement Title
  + Start Date
  + End Date
  + Type of Agreement
  + Other relevant details as specified
* **AI Assistant**
  + User can ask any question related to specific contract.
* **Approach**:
  + Develop a user-friendly interface for uploading documents and displaying extracted information and allow to do Q&A on specific contract.
  + Perform text extraction from uploaded PDF.
  + Use Natural Language Processing (NLP) techniques to identify and extract relevant information.

#### **2. Contract Search and Query**

* **Objective**: Enable users to search for specific contracts and query related information.
* **Functionality**:
  + Implement a search feature to find contracts based on user queries.
  + Use a Large Language Model (LLM) to interpret user queries and retrieve relevant contract information.
  + Provide concise and accurate answers based on the content of the contracts.
* **Approach**:
  + Index all contracts in a searchable database.
  + Integrate LLM to understand and process user queries.
  + Develop an interface for users to input queries and view results.

#### **3. Model Training (TBD)**

* **Objective**: Train a custom model using contract data.
* **Approach**:
  + Use a foundational model as the base (e.g., GPT-4, BERT).
  + Fine-tune the model with contract-specific data to improve accuracy and relevance.
  + Continuously update the model with new data to enhance performance.
* **Implementation**:
  + Collect and preprocess contract data.
  + Train the model using supervised learning techniques.
  + Evaluate and refine the model to ensure high accuracy.

### **Deliverables**

* A system for extracting key information from contracts and user should perform Q&A for specific contract.
* A searchable database of contracts with an integrated LLM for query processing.
* A custom-trained model based on contract data.
* At the conclusion of POC, Radix AI portfolio will expand to include 4 AI developers and 1 AI Manager.

### **Timeline**

|  |  |  |
| --- | --- | --- |
| Phase 1 | Q&A & Information Extraction | [Sprint 01](https://projects.radixweb.in/projects/ai-documentintelligence-poc-internal/issues?utf8=%E2%9C%93&set_filter=1&sort=id%3Adesc&f%5B%5D=status_id&op%5Bstatus_id%5D=*&f%5B%5D=agile_sprints&op%5Bagile_sprints%5D=%3D&v%5Bagile_sprints%5D%5B%5D=1078&f%5B%5D=tracker_id&op%5Btracker_id%5D=%3D&v%5Btracker_id%5D%5B%5D=2&f%5B%5D=&c%5B%5D=parent&c%5B%5D=status&c%5B%5D=priority&c%5B%5D=subject&c%5B%5D=author&c%5B%5D=assigned_to&c%5B%5D=updated_on&c%5B%5D=category&c%5B%5D=fixed_version&c%5B%5D=start_date&c%5B%5D=due_date&c%5B%5D=estimated_hours&group_by=&t%5B%5D=) |
| Phase 2 | Contract Search and Query | [Sprint 02](https://projects.radixweb.in/projects/ai-documentintelligence-poc-internal/issues?utf8=%E2%9C%93&set_filter=1&sort=id%3Adesc&f%5B%5D=status_id&op%5Bstatus_id%5D=*&f%5B%5D=agile_sprints&op%5Bagile_sprints%5D=%3D&v%5Bagile_sprints%5D%5B%5D=1086&f%5B%5D=tracker_id&op%5Btracker_id%5D=%3D&v%5Btracker_id%5D%5B%5D=2&f%5B%5D=&c%5B%5D=parent&c%5B%5D=status&c%5B%5D=priority&c%5B%5D=subject&c%5B%5D=author&c%5B%5D=assigned_to&c%5B%5D=updated_on&c%5B%5D=category&c%5B%5D=fixed_version&c%5B%5D=start_date&c%5B%5D=due_date&c%5B%5D=estimated_hours&group_by=&t%5B%5D=) |
| Phase 3 | Model Training |  |

### **Resources Needed**

|  |  |  |
| --- | --- | --- |
| Sample Contracts | Sample Contracts collected from [Sam.Gov](https://sam.gov/search/?page=1&pageSize=25&sort=-modifiedDate&index=opp&sfm%5BsimpleSearch%5D%5BkeywordRadio%5D=ALL&sfm%5BsimpleSearch%5D%5BkeywordTags%5D%5B0%5D%5Bkey%5D=construction&sfm%5BsimpleSearch%5D%5BkeywordTags%5D%5B0%5D%5Bvalue%5D=construction&sfm%5BsimpleSearch%5D%5BkeywordEditorTextarea%5D=&sfm%5Bstatus%5D%5Bis_active%5D=true&sfm%5Bstatus%5D%5Bis_inactive%5D=false) | Collected contracts available [here](https://radixweb0-my.sharepoint.com/my?id=%2Fpersonal%2Fnirav%5Fshah%5Fradixweb%5Fcom%2FDocuments%2FAI%2DDocumentIntelligence%2DPOC%2FSample%20Contracts&login_hint=nirav%2Eshah%40radixweb%2Ecom&source=waffle) |
| IDE | Visual Studio Code |  |
| Source Control | [Gitlab](https://rxgit.radixweb.in/rxprojects/aidocumentintelligencepoc/dotnet/aidocumentintelligencepoc/) |  |
| Project Management | [Redmine](https://projects.radixweb.in/projects/ai-documentintelligence-poc-internal) |  |
| Software | Python, Mongo DB |  |

### **Technology Stack**

|  |  |
| --- | --- |
| Front-end | React |
| Backend | Python |
| Database | Mongo DB |
| Extract text from PDF | PyPDF (Python library) |
| AI Model | To be decided |
| AI Library | Transformers |

### **Out of Scope**

1. Only PDF will be supported
2. Currently OCR will not be performed on scanned PDF.