Evidenze
Knowledge in Health

# "AI-Based Review, Documentation Generation, and Automatic Assistance System for CROs"

## Impact and Benefits

The project expects to have a significant impact nationally and internationally on CROs, improving operational efficiency, reducing time and errors in document management, and ensuring regulatory compliance. In addition, it will contribute to environmental sustainability by reducing the carbon footprint associated with manual processes.

## Key Innovations

- **AI Document Analysis**: Upload documents in multiple formats, extract key data, and verify consistency and compliance.
- **Visualization Dashboard**: Centralized dashboard that shows the status of documents and generates compliance alerts.
- **Automatic Document Generation**: Use of pre-designed templates and intelligent context for the creation of new documents.
- **Automatic Assistance with On-Premise LLM**: Provision of answers to frequently asked questions and real-time assistance with continuous model adjustment.

## Expected Results

- **Reduced Review Time**: Automation that significantly decreases the time spent manually analyzing documents.
- **Consistency and Compliance**: Assurance that all documents comply with regulations and protocols.
- **Efficient Documentation Generation**: Fast and consistent creation of new documents.
- **Automated Expert Assistance**: Provision of accurate answers and contextual assistance in real-time.
- **Operational Efficiency**: Saving time in the search for information and improving decision-making.
- **Carbon Footprint Reduction**: Reduction of environmental impact through digitalization and optimization of processes.

This AI system will provide CROs with an advanced tool to improve their productivity and reduce their ecological impact, contributing to sustainability in the era of environmental awareness.

# 2. OBJECTIVES AND INNOVATIONS

## 2.1. Suitability of the project

### Description

A Contract Research Organization (CRO) is a company that provides support services to the pharmaceutical, biotechnology, and medical device industry in the form of contract research. CROs offer a wide range of services that may include:

1. Clinical Research: Design and execution of clinical trials to test the safety and efficacy of new drugs or treatments.
2. Drug Development: Support in the discovery and development phases of new pharmaceutical compounds.
3. Regulation and Compliance: Assistance in the preparation and submission of documents necessary to obtain regulatory approval for new drugs and medical devices.
4. Clinical Trial Monitoring: Supervising clinical trials to ensure that they are conducted in accordance with established protocols and relevant regulations.
5. Data Management: Collection, analysis, and reporting of data obtained during clinical studies.
6. Laboratory Services: Performing tests and analyses necessary to support clinical trials and research studies.
7. Scientific Consulting: Providing expert advice on various aspects of drug development and clinical research.

CROs allow pharmaceutical and biotech companies to outsource parts of their research and development process, which can be more efficient and cost-effective. In addition, CROs can bring specialized expertise and resources that may not be available internally at companies developing new treatments and devices.

The "AI-based Review, Documentation Generation and Automatic Assistance System for CROs" is perfectly suited to today's technological challenges in the field of clinical research and document management. CROs face significant challenges in terms of efficiency, consistency, and regulatory compliance in document management. This project addresses these challenges by implementing advanced artificial intelligence solutions that optimize and automate critical document review, generation, and assistance processes.

### Project Justification

The rationale for the "AI-based Automatic CRO Review, Documentation, and Assistance System" is based on the critical need to improve efficiency, consistency, and compliance in document management within CROs. Current systems do not fully address these needs, and implementing advanced AI solutions can provide significant improvements in these aspects.

**Operational Efficiency:**

- Manual document review is a laborious and error-prone process. Automating this process using AI can significantly reduce the time needed to review large volumes of documentation, allowing employees to focus on more strategic tasks.

**Consistency and Compliance:**

- Ensuring that documents comply with all regulations and protocols is crucial in the field of clinical research. Automatic consistency and compliance verification can minimize the risk of human error and ensure that documents are compliant with current regulations.

**Efficient Documentation Generation:**

- The creation of new documentation based on pre-built templates and AI-based contextual suggestions ensures consistency and quality in the generated documents. This is especially important in an environment where accuracy and clarity are essential.

**Real-Time Expert Support:**

- Providing automated, contextual support in real-time using a large language model tuned specifically for CROs' needs can improve decision-making and operational efficiency, reducing reliance on manual document queries.

**Environmental Sustainability:**

- The digitization and optimization of manual processes not only improves operational efficiency but also contributes to environmental sustainability by reducing the use of physical resources such as paper and electricity.

## General Objective

The main objective of the project is to research and implement a comprehensive AI-based system that optimizes documentation management and provides automated assistance in CROs, guaranteeing data security and privacy through its on-premise implementation.

## Specific Objectives

1. **Automate document review:**
    - **Description:**
        - The goal is to research and implement an automated document review module that uses natural language processing (NLP) and machine learning techniques to analyze large volumes of documentation. This module will identify and extract key data, verify internal consistency and compliance with specific regulations and protocols, and generate alerts for any discrepancies or errors detected.
    - **How we do it:**
        - **Data collection:** Compilation of documents from different CROs to train the system. Data mining techniques will be used to identify common patterns and features in documents.
        - **Development of algorithms:** Implementation of NLP and machine learning models. Supervised and unsupervised models will be developed that are capable of learning from historical data and applying it to new documents.
        - **Validation:** Extensive testing with real documents to fine-tune and optimize algorithms. Continuous iterations will be performed to improve the accuracy and efficiency of the system.
2. **Ensure document consistency and compliance:**

- **Description:**
  - Research and implement automatic verification tools that compare documents against a set of established standards and protocols, ensuring they comply with all relevant regulations. This objective includes the integration of regulatory databases and the continuous updating of regulations.
- **How we do it:**
  - **Development of a database of regulations:** Compilation and digitization of regulatory regulations. Automatic processes will be implemented for the continuous updating of these regulations.
  - **Comparison algorithms:** Research and implementation of document comparison and analysis techniques. Algorithms will be implemented that can identify discrepancies and generate alerts automatically.
  - **Continuous updating:** Constant maintenance and updating of the regulations database. Mechanisms will be created for the incorporation of new regulations and the elimination of those that have been repealed.

3. **Facilitate the generation of new documentation:**

- **Description:**
  - Research and implement an automatic document generation system that uses pre-built templates and suggests content based on document context and historical data. This system will allow users to create high-quality documents quickly and efficiently.
- **How we do it:**
  - **Template development:** Creating standardized templates for different types of documents. Experts in each area will be worked on to ensure that the templates meet the necessary requirements.
  - **Suggestion algorithms:** Research and implement algorithms that analyze context and suggest relevant content. Natural language processing techniques will be used to understand the context and provide appropriate suggestions.
  - **User Interface:** Design of an intuitive interface that allows users to select templates and customize them easily. Usability tests will be conducted to ensure that the interface is user-friendly and efficient.

4. **Provide real-time automated support:**

- **Description:**
  - Research and implement a large language model (LLM)-based support module tuned specifically for CROs' needs. This module will provide quick and accurate answers to frequently asked questions and offer real-time contextual assistance.
- **How we do it:**
  - **Model training:** Tuning the LLM with CRO-specific data. Historical data will be collected and used to train the model.
  - **Integration with the system:** Connection of the assistance module with the document management system. APIs will be created that allow communication between the assistance module and other components of the system.
  - **Continuous improvement:** Continuous updating and adjustment of the model based on user feedback. Mechanisms will be implemented for the collection of feedback and its incorporation in future iterations of the model.

5. **Implement a centralized dashboard for viewing and managing documents and compliance alerts:**

    - **Description:**
        - Develop a centralized dashboard that provides an overview of all documents managed by the system, including compliance indicators, alerts, and notifications. This dashboard will allow users to monitor the status of their documents efficiently and make informed decisions.
    - **How we do it:**
        - **Dashboard design:** Creating an intuitive and easy-to-use design. Workshops will be held with users to understand their needs and preferences.
        - **Data integration:** Connection of the dashboard with all the modules of the system. It will ensure that the dashboard has access to the necessary data and that it is presented in a consistent and understandable manner.
        - **Alerting functionalities:** Implementation of alerts and notifications for documents that require attention. Algorithms will be developed that identify critical events and notify users proactively.
        -

## 2.2. Detailed description, proposed architecture and scope of the work plan

### 2.2.1. Proposed architecture

#### Main Components

1. **Document Upload Module:**

    - **Functionality:** Allows the upload of documents in different formats (PDF, Word, Excel).
    - **Deployment:** Use a secure on-premise storage system to ensure data privacy.
    - **Technologies:** Local servers, secure databases, document upload interfaces, REST APIs for integration.

2. **Documentation Analysis Module:**

    - **Functionality:** Uses natural language processing (NLP) techniques to extract key data from uploaded documents.
    - **Implementation:** Verifies the consistency and compliance of documents with established regulations and protocols.
    - **Technologies:** NLP models, machine learning algorithms, regulatory databases, text analysis services.

3. **Visualization Dashboard:**

    - **Functionality:** Presents an overview of all the documents analyzed, showing compliance indicators and generating alerts.
    - **Implementation:** Centralizes document management and offers monitoring and analysis tools.
    - **Technologies:** Graphical user interfaces (GUIs), notification systems, tracking databases, data visualization frameworks such as D3.js or Tableau.

4. **Automatic Documentation Generation Module:**

- **Functionality:** Use pre-built templates and intelligent context for the creation of new documents.
- **Implementation:** Provides contextual suggestions based on analysis of existing protocols and documents.
- **Technologies:** Content generation systems, document templates, context analysis algorithms, template engines such as Apache FreeMarker.

5. **Automatic Assistance Module with On-Premise LLM:**

    - **Functionality:** Train and tune a large language model with CRO-specific data.
    - **Implementation:** Provides fast, accurate answers to frequently asked questions and offers real-time support.
    - **Technologies:** Large language models, knowledge databases, user interfaces for assistance, APIs for integration of language models such as GPT.

## Infrastructure

The infrastructure is designed to withstand the intensive processing and secure storage demands required by the system. It includes local servers for on-premise data storage, ensuring the security and privacy of sensitive data. The infrastructure is scalable to accommodate the growth in the volume of documents and users.

- **Servidores On-Premise:**

    - **Functionality:** Secure data hosting and local processing.
    - **Security:** Implementation of firewalls, intrusion detection systems (IDS), and encryption of data at rest and in transit.
    - **Scalability:** Ability to add additional servers as compute and storage needs increase.

- **Secure Communications Network:**

    - **Functionality:** Secure connection between the different modules and users.
    - **Technologies:** Virtual Private Networks (VPNs), Secure Sockets Layer (SSL), and Transport Layer Security (TLS) for encrypted communications.

## Safety

Security is a critical aspect of the system, as it handles sensitive and confidential information. Multiple layers of security are implemented to protect data and infrastructure.

- **Encryption:**

    - **Data at Rest:** Encryption of stored data using algorithms such as AES-256.
    - **Data in Transit:** Encryption of communications using SSL/TLS.

- **Authentication and Authorization:**

    - **User Authentication:** Use of multi-factor authentication (MFA) to ensure that only authorized users access the system.
    - **Granular Authorization:** Implementation of role-based access (RBAC) policies to control access to different parts of the system.

- **Monitoring and Auditing:**

- **Continuous Monitoring:** Use of monitoring tools to detect and respond to potential threats in real-time.
- **Audit:** Record of all system activities to ensure traceability and regulatory compliance.

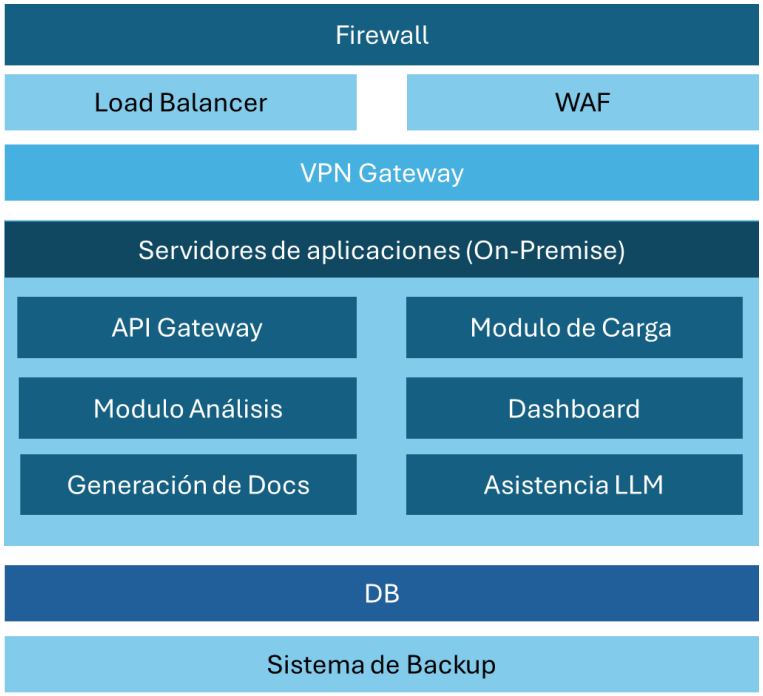


Figure: Security Infrastructure. Security Measures

## 2.3. Innovations

This project aims to carry out a comprehensive solution aimed at optimizing document management and providing automated assistance in Contract Research Organizations (CROs). This system, based on artificial intelligence, not only reviews and generates documentation, but also answers questions about specific processes and protocols, guaranteeing the security and privacy of data through its on-premise implementation. In this way, the main innovations of the project are the following:

- **AI Documentation Analysis**: Utilizing advanced natural language processing (NLP) techniques for key data extraction and consistency and compliance verification. This includes the ability to analyze large volumes of documents to identify and extract critical information, ensuring that all documents comply with applicable regulations and protocols.
- **Visualization Dashboard**: Centralization of document management and visualization of compliance indicators through a centralized dashboard that shows the status of documents and generates compliance alerts. This tool facilitates the monitoring and management of documentation, providing a clear and consolidated view of the status of documents.
- **Automatic Documentation Generation**: Creation of new documents by using pre-designed templates and contextual suggestions based on artificial intelligence. This process ensures that the documents generated are consistent and of high quality, significantly reducing the time and effort required for document creation.

- **Automatic Assistance with On-Premise LLM**: Provision of real-time assistance using a large language model tuned specifically for each CRO. This assistance provides quick and accurate answers to frequently asked and contextual questions, improving decision-making and operational efficiency.

Each of these innovative elements brings significant improvements over the current state of the art. AI-powered process automation reduces the time needed to review and generate documentation, allowing employees to focus on more strategic and higher-value tasks. Likewise, the use of AI techniques for consistency and compliance verification minimizes the risk of human error, ensuring that all documents comply with applicable regulations and protocols. Similarly, the centralization of document management through a dashboard provides a consolidated and clear view of the status of documents, facilitating effective monitoring and management of documentation. And finally, the provision of real-time automatic assistance improves decision-making by providing quick and accurate answers to contextual and frequently asked questions.

It is also worth highlighting the importance of these innovations with respect to the Current State of the Art:

1. **Document Processing**: This process is currently being done manually. The improvement in the speed and accuracy of document analysis using AI techniques represents a significant advance compared to traditional manual methods, which are laborious and error-prone.
2. **Document Generation**: Currently this process is being carried out manually. Streamlining document generation through the use of pre-built templates and AI-based contextual suggestions ensures that documents are consistent and of high quality, something that is not always achieved with manual methods.
3. **Automatic Assistance**: This process is currently being carried out manually. The development of large language models tuned to provide contextual and accurate assistance in real-time is a significant improvement over traditional assistance systems, which can be slow and less accurate.

Finally, in reference to the scientific and technological increases of the project, the following should be highlighted:

- **Improved Document Processing**: The implementation of advanced natural language processing and machine learning techniques improves the speed and accuracy of document analysis, enabling the identification and extraction of key data more efficiently and accurately.
  There are currently collaborative platforms such as Microsoft SharePoint that facilitate document management and workflow, but with limited capabilities in automating document review and generation. There are also document management and workflow automation solutions, such as DocuWare, focused more on organization and storage than on intelligent review and automated document generation. And there are also systems for the management of large volumes of documents with some automation capabilities, such as IBM FileNet, but their implementation is complex and expensive, not always adapted to the specific needs of CROs.
  But our solution will differentiate itself from these by employing advanced natural language processing techniques for key data extraction and consistency and compliance verification, surpassing the basic textual analysis capabilities of current systems. And similarly, it will implement machine learning algorithms to analyze large volumes of documentation, significantly reducing the time needed compared to current solutions that rely on manual intervention.

- **Document Generation Optimization**: Automatic documentation generation using pre-built templates and AI-based contextual suggestions improves document consistency and quality, ensuring they are consistent and of high quality.
  Tools such as Google Cloud Natural Language API currently exist for text analysis and information extraction, but it requires complex integration and is not specifically tailored to the review of clinical and regulatory documents. Similarly, there are also NLP services such as Amazon Comprehend that allow data extraction and sentiment analysis, useful for certain analyses, but not optimized for clinical review and documentation.
  Our solution will differentiate itself from these by employing standard templates and contextual algorithms to generate new documentation, improving consistency and quality compared to existing solutions that offer limited auto-generation capabilities. And it will also incorporate automatic document generation using AI ensuring that documents are consistent and of high quality, something that current solutions do not always achieve due to a lack of personalization and context.
- **Advances in Automatic Assistance**: The development of large language models tuned to provide contextual and accurate assistance in real time represents a significant advance in the field of automatic assistance, improving decision-making and operational efficiency.
  There are currently AI-based virtual assistants such as IBM Watson Assistant that can be trained to provide answers to frequently asked questions, but it needs significant customization for the specific context of CROs. There are also large language models such as OpenAI GPT with advanced text generation capabilities and contextual responses, but it faces challenges in terms of customization and on-premise implementation to ensure data privacy and security.
  Instead, our solution will provide real-time assistance using a large language model tuned specifically for each CRO, improving the accuracy and relevance of responses compared to the general models currently used; and it will also provide contextual and accurate assistance, which significantly improves decision-making and operational efficiency, exceeding the capabilities of current systems that are not specifically adapted to the context of CROs.

In this way, the scientific-technological increases of the SGDA-IA project represent significant improvements over the current state of the art, providing more advanced and personalized tools for document management and automatic assistance in CROs. These improvements not only optimize internal processes and reduce errors, but also ensure regulatory compliance and improve the operational efficiency of CROs,