Project Goals and Use Cases

**Project**: Biomedical LLM Information Extraction Tool

**Version**: 0.1

**Author**: Elena Jolkver

**Date**: 25.07.2025

# 1. Introduction

This document outlines the goals, intended users, and primary use cases for the Biomedical LLM Information Extraction Tool, a Streamlit-based application for extracting and summarizing key information from ClinicalTrials.gov documents using Large Language Models (LLMs).

# 2. Project Goals

1. **Automate Key Information Extraction:**
   * Enable efficient and accurate extraction of key structured components from unstructured clinical trial data (e.g., PICO: Population, Intervention, Comparator, Outcomes) to reduce manual effort.
2. **Facilitate Biomedical Literature Review:**
   * Provide concise, machine-generated summaries and relevant metadata from clinical trial documents to accelerate the literature review process for researchers and clinicians.
3. **User-Friendly Interface:**
   * Offer a simple, intuitive web-based tool requiring no coding experience, targeting domain experts (not just data scientists).
4. **Security & Privacy:**
   * Allow for local/private deployment (including via Docker), ensuring sensitive unpublished data does not leave the organization's environment.
5. **Extensibility:**
   * Build a modular backend that supports further customization (e.g., new extraction types, translation, Question Answering) and integration with downstream workflows.

# 3. Target Users & Stakeholders

* **Clinical Researchers**
  + For rapid review and structuring of clinical trials relevant to their field.
* **Medical Affairs & Regulatory Teams**
  + To support due diligence, preparation of regulatory submissions, and competitive intelligence.
* **Healthcare Data Scientists/Bioinformaticians**
  + For pre-processing clinical trial data in machine learning or meta-study pipelines.
* **Pharmaceutical/Biotech Organizations**
  + Interested in automating the review and summary of public/private clinical studies for R&D decision-making.
* **Educators and Trainers**
  + For demonstration of automated information extraction techniques in bioinformatics and clinical research education.

# 4. Primary Use Cases

Below are listed primary use cases for version v0.1.

## 4.1 Rapid Clinical Trial Screening

**Description**:

* A researcher uploads a batch of ClinicalTrials.gov XML files or exported summaries.

**Outcome**:

* Tool extracts structured PICO elements and summary tables to quickly assess study relevance for a systematic review.

## 4.2 Automated Literature Summarization

**Description**:

* A medical affairs professional needs quick, structured summaries of recent published trials on a target disease.

**Outcome**:

* Tool ingests texts or abstracts and returns concise, LLM-generated key point summaries.

## 4.3 Private On-Premise Extraction

**Description**:

* A pharmaceutical company processes proprietary clinical trial documents.

**Outcome**:

* Tool is deployed via Docker for use within internal secure networks, ensuring no data leaves the company.

## 4.4 Custom Integration for NLP Workflows

**Description**:

* A data scientist wants to integrate the extraction logic as a component in a larger pipeline.

**Outcome**:

* Tool’s backend can be called programmatically or via API, and the modular architecture supports future automation or integration.

# 5. Scope and Limitations

**Current Scope:**

* Extraction/summarization from English-language ClinicalTrials.gov data using open-source LLMs; local user upload via Streamlit UI; prototype stage.

**Not in Scope (for v0.1):**

* Real-time processing of very large datasets or full databases.
* Handling of non-English languages.
* Production-level multi-user authentication or detailed audit logging.

# 6. Success Criteria

* End-users can successfully upload and process at least 10 clinical trial files and receive interpretable, structured outputs.
* Extraction quality is comparable to or better than manual efforts in sample tasks.
* Tool can be deployed locally (including via Docker) with minimal configuration.

# 7. Future Directions (Optional)

* Expand to support PubMed or other literature sources.
* Add feedback/correction mechanisms to improve model output quality.
* Implement advanced QA (question answering) and user interaction features.
* Support multi-language and multi-modal data.

# 8. References

* [ClinicalTrials.gov](https://clinicaltrials.gov/)
* [HuggingFace Transformers documentation](https://huggingface.co/docs/transformers/index)
* [Streamlit documentation](https://streamlit.io/)