Paper ML01

Accelerating Clinical Study Startup: Automated eCRF Generation from Clinical Protocol Document

Saiteja Nalla, Lifio.ai, Bangalore, India
Kedar Deshpande, Lifio.ai, Bangalore, India

ABSTRACT

Designing electronic Case Report Forms (eCRFs) from clinical protocols remains a manual and resource-intensive process, often challenged by the variability and ambiguity of protocol structures. We present a novel AI-driven system to automate eCRF prediction from unstructured clinical protocol documents. Our approach combines large language models (LLMs), classical machine learning algorithms, and semantic similarity techniques to extract data from variably structured protocol content indicative of data collection needs. The system predicts relevant forms and granular data fields using a hybrid architecture that integrates generative models with classification techniques. These are then mapped to standardized eCRF components through ontology-driven alignment and semantic normalization. The proposed system ensures adaptability to different protocol formats and accurately captures the nuances of clinical data collection. This solution addresses a longstanding bottleneck in clinical study design, enabling faster, consistent, and cost-effective eCRF development within existing electronic data capture (EDC) systems.

INTRODUCTION

Clinical trials depend on carefully designed electronic Case Report Forms (eCRFs) to capture study data in a way that is accurate, consistent, and compliant with regulatory expectations. eCRFs serve as the practical implementation of a clinical protocol, translating study objectives, endpoints, and assessments into structured fields that are collected through Electronic Data Capture (EDC) systems. Despite their importance, eCRF design is still largely a manual activity, requiring substantial time, repeated interpretation of protocol text, and extensive input from experienced clinical data professionals.

Protocol documents can vary substantially from one study to another. Differences in layout, terminology, formats, and level of detail are common across therapeutic areas and trial phases. The same clinical concepts such as objectives, endpoints, safety assessments, or the schedule of activities may appear in multiple sections of a protocol and is not always described in the same way. In some cases, expected data collection is only implied through study procedures rather than stated directly. Because of this, designing eCRFs often requires repeated reviews of the protocol, follow-up discussions between teams, and several rounds of revision. These back-and-forth cycles slow study startup and increase the likelihood of mistakes introduced through manual interpretation.

Various approaches have been explored to reduce this effort, but none have fully addressed the problem. Rule-based and template-driven systems generally work only when protocols follow a predictable structure, and they tend to break down when documents deviate from those assumptions. NLP-based methods have shown value in identifying terms or labelling sections, but they rarely produce outputs that can be used directly for eCRF design without further manual work. More generative solutions offer greater flexibility, but they are not reproducible, deterministic, can be difficult to control and to audit, which is a significant concern in regulated clinical data environments. As a result, automation in this area has seen limited adoption in real-world study workflows

In this paper, we propose an end-to-end system for automated eCRF generation directly from clinical trial protocol documents that addresses these limitations. Our approach is based on the observation that translating a protocol into eCRFs is neither purely a language task nor a purely rules-based one.

Instead, it requires a combination of semantic understanding and structured alignment with established data collection standards. To support this, we integrate large language models with classical machine learning techniques and semantic similarity methods within a modular and transparent architecture.

The system first converts protocol content into a structured digital representation, covering both narrative text and tabular schedules of activities. It then identifies protocol sections likely to correspond to eCRF forms using interpretable classification models, followed by targeted generation of candidate form names. Instead of treating these outputs as final, the system grounds predicted forms and data items against standardized or study-specific eCRF libraries using a combination of lexical and semantic matching techniques. This step ensures alignment with existing standards, such as CDISC-based and study specific libraries, while allowing flexibility for sponsor-defined requirements.

By handling form selection and individual data element definition as separate steps, and by pairing generative outputs with explicit similarity checks against existing libraries, the approach allows flexibility without sacrificing control. The eCRFs produced through this process can be carried forward directly into EDC configuration, with minimal rework. In practice, this approach materially shortens the transition from protocol approval to study setup by reducing the amount of manual interpretation required upfront. Across repeated or closely related studies, it produces more consistent eCRF designs and reduces rework caused by interpretation differences between teams. The reduction in manual handling also lowers the incidence of avoidable specification errors. At scale, the same workflow supports concurrent development across multiple protocols without increasing operational overhead, while continuing to adhere to established regulatory, compliance, and data quality requirements expected in clinical trial execution.

DATA EXTRACTION AND PROTOCOL DIGITIZATION

Clinical trial protocols are usually written as lengthy documents intended for review by clinical writers, subject matter experts, reviewers, and study teams, not for direct computational use. They are commonly present as PDFs or word-processed files with varied formats, structures and combine free-text descriptions with structured sections, tables, cross-references, and formatting elements. While this format works well for human readers, it creates challenges when the content needs to be processed programmatically. For automated eCRF generation, it is essential that the protocol be converted into a digital representation that preserves both what is written and where it appears within the document.

For this reason, protocol digitization is treated as a core preparatory step in our workflow. Rather than viewing it as a simple text extraction task, we approach it as a process of reconstructing the protocol in a form that downstream models can reliably interpret.
The protocol documents are processed at the outset to convert them into a form that can be used for analysis while still reflecting how they were originally written. Instead of treating the protocol as plain text, the parsing step preserves the logical and overall organization of the document. Section and subsection headings, numbering, and paragraph structure are kept so that the extracted content follows the same flow and layout as the source file.

During this process, each extracted passage is linked to the section of the protocol in which it appears. This is necessary because the same terminology is often reused across different parts of a protocol with different intent. For example, a laboratory test described in the safety section has a different role than when it appears in the context of efficacy assessments. Keeping track of where each passage comes from allows these differences to be handled explicitly in later stages, instead of assuming that similar wording always implies the same requirement.

After the protocol text has been normalized, it is split into smaller portions so that it can be examined in a more controlled way. The boundaries of these portions are not defined by an arbitrary length.

Instead, they follow the way the protocol itself is organized, using paragraph breaks, subsection changes, and numbered descriptions of study procedures as natural cut points. Each portion is then tagged with the full section reference from which it originated (for example, 6.2 Study Assessments → 6.2.3 Laboratory Evaluations), so its placement within the protocol is preserved. Keeping this information allows similar statements to be interpreted differently when they appear in different parts of the document, rather than being treated as interchangeable based on wording alone.

Once the protocol text has been extracted, it is reviewed and standardized to account for differences in how protocols are written and formatted. Section names and numbering are aligned to a common convention, which helps avoid inconsistencies introduced by sponsor or study-specific styles. Material that does not add clinical or operational meaning, such as page headers, footers, page numbers, and repeated boilerplate language is removed at this stage. At the same time, care is taken to retain elements that convey study intent or execution details, including structured lists, annotations, and references to specific procedures. This selective cleanup reduces background clutter while preserving the information required for downstream processes and subsequent analysis.

TABLE EXTRACTION

Clinical trial protocols rely heavily on tables to communicate operational and clinical requirements that are not always described in full detail in the narrative text. In practice, many protocol decisions, particularly those related to visit structure and assessment timing are defined almost entirely within tabular sections. The Schedule of Activities (SoA) tables are especially important, as they specify visit names, timing, assessment frequency, and allowable visit windows. Other tables frequently describe laboratory collections, dosing procedures, eligibility conditions, and safety monitoring plans.

For this reason, table extraction is treated as a central element of the protocol interpretation process rather than as a supplementary step. The system identifies and processes all tables contained in the protocol document, including but not limited to SoA tables. A table specific transformer model, fine-tuned on clinical document layouts, is used to identify table boundaries, reconstruct row and column structure, and extract individual cell content. This approach allows the system to handle tables with complex formatting, such as merged cells, multi-level headers, embedded footnotes, and visit dependent annotations, which are common in clinical trial protocols.

SoA tables receive additional processing to derive a structured representation of the study's visit framework. Information such as visit labels, study days or weeks, cycles, visit windows, and assessment indicators is extracted and normalized into a consistent format. Common symbols used in SoA tables, including markers such as "X," filled circles, or footnote references, are interpreted in context and converted into explicit indicators of whether an assessment is required, optional, repeated, or conditionally performed. The resulting structured SoA representation provides a stable temporal reference that supports downstream form identification and scheduling alignment, helping ensure that generated eCRFs accurately reflect the protocol defined visit plan.

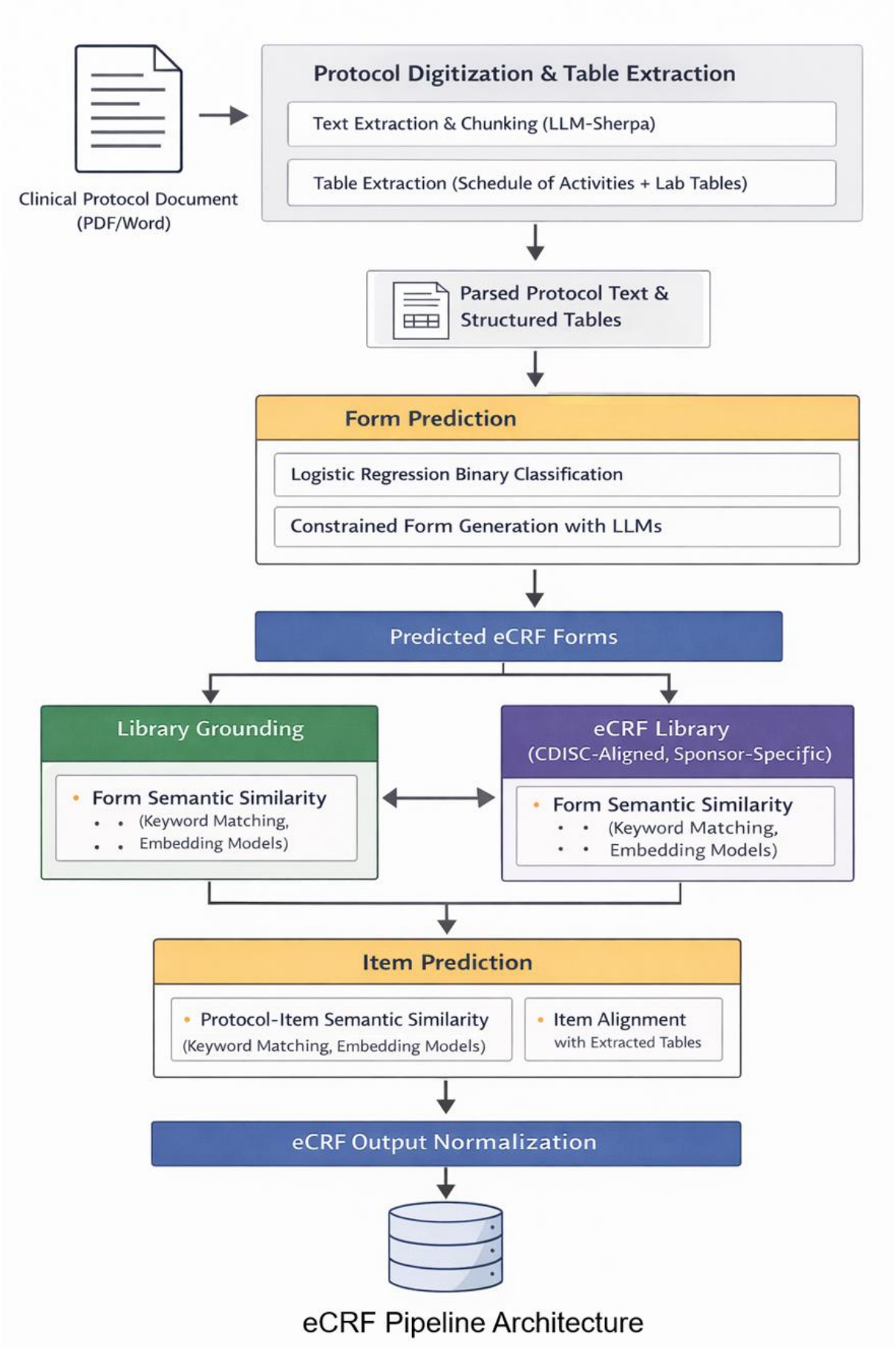
Beyond the SoA, the extraction pipeline also processes other clinically relevant tables distributed throughout the protocol. These include laboratory assessment tables, specimen collection schedules, dosing and administration summaries, and safety or pharmacokinetic sampling matrices. Such tables frequently contain detailed data collection requirements, such as specific laboratory analytes, specimen types, collection timing, fasting conditions, or handling instructions, that may be only partially described in the surrounding text. Extracting these details directly from tables reduces the risk of missing required data elements or relying on incomplete narrative descriptions.

After extraction, table content is represented in a single machine-readable format, but without stripping away the context in which the table appears. Along with the raw table values, the system retains information such as the protocol section where the table is located, the table's title, any accompanying footnotes, and the practical meaning implied by its rows and columns. This

surrounding context is essential for interpreting the table correctly and for associating its contents with the appropriate eCRF forms and data fields. As an example, laboratory panels specified in protocol tables can be linked directly to laboratory eCRF sections, while visit-level indicators can be interpreted in light of the scheduling logic applied at the form level.

By consistently extracting and organizing information from both Schedule of Activities tables and other protocol tables, this stage helps ensure that protocol requirements are captured even when they are only implicitly defined or confined to tabular formats. Using this structured table data alongside narrative text analysis, more complete, consistent, and dependable eCRF forms and item generation can be generated.

SYSTEM ARCHITECTURE



eCRF Pipeline Architecture

FORMS PREDICTION

The identification of eCRF forms referenced within a clinical protocol represents a critical intermediate step between the protocol interpretation and structured data collection in the overall study design. Protocols rarely enumerate eCRFs explicitly or consistently, instead, form level requirements are often implied through Schedule of Activities table, descriptions of assessments, procedures, or evaluations scattered across multiple sections. To address this challenge, the proposed system adopts a multi-stage form prediction strategy that combines lightweight classification with targeted generative modelling and library grounding.

Following protocol digitization, the section-aware text chunks serve as the primary input for form detection. Each chunk is first evaluated using a supervised binary classification model based on logistic regression to determine whether it contains information indicative of an eCRF form. This machine learning model is trained to classify if the text chunk has any text that is indicative of a data capture element (such as clinical assessments, laboratory evaluations, study specific data or safety monitoring) or an CRF description mentioned either contextually or directly. The choice of a classical, interpretable classifier at this stage provides robustness, low latency, and stable performance across diverse protocol styles.

Text chunks classified as form-relevant are then passed to a fine-tuned large language model that generates candidate eCRF form names. Rather than performing open-ended text generation, the model operates in a constrained, task-specific manner, extracting or inferring concise form names grounded in the clinical context of the chunk. This step allows the system to capture implicit form concepts even when no explicit form terminology is used in the protocol, for example, inferring a “Vital Signs” or “Adverse Events” form from descriptive assessment language.

The initial set of candidate form names is further refined by incorporating information extracted from the Schedule of Activities and other tables in the protocol. Tabular signals such as visit-specific indicators, assessment frequency, and procedure groupings are used to confirm, merge, or adjust candidate forms. In practice, this refinement step reduces unnecessary duplication and helps ensure that the resulting forms align with the way assessments are planned and repeated across study visits. Rather than relying solely on narrative descriptions, the approach considers how procedures are represented in tables and scheduled over time. Taking both sources into account allows the identified forms to reflect not just what the protocol describes in words, but how study activities are actually carried out during the conduct of the trial.

After this consolidation, the resulting set of form candidates is compared against an established eCRF library to bring the naming into line with existing standards and sponsor conventions. Depending on the study context, this reference library may be based on widely used industry standards, such as CDISC-aligned eCRFs, or on a curated catalogue maintained by a specific sponsor or organization. Alignment is carried out using a combination of complementary matching strategies rather than a single similarity measure. These include exact string comparisons, fuzzy and edit-distance based matching, as well as cosine similarity computed over sentence-level embeddings produced by an ensemble of transformer models. The matching thresholds for each method are determined empirically, with the aim of balancing recall and precision while preserving robustness when applied to previously unseen protocols.

Through this hybrid process, the proposed approach produces a normalized and de-duplicated set of eCRF forms that are both protocol relevant and library aligned. By separating form relevance detection, form name generation, and library grounding into distinct but tightly coupled stages, our system achieves high adaptability to protocol variability while preserving the determinism and traceability required for clinical data management workflows.

ITEMS PREDICTION

After protocol relevant eCRF forms have been identified and mapped to the form library, the next task is to determine which individual data items within each form are required by the protocol. Compared with form-level identification, item-level prediction is more detailed and often more challenging. Protocols rarely specify data items in the form of explicit questions. Instead, item requirements are commonly embedded in narrative descriptions, conditional language, or annotations within tables. To address this, the proposed system adopts a context-aware, similarity-driven approach to item prediction.

For each predicted and grounded form, the system first retrieves protocol sections that are most relevant to that form. This is achieved by computing semantic similarity between the normalized form name and the full set of section-aware text chunks extracted during protocol digitization. Only text segments that show a clear relationship to the form are kept and used as evidence for item identification. Focusing on these form-relevant passages helps avoid unnecessary noise and prevents unrelated parts of the protocol from affecting item selection.

The retained protocol text is then reviewed in relation to all items defined for the corresponding form in the eCRF library. Each item is considered in terms of its standard name as well as the question or prompt used to collect the data, together with any supporting metadata such as data type or controlled terminology, when available. To determine whether a particular item is mentioned or implicitly required by the protocol, the system applies a combination of similarity measures that compare the protocol text with these item representations. These measures include semantic similarity based on transformer generated embeddings, as well as non-semantic techniques such as lexical overlap, fuzzy matching, and edit-distance comparisons.

To improve robustness to variation in clinical language, embeddings are generated using an ensemble of sentence transformer models, each capturing different aspects of semantic similarity. The similarity thresholds used at the item level are adjusted based on observed behaviour across protocols, reflecting differences in how clearly individual items are described and how commonly they occur. This combined matching approach makes it possible to identify required items even when the protocol does not mention them directly, such as when a clinical concept, measurement, or condition is described without using the exact terminology found in the eCRF library.

Item identification also draws on information extracted from protocol tables. This is especially relevant for highly structured areas, including laboratory testing, vital signs collection, and pharmacokinetic sampling, where key data requirements are often defined primarily in tabular form rather than in narrative text. Table derived elements, such as laboratory tests, analyte names, specimen types, or collection conditions are matched against library items to ensure that all required data elements are captured. Incorporating this tabular evidence helps improve coverage for forms that depend heavily on structured protocol content.

The outcome of this stage is a form-specific set of predicted data items, each linked explicitly to a corresponding definition in the eCRF library. Keeping a clear connection between the protocol text, the items identified from that text, and the corresponding entry in the standardized library makes the results easier to trace and review. This linkage also supports validation activities by allowing reviewers to see how each data item was derived. By grounding item selection at this level, the system converts unstructured protocol descriptions into structured eCRF specifications, supporting automated form generation while still ensuring alignment with the protocol and established organizational standards.

EVALUATION AND PERFORMANCE

The system was evaluated using a separate set of clinical trial protocols that were not used during development. To reflect real-world use, the outputs were reviewed manually by experienced clinical data professionals, including eCRF designers, who assessed whether the predicted forms and data

items aligned with what would be expected from a standard protocol driven build. For comparison, reference eCRF structures were created through manual protocol review and alignment with existing form libraries, following established study design practices.

At the form level, the system showed strong performance across a range of protocol formats and therapeutic areas. At the form level, the system was able to identify most of the forms that would normally be created during manual eCRF design. From a quantitative standpoint, the measured precision, recall, and F1-score were all 0.95, with an overall accuracy of 95%. This level of performance was not limited to a small subset of studies. Similar results were seen across protocols and therapeutic areas that differed in format and wording, indicating that form-level requirements were identified reliably even when studies were described in different ways.

Item level performance showed a more moderate trend, which aligns with the nature of the task. Defining individual data elements often depends on subtle or implied language in the protocol, rather than on explicit statements. At this level, the system achieved a precision, recall, and F1-score of 0.85, with an overall accuracy of 85%, with most discrepancies arising in cases where protocol wording left room for interpretation. Most mismatches were traced back to ambiguous language in the protocol or to data collection needs that were highly specific to individual studies and not consistently described across documents.

Based on review by subject matter experts and eCRF designers, the generated forms and data items were generally consistent with what would be expected from library aligned, manually designed eCRFs. In many cases, the outputs could be carried forward into study build with little or no modification. Reviewers also noted a clear reduction in the amount of manual work needed during form design, particularly in the early stages of study setup. As a result, study timelines could be shortened while maintaining the same expectations around data quality, standards compliance, and regulatory review.

CONCLUSION

In this study, we describe a complete approach for generating eCRFs directly from clinical trial protocols, with the aim of addressing a well-known bottleneck in study startup. The system brings together protocol digitization, structured table extraction, machine learning based classification, language modelling, and similarity-based matching to translate protocol text into eCRF forms and data items that are aligned with existing libraries. Rather than treating protocols as unstructured text alone, the approach preserves document structure and context, which proved critical for producing usable outputs.

An important design choice was to keep the system modular rather than treating eCRF generation as a single end-to-end step. Digitization, table extraction, form identification, form inference, library matching, and item-level alignment are handled separately, but in a coordinated way. This separation made it easier to accommodate differences in protocol structure and writing style, while still retaining the level of oversight, consistency, and traceability that is expected in regulated clinical data workflows. Evaluation on previously unseen protocols showed strong performance at both the form and item levels, and expert review confirmed that the outputs closely matched manually designed eCRFs while requiring substantially less manual effort.

Taken together, these results suggest that AI can be applied in a practical and controlled way to support core clinical operations, rather than being limited to exploratory text analysis. By taking over some of the more repetitive and time-consuming aspects of eCRF design, this approach can shorten the early phases of study setup and help bring greater consistency to how data are collected across related trials. This, in turn, allows clinical and data management teams to spend more time on study-specific decisions that benefit from experience and judgment rather than manual transcription. Ongoing work will focus on integrating agentic-ai architectures directly into EDC build and

configuration workflows and on adapting the approach to studies with more specialized or less conventional designs.

REFERENCES

1. Wang, Y., Afzal, N., Fu, S., Wang, L., Shen, F., Rastegar-Mojarad, M., & Liu, H. *MedSTS: A Resource for Clinical Semantic Textual Similarity*. arXiv. Preprint. 2018.
   This work introduces a large resource for clinical semantic textual similarity, which is foundational for semantic comparison and information extraction tasks.
2. Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., & Hurdle, J. F. *Extracting information from textual documents in the electronic health record: a review of recent research*. Yearbook of Medical Informatics. 2008.
   A foundational review of clinical text extraction methods, useful for contextualizing clinical NLP tasks.
3. Ye, C. *Leveraging medical context to recommend semantically similar terms for chart reviews*. BMC Medical Informatics and Decision Making. 2021;21:353.
   This study illustrates semantic similarity techniques in clinical language tasks, relevant to similarity modeling in form and item matching.
4. Meystre, S. M., Lovis, C., Bürkle, T., Tognola, G., Budrionis, A., & Lehmann, C. *Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress*. IMIA Yearbook of Medical Informatics. 2017;38:38–52.
   A review of clinical data reuse and information extraction, useful for positioning automated extraction tasks in broader health informatics.
5. *Clinical Data Interchange Standards Consortium (CDISC)*. Wikipedia.
   Overview of CDISC standards, including operational models that define eCRF metadata and semantic structures relevant to standard-aligned eCRF generation.
6. *Case Report Form (CRF)*. Wikipedia.
   Defines essential eCRF concepts and their role in clinical trials, grounding the paper in standard clinical terminology.
7. Trial Bank Project: Chapman, et al. *Automated Information Extraction of Key Trial Design Elements.* BioMed Central / PMC.
   Demonstrates a text classification + pattern matching approach for extracting clinical trial design information, highlighting tasks comparable to protocol interpretation.
8. Kumar, A., et al. *LLM-Sherpa: Layout-Aware Document Parsing and Chunking for Large Language Models*. Open-source project. https://github.com/nlmatics/llm-sherpa
9. Smock, B., Pesala, R., & Abraham, R. (2022). *PubLayNet and Table Transformers for Structured Table Extraction*. Proceedings of ICDAR 2022. https://github.com/microsoft/table-transformer
10. OpenAI. (2025). ChatGPT (GPT-4o, 2 July version) [Large language model]. https://chat.openai.com/

CONTACT INFORMATION

Author Name: Saiteja Nalla
Company: Lifio.ai
Email: saiteja.nalla@lifio.ai
Website: https://lifio.ai/

Co-Author Name: Kedar Deshpande
Company: Lifio.ai
Address: Bangalore, India
Work Phone: +91 9845242861
Email: kedar.deshpande@lifio.ai
Website: https://lifio.ai/