Paper ML26

# Gen-AI for Statistical Reporting and Clinical Study Report (CSR) Writing Leveraging CDISC Analysis Result Standard (ARS) Model

Navin Dedhia, Clymb Clinical, Burlington, MA, United States

Bhavin Busa, Clymb Clinical, Burlington, MA, United States

## ABSTRACT

Imagine a future where writing Clinical Study Reports (CSRs) is faster, more accurate, and nearly effortless — this is the promise of our Gen-AI-driven model leveraging CDISC Analysis Results Standard (ARS). In this paper, we present a model fine-tuned on ARS and Analysis Results Data (ARD) metadata that streamlines the interpretation of analysis results and generates narrative summaries for the CSR's safety and efficacy sections. By processing pre-structured ARD datasets, the model identifies key findings and produces detailed summaries, significantly reducing manual effort. This approach enhances efficiency, saving time for biostatisticians, clinicians, and medical writers, while ensuring accuracy and consistency across Tables, Figures and Listings (TFL) outputs and CSR text. We also discuss the pros and cons of building this solution, including scalability, regulatory alignment, and data privacy challenges.

## INTRODUCTION

Clinical Study Reports (CSRs) authoring is a critical yet time-consuming task in clinical research. These reports provide a comprehensive summary of a clinical trial's design, methodology, results, and conclusions, serving as a key document for regulatory submissions. Traditionally, CSR writing involves extensive manual effort, requiring biostatisticians, medical writers, and clinicians to interpret statistical outputs from Tables, Figures and Listings (TFLs) and translate them into structured narratives.

With advancements in Generative AI (Gen-AI) and structured clinical data models like CDISC's Analysis Results Standard (ARS)[1, 2, 3], there is an opportunity to automate large parts of this process. This paper presents a Proof of Concept (POC) Gen-AI-driven prompt engineering model trained on ARS and Analysis Results Data (ARD). The model can generate accurate, consistent CSR narratives while significantly reducing manual workload.

### Background

Clinical trials generate vast amounts of data, typically structured in compliance with CDISC standards such as the Study Data Tabulation Model (SDTM) and Analysis Data Model (ADaM). These datasets are analyzed to produce TFLs, which form the basis of CSR sections such as safety and efficacy summaries.

Despite the availability of structured datasets, translating TFLs into CSR text remains largely manual, repetitive, and prone to inconsistencies. Biostatisticians and medical writers often interpret statistical outputs in RTF or PDF format, identify significant findings, write structured summaries, and ensure consistency across different sections of the CSR. This manual process is slow, resource-intensive, and susceptible to human error.

Gen-AI provides a powerful solution for automating CSR narrative generation by leveraging ARD datasets, which contain machine-readable analysis results paired with descriptive metadata. This metadata delivers essential context, enabling the model to interpret statistical outputs accurately and produce structured, high-quality narratives. ARDs, as defined by CDISC standards, offer a reliable foundation by including both the results of statistical analyses and the necessary contextual information, such as population, parameters, and timepoints. This structured, metadata-rich input streamlines the process, enhances traceability, and ensures consistency across TFLs and regulatory submissions.

## THE GEN-AI MODELS USED BY CLYMB

Generative AI can be implemented in various ways to assist in creating Clinical Study Reports (CSRs). For the scope of our POC projects, we have limited it to the below listed methods.

### Prompt Engineering

This involves crafting specific instructions or questions that guide the AI in generating accurate and relevant text. In this method, prompts can be designed for different CSR sections, such as summarizing safety outcomes based on ARD datasets. This approach is particularly beneficial due to its ease of implementation and flexibility, allowing for quick fine-tuning based on the specific section being generated. However, the quality of the output is highly dependent on the design of the prompts, making it essential to craft them carefully to avoid irrelevant or incomplete content.

### Embedding Models

Another essential approach is the use of embedding models, which convert textual data into numerical vectors. This transformation allows AI systems to understand and process complex relationships between data points. Embedding models are especially useful when working with large ARD datasets as they help the AI identify and link key findings, such as connecting specific adverse events to patient groups. While embedding models enhance the AI's comprehension and improve the accuracy of the generated text, they require high-quality training data and substantial computational resources.

### Open-source Models

Open-source language models, such as those based on GPT architectures, provide another promising approach. These models offer a flexible foundation that can be fine-tuned using clinical data, including CDISC ARS and ARD metadata. The primary advantage of using open-source models is their cost-effectiveness and the ability to customize them according to specific needs without the need to build models from scratch. However, this approach requires technical expertise for fine-tuning and maintaining model performance while ensuring data privacy and regulatory compliance.

## INSTRUCTIONS, CONTEXT AND USER INPUTS

Our first Proof of Concept employs Prompt Engineering within a Gen-AI model (GPT-4o) to automate CSR narrative generation. In this method, instructions, context, and user inputs are essential components.

As part of these instructions, we provide CDISC standards, including models and implementation guides for SDTM, ADaM, ARS, and the corresponding user guide. Additionally, we incorporate example protocols, Statistical Analysis Plans (SAPs), and CSRs. These resources serve as context for the AI during narrative generation.

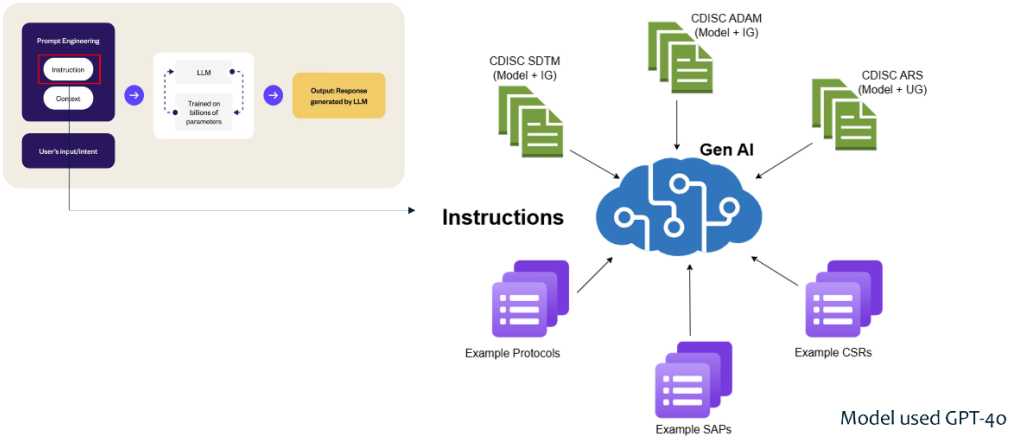
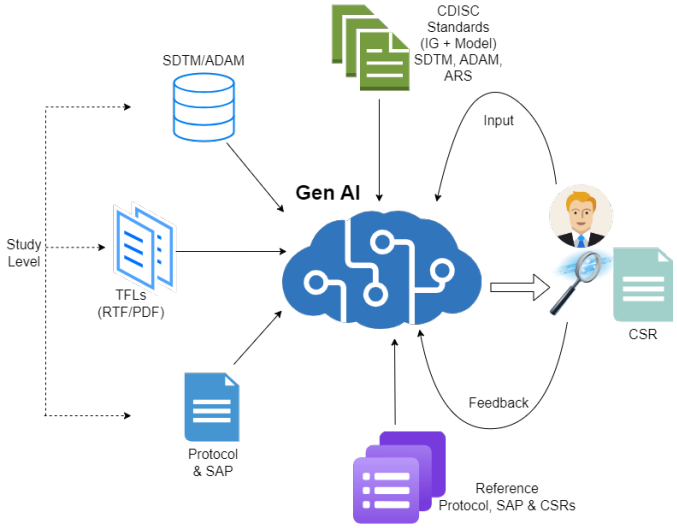



Figure 1: Providing Instructions (CDISC documentation and examples) to Gen AI Model [4]

The model is then supplied with study-level documents and data by the user providing context, including the protocol, SAP, SDTM, ADaM, and summarized results. The process of providing summarized results to the Gen-AI model follows two distinct approaches:

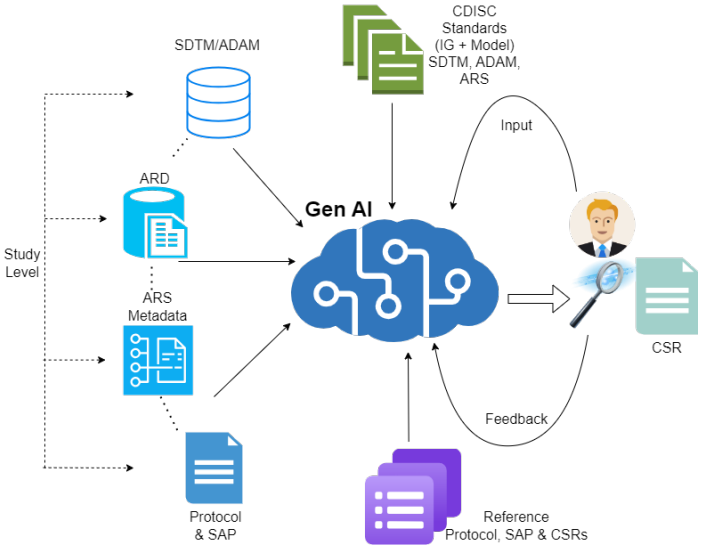
### Approach 1: TFLs in RTF/PDF Formats

These files, commonly used in traditional CSR writing, present challenges due to their unstructured nature. The AI model must parse and interpret text-based tables, which requires complex logic to handle formatting variations. This process increases the risk of errors and inconsistencies when extracting key findings and integrating them into cohesive narratives.




### Approach 2: ARS-Metadata and ARD Datasets

This approach utilizes structured, machine-readable datasets aligned with CDISC ARS standards. ARD datasets provide results with associated metadata that offers context and ensures consistency across CSR sections. This method simplifies narrative generation, enhances accuracy, and supports efficient scaling across multiple studies.




Comparing these approaches, we observed that ARS-metadata and ARD datasets deliver more reliable, accurate, and consistent results with less preprocessing effort, with the support of structured metadata. This context ensures that the AI model can accurately interpret statistical outputs and generate reliable narratives, making ARS-metadata and ARD the preferred input method for Gen-AI-driven CSR automation.

Furthermore, providing ARD along with linked metadata enables a more interactive approach, allowing users to drill down into contributing patient-level data and ask specific questions about the summarized results. This approach ensures that the model can retrieve and present patient-level insights more reliably when needed.

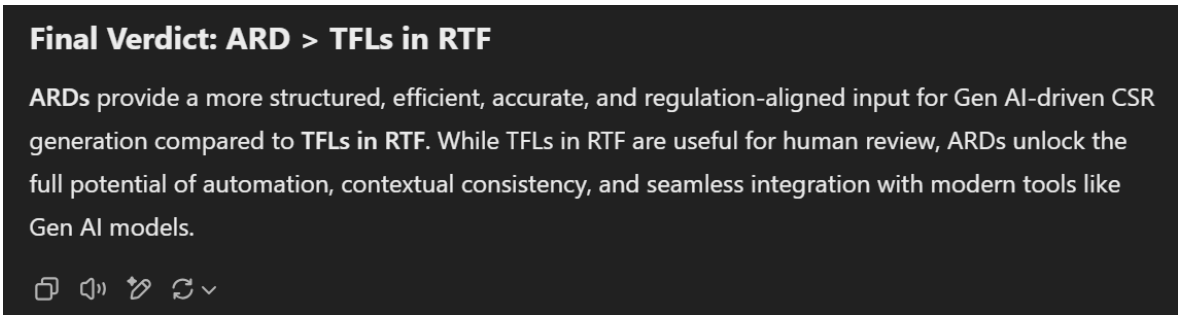



Figure 2: "Final Verdict" - a Summary by the Gen AI Model following provided Instructions and Context

We then asked the model: 'Why prioritize summarized results (ARD)? Can't we simply use patient-level data (SDTM/ADaM)?'

The response highlighted that ARD enhances processing efficiency by 3–5 times compared to patient-level data, minimizes risks of misinterpretation, and improves narrative quality through clear, objective-driven outputs. Additionally, ARDs allow for human intervention, ensuring critical safety signals are not missed and methodologies are correctly applied, resulting in reliable and high-quality CSR outputs. Consequently, ARDs are recommended as the primary input for AI-driven CSR automation, with patient-level data reserved for exploratory analyses. From a regulatory perspective, this supported the summary of the verdict as shown in Figure 3 below.

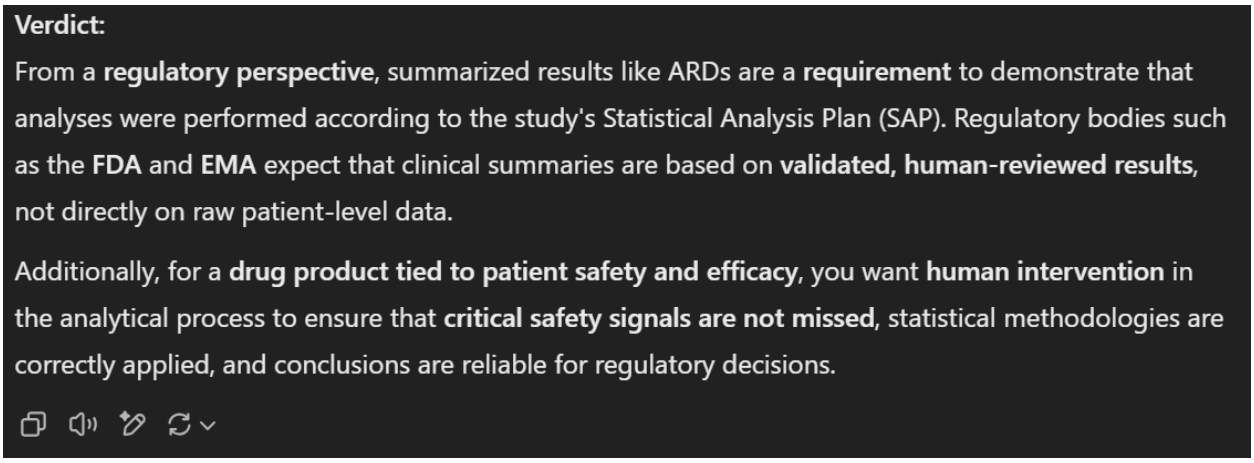



Figure 3: Regulatory Perspective - Gen AI Model's Summary on the Use of ARD Compared to Patient-Level Data

Additionally, prompt engineering is employed to integrate human-in-the-loop validation within the workflow. Prompts are used to generate CSR sections that can be easily reviewed and refined by medical writers and statisticians, ensuring that the final CSR narratives uphold accuracy, regulatory standards, and contextual relevance.

## GEN-AI MODEL PERFORMANCE ACROSS CSR SECTIONS

Our Gen-AI-driven model utilized inputs from the Study Protocol, Statistical Analysis Plan (SAP), ADaM datasets, Analysis Results Data (ARD), and Analysis Results Standard (ARS) metadata to automate several key sections of a Clinical Study Report (CSR).

For this Proof of Concept (POC), we utilized the CDISC Pilot Study package [5], which provided the necessary study documentation and ADaM datasets. Additionally, we accessed ARS-compliant packages through the CDISC eTFL portal [6, 7], which included ARS metadata and ARD datasets derived from the same CDISC Pilot Study.

The level of effort required varied across sections, with some needing more prompts than others for accurate generation.

**1. Synopsis (Section 1):**
The model generated summaries of study objectives, key endpoints, and overall study results by processing information from the Protocol, SAP, ARD, and ARS. Minimal prompt adjustments were needed due to the straightforward structure of these summaries.

**2. Introduction (Section 2):**
The study rationale and background, derived primarily from the Study Protocol, were automated with minimal manual input, ensuring consistency with the study design.

**3. Study Objectives (Section 3):**
Primary, secondary, and exploratory objectives were extracted directly from the Study Protocol and SAP with high accuracy.

**4. Study Design (Section 4):**
The model produced descriptions of the study population, treatment groups, and randomization procedures using details from the Protocol and SAP.

**5. Efficacy Evaluation (Section 8):**
Efficacy-related information was not included in this POC, but the model is designed to generate narratives for primary and secondary endpoints, statistical methods, and relevant tables using ARD and ARS metadata when provided.

**6. Safety Evaluation (Section 9):**
Safety narratives covering adverse events, laboratory results, vitals, and ECG findings were generated using ARD, ARS, and ADaM datasets.

**7. Statistical Methods (Section 11):**
Statistical analysis procedures were automated using information from the SAP and ARS metadata with minimal manual intervention.

**8. Results (Section 12):**
The model successfully summarized demographic and baseline characteristics, subject disposition, and safety outcomes by processing ARD, ARS, and ADaM datasets.

**9. Conclusions (Section 13):**
While the model generated preliminary conclusions based on the safety findings, this section required significantly more prompts than others. As a result, we recommend increased human oversight during final validation to ensure accuracy and regulatory compliance.

This variation in prompt requirements underscores the importance of carefully tailoring instructions and metadata inputs for each section, particularly for complex or interpretation-heavy components of the CSR.

**Takeaway:** Gen AI can generate good-quality CSR narratives, but *Gen AI can make mistakes.* Human-in-the-loop validation is necessary to maintain clinical accuracy and regulatory compliance.

## KEY CONSIDERATIONS FOR GEN AI-BASED CSR WRITING

The automation of CSR writing using Gen-AI involves a structured, multi-step approach that integrates structured clinical trial datasets, statistical analysis results, and AI-driven natural language processing (NLP) techniques. This approach ensures that AI-generated narratives maintain accuracy, consistency, and regulatory compliance while reducing the manual effort required for CSR development.

To fine-tune a model specifically on clinical and regulatory data, especially for tasks like generating Clinical Study Reports (CSRs), it's essential to curate a training dataset rich in domain-specific content that reflects both the structure and language of clinical documentation. Here's a step-by-step approach you could use:

**1. Data Collection**

- Clinical Study Reports (CSRs): Collect de-identified CSR examples, especially those including the Analysis Results Standard (ARS) and Analysis Results Data (ARD) tables and summaries.
- CDISC-Compliant Datasets: Ensure you have CDISC-based datasets, including SDTM (Study Data Tabulation Model) and ADaM (Analysis Data Model) formats, as these are the standard structures for clinical data submission to regulatory bodies.
- Regulatory Guidelines: Use publicly available guidelines from regulatory authorities like the FDA and EMA, which often outline expectations for CSR formats, statistical methods, and interpretation guidelines.

**2. Data Preprocessing**

- Text Extraction: Extract text from the CSR documents, focusing on structured sections like safety summaries, efficacy summaries, and statistical methods. Break down content by ARS and ARD components, highlighting patterns the model should learn to recognize.
- Standardize Language and Terminology: Ensure terminology is consistent with regulatory language. Create a glossary for terms often used in CSRs and regulatory guidelines.
- Tokenization and Formatting: Preprocess the text into tokenized segments, where tables and figures are annotated for easy identification. Special formatting tags can help the model learn how to interpret and generate structured text.

**3. Create Training Pairs (Input-Output Examples)**

- Annotation of Input Data: Pair ARD tables and ARS metadata as inputs with corresponding narrative summaries as outputs. This mapping will help the model learn how data translates into narrative text.
- Examples for Different Sections: Provide separate examples for each section of the CSR, like safety results, efficacy results, and patient disposition, so the model learns to generate context-appropriate summaries.
- Template-Based Fine-Tuning: Use templated CSR sections with placeholders for specific data points, which can make the model more adaptable to different studies and datasets.

**4. Fine-Tuning Strategy**

- Supervised Fine-Tuning: Use supervised learning to fine-tune the model, feeding it annotated CSR examples and fine-tuning it to predict CSR text from ARD/ARS inputs.
- Language Models: If available, start with a pre-trained language model already fine-tuned on medical or clinical data, such as BioBERT, as the base model. These models are more attuned to medical terminology and will perform better on clinical text. After exploring multiple models including BioBERT, Llama and GPT-40, we decided to continue building the POC on GPT-40. As an alternate approach, use of embeddings or vectorized data representations can be utilized. After storing the ARD/ARS into a vector database, they can be retrieved and fed to the model when generating a CSR. Being a POC, this approach relied on prompt engineering to achieve desired results
- Evaluation Metrics: Use specific metrics like BLEU for language similarity and compliance checks for adherence to regulatory terminology.

**5. Validation and Testing**

- Internal Review by Clinical Experts: Clinical experts should review generated CSRs to assess the quality of interpretations and summaries.
- Comparison Against Existing Reports: Compare AI-generated outputs to existing CSRs for the same dataset to ensure that interpretations align.
- Regulatory Compliance Testing: Run tests to ensure output meets regulatory standards and CDISC compliance.

**6. Continuous Refinement**

- Feedback Loops: Set up a feedback loop from clinical reviewers to further fine-tune the model.
- New Data Incorporation: As new CSRs and guidelines become available, incorporate them into the dataset for ongoing model improvement.

This approach will help the model accurately generate structured and compliant CSR narratives, improving both the speed and quality of CSR development.

By combining structured clinical data, AI-driven statistical interpretation, NLG-based text generation, and human validation, this approach streamlines CSR development while ensuring compliance, accuracy, and efficiency. This hybrid AI-human collaboration enables institutions to produce high-quality CSRs faster, reducing time-to-submission and enhancing regulatory readiness.

## IMPLEMENTATION AND KEY BENEFITS

**Increased Efficiency**

By automating CSR writing, manual effort is significantly reduced, allowing statisticians and writers to focus on data interpretation rather than repetitive text generation. This results in faster report generation, improving overall efficiency.

**Improved Consistency**

AI ensures uniform interpretation of statistical results across different CSR sections and reduces the risk of inconsistencies between TFLs and narrative text.

**Enhanced Accuracy**

AI-based interpretation eliminates human errors, such as rounding inconsistencies or misreported p-values, ensuring the reliability of the generated narratives.

**Scalability**

This approach can be scaled across multiple studies and therapeutic areas. AI models can also be fine-tuned for specific therapeutic domains, improving contextual relevance and applicability across different trials.

## ENSURING SECURITY AND DATA PRIVACY

The integration of Generative AI (Gen-AI) into Clinical Study Report (CSR) writing introduces critical considerations for data security and regulatory compliance. Given the sensitive nature of clinical trial data, it is imperative to implement a robust security framework that protects Protected Health Information (PHI) and Personally Identifiable Information (PII) while maintaining adherence to global regulatory standards such as HIPAA, GDPR, and 21 CFR Part 11.

To ensure secure data handling, AI models should operate on de-identified and anonymized datasets, where patient identifiers are masked, demographic details are generalized, and direct identifiers such as

names and addresses are removed. This ensures that AI processes statistical insights rather than identifiable patient-level data, reducing the risk of privacy breaches.

Deployment models must also align with organizational security requirements, with AI systems being hosted in on-premises secure infrastructure or regulatory-compliant cloud environments (e.g., AWS, Azure, Google Cloud). Additionally, role-based access control (RBAC) mechanisms should be enforced to limit AI-generated outputs to authorized personnel, while comprehensive audit trails track all data access and modifications, ensuring compliance with regulatory requirements.

To further protect sensitive clinical data, organizations should employ state-of-the-art encryption mechanisms, including AES-256 encryption for stored data and TLS encryption for data transmissions. These measures ensure that data always remains protected, whether at rest or in transit. Additionally, federated learning offers a secure AI training approach, allowing models to learn from distributed clinical trial datasets without requiring centralized data storage. This method enables AI to gain insights while ensuring patient data never leaves its originating institution, mitigating the risk of unauthorized access.

Ensuring regulatory compliance and transparency is paramount in AI-driven CSR writing. AI models should be designed to align with ICH E3 guidelines, FDA and EMA submission standards, and ISO 27001 security protocols. To maintain regulatory trust, AI-generated outputs must be fully auditable, enabling reviewers to trace narrative outputs back to the statistical datasets they are derived from. A human-in-the-loop (HITL) validation process is essential, ensuring that biostatisticians and medical writers verify AI-generated content before submission.

By implementing these security best practices, AI-driven CSR automation can revolutionize regulatory reporting while ensuring data integrity, privacy, and compliance. These measures allow pharmaceutical companies, CROs, and regulatory agencies to leverage AI's efficiency while maintaining strict security and governance controls, ensuring that AI-generated CSRs are both reliable and secure for regulatory submission.

**Addressing Client Concerns About AI Security**

Many institutions are hesitant to adopt AI due to data privacy risks. Here's how those concerns can be addressed:

| Concern | Proposed Solution |
|---|---|
| AI models exposing patient data | Use de-identified datasets and federated learning to ensure raw patient data never leaves the secure environment. |
| Lack of regulatory acceptance | Ensure AI outputs follow ICH, FDA, and EMA CSR writing guidelines and keep an audit trail for every generated report. |
| Bias in AI-generated text | Train models using diverse clinical datasets and require human oversight in all AI-generated narratives. |
| AI-generated reports not aligning with study-specific nuances | Fine-tune AI models on study-specific protocols and allow real-time human edits to AI-generated text. |

## CHALLENGES AND CONSIDERATIONS

While Gen-AI presents a promising solution, several challenges must be addressed for broader industry adoption.

| Challenge | Considerations |
|---|---|
| Regulatory Compliance | AI-generated narratives must align with ICH, FDA, and EMA guidelines. Human oversight remains essential. |
| Data Privacy | AI models require access to sensitive clinical data, necessitating strict data security measures. |
| Model Interpretability | AI-generated results must be transparent and explainable for regulatory review. |
| Adaptability | Models must handle study-specific nuances, such as variations in endpoints and statistical methods. |

## CONCLUSION AND FUTURE DIRECTIONS

AI-driven CSR writing offers a transformational opportunity to enhance the efficiency, accuracy, and consistency of clinical study reporting. By leveraging CDISC ARS metadata and structured ARD datasets, our model streamlines statistical interpretation and narrative generation while reducing manual workload.

In the future, further advancements could include:

- Integration with regulatory submission platforms to automate CSR preparation.
- AI-enhanced quality control tools to detect inconsistencies before submission.
- Adaptive learning models that improve based on user feedback and historical study reports.

Institutions can confidently leverage AI while ensuring:

- Patient data is never compromised.
- AI-generated outputs are fully traceable.
- Regulatory bodies accept AI-driven reporting.

As proven by this proof of concept, with the right safeguards in place, Gen-AI can revolutionize CSR writing while maintaining trust and compliance. While AI cannot fully replace human expertise, it serves as a powerful assistive tool for biostatisticians, medical writers, and regulatory teams.

As the next step in our CSR automation, we plan to explore the use of embedding models to enhance the AI's ability to understand and process complex clinical data. Embedding models will allow us to convert large volumes of textual and numerical data into dense vector representations, enabling better contextual comprehension and semantic analysis.

By integrating embedding models, our AI system will improve its capacity to identify relationships between study variables, detect subtle patterns in ARD/ARS metadata, and generate more accurate and cohesive narratives. This approach will also enhance the AI's retrieval and ranking capabilities, supporting more precise extraction of relevant data for CSR sections. Ultimately, leveraging embedding models will enable us to create a more intelligent, flexible, and scalable automation framework that maintains high accuracy and regulatory compliance while reducing manual effort.

## REFERENCES


1. CDISC ARS Model: https://cdisc-org.github.io/analysis-results-standard/
2. CDISC ARS User Guide version 1.0: https://wiki.cdisc.org/display/ARSP/Analysis+Results+Standard+User+Guide+v1.0

3. Busa, Bhavin; Marshall, Richard; LeRoy, Bess; 'All You Need to Know about the New CDISC Analysis Result Standards!' at PharmaSUG (May 2023). Paper Link
4. Prompt Engineering image reference: https://yellow.ai/blog/how-yellow-ai-works-around-the-pitfalls-of-prompt-engineering/
5. CDISC Pilot Study package: https://github.com/cdisc-org/sdtm-adam-pilot-project
6. eTFL Portal: https://cdisc.org/kb/etfl
7. Busa, Bhavin; LeRoy, Bess; 'The Development of the CDISC eTFL Portal to Support the Adoption and Implementation of the Analysis Result Standard' at PHUSE EU Connect (Nov 2024). Paper Link & Presentation Link

## ACKNOWLEDGMENTS

We thank the contributors involved in AI model development, validation, and data processing. Special recognition goes to the clinical data teams for structuring ARD datasets.

We would also like to acknowledge the use of OpenAI's ChatGPT (GPT-4o model) for supporting the development of this work, including text generation, content structuring, and narrative assistance. The model's capabilities in processing ARS metadata and ARD datasets were integral to our Proof of Concept (POC) implementation.