Paper ML31

LLMs in clinical trials: Overcoming challenges related to patient confidentiality and intellectual property

Amit Javkhedkar, Sycamore Informatics India Private Limited, Pune, India
Samip Shah, Sycamore Informatics India Private Limited, Pune, India

ABSTRACT

In the rapidly evolving landscape of clinical trials, Large Language Models (LLMs) will play a pivotal role in radically changing the way clinical trial data is collected, transformed, analyzed and submitted to regulatory bodies. These Large Language Models can be designed to understand, interpret, and generate human-like text, thereby simplifying and speeding up the process leading up to submissions. Although LLMs offer promising solutions and show exciting prospects, there are some inherent challenges. Key among those include setting up sophisticated technical infrastructure for LLMs and ensuring data privacy, particularly when training with Personal Health Information (PHI) and organization-specific standards. This paper discusses how to potentially apply LLMs to visionary use cases that will empower the industry and possibly revolutionize the various aspects of the drug discovery process.

INTRODUCTION

The Current Landscape of Clinical Trials

Clinical trials are a pivotal phase in the drug development process, providing essential evidence on the safety and efficacy of new treatments. These trials require significant investment in terms of time, resources, and highly specialized expertise to ensure that the various stages are conducted effectively. Despite their importance, the clinical trial process faces numerous challenges, including long timelines, high costs, and increasingly complex regulatory requirements. Traditional methods, often dependent on manual data collection, siloed systems, and labor-intensive analysis, can lead to inefficiencies and hinder overall progress.

In recent years, advancements in technology, particularly in Artificial Intelligence (AI), have begun to offer solutions to many of these challenges. Large Language Models (LLMs), a branch of AI that excels at processing and generating human-like text, have shown promise in streamlining several aspects of clinical trials. By automating tasks that rely heavily on language, such as extracting insights from unstructured data and predicting patient eligibility from health records, LLMs can drive significant improvements in clinical trial efficiency and accuracy.

Introduction to Large Language Models

Large Language Models (LLMs), such as GPT (Generative Pre-trained Transformer), are AI models trained on vast amounts of text data. These models excel at understanding and generating human-like language, making them versatile tools for a variety of applications, including natural language processing, content generation, and data extraction. In clinical trials, the integration of LLMs offers a transformative opportunity to address inefficiencies in data collection, patient engagement, and regulatory reporting.

LLMs can analyze vast volumes of unstructured text data, synthesize key insights, and generate detailed outputs such as regulatory reports, trial summaries, and eligibility criteria. They also offer the potential to automate tedious, time-consuming tasks, such as reviewing scientific literature, ensuring that clinical trial data adheres to compliance standards, and generating real-time trial insights.

However, integrating LLMs into clinical trials requires overcoming several hurdles. These models need tailored training data, secure technical infrastructure, and a focus on data privacy. Additionally, issues related to bias and regulatory compliance must be addressed to ensure fair, ethical, and legally compliant AI usage.

This paper explores the potential applications of LLMs in clinical trials, highlighting key challenges, providing solutions, and offering visionary use cases that illustrate the future impact of these technologies on statistical computing and clinical trial efficiency.

CHALLENGES IN ADOPTING LARGE LANGUAGE MODELS (LLMS) IN A CLINICAL TRIAL

Domain-Specific Training Data

Although LLMs have demonstrated impressive capabilities in general text processing, they are not inherently equipped to handle the specialized terminology and nuances present in clinical trials. Without access to domain-specific training data, such as clinical guidelines, medical literature, and trial protocols, LLMs may struggle to provide accurate and contextually relevant outputs. This is particularly important in clinical trials, where precision is critical to ensure the validity of results and compliance with regulatory standards.

To address this issue, it is essential to create customized training datasets that reflect the specific needs of the clinical trial environment. These datasets should include medical terminologies, regulatory frameworks, and clinical trial design documents, ensuring that LLMs can process and generate outputs tailored to the needs of the industry.

Continuation of body – after source code.

Data Privacy and Security Concerns

The integration of LLMs into clinical trials presents significant data privacy and security concerns. Clinical trial data, especially patient health information, is highly sensitive and governed by strict privacy regulations, such as the Health Insurance Portability and Accountability Act (HIPAA) in the U.S. and the General Data Protection Regulation (GDPR) in the European Union. The use of LLMs requires careful consideration of how data is collected, stored, and processed to protect patient confidentiality and comply with these regulations.

Organizations must ensure that any LLMs deployed in clinical trials are trained and operated within secure environments, with strong encryption protocols and data access controls in place. Furthermore, it is essential to implement robust data anonymization techniques to protect patient identities while still allowing for the analysis of large datasets.

Lack of Expertise

LLMs are powerful tools, but they still require specialized knowledge to be effective in clinical trials. While LLMs excel at processing language, they do not inherently understand the medical, scientific, and regulatory context that is crucial to clinical research. Clinical trial processes involve complex tasks such as patient enrollment, data analysis, regulatory reporting, and adherence to ethical standards, all of which require a high level of domain-specific expertise.

To bridge this gap, LLMs must be integrated with domain experts who can validate the outputs generated by the models. Additionally, training LLMs with clinical trial-specific datasets can help improve their contextual understanding and make them more effective in real-world applications.

Infrastructure Requirements for Deploying LLMs

Deploying LLMs in clinical trials requires robust infrastructure to support the computational demands of these models. LLMs are typically resource-intensive, requiring high-performance computing systems with powerful GPUs or TPUs for model training and inference. Storage systems must also be scalable to handle the large volumes of data generated in clinical trials, including patient records, regulatory documents, and scientific literature.

Data privacy and security remain top priorities, especially when handling sensitive health information. Organizations must invest in secure cloud platforms or on-premises solutions that comply with privacy regulations like HIPAA and GDPR. Furthermore, the integration of LLMs with existing clinical trial systems, such as Electronic Data Capture (EDC) platforms or Clinical Trial Management Systems (CTMS), may require custom APIs and interoperability solutions.

OVERCOMING CHALLENGES IN THE USAGE OF LLMS IN CLINICAL TRIALS

By recognizing the limitations and challenges associated with large language models (LLMs) in clinical trials, life sciences companies can take proactive steps to harness the full potential of AI while addressing potential risks. This approach ensures that LLMs are integrated responsibly and ethically into clinical research, advancing efficiency and innovation without compromising compliance, accuracy, or patient trust. Here are some key recommendations to overcome the challenges in using LLMs in clinical trials:

1. "Local" Large Language Models

Local Large Language Models (LLMs) are AI systems specifically designed to run on local hardware or private servers, eliminating the reliance on remote cloud-based infrastructure. These models provide similar capabilities to cloud-based LLMs, such as text generation, question answering, and natural language processing. However, the critical distinction is that all processing occurs directly on the user's device or within the organization's own infrastructure, offering several advantages in terms of privacy and security.

Local LLMs provide complete control over data, ensuring that sensitive or proprietary information never leaves the organization's network. This level of privacy is particularly important in clinical trials, where patient data, trial results, and other confidential information must be rigorously protected. By processing information directly on local systems, organizations can more easily comply with regulatory requirements, such as HIPAA and GDPR, while maintaining the advanced capabilities of AI tools. Furthermore, the use of local LLMs reduces dependency on external vendors, enabling organizations to maintain tighter control over their AI operations.

2. Building the Right Infrastructure

Deploying LLMs in clinical trials requires careful consideration of the underlying infrastructure. Setting up the necessary infrastructure to support LLMs involves several key components:

- **High-Performance Computing:** LLMs require substantial computational resources, including powerful Graphics Processing Units (GPUs) or Tensor Processing Units (TPUs), to handle the complex algorithms and vast datasets involved in processing medical and clinical trial data. High-performance computing systems, either on-premises or through secure cloud platforms, are essential to ensure the efficient operation of LLMs.
- **Data Storage and Scalability:** Clinical trial data is often vast and varied, requiring robust storage solutions that are scalable. Cloud-based solutions are popular for their flexibility, but organizations working with sensitive health information may prefer on-premises infrastructure for greater data control. The infrastructure must be designed with data security, redundancy, and regulatory compliance in mind to minimize the risk of data breaches and ensure the integrity of trial data.
- **Security and Compliance:** LLMs used in clinical trials must be deployed in secure environments that comply with relevant privacy regulations. This involves implementing strict access controls, encryption

mechanisms, and continuous monitoring to safeguard against unauthorized access and ensure that patient data remains protected at all times.

3. Training the AI Model

Training large language models for clinical trial applications is a critical process that requires a systematic and thorough approach to ensure accuracy, relevance, and ethical compliance. Some of the key aspects involved in the training process include:

- **Curated and Domain-Specific Datasets**: LLMs must be trained on high-quality, domain-specific datasets that reflect the specialized needs of clinical trials. This includes medical literature, clinical trial protocols, patient health records (with appropriate de-identification), regulatory documents, and other relevant resources. The datasets should encompass a wide range of medical terminology, clinical procedures, regulatory guidelines, and trial-specific contexts to ensure the model can understand and generate accurate outputs.
- **Fine-Tuning for Specific Tasks**: While pre-trained LLMs are powerful, they may require fine-tuning to improve performance on specific clinical trial tasks, such as generating regulatory reports, analyzing patient data, or interpreting clinical trial results. This fine-tuning ensures that the model can understand the nuances of clinical trial processes, patient care, and medical terminology, ultimately improving its relevance and precision.
- **Ethical Considerations and Privacy Compliance**: Ethical considerations are a cornerstone of training LLMs for clinical trials. The AI models must be trained with a strong emphasis on privacy, adhering to regulations such as HIPAA and GDPR. Patient data used in training must be anonymized to prevent the model from inadvertently exposing sensitive information. Additionally, the AI training process must include human-in-the-loop evaluations, with clinicians and researchers validating outputs and identifying any biases, inaccuracies, or ethical concerns.
- **Ongoing Monitoring and Retraining**: As medical knowledge and regulatory standards evolve, LLMs must be continually updated to remain effective and compliant. Ongoing monitoring of model performance and periodic retraining with new datasets are essential for ensuring that the model stays current and reliable. This process also helps address emerging biases and ensures that the model's outputs continue to align with the latest medical standards and ethical guidelines.

4. Addressing Biases and Ensuring Fairness

One of the significant challenges in AI adoption, particularly in healthcare, is the potential for model biases. LLMs, like other AI systems, can inadvertently perpetuate biases present in their training data. In clinical trials, biased outputs can have serious implications, leading to inequitable treatment of patients, incorrect data analysis, and flawed conclusions. To mitigate these risks, it is important to:

- **Conduct Bias Audits**: Regular audits should be conducted to identify and mitigate biases in the model. This involves reviewing both the training data and the model's outputs to ensure that they are free from discriminatory patterns based on race, gender, socioeconomic status, or other protected characteristics.
- **Diverse and Inclusive Datasets**: To reduce the risk of biases, datasets used for training must be diverse and representative of the entire patient population. This includes accounting for a variety of demographic factors, medical conditions, and treatment protocols.
- **Human Oversight**: LLMs should not be used in isolation without human oversight. Clinical experts should continually evaluate the model's outputs to ensure that they are accurate, fair, and aligned with clinical trial objectives. This oversight helps prevent the model from making inappropriate recommendations or generating harmful conclusions.

5. Collaboration with Domain Experts

For LLMs to be truly effective in clinical trials, they must be integrated into workflows alongside domain experts. Collaboration with medical professionals, statisticians, clinical researchers, and regulatory specialists ensures that the model's outputs are clinically relevant and scientifically valid. This human-in-the-loop approach helps address the model's limitations and ensures that it adds value to the clinical trial process.

Domain experts can help guide the training process, identify the most important data sources, and validate the model's outputs. This collaboration fosters trust in the system and allows for continuous improvement, ensuring that LLMs contribute to the efficiency, accuracy, and success of clinical trials.

6. Establishing Clear Governance and Ethical Frameworks

To guide the responsible use of LLMs in clinical trials, organizations must establish clear governance frameworks. This includes setting policies around data use, model transparency, ethical considerations, and regulatory compliance. Governance frameworks should address the following:

- **Transparency**: Ensuring that the LLM's decision-making processes are transparent and understandable to end-users.
- **Accountability**: Clearly defining who is responsible for the model's outputs and ensuring that organizations have mechanisms in place to address any issues that arise.
- **Ethical Guidelines**: Adopting industry-standard ethical guidelines that prioritize patient safety, data privacy, and fair treatment across all stages of clinical trial design, execution, and analysis.

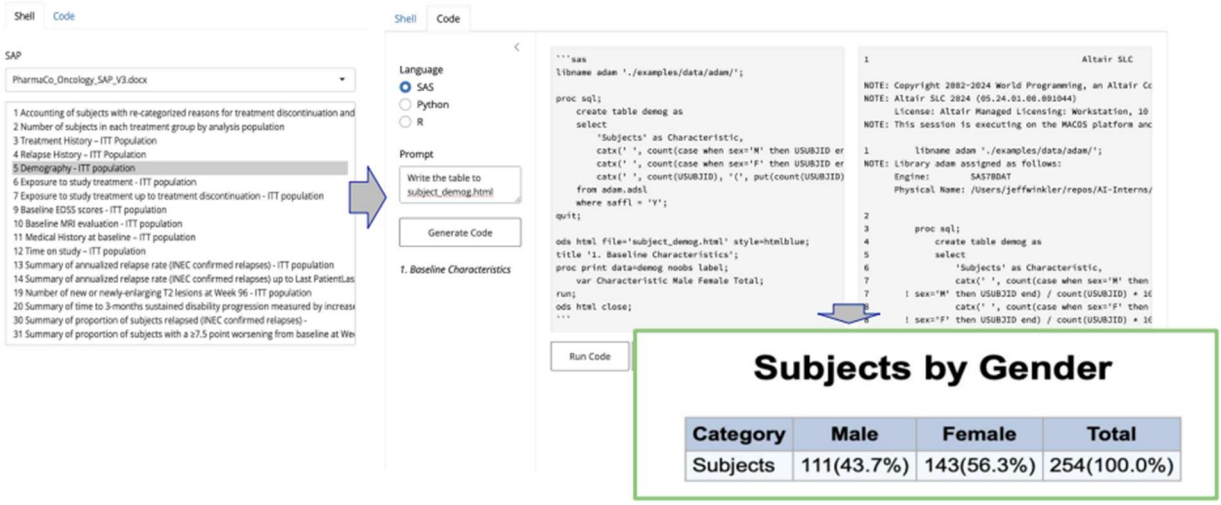
VISIONARY USE CASES

The integration of next-generation Statistical Computing environments with Large Language Models (LLMs) has the potential to revolutionize the way statistical programmers work. By harnessing the power of AI, statistical programming tasks can be automated, streamlined, and optimized for faster results and improved accuracy. Below are some visionary use cases demonstrating how LLMs can be leveraged to enhance the statistical programming workflow, particularly in the context of generating tables, listings, and figures (TLFs), among other tasks.

1. Generating Code for TLFs with LLMs

A groundbreaking use case involves using LLMs to automatically generate SAS code for TLFs (Tables, Listings, and Figures) from Statistical Analysis Plans (SAPs) and mock shells.

- **Process Overview:**
  With the LLM in place, statistical programmers can input an SAP, along with a series of mock shells outlining the desired structure of tables, and the system will automatically generate the corresponding SAS program. Initially, this can be done using dummy data or sample data, which is later replaced with real clinical trial data as it becomes available.
  The LLM analyzes the structure, content, and formatting of the mock shell, then generates the necessary SAS code to populate the table with relevant statistical outputs. As data begins to flow in, the model continually refines its code generation, ensuring that the SAS program stays up to date with the latest information and regulatory guidelines.
- **Business Value:**
  - **Accelerated Program Generation:** The ability to automatically generate SAS code from structured inputs dramatically reduces the time spent writing boilerplate code for TLFs. Statistical programmers can focus more on the analysis and interpretation of the data rather than on repetitive programming tasks.
  - **Reusability of Macros and Coding Conventions:** The model can leverage existing coding conventions, macros, and standards used by the organization. This reduces errors, maintains consistency across projects, and ensures that the generated code adheres to best practices and regulatory requirements.
  - **Enhanced Focus on Analysis:** By automating the code generation process for TLFs, statistical programmers can redirect their energy and expertise toward the core tasks of data analysis, hypothesis testing, and interpretation of results, thus driving more value from their time and skills.
  - **Scalability and Efficiency:** With the LLM handling routine tasks, the team can scale more easily, especially when managing multiple trials or complex datasets. The ability to quickly generate accurate code for new TLFs allows for faster trial progress without compromising data integrity.



2. Automating Quality Control (QC) of Statistical Programs

LLMs can play a pivotal role in improving the QC processes in statistical programming. The system can automatically review SAS programs to identify errors, inefficiencies, and potential discrepancies in the code.

- **Process Overview:**
  Once a SAS program is written, the LLM can be tasked with running a quality check on the code. The model will assess the program against predefined standards, including compliance with regulatory guidelines (e.g., CDISC), checking for missing variables, incorrect coding patterns, and alignment with the SAP and mock shells. It can even provide feedback and suggestions for optimization.
- **Business Value:**

    - **Enhanced Accuracy and Reliability:** With automated QC, human errors can be significantly reduced, ensuring that the statistical programs meet the highest standards before being submitted for regulatory review.
    - **Faster Turnaround Time:** By automating the quality control process, statistical programming teams can expedite the preparation of submission-ready programs, speeding up clinical trial timelines.
    - **Continuous Improvement:** As the LLM analyzes more programs, it can learn and adapt to improve its error detection and offer more nuanced insights into complex programming issues.

3. Real-time Data Analysis and Reporting

LLMs can assist statistical programmers in creating real-time reports as clinical trial data becomes available, allowing for faster decision-making.

- **Process Overview:**
  Once data is available, the system can use LLMs to generate initial reports, summaries, and statistical analyses without the need for manual programming. The LLM can pull data directly from clinical databases or analysis datasets, process it in real-time, and generate the necessary TLFs, figures, and summaries according to the SAP. The reports are automatically populated with real data, and the system can continuously update them as new data flows in.
- **Business Value:**
    - **Faster Decision-Making:** With automated real-time reporting, trial stakeholders, including clinical researchers and regulatory bodies, can access up-to-date analysis and insights at any point during the trial, leading to quicker decision-making and more agile trial management.
    - **Reduced Time to Market:** The ability to rapidly generate and revise reports based on live data accelerates the drug development process, helping to bring products to market faster.
    - **Adaptive Reporting:** The LLM can dynamically adapt to changes in trial parameters or data structure, ensuring that reports are always aligned with the evolving needs of the trial.

4. Standardizing and Personalizing Reporting Outputs

While standardization of statistical reporting is essential for regulatory compliance, personalization can make the data more accessible and understandable for various stakeholders. LLMs can support both standardization and customization.

- **Process Overview:**
  The LLM can be trained to understand the specific preferences and requirements of different audiences—whether they be internal stakeholders, regulatory agencies, or external collaborators. It can then automatically customize the output format while ensuring it adheres to the necessary regulatory standards. For instance, the LLM can generate both high-level executive summaries for stakeholders and more detailed, data-heavy reports for clinical researchers.
- **Business Value:**
    - **Enhanced Stakeholder Communication:** Personalized reports improve the communication of complex statistical results to non-technical stakeholders, ensuring they can make informed decisions based on clear, accessible insights.
    - **Consistency and Compliance:** Despite customization, the LLM ensures that the reports remain consistent with industry standards and regulatory requirements, reducing the risk of compliance issues.
    - **Flexibility and Scalability:** The model can generate multiple versions of reports for different audiences without additional effort from the statistical programming team, saving time and resources while maintaining flexibility.

5. Knowledge Sharing and Collaboration Across Teams

LLMs can facilitate knowledge sharing and collaboration among statistical programming teams by acting as a knowledge repository and a tool for real-time assistance.

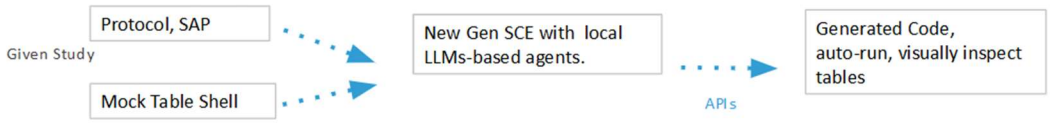
- **Process Overview:**
  LLMs can be integrated with collaborative platforms or knowledge management systems, where they store a vast repository of programming knowledge, best practices, common solutions, and regulatory guidelines. When a programmer encounters a challenge or needs assistance, the LLM can provide recommendations, example code snippets, or point them to relevant documentation.
- **Business Value:**
    - **Faster Onboarding:** New team members can quickly get up to speed by accessing AI-generated suggestions and resources, improving onboarding efficiency.
    - **Cross-Team Collaboration:** Statistical programmers across different teams or geographies can collaborate more effectively, sharing insights, code, and strategies without barriers.
    - **Continuous Learning:** By learning from previous projects, the LLM continually refines its knowledge base, ensuring that the team has access to the most up-to-date information and techniques.

6. Predictive Analytics and Trend Identification
LLMs can analyze past clinical trial data to identify trends, outliers, and potential issues, offering predictive insights that can guide future programming tasks.

- **Process Overview:**
  By processing historical clinical trial data, LLMs can identify patterns in how different datasets impact the generation of TLFs. The system can predict potential roadblocks, such as unexpected data distributions or regulatory changes, allowing programmers to proactively address issues before they arise.
- **Business Value:**
  - **Proactive Issue Resolution:** Predictive capabilities allow statistical programmers to mitigate risks early, reducing the likelihood of costly delays or errors.
  - **Better Trial Design:** By recognizing trends in data from previous trials, LLMs can offer insights that help optimize the design of future trials, ensuring more reliable and accurate results.
  - **Efficient Resource Allocation:** Predictive analytics help allocate resources more efficiently, ensuring that statistical programming teams are focused on the most critical tasks at any given time.

7. Generating SAP (Statistical Analysis Plan) with LLMs
Creating a robust SAP is one of the most critical tasks in the planning phase of clinical trials. By using LLMs, this process can be automated, ensuring consistency, accuracy, and adherence to regulatory standards.

- **Process Overview:**
  An LLM can generate the initial draft of an SAP by analyzing historical SAPs, trial design documents, and regulatory guidelines. Given specific trial parameters, such as study objectives, endpoints, and sample size, the system can create a comprehensive SAP that outlines the statistical methods to be used, the types of analyses, data handling procedures, and more. The generated SAP can include tables, figures, and listings templates, clearly defining the structure and content of statistical outputs expected during the trial.
  The LLM can prompt users to confirm key details, such as treatment groups, covariates, or randomization methods, further refining the plan. The system can continuously update the SAP as trial conditions change, incorporating new data or regulatory updates to ensure the plan remains aligned with the latest requirements.
- **Business Value:**
  - **Time Savings:** Automating the generation of the SAP drastically reduces the time and effort involved in drafting a plan. The initial framework can be generated in minutes, allowing statisticians to focus on fine-tuning the plan rather than starting from scratch.
  - **Consistency and Compliance:** The LLM can ensure that the SAP adheres to industry standards, regulatory guidelines (such as CDISC), and best practices. It can also ensure that templates are used consistently across trials, reducing human error and variability.
  - **Adaptability:** As trial parameters evolve or as new regulatory guidelines emerge, the LLM can help adjust the SAP in real-time, ensuring that the plan remains current and compliant with the latest changes.
  - **Collaboration and Sharing:** The automated generation of SAPs enables teams to collaborate more effectively, with a standardized starting point that can be easily reviewed and revised across departments, making sharing and communication with cross-functional teams simpler.



CONCLUSION

The integration of next-generation statistical computing environments with LLMs opens up a world of possibilities for statistical programmers in clinical trials. By automating repetitive tasks, ensuring high standards of code quality, providing real-time data analysis, and enhancing collaboration, LLMs can significantly improve efficiency, accuracy, and speed in the drug development process. This vision of AI-powered statistical programming promises to not only streamline operations but also drive innovation and excellence in clinical trials, ultimately accelerating the path to new treatments and therapies.

CONTACT INFORMATION (HEADER 1)

Your comments and questions are valued and encouraged. Contact the author at:
Author Name: Amit Javkhedkar
Company: Sycamore Informatics India Private Limited
Address: No. 54, 6th Sector, Behind BDA Complex, HSR Layout, Bangalore South, Bangalore-560102, Karnataka.
Work Phone: +91 86000 45798
Email:amit.javkhedkar@sycamoreinformatics.com
Website: https://www.sycamoreinformatics.com/