PharmaSUG China 2023 - Paper AA-188

AI+HI: Accelerating Biometrics Activities with the Power of Large Language Models (LLMs)

Hao (Harry) Chen, AstraZeneca China Biometrics, Shanghai, China

ABSTRACT

Within the Biometrics group, our current process heavily relies on human review and comprehension for an extensive range of artifacts, including documentation, algorithm specifications, programming codes, and output tables, etc. However, the rapid advancement of AI technology, particularly Large Language Models (LLMs), empowers the machine to understand and transform these human-readable elements into machine-readable structured data, thereby promoting more automated or semi-automated processes. By seamlessly integrating artificial intelligence and human intelligence (AI+HI), we are not only redefining our work methodology but also enhancing overall efficiency. Importantly, the human element (HI) serves as a vital supervisory layer, effectively mitigating potential risks associated with AI and ensuring the robustness and reliability of the whole processes.

INTRODUCTION

Traditional clinical trials' data flow is a complex and non-linear procedure, demanding considerable manual effort and often resulting in rework and inefficiency. The creation of artifacts involves manually transcribing data from documents and systems, leading to inconsistencies and inaccuracies. Despite the frequent reuse of data components across trials, redundancies are common. To address these inefficiencies, it is crucial to use AI to power digital data flow to enhance the efficiency and effectiveness of clinical trial processes, resulting in improved patient outcomes.

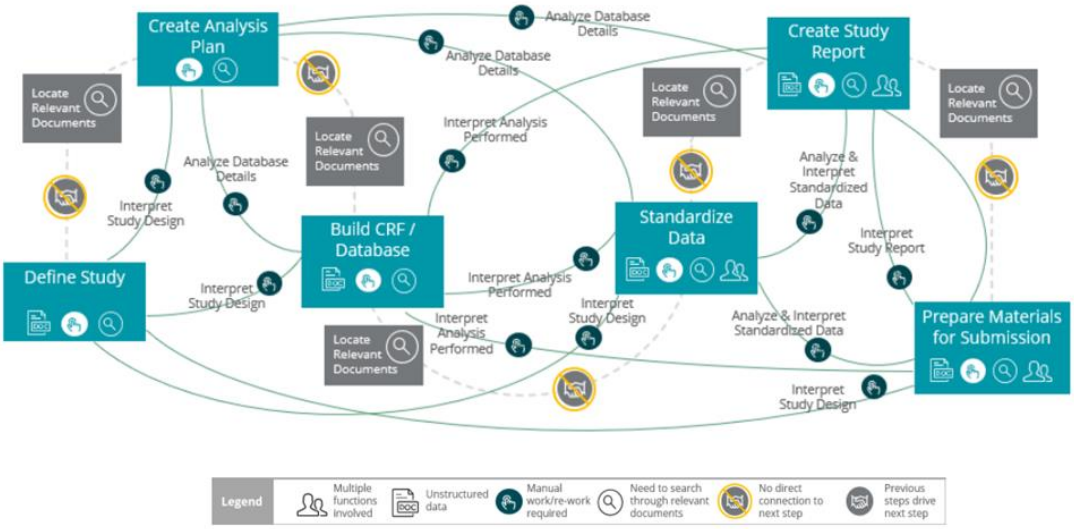

Figure 1. Clinical Data Analysis Flow: Current State（Deloitte,2022）

In clinical research, digitalization of trial artifacts, including documentation, programming codes, data, output tables, and complex processes, is critical for achieving FAIRness (Findable, Accessible, Interoperable, and Reusable). By converting these artifacts into metadata and systematically managing them, clinical data analysis can be conducted more efficiently, leading to better decision-making.

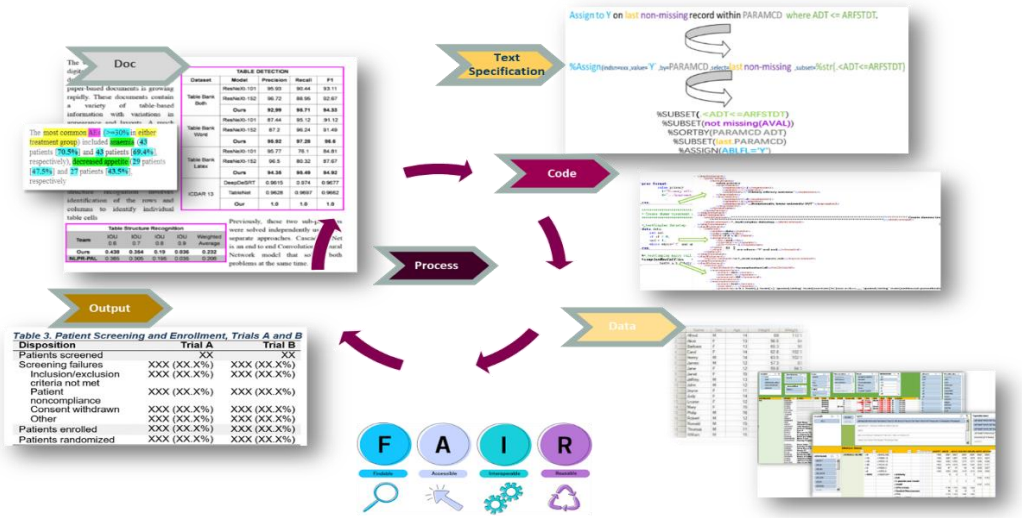

Figure 2. Digitalization of trial artifacts

AI & LLM (LARGE LANGUAGE MODELS)

HOW AI CAN HELP

AI plays a pivotal role in converting unstructured data into structured data, thereby transitioning most analysis processes from being human-driven to machine-driven – a process also known as digitization.

The evolution of AI has journeyed through several transformative stages. It started with Intelligent Systems, which use databases and control systems to implement pre-set rules to problems, forming the backbone of expert systems but lacking autonomous learning capabilities. The next progression was to Machine Learning, which applies logical reasoning to extend rules and increase databases. It leverages statistical relationships, as seen in Hidden-Markov-models used in Siri, and uses complex probability calculations like Bayesian networks for making predictions. Further advancement led to Deep Learning, which employs inductive learning, applying rules of generalization and specialization. It utilizes reinforcement learning for strategy optimization via trial and error. Deep learning involves neural networks, like Google's DeepMind, that dynamically change connection weights while learning. The most recent development is Large Language Models (LLMs), which are deep learning applications for natural language processing. It can learn the grammar and semantics of natural language, enabling it to generate human-readable text. LLMs are typically trained using large-scale corpora, such as the vast amount of text data available on the Internet. These models usually possess billions to trillions of parameters, allowing them to handle various natural language processing tasks, such as natural language generation, text classification, text summarization, machine translation, and speech recognition.

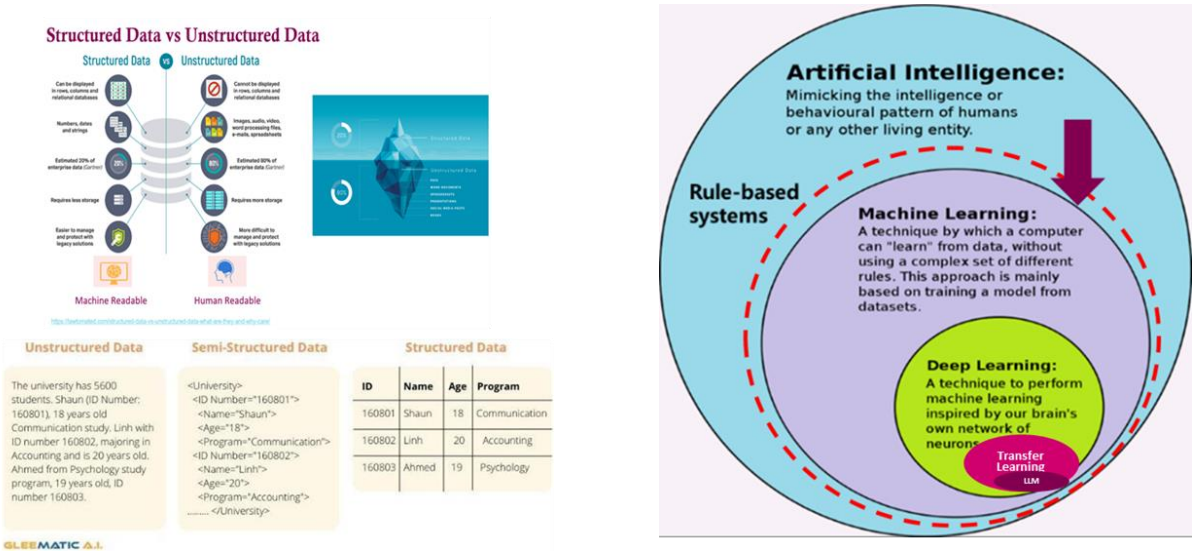

Figure 3. AI to bridge Structured and Unstructured data

CHATGPT AND LARGE LANGUAGE MODELS (LLMS)

Since OpenAI introduced GPT-3 in 2020, Large Language Models (LLMs) have consistently gained traction in the field of generative artificial intelligence. The launch of ChatGPT in November 2022, further boosted interest in LLMs. ChatGPT is an embodiment of generative AI, creating new content from existing data. It learns through Reinforcement Learning from Human Feedback (RLHF) and supervised fine-tuning, with human AI trainers assisting in crafting responses. Much like autocomplete systems, ChatGPT uses language models to generate narratives by predicting subsequent words based on word patterns. These models are typically voluminous, leveraging vast text data for training, and display remarkable learning capacity due to their large parameter counts.

The domain of Large Language Models (LLMs) is expansive and not limited to OpenAI. There are significant contributions like LLaMA, the first high-performance open-source LLM, and Google's Bard chatbot. The field continues to evolve with constant innovations, including OpenAI's GPT-4 and Claude from Anthropic. These advancements in AI technology are rapidly pushing the boundaries and reshaping the future of artificial intelligence.

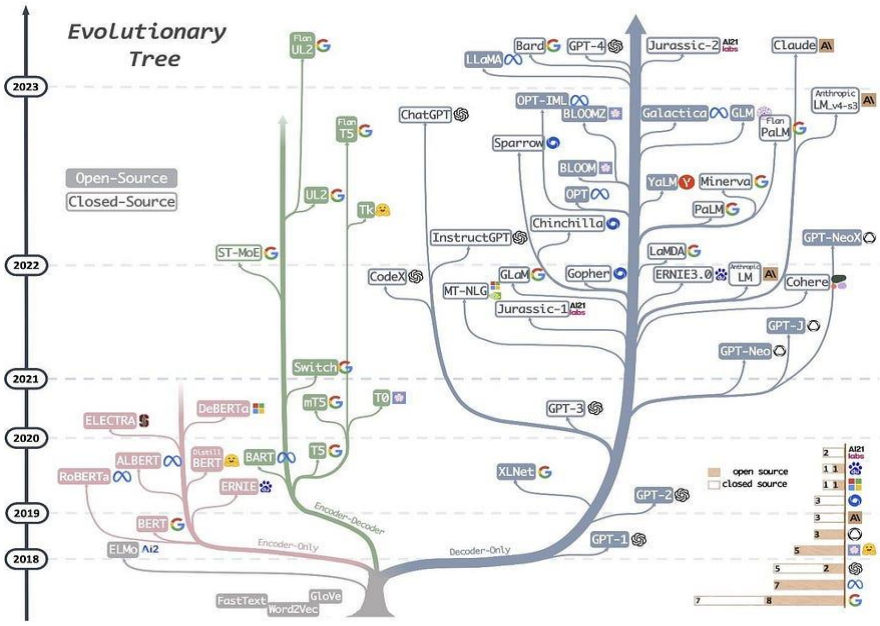

Figure 4. The GPT history and evolutionary tree of modern LLMs

Large Language Models (LLMs) have applications extending far beyond just language processing, including automating routine aspects of creative exercises. Key applications include:

- Question-Answering and Chatbots: Responding to user queries in a natural, human-like manner.
- Text Extraction and Summarization: Drawing key points from large bodies of text.
- Auto-Correct: Fixing typographical and grammatical errors in real-time.
- Translation: Translating between languages, maintaining semantic consistency.
- Classification: Categorizing text into predefined groups based on its content.
- Natural Language Generation: Creating human-like text, such as articles or reports.
- Embedding: Representing words as vectors to capture semantic relationships, which is essential in tasks like sentiment analysis and document search.
- Code Completion: Predict and suggest the next segment of code based on the current context, which leads to more efficient coding by assisting in generating, editing, and explaining code.

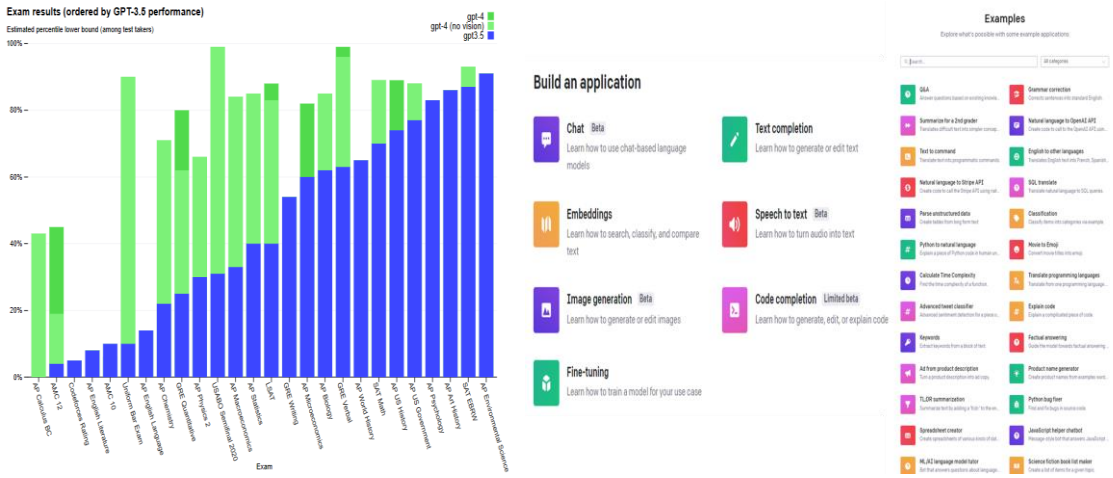

Figure 5. Capability and Application of the GPT model

TESTING CASES

As a case sharing exercise, we have tested the LLM solutions on three tasks: 1) Algorithm translation 2) Code Generation 3) Data Visualization. Through this case sharing exercise, we aim to demonstrate the potential of LLM to change traditional biometric process, and enabling more effective and efficient decision-making in the field.

Algorithm translation

We typically need to define algorithms for deriving variables, and translate these into Chinese for submission of eCRT packages to the China Center for Drug Evaluation (CDE). However, most translation engines struggle with the complexity of algorithms that combine free text and pseudo SAS code. Leveraging LLMs' capabilities in both language and code, we can greatly improve translation quality.

For example, the following algorithm contains many intricate logical details, but GPT models can handle these nuances well.

*If ADY ≥ 0 then AWTDIFF = ADY – AWTARGET; else if .<ADY < 0 then AWTDIFF = ADY - (AWTARGET -1); else AWTDIFF = .; select analysis visit flag ANL01FL= **"Analysis Record Flag 01"**; sort BY USUBJID PARAMCD AVISITN abs(AWTDIFF); if **first.**AVISITN then ANL01FL = "Y";*

Figure 6. Comparison of algorithm translation result by different Machine Translation engines

Code Generation

SAS is the primary programming language for clinical data analysis. In our quest to generate SAS code based on text algorithms, we conducted a series of tests. It was observed that R code outperforms SAS code. This superiority can be attributed to R's access to a broader range of open-source training data. An intriguing discovery was that translating code into R before rendering it into SAS could enhance the output quality.

However, achieving accurate results can necessitate multiple rounds of refinement and adjustments. It can also be challenging to control the quality of the outputs due to the black-box nature of LLM. Despite these challenges, LLM stands out for its ability to effectively process highly complex text, surpassing the capabilities of many proprietary systems.

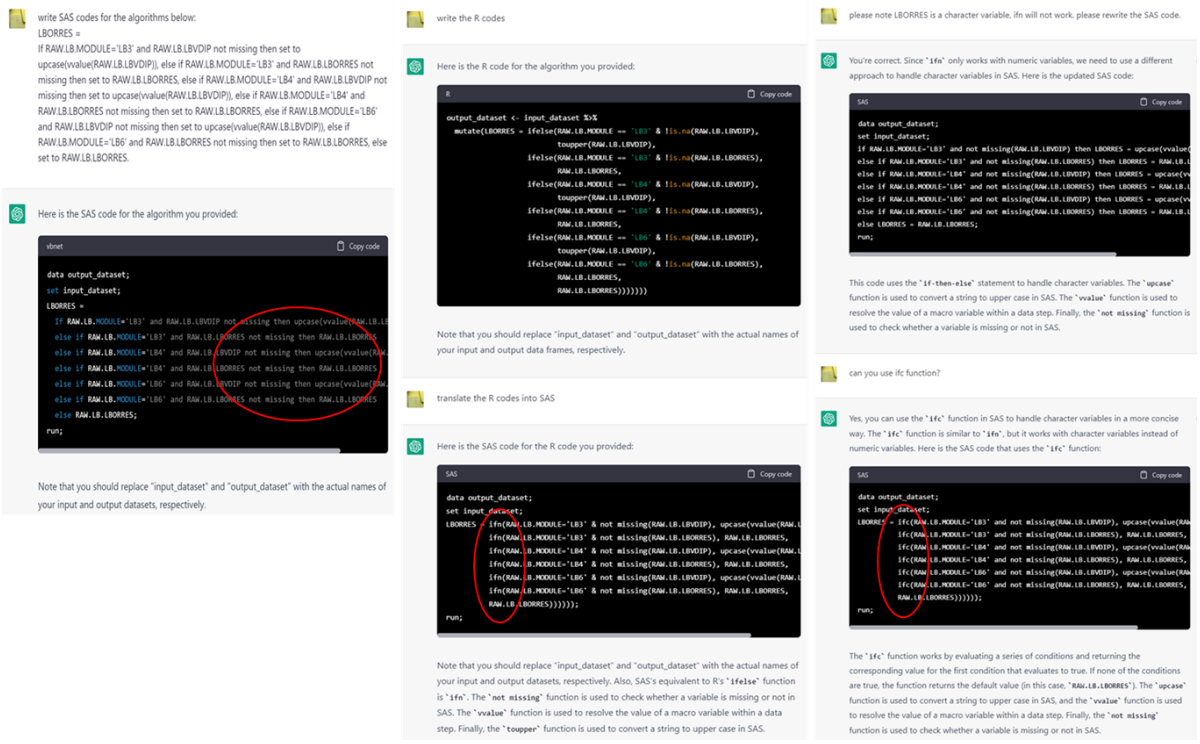

Figure 7. SAS code generation from text Spec

Data Visualization

Real-time data visualization is a powerful tool that significantly facilitates the transformation of data into insights. The R programming language stands out in this regard due to its strength in both analysis and plotting. To test these capabilities, we conducted an experiment with RShiny to create a plot.

In our experiment, we uploaded two test data sets—AE and DM—and provided the LLM with basic metadata, including data name, variable name, and type. We then requested bar charts with numerous specifications. Remarkably, the LLM was able to comprehend both the metadata and the human language request, and it swiftly generated the necessary R code within seconds. Once executed, this code produced the desired graph immediately.

Typically, a programmer might need more than 10 minutes to understand all the details, write the code, and debug it. However, this experiment demonstrates the efficiency and effectiveness of using the LLM in conjunction with R for data visualization.

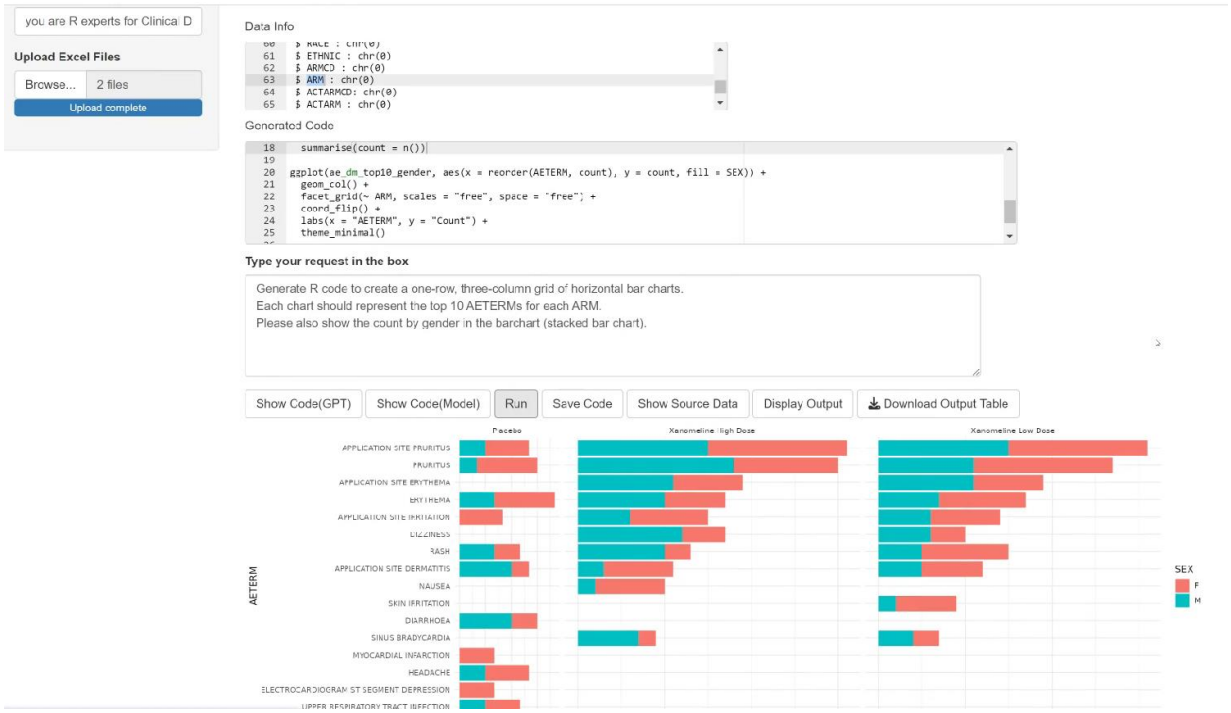

Figure 8. Real-time data visualization powered by LLM

Table Generation

In our experiment, we explored the potential of R in transforming raw data into a structured table. We provided the Language Model (LM) with two datasets—DM and AE—and a mockup table template. The task required the LM to interpret the template, comprehend the structure of the desired table, and extract the implicit analytical requirements.

The LM's performance was outstanding. It skillfully decoded the task and rapidly generated the corresponding R code. This step involved the creation of a comprehensive script that adheres to the structure and analytical needs specified in the mockup table, all without the need for hard-coded values. This ensures the adaptability of the code to similar datasets, underlining the flexibility of the LM's understanding and translation capabilities.

Upon execution, the R code produced the desired table that accurately mirrored the format of the provided mockup. This demonstrated the remarkable effectiveness of using the LM alongside R for efficient and precise data tabulation. Typically, a data analyst might spend a considerable amount of time completing such a task. The LM, however, accelerates this process significantly, underscoring its transformative power in the realm of data analysis.

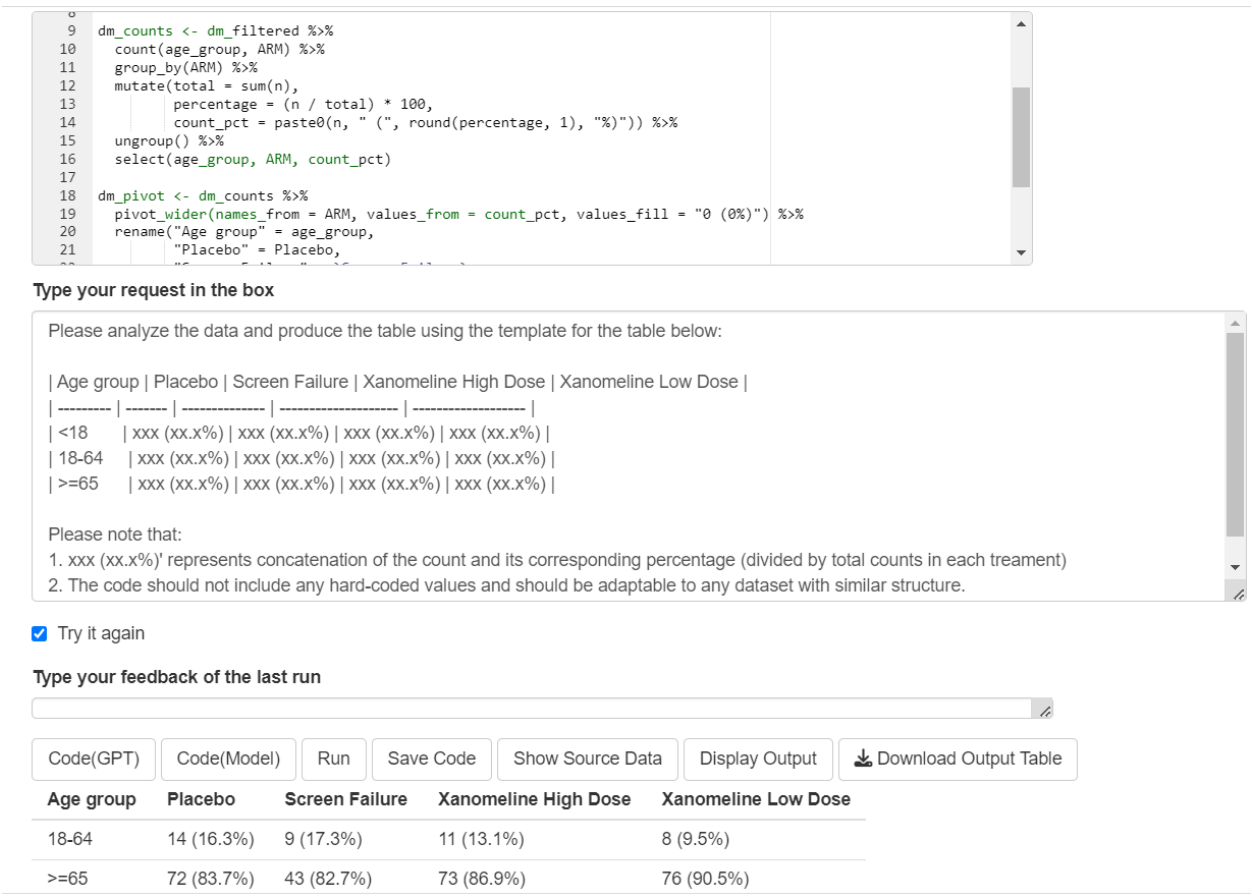

Figure 9. Real-time table generation powered by LLM

Other Typical Tasks

Large Language Models (LLMs) can be effectively employed for tasks such as SDTM, ADaM, TLG, and CSR generation - areas where traditional NLP solutions have been previously evaluated. The capabilities of LLMs far outstrip those of traditional NLP, enabling a significant acceleration in automation these tasks. Leveraging LLMs, the system can now generate output directly from the input, bypassing the need for Named Entity Recognition (NER), Intent Recognition, and the less reliable similarity calculations found in traditional NLP processes. Please note that ADaM still poses a challenge due to its inherent complexity and dynamism.

- SDTM: This task is simple, as it involves short sentence and simple logic with limited number of patterns. It is easy for LLM to both recommend the mapping and convert the mapping spec into executable codes.
- ADaM: This task involves long, complex sentences and can be challenging for non-standard specifications. Standard macros/functions and knowledge databases are required for support. Detailed logic is still challenge to machine and human intelligence
- TLG: Short and relatively standard sentences are involved in this task, requiring support from standard macros/functions as well as table layout analysis from mockup tables.
- CSR: The LLM solution is effective in key information extraction, similarity calculation, and text summarization, and can predict and identify key information in CSRs and link it to different documents and sections.

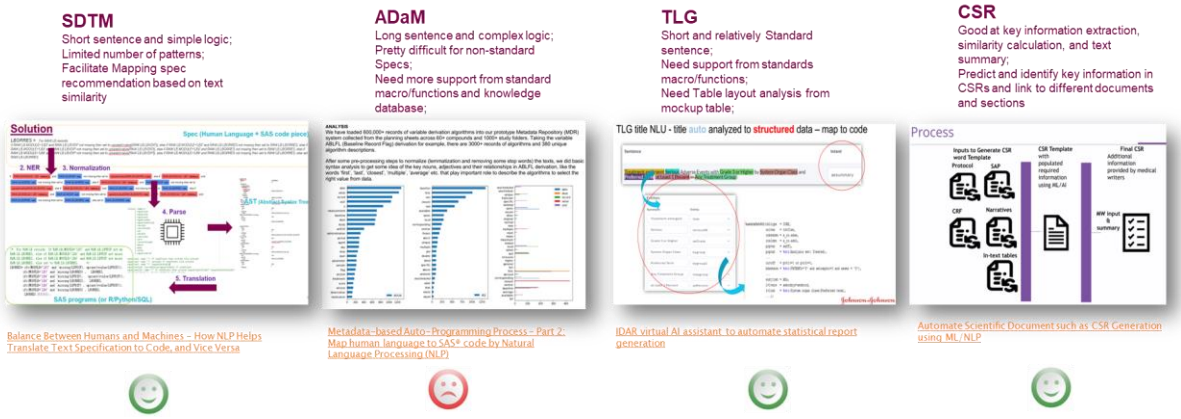

Figure 10. Traditional NLP solutions across typical tasks

CONCLUSION

This study tackles the challenges of traditional biometrics and suggests solutions using artificial intelligence (AI) and Large Language Models (LLMs). LLMs are excellent at analyzing documents and codes, which are central to biometrics. Therefore, it's beneficial to use LLMs to automate routine tasks, freeing up people to focus on more complex work.

In our case study, we showed that LLMs can understand documents, metadata, and codes, tasks that were once only possible by humans. However, LLMs have limitations. It's important for people to check the work done by these models, provide feedback, and continually train them for more complex tasks.

As LLMs continue to improve, we expect AI's ability to analyze data, documents, and codes to grow, changing how we approach biometrics. Combining AI with human intelligence (AI+HI) can create new ways of working and improve efficiency. Importantly, human involvement helps manage the risks of AI and ensure our path from data to insight is robust. The mix of AI and human intelligence (AI+HI) is revolutionizing how we work, learn, and solve problems in biometrics. It's essential for our industry to keep up with advances in AI and LLMs. These tools offer many benefits, but also come with risks and limitations. Proper testing is needed to ensure these systems are reliable and safe. By addressing these risks, we can use AI and LLMs effectively and responsibly in our industry.

This paper has been composed with assistance from GPT-4 and Claude 2.

REFERENCES

1. AI/ML – Challenges and Opportunities in Implementing Analytics in the Pharma Industry. https://phuse.s3.eu-central-1.amazonaws.com/Archive/2021/SDE/APAC/Virtual%20-%20India%20Winter/PRE_India06.pdf

2. Automate Scientific Document such as CSR Generation using ML/NLP. https://www.lexjansen.com/phuse-us/2021/ml/PRE_ML01.pdf

3. ChatGPT: Optimizing Language Models for Dialogue. https://openai.com/blog/chatgpt

4. Metadata-based Auto-Programming Process. https://www.lexjansen.com/phuse-us/2018/si/SI10.pdf

5. Metadata-based Auto-Programming Process – Part 2: Map human language to SAS® code by Natural Language Processing (NLP). https://www.lexjansen.com/phuse-us/2019/ml/ML05.pdf

6. IDAR virtual AI assistant to automate statistical report generation. https://www.lexjansen.com/phuse-us/2020/ml/ML01_ppt.pdf

ACKNOWLEDGMENTS

We would like to express our gratitude to Meng Chen for her generous support in sponsoring this project. Special thanks to Ning Zheng, Yiwen Wang, Zhanjian Dong, Xiaolin Zhu and Randy Dai for their significant support to designing, testing and providing feedback on the business cases.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Hao (Harry) Chen
AstraZeneca R&D China Biometrics
88 Xizang North Road, Jing'an District
Shanghai, China, 201210
Email: hao.chen15@astrazeneca.com