Paper SI07 Standards Implementation

Navigating the Clinical Research Landscape:

The Role of Regulatory Intelligence and LLMs

Author:

Federica Citterio, Global Technology Practice, SAS INSTITUTE

Soundarya Palanisamy, Global Health & Life Sciences Advisory, SAS INSTITUTE

Strasbourg - PHUSE EU Connect 2024

OVERVIEW

In the rapidly evolving landscape of clinical research, the need for efficient and accurate interpretation of regulatory documents is paramount. We will explore a novel approach to this challenge, termed ‘Regulatory Intelligence’, to extract and standardize information from such documents.

The Clinical Data Interchange Standards Consortium (CDISC) provides a comprehensive set of guidelines and standards for the collection, exchange, reporting, and archiving of clinical research data. While these standards are crucial for ensuring data quality and interoperability, their breadth and depth can be daunting, particularly for organizations with limited resources or experience.

We will discuss how LLMs can be used to assist in interpreting and applying CDISC standards. By extracting and standardizing information from regulatory documents, LLMs can guide organizations through the compliance process more efficiently and effectively. This approach not only reduces the burden of compliance but also enhances the quality and consistency of data reporting.

The CDISC Case: Navigating Complexity and Reaping Benefits

Importance and Challenges

Complexity of the Standard: The Clinical Data Interchange Standards Consortium (CDISC) provides comprehensive guidelines for the collection, exchange, reporting, and archiving of clinical research data. However, the extensive and intricate nature of these standards can be daunting, particularly for organizations with limited resources or experience.

Challenges in Revisions: The CDISC standards are continually evolving, making it challenging for companies to stay compliant without ongoing external support. Ensuring data compliance can be a time-consuming and error-prone process.

Benefits

Speed of Standard Adoption: Large Language Models (LLMs) can help interpret and apply CDISC standards, streamlining the compliance process for organizations. Their ability to process vast amounts of text accelerates the review of documentation and identification of non-compliance areas.

Reduction of Errors: Automation and the precision of LLMs can significantly reduce human errors in the mapping and application of standards. Generative AI can offer corrections and improvements based on extensive knowledge, enhancing data quality.

Democratization of Access: LLM-based tools make CDISC compliance more accessible to a wider range of organizations, lowering barriers for entities with fewer resources to meet required standards.

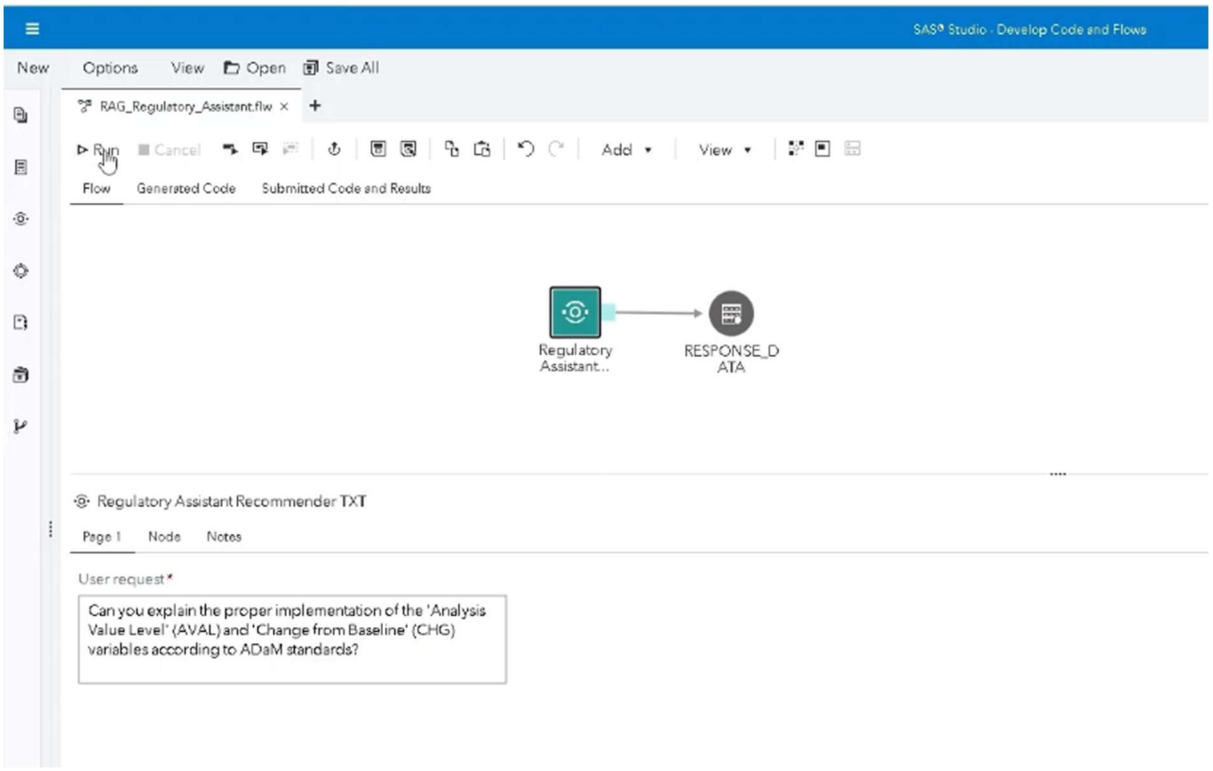
Decision Support: By providing data-driven analysis and recommendations, LLMs assist researchers and professionals in making informed decisions about clinical data management, thereby improving the effectiveness of studies.

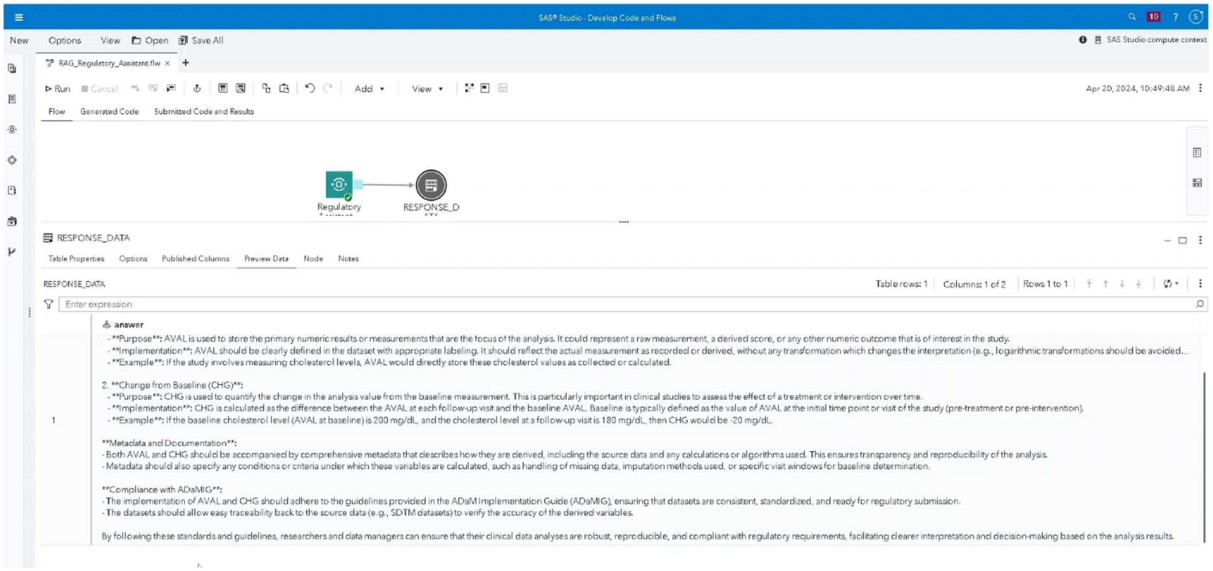
Accelerating CDISC Compliance: Empowering Compliance with Regulatory Assistant

Advanced Technology Integration: The Regulatory Assistant provides direct access to comprehensive CDISC documentation for SDTM and ADaM. It utilizes a Retrieval-Augmented Generation (RAG) system for enhanced information retrieval and leverages LLMs to filter and synthesize pertinent information.

Strategic Advantages of Document Assistant: This tool simplifies the process of adhering to CDISC standards, making the path to compliance more straightforward. It enhances users' understanding of CDISC standards, supporting informed decision-making. By integrating vast amounts of data, it provides comprehensive insights and recommendations, enhancing decision-making processes.

Enhanced Data Quality and Compliance: The Regulatory Assistant ensures that organizations can meet CDISC requirements with ease, minimizing manual efforts in navigating and applying compliance guidelines. This empowers users to focus on research outcomes rather than compliance processes.





We will also discuss a use case which is aimed at creating a Unified Studies Definition Model (USDM) by translating diverse human-readable protocols into a standardized format. The envisaged solution combines the versatility of Excel spreadsheets and the efficiency of Natural Language Processing (NLP) techniques to bridge the gap between heterogeneous study designs. The primary objective of this use case is to enhance the clarity, accessibility, and interoperability of research protocols which may lead to development of a standardized, machine-readable representation USDM.

LLMs vs TRADITIONAL NLP TECHNIQUES

The practice of information extraction dates back to the early 1970s and builds on solid foundations of natural language processing (NLP) and linguistics. Such an approach is robust and replicable, but depending on the complexity of the pattern to extract it might hit some barriers. For example, describing contextual dependencies with rules might be a daunting task if there is no standardization in the text documents. Moreover writing linguistic rules requires knowledge about the language and might not scale well in global scenarios.

Generative AI (GenAI) with large language models (LLMs) are now unlocking new levels of accuracy and scalability by providing a flexible and almost ready-to-go solution for extracting information. LLMs can be instructed with ad-hoc prompts to extract specific information in context, without or with limited pretraining effort.

But still LLMs are not the solution to all problems, mostly because of:

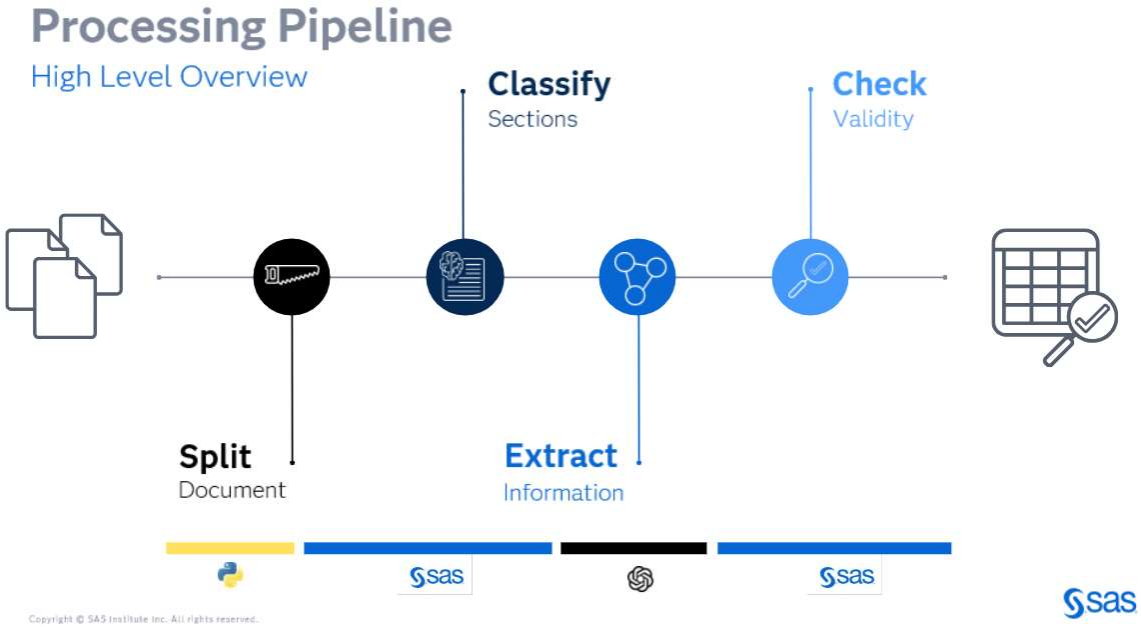
- **Cost**: using commercial LLMs results in costs that are dependent on the length of the text. When it comes to SMPCs, we estimated that 1'000 documents cost 105-240$ for the input only. Output is often 2-3x the price per token as the input.
  Using open source LLMs is not free either, because it requires spinning up dedicated environments with high computational requirements.
- **Hallucinations**: LLMs are generative models in nature, so they can't be fully trusted. In fact, they tend to generate novel content, or reproduce content that was learned during the training process but that might not be relevant to the case at hand. Especially if the extraction impacts downstream tasks, LLMs should not work unchecked.

In the next paragraph we will show how we can combine the best of the two worlds, to get a solution that automates most of the time-consuming tasks, keeps cost under control and provides results accompanied by a confidence score.

COMBINING THE BEST OF TWO WORLDS

We designed a pipeline that processes SMPC documents (in pdf form) and extracts information in a structured way. There are 4 main steps in the pipeline:

1) **Split** the pdf document into paragraphs and sections that are semantically connected.
2) **Classify** each section according to a set of predefined labels (e.g. "Composition", "Method of Administration", …) that are standardized across languages
3) **Extract** information from the relevant section. Here we have a combination of LLM-based extractions and rule-based extractions.
4) **Check** the validity of the LLM-based extraction by means of an automated process that applies NLP techniques and calculates a confidence score.



In the next paragraph we will focus on steps 2) and 4) and highlight how traditional NLP techniques can be used in combination with LLMs to build trust in the results.

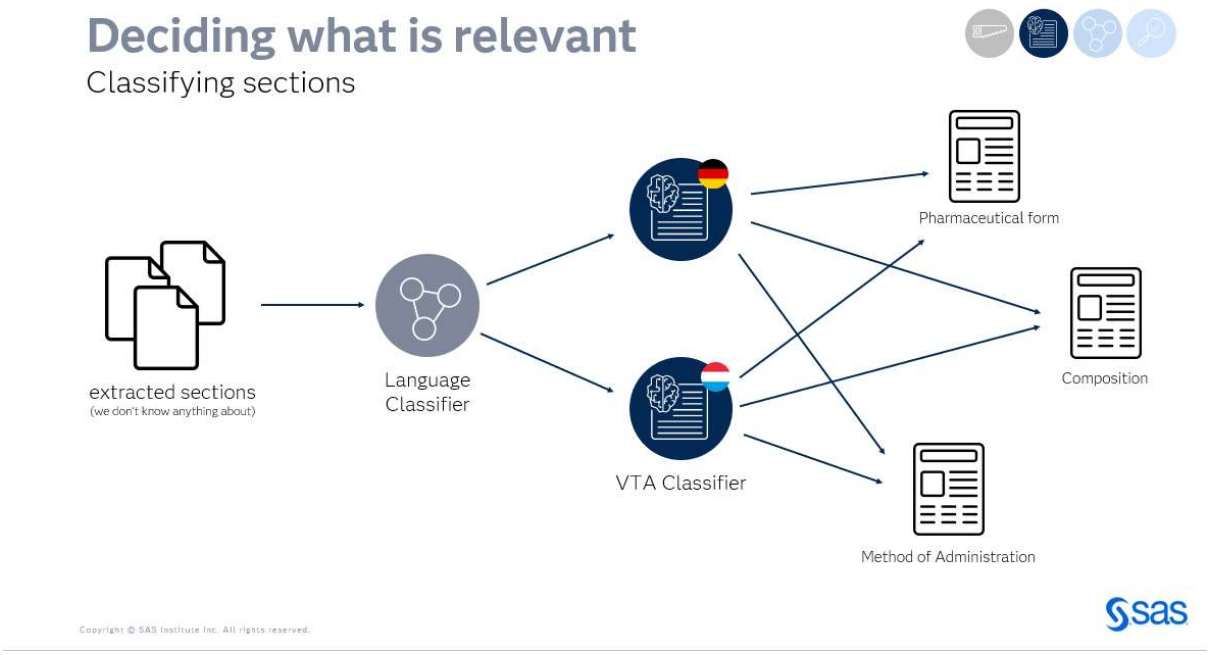
BUILDING TRUST IN THE RESULTS

Pre-Filtering Content

One way to improve the accuracy of LLMs and reduce the risk on hallucinations, is to provide them only with the most relevant content. For example, if you are using an LLM to extract adverse events from an SMPC document, it's better to provide to the LLM only the paragraph describing undesirable effects and contraindications, rather than the entire document.

To achieve this, we split the document into paragraphs and trained a machine learning classifier to label each portion. We then use it as a pre-filter to select the chunks to send to the LLM.

SAS provides tools to automatically detect the language of the document and train one classifier per language, based on topics embeddings and gradient boosting models.



Checking the Results' Validity

After we get the extractions from the LLM, we validate them using NLP techniques.

The flow is fully automated and consists of 5 steps:

1) **Tokenization and lemmatization**: The extraction from the LLM is tokenized and lemmatization is applied to match each term with its lemma (see _Parent_ column). Moreover, a part-of-speech message is attached to each token (see _Role_ column)

	Term	_Parent_	_Role_
1	liver transaminases	liver transaminase	nlpNounGroup
2	increase	increase	V
3	in	in	PPOS
4	liver	liver	N
5	transaminases	transaminase	N

2) **Filter irrelevant or non-important terms**: We used stoplists to filter out terms that can be ignored since they don't carry specific information (articles, prepositions, etc.). We also leverage part-of-speech analysis, to filter out terms based on their role.

	Term	_Parent_	_Role_
1	liver transaminases	liver transaminase	nlpNounGroup
2	increase	increase	V
3	in	in	PPOS
4	liver	liver	N
5	transaminases	transaminase	N

3) **Leveraging noun group identification**: We extract noun groups, such as head nouns and closely tied modifiers. In this setting, it is very useful to detect specific clinical concepts without prior knowledge (e.g., “liver transaminases” carry more information than “liver” or “transaminases” alone). In this step, the tokens that are part of a noun group are dropped when considered separately and just considered as part of the noun group.

	Term	_Parent_	_Role_
1	liver transaminases	liver transaminase	nlpNounGroup
2	increase	increase	V
3	in	in	PPOS
4	liver	liver	N
5	transaminases	transaminase	N

4) **Automatic creation of linguistic rules**: At this point, for each LLM extraction, we have subsets of only the most relevant terms, and we can build a weighted linguistic rule that looks for those terms in a broader context. For more information about Language Interpretation for Textual Information (LITI) rules, see the SAS documentation link. Note that the LITI rule is automatically generated by concatenating the proper operators and the terms extracted in the previous steps.

	rule
1	1:rule_1:(OR[NW],(AND[0.5],"increase@"),(AND[0.5],"liver transaminase@"))

5) **Rule inference and confidence score calculation:** Finally, the same corpus that was used to extract information with the LLM can be scored against the LITI rule to check if the information is actually present. A confidence score is calculated based on the number of relevant terms matched by the rule. Confidence is a number between 0 and 1, where a higher value indicates a better quality of the extraction.

doc_id	text	match	confidence
1	Elevated liver transaminase levels	liver transaminase	0.5
2	Increased liver transaminase levels	Increased liver transaminase	1
3	Increasing liver transaminases after consumptions	Increasing liver transaminases	1
4	It has been reported that liver transaminases increased in 0....	liver transaminases increased	1

This method allows to generate, in an automated way, a quantitative measure of confidence in the results. The same method could be used to compare performance of different LLMs at the information extraction task.

Benefits of integrating SAS NLP and Gen AI

1. **Avoiding Hallucinations:** The NLP and text analytics pre-filtering process assimilates the most relevant source data from various documents, ensuring the outputs are more accurate and reliable.
2. **Enhancing time to value**: By pre-filtering the data, a smaller LLM can handle GenAI tasks more efficiently, leading to quicker results. Providing more focused context to an LLM significantly enhances output quality, especially for weaker models.
3. **Ensuring privacy and security**: Using a local vector database for fine-tuning generative models is possible. This gives users only relevant embeddings to the LLMs via APIs or localized instances of the LLM, ensuring the privacy and security of sensitive data.
4. **Reducing costs**: Text analytics and NLP significantly reduce the amount of information sent to the LLMs. In some cases, only 1 – 5% of the overall data is used for answers. This eliminates the need for excessive external API calls and reduces the computational resources required for localized LLMs.
5. **Traceability**: SAS® Viya® enables end-to-end verification and traceability of results, helping users to verify information and trace it back to the statements from which the summaries were derived, potentially thousands of statements. This traceability feature enhances transparency and trust in the generated outputs.
6. **Establishing guard rails**: establish guard rails to control the information that's sent to the LLMs and also analyse the output – quality checks. Adopting a "trust but verify" approach ensures that LLMs' extractions, which can impact downstream tasks, are checked and validated to prevent unchecked errors.

Conclusion: So What?

- **Transforming Clinical Trial Protocols**: This work underscores the transformative potential of advanced data management techniques in clinical trial protocols.
- **Efficiency and Accuracy**: By adopting innovative technologies and fostering collaborative efforts, the industry can significantly enhance efficiency and accuracy in clinical research.

- **Unlocking Potential**: These examples merely scratch the surface of the vast potential unlocked when combining the precision of linguistic methods in SAS NLP with LLMs.
- **Quality and Control**: These techniques not only address quality issues in text data but also integrate subject matter expertise, providing organizations with substantial control over their corpora.
- **Reducing Corpus Size**: In some cases, the corpus size for fine-tuning can be reduced by up to 90%.
- **Improving LLM Responses**: By curating higher-quality data for fine-tuning, more accurate responses from LLMs can be achieved, reducing hallucinations and establishing a method for validating responses.

These conclusions highlight the significant impact and future potential of integrating advanced data management and AI techniques in clinical research.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Please contact the authors at:

Author Name: **Federica Citterio**
Company: SAS Institute
Address: Via Carlo Darwin, 20/22, 20143 Milano, Italy
Email: Federica.Citterio@sas.com
https://www.linkedin.com/in/federica-citterio/

Author Name: **Soundarya Palanisamy**
Company: SAS Institute
Address: In der Neckarhelle 162, 69118 Heidelberg, Germany
Email: Soundarya.Palanisamy@sas.com
https://www.linkedin.com/in/soundaryap/