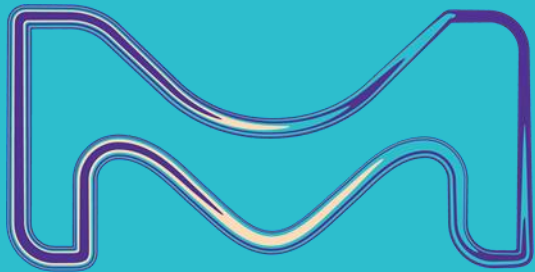


SAMBA Bridging the Gap

How GenAI Transforms Analysis Plans into ADaM Specs

Thiago Figueiredo, Richard Becker

PHUSE EU Connect 2025



MERCK

Overview of the presentation

1. Motivation
2. PoC - Phase I and Phase II
3. Architecture details of SAMBA
4. Outlook
5. Conclusion

Background

Statistical programmers play a crucial role in supporting analyses for regulatory Submissions on clinical trials. These analyses are based on the Statistical Analysis Plan (SAP), developed together with the biostatisticians, quantitative pharmacologists and the clinical team.

A significant challenge is converting these SAPs into *programming specifications or specs* adhering to the ADaM (Analysis Data Model)

What is our goal?

- Reduce manual effort in creating ADaM specifications
- Support programmers by connecting data sources like SAP and specs and making them available in one place



Introduction

What we planned to do?

- Create a tool to leverage LLMs for (partially) automating the task of mapping information from the SAP to the ADaM specifications
- Reduce effort and amount of manual work during specs creation
- Use-cases in scope:
 - Create specs for a new study based on existing specs
 - Update specs for ongoing study based on SAP / protocol updates.

Which input sources were used?

- Statistical Analysis Plan (SAP)
- SME domain-specific knowledge built to support the tool (dataset and domain descriptions, labeled examples, etc)

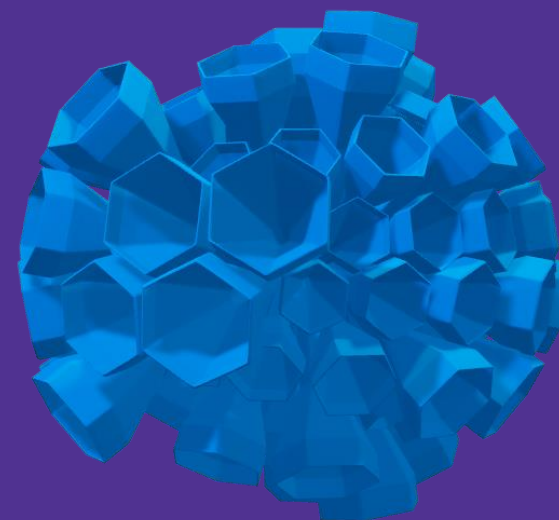


Phase I: **POC with External provider**

The POC was done in a timeframe of 5 weeks.

Approach taken

- SAP markdown files used for text chunking
- Identify chunks related to a domains[pre-defined Domain to Variable mapping]
- LLM based Variables extraction from these domain's chunks based on dependency graph [pre-defined extraction queries for a curated dataset]



External POC Evaluation results

PROS

Scaffolding: Provide infrastructure code to manage LLM calls, model settings and document processing (e.g. chunking, indexing).

Initial search/extraction workflow: Set up initial mechanism for search and extraction based on manually defined queries

Known variables: Current design: would be best used to answer specific questions on known variables

Overall Impression: The tool has significant potential for aiding programming tasks but requires enhancements in accuracy, scope, and technical capabilities to maximize its effectiveness and user friendliness

CONS

Approach: Searching for variables but iAP describes analyses. Excessive reliance on model knowledge for content creation.

Limited accuracy: Actual useful generated content was low.

Blank template specs as starting point: Tool starts from blank specs, does not reflect actual user process.

Manual effort: Document processing and correction was outside of the scope of the tool and led to high burden on the user.

Phase II: Development & Achievements



.docx iAP upload



Analysis Discovery



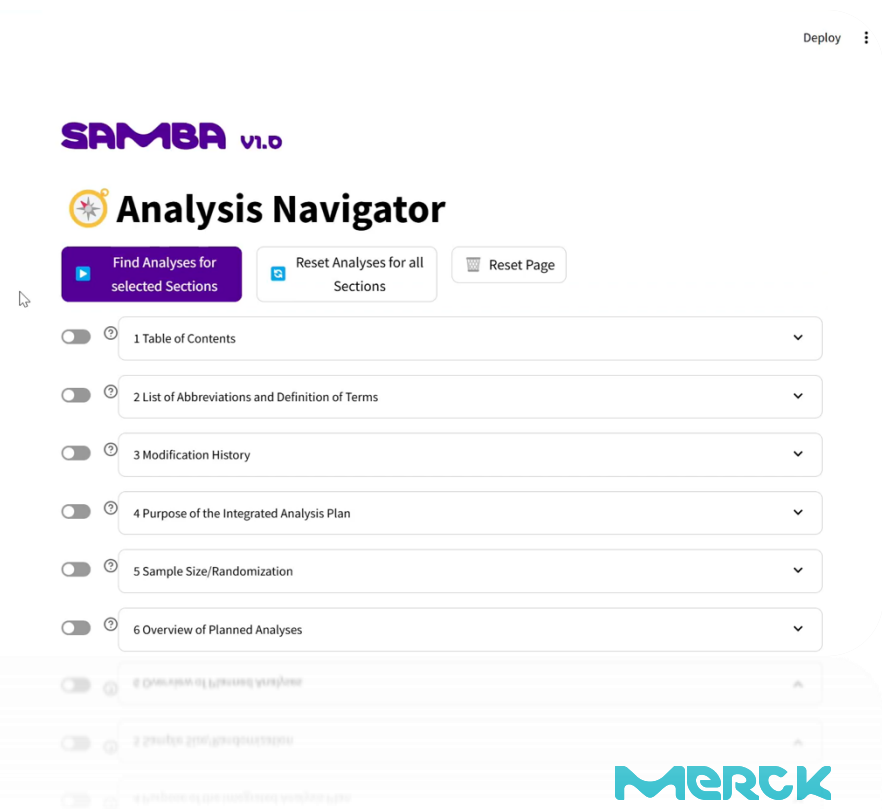
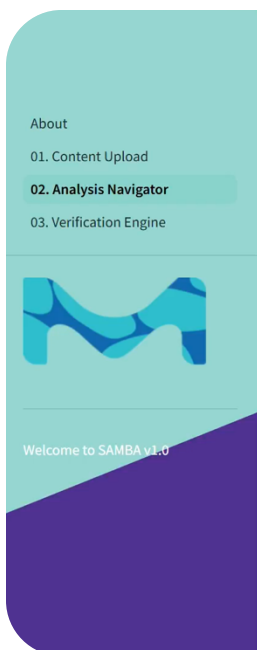
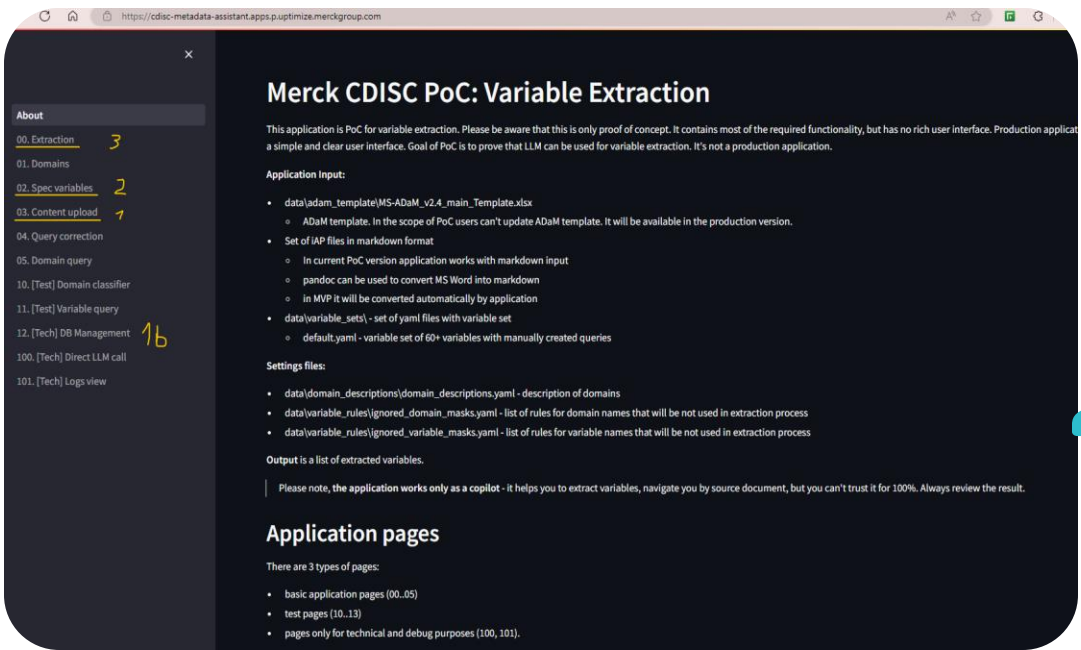
Specs Verification



Query Optimizer

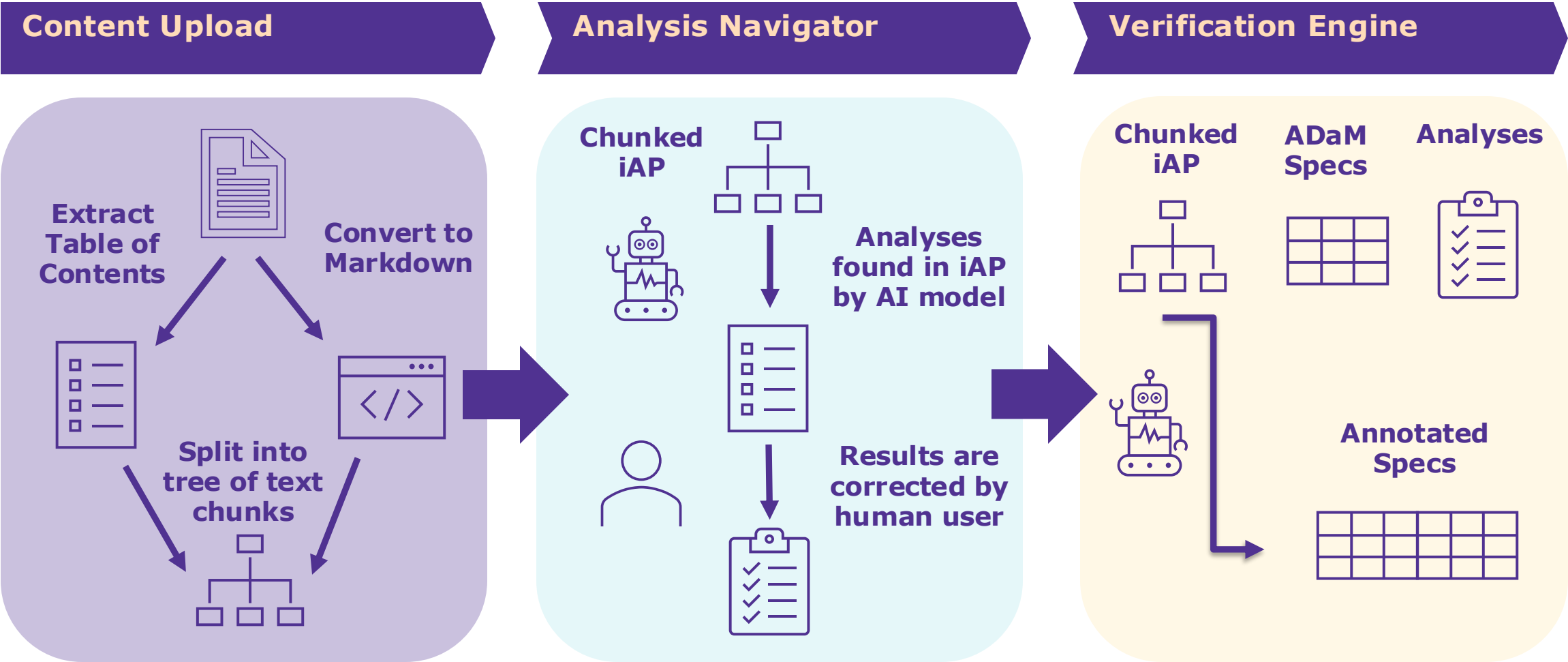


UI Enhancements



Architectural Details

SAMBA Components



Analysis Navigator

About

01. Content Upload

02. Analysis Navigator

03. Verification Engine

04. Query Optimization

[Tech] Analysis Navigator Evaluation

[Tech] Direct LLM call

[Tech] Logs view

[Tech] SQL Query Editor

Quick Start Guide

Support



Analysis Navigator

Find Analyses for selected Sections

Reset Analyses for all Sections

Reset Page

☐

1 Table of Contents

▼

☐

2 List of Abbreviations and Definition of Terms

▼

☐

3 Modification History

▼

☐

4 Purpose of the Integrated Analysis Plan

▼

☒

5 Objectives and Estimands

▼

☒

6 Overview of Planned Analyses

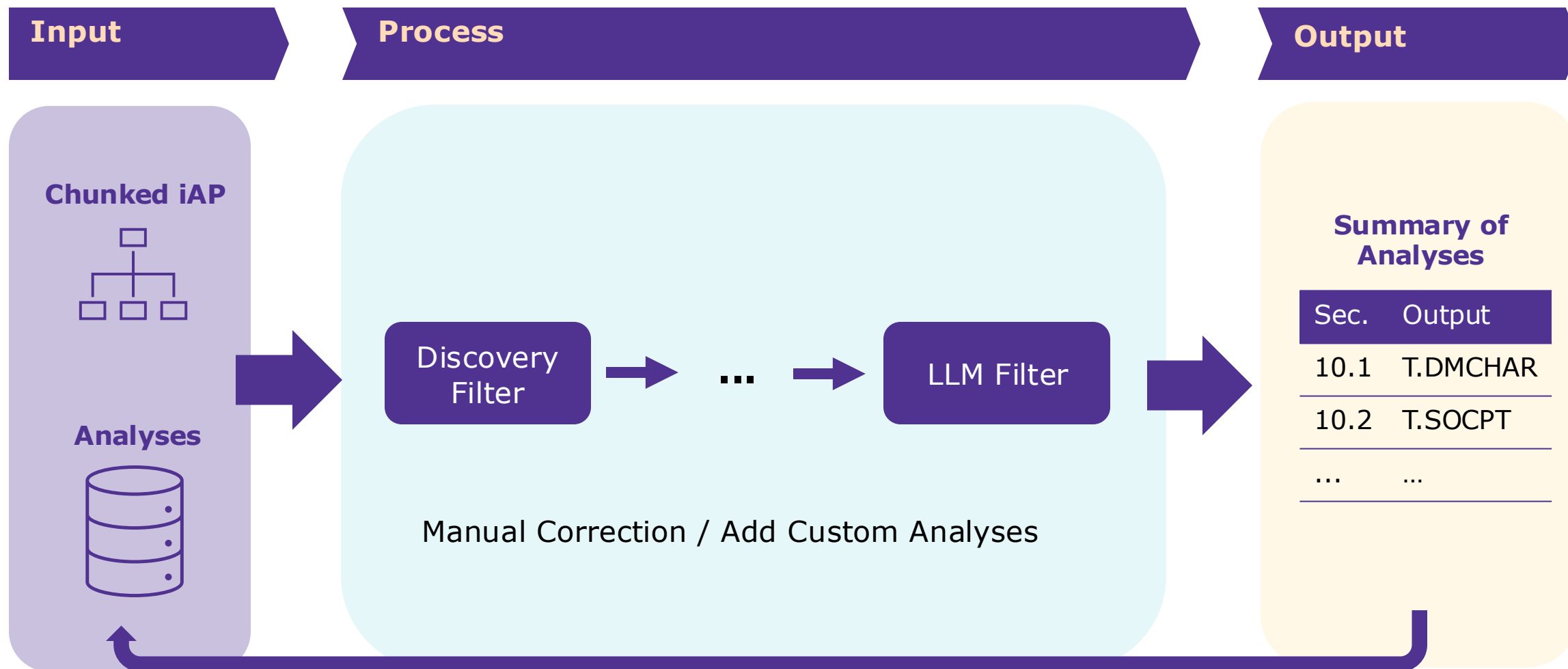
▼

☐

7 Changes to the Planned Analyses in the Clinical Study Protocol

▼

Analysis Navigator



Selection of specs file for verification

About

01. Content Upload

02. Analysis Navigator

03. Verification Engine

04. Query Optimization

[Tech] Analysis Navigator Evaluation

[Tech] Direct LLM call

[Tech] Logs view

[Tech] SQL Query Editor

Quick Start Guide

Support

Go Back

✓

2

3

4

Analysis Plan Specifications Variable Set Verification

MS201923_0050-...

specification sheet source

Select from study

Upload your own

Study Number:

Program Name:

Year of Regulatory Approval:

Therapeutic Area:

Study Phase:

Xev

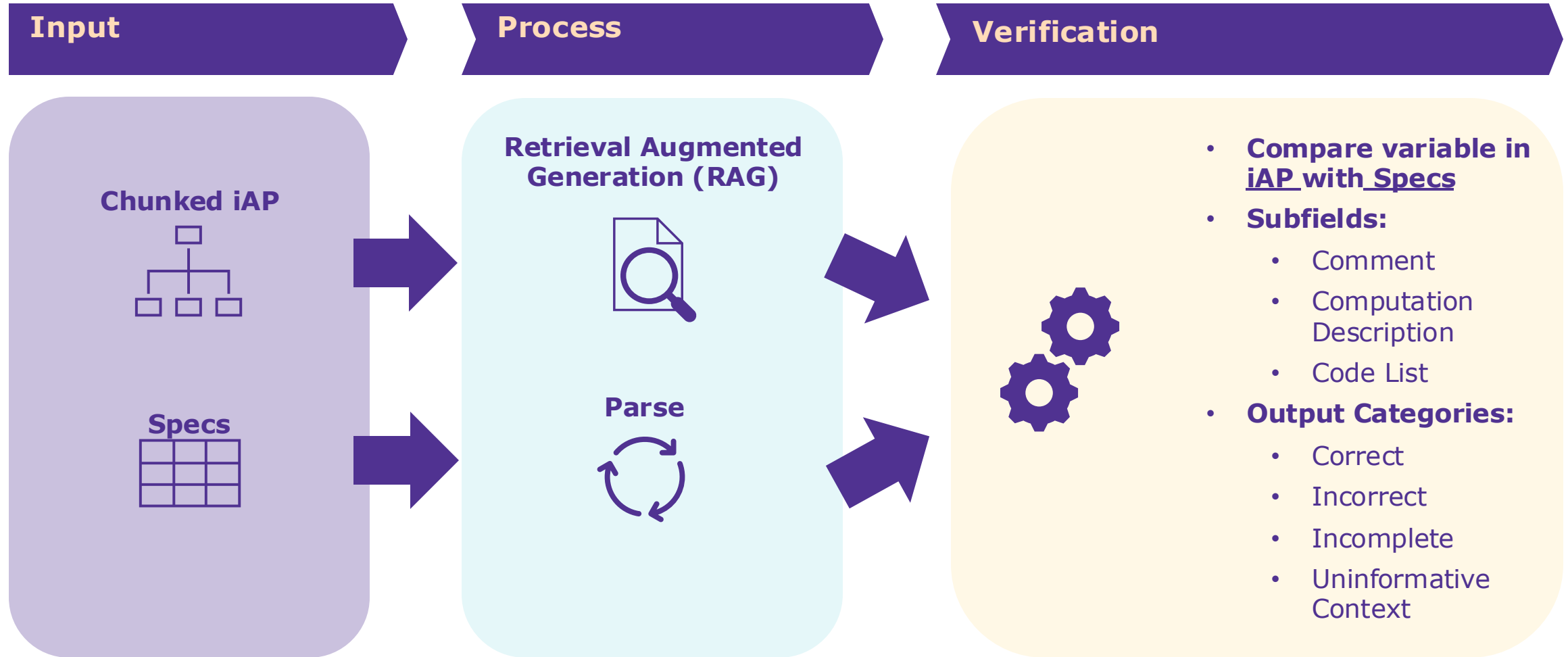
O... x x v

Any v

Select	Study Number	Program Name	Year of Regulatory Approval	Therapeutic Area	Study Phase	id
☒	MS202359_0004	Xevinapant	2022	Oncology	Ph I	https://merckg

Continue

Verification Engine



What do users get as output of the tool

1. Table of Contents

(Analysis Navigator)

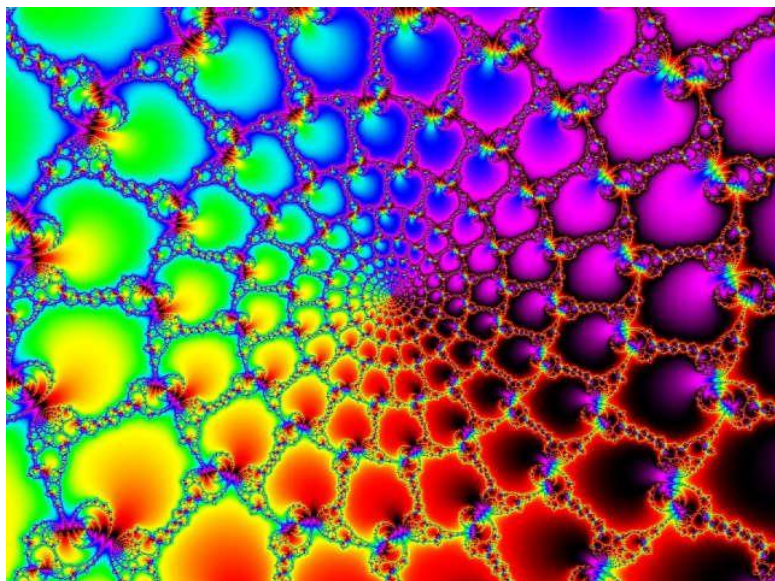
2. Annotated ADaM Specs

- Verification Engine Classifications
- Required Variables per Analysis Navigator
- Added placeholder variables for required variables not contained in given specifications

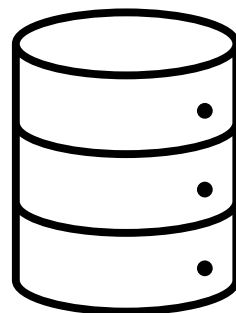
Dataset	Variable	Verification of Computation Description	Explanation	Reference Sections
ADSL	TRT01P	Correct	The answer correctly describes the study intervention presentation for the main study and safety run-in dose levels. The information provided in the chunks supports the answer. Specifically, Chunk 2 details the dose levels and the combination of Compound 1 and Compound 2 for DL1 and DL2.	## Objectives and Estimands > Safety Run-in Part in Japan
ADSL	AENDT	Correct	The answer correctly describes the method used to compute the end date for adverse events (AE) relative to the study start. It mentions that the end date-time of the Adverse Event Occurrence is the one of the last record in the chain of records that are part of that Adverse Events Occurrence, ordered by increasing start date-time	## Safety Analyses > Adverse Events (Secondary Endpoint)

Hallucinations can be mitigated by providing clear links to original data and human in the loop

Hallucinations



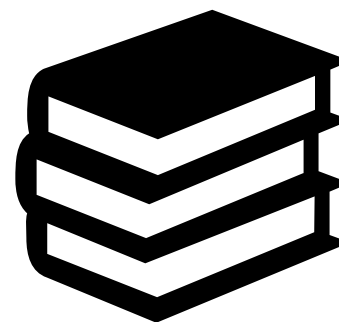
Structured Data



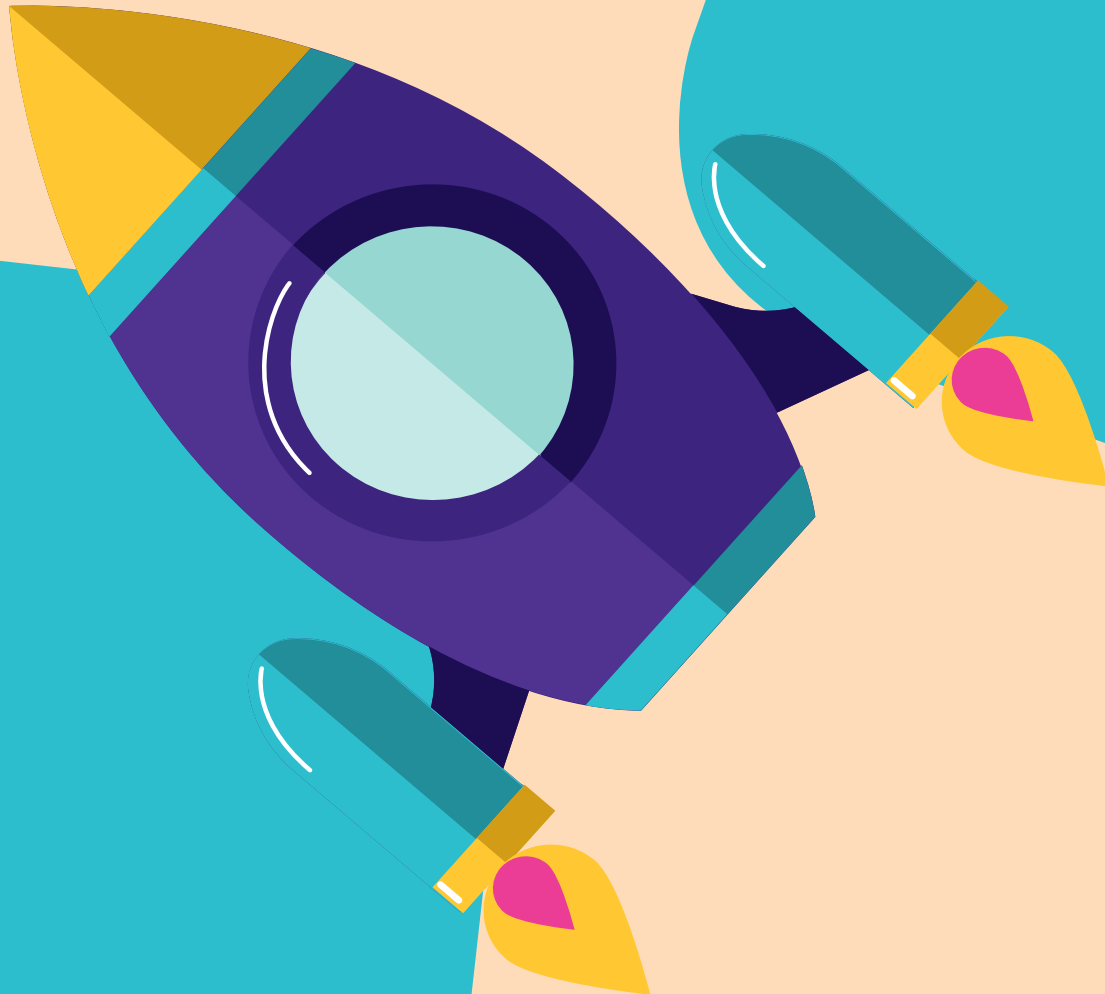
Human in the Loop



Link to original documents



Outlook



Next Steps



Ideas currently in evaluation:

- Improve integration into programmer's workflow and with MS-ADaM Standards
- Creation of module for content generation / recommendation based on verification step
- Evaluate integration with other tools and data sources (e.g. protocol)



Conclusion

- SAMBA has shown that LLMs can be used to extract information from the iAP to fill in ADaM specs
- Structured input data and labelled data are crucial for improving performance of the tool
- LLMs tend to hallucinate information and direct links to original data are essential for traceability
- Human in the loop is and will remain the most important element driving automation

Contact Details



THANK
YOU!

