



Revolutionising Clinical Data Transformation: Harnessing GenAI and Open Source for Automated SDTM Conversion

Hao Chen, Daniel Huang
Scientific Programming
Global Statistics and Data Science (GSDS)
BeOne Medicines Co., Ltd.



Disclaimer

The content, ideas, and conclusions presented in this presentation are solely those of the author(s) and do not necessarily represent or reflect the views, policies, or official positions of BeOne Medicines. While every effort has been made to ensure the accuracy and reliability of the information provided, BeOne Medicines does not endorse or assume responsibility for any statements, interpretations, or recommendations made herein. The material is intended for informational purposes only and should not be construed as the official opinion or guidance of BeOne Medicines.

The Challenge in Clinical Development

What We Expect



Tesla Gigafactory 3 Shanghai
<https://www.youtube.com/watch?v=-Ds1xV7M2gI>

- Standardized, accurate, and repeatable
- Well-defined, controlled processes
- Errors easily detected and corrected

What We Face



Pony.ai self-drive
<https://www.youtube.com/watch?v=jKFV7qjWaD4>

- Dynamic, ever-changing, and complex
- Human factors and real-world variability
- Rare events and exceptions are critical

SDTM Transformation: A Critical AI Opportunity

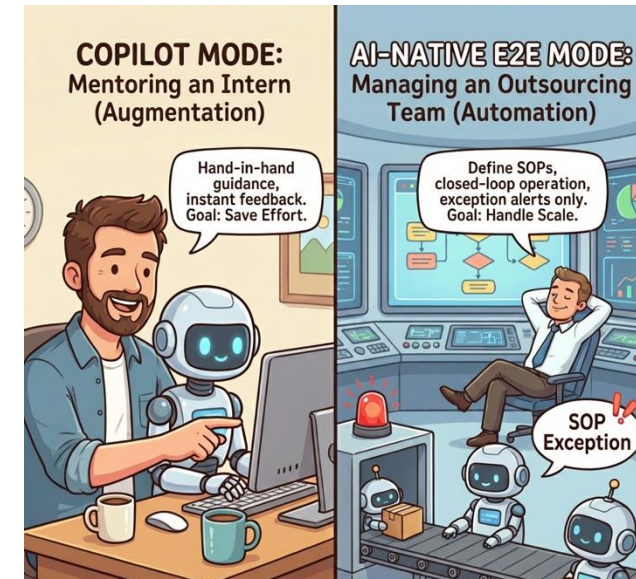
- | | |
|-----------------------------|---|
| 1 Regulatory-critical | Mandatory CDISC standard for global regulatory submissions |
| 2 Data complexity explosion | Rapid growth in data volume, velocity, and diversity across clinical trials |
| 3 Operational bottleneck | Traditional SDTM transformation is labor-intensive, slow, and error-prone |
| 4 Decision lag | Conventional BI and rule-based automation cannot enable real-time, adaptive decision-making |

Strategic inflection point:

AI is the only viable path to simultaneously improve efficiency, quality, and speed at scale

Evolution of SDTM Mapping

- 1 Manual Programmer-Centric**
Study-specific programs; limited reuse; long cycle times
- 2 Metadata-Driven**
Central repositories store mapping rules; specs auto-generated (copied)
- 3 AI/ML-Assisted Mapping (NLP, Classification)**
ML models suggest mappings
- 4 Generative-AI Copilots**
LLM assistants draft code and explain standards with human review
- 5 Generative-AI End-to-End**
Orchestrated workflows from Raw to SDTM under human governance



The Power of R

- **Great Language for Data Analysis & Reporting**

- Mature, production-grade platform for statistical computing, analysis, and reporting (Posit-maintained).

- Readable syntax and strong ecosystem for pharma use (e.g., tidyverse/ggplot2; broad adoption).

- **Open-Source**

- Large package ecosystem enables reuse, extensibility, and rapid innovation.

- Enables AI-assisted R development (code understanding/generation) and supports advanced AI-enabled applications.

- **R Shiny UI**

- Robust framework for interactive analytics apps and dashboards.

- Supports human-in-the-loop (AI + HI) workflows and near real-time reporting via web-deployed R logic.



SDTM Automated Transformation Assistant (Sata)

- **Virtual agent** that transforms raw clinical data into SDTM datasets
- **Automated mapping** using standard rules and study documentation (protocol, CRFs)
- **CDISC compliance** through repeatable logic and structured mappings
- **Manager-led oversight** combines AI automation with human governance
- **Intelligent system** for understanding research, supporting analysis, and enabling intelligent insights

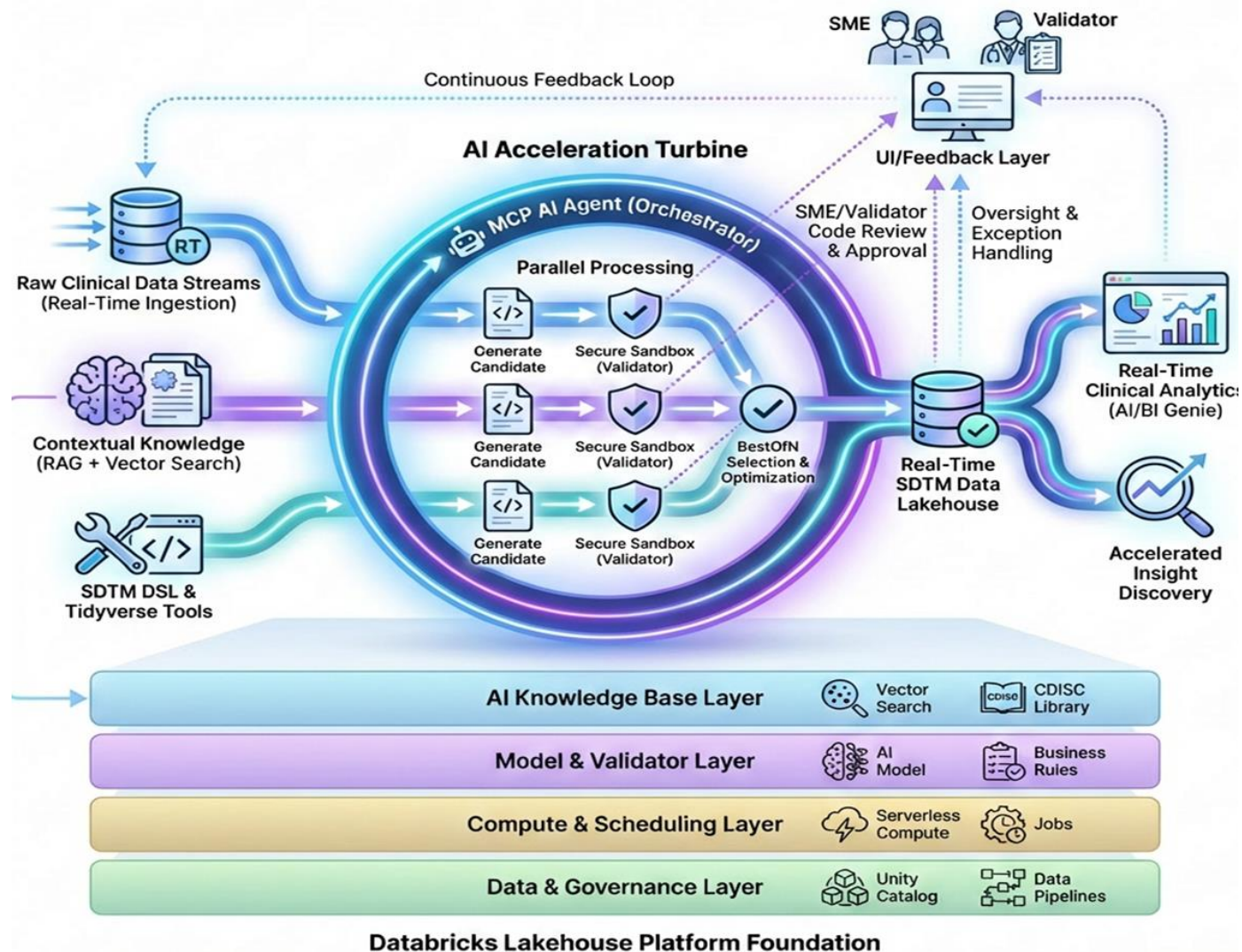


Core Benefits

- **Faster SDTM delivery**
- **AI + cloud-scale compute**
- **Less manual work, fewer errors**
- **Higher data quality**
- **Earlier, better decisions**
- **Accelerated downstream workflows**
- **Faster time-to-market**
- **Single source of truth**

Core Design

- AI Prediction
- R Code Generation
- Secured Sandbox
- Automatic Testing
- Human-in-the-loop Validation



Key Features

- All-in-One integrated web application
- Dramatic efficiency gain - Initial release reduces SDTM mapping time from hundreds of hours to hours
- End-to-end acceleration - Full version delivers complete SDTM results in several days
- Broad data support - Supports Both corporate and industry standard raw data

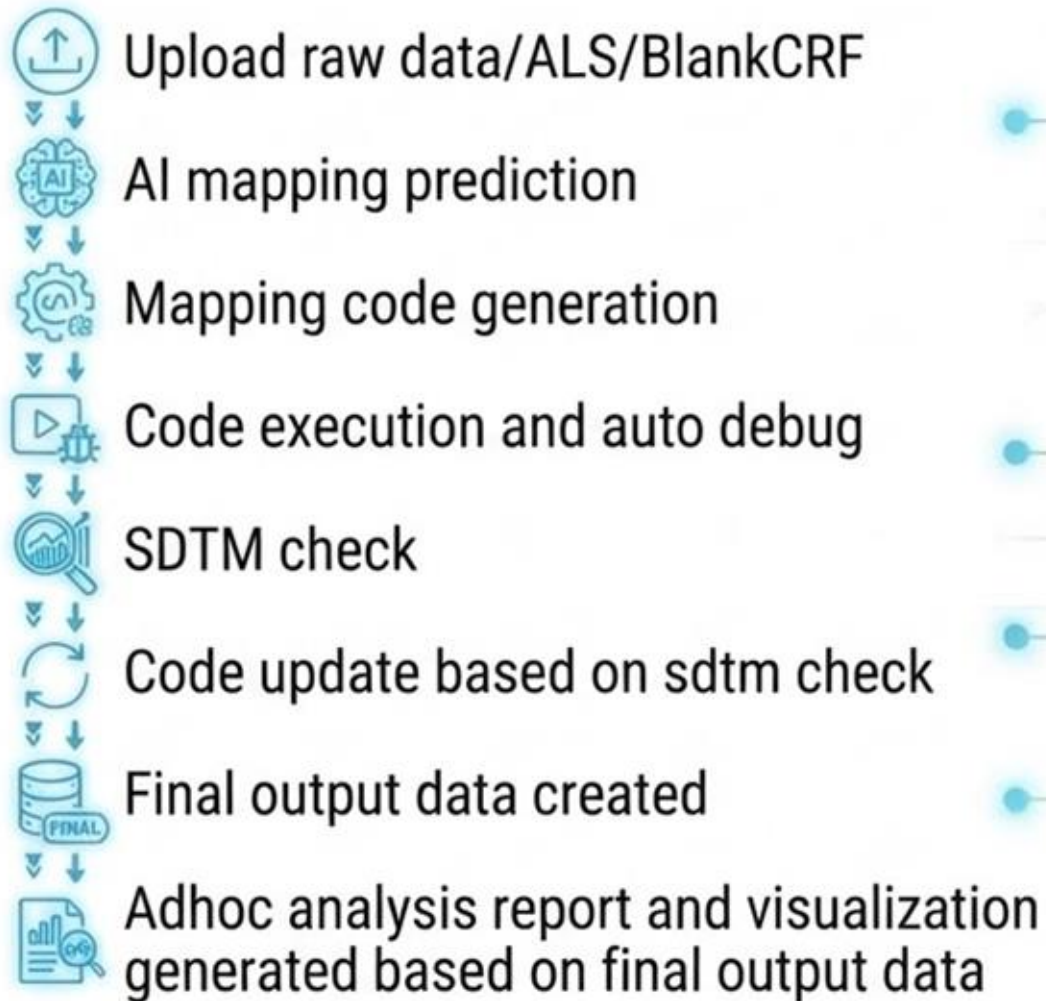


Engines

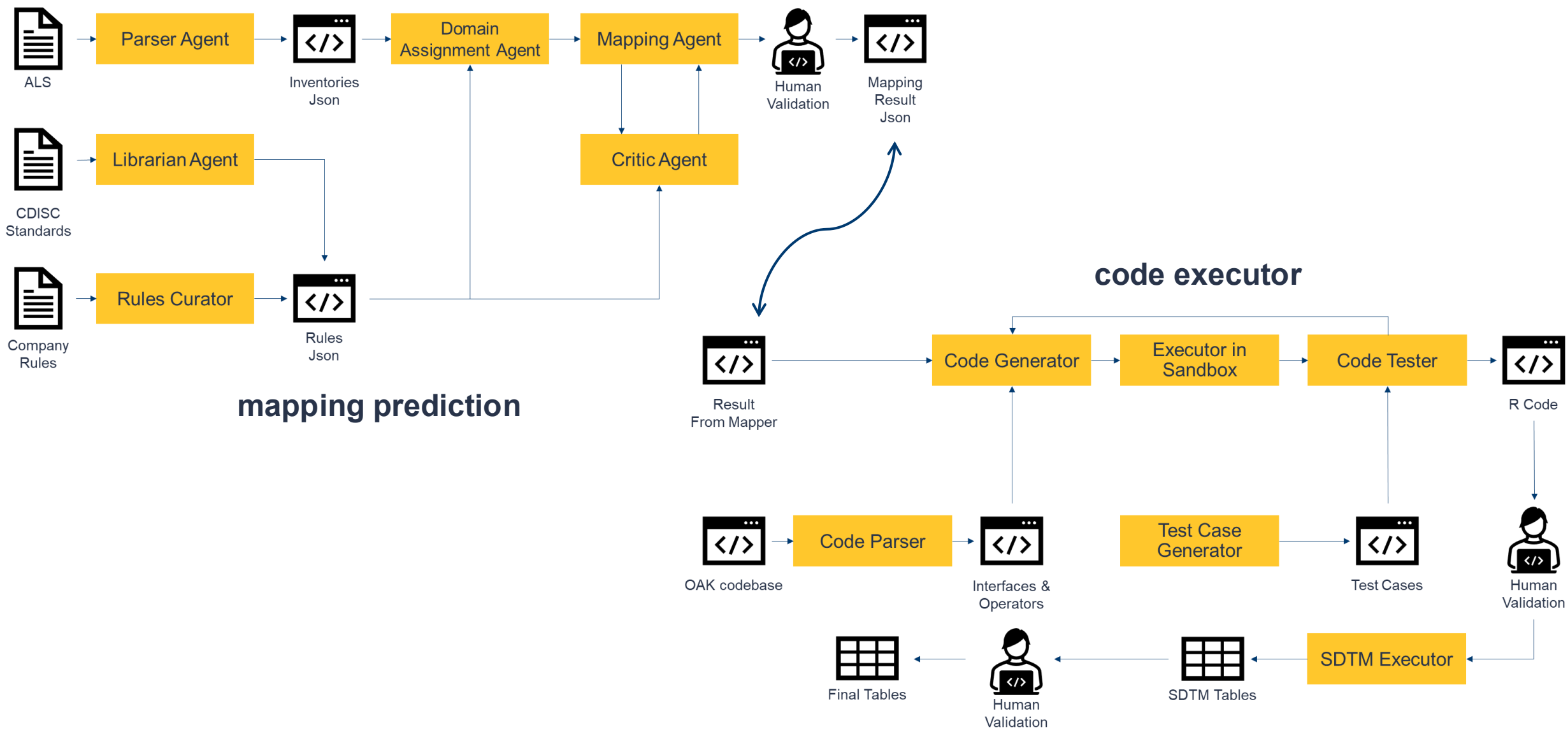
- No mapping spec needed, Only metadata from Raw data / ALS / BlankCRF /Data Description
- Real time analysis and Full traceability
- On-demand review/partial update/rerun
- Seamless AI+HI integration



End-to-End Workflow



Mapping Prediction and Code Executor agents



Prediction Methods

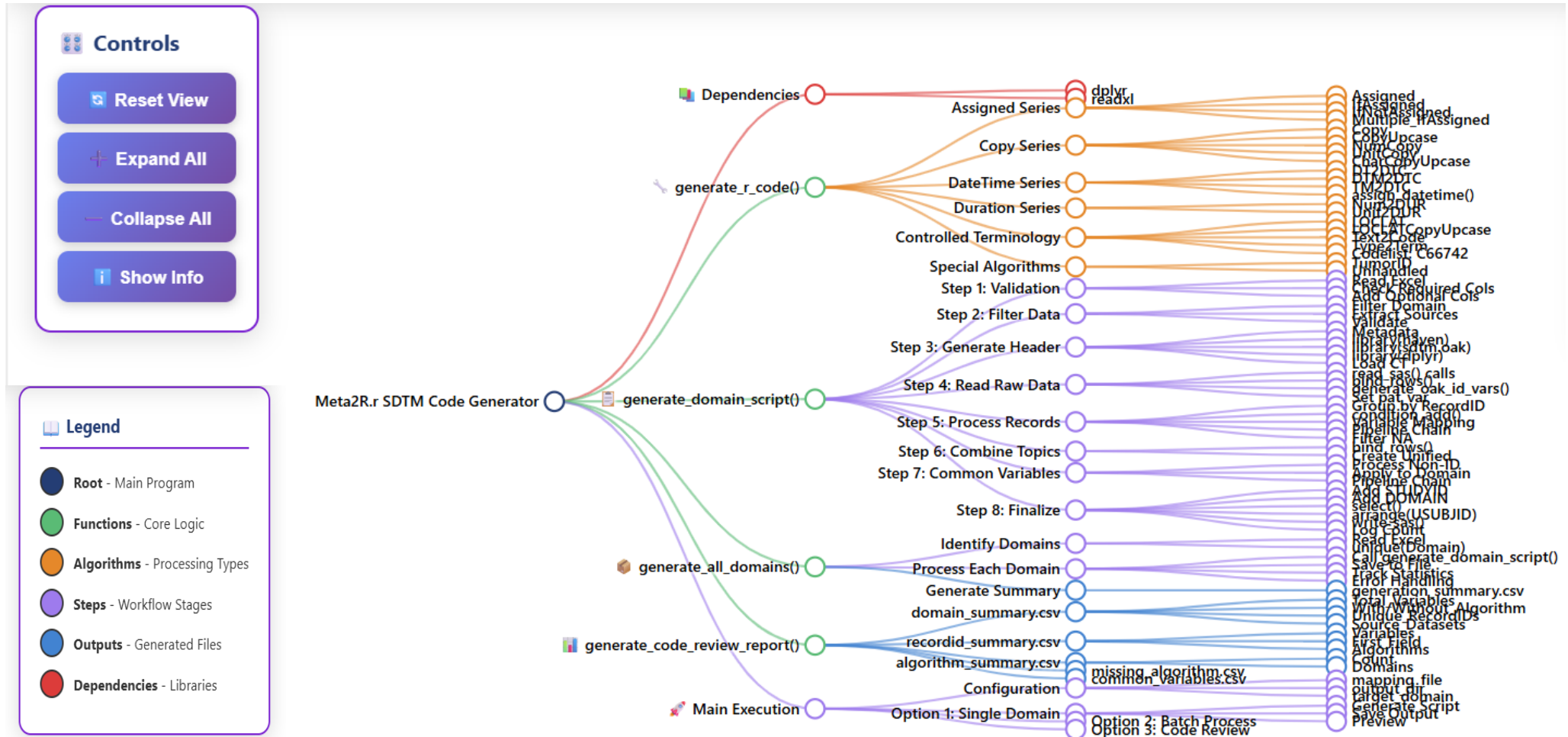
METHOD TYPE	TRACEABILITY	DEEP THINKING	DESCRIPTION
Exact Match (100%)	Good	No	Direct alignment of labels/definitions
Fuzzy Match (>95%)	Good	No	High-confidence similarity
Fuzzy + Reasoning (70–95%)	OK	Yes	Batch of candidates with rationale and workflow checks
Deep Reasoning (Search)	Not good, Need post-processing	Yes	Concept-level inference requires workflow controls, Understands CDISC IG and historical mappings

FormOID	FormName	FieldName	FieldLabel	Predict_Annotation	GroundTruth			
SUBJECT	Subject	Form	Subject	DC = Demographics as Collected	DC = Demographics as Collected			
SUBJECT	Subject	SUBIID	Subject number	SUBIID	SUBIID			
SUBJECT	Subject	COUNTRY	Country	COUNTRY	COUNTRY			
SV	Date of Visit	Form	Date of Visit	SV = Subject Visits	SV = Subject Visits			
SV	Date of Visit	SVOCCUR	Did visit occur?	SVPRESP = Y; SVOCCUR = Y; SVOCCUR = N	SVPRESP = Y; SVOCCUR = Y; SVOCCUR = N			
SV	Date of Visit	SVSTDAT	Date of visit	SVSTDTC	SVSTDTC			
DS_IC	Informed Consent	Form	Informed Consent	DS = Disposition; DSCAT = PROTOCOL MILESTONE	DS = Disposition; DSCAT = PROTOCOL MILESTONE			
DS_IC	Informed Consent	DSSTDAT	Date informed consent	DSSTDTC where DSDECOD = INFORMED CONSENT	DSSTDTC where DSDECOD = INFORMED CONSENT			
DS_IC	Informed Consent	ICPV	Protocol version	SUPPDS.QVAL where QNAM = DSPTOTVS	SUPPDS.QVAL where QNAM = DSPTOTVS			
DS_OIFC	Optional Informed Consent	Form	Optional Informed Consent	DS = Disposition	DS = Disposition			
DS_OIFC	Optional Informed Consent	OIFC_DSCTYPE	Type of consent	DSTERM	DSTERM			
DS_OIFC	Optional Informed Consent	OIFC_DSYN	Was consent obtained	DSCAT = PROTOCOL MILESTONE if Yes; DSDECOD = INFORMED CONSENT if No	DSCAT = PROTOCOL MILESTONE if Yes; DSDECOD = INFORMED CONSENT if No			
DS_OIFC	Optional Informed Consent	OIFC_DSSTDAT	Date of consent	DSSTDTC	DSSTDTC			
DM	Demographics	Form	Demographics	DC = Demographics as Collected; RP = Reproductive System	DC = Demographics as Collected; RP = Reproductive System			
DM	Demographics	BRTHYY	Year of birth	BRTHDTC	BRTHDTC			
DM	Demographics	AGE	Age	AGE; AGEU = YEARS	AGE; AGEU = YEARS			
DM	Demographics	SEX	Sex	SEX = M; SEX = F	SEX = M; SEX = F			

Prediction Metrics

Accuracy (%)	Study A	Study B	Study C	Study D	Average
Rule Accuracy	96.84	93.45	93.64	94.53	93.94
Field Accuracy	93.83	89.13	86.75	90.71	89.29
Rule Acc (ex-SUPP)	97.95	92.88	93.93	96.52	94.93
Field Acc (ex-SUPP)	95.98	89.56	87.80	94.13	91.36

R DSL/Engine – SDTM.OAK



[pharmaverse/sdtm.oak](#): An EDC and Data Standard agnostic SDTM data transformation engine that automates the transformation of raw clinical data in ODM format to SDTM based on standard mapping algorithms

SDTM Code Generator

SDTM Mapping

Step 1: Raw Data

✓ Loaded: 85 files

Browse... Upload ac

Step 2: Mapping File

✓ Loaded:

BGBA317A445_101EXP_V1_CRF2SDTMSpecs.xlsx

Domains: 17

Browse... Upload di

Step 3: Select Domain

✓ Available: 17 domains

PR

Generate Code

Run Generated Code

Source Code

Download Code Run

```
1 # =====
2 # SDTM Domain: DS
3 # Generated: 2025-12-08
4 # =====
5
6 library(haven)
7 library(sdtm.oak)
8 library(dplyr)
9
10 # Read CT
11 study_ct <- read.csv("rawdata/sdtm_ct.csv", na.strings = c(""))
12 study_ct <- study_ct %>%
13   filter(!is.na(codelist_code), !is.na(term_value))
14
15 ### Read Raw Data
16
17 ds_ic_raw_ds <- read_sas("rawdata/ds_ic.sas7bdat")
18 ds_enr_raw_ds <- read_sas("rawdata/ds_enr.sas7bdat")
19 swt_raw_ds <- read_sas("rawdata/swt.sas7bdat")
20 ds_eot_raw_ds <- read_sas("rawdata/ds_eot.sas7bdat")
21 ds_eot_317_raw_ds <- read_sas("rawdata/ds_eot_317.sas7bdat")
22 ds_eot_1_raw_ds <- read_sas("rawdata/ds_eot_1.sas7bdat")
23 ds_eot_2_raw_ds <- read_sas("rawdata/ds_eot_2.sas7bdat")
24 ds_eos_raw_ds <- read_sas("rawdata/ds_eos.sas7bdat")
25 reic_raw_ds <- read_sas("rawdata/reic.sas7bdat")
26 ds_combined <- bind_rows(ds_ic_raw_ds, ds_enr_raw_ds, swt_raw_ds, ds_eot
27
28
29 # Create oak_id_vars
30 ds_raw <- ds_combined %>%
31   generate_oak_id_vars(
32     pat_var = "Subject",
33     raw_src = "ds"
34   )
35
36
37 # =====
38 # Topic: DSSTDAT
39 # =====
```

Result

Console Log Raw Data Output Dataset

Search:

Multiple datasets generated (20 total) - Click row to view details

Dataset	Rows	Columns
ds	4830	14
ds_combined	690	149
ds_dsstdat	690	7
ds_enr_dsstdat	690	7
ds_enr_raw_ds	132	51
ds_enryn	690	8
ds_eos_raw_ds	92	51
ds_eot_1_raw_ds	12	41
ds_eot_2_raw_ds	12	41

Showing 1 to 20 of 20 entries

Previous 1 Next

SDTM Code Executor

Session reset. Signed in as: hao1.chen@beigene.com

Language

R

Cluster

SDTM-Coder (1119-062443-q753chwi)

Number of Threads

10

Max Retries per Thread

10

Attachments

Upload Files

No files uploaded yet.

Chat Prompt

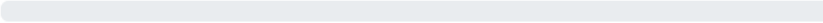
Explain the task you want the LLM to generate code for...

Reset conversation

Generate code

Conversation History

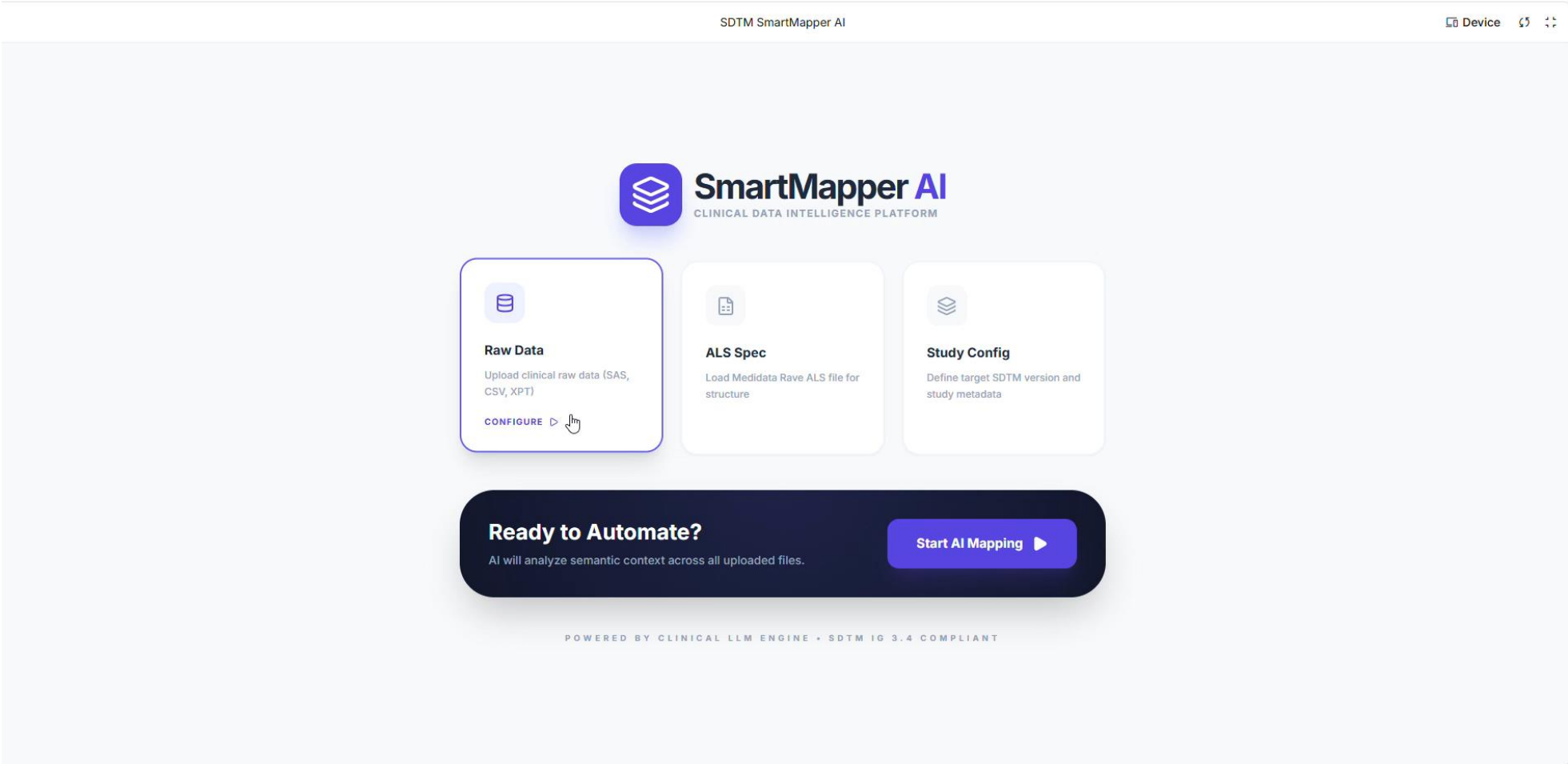
Multi-thread Execution Progress (10 threads)



Success: 0 | Ready: 0 | Failed: 0 | Running: 0 | Idle: 10

Thread 0 Temp: 0.50 Retry: 0/10 Idle
Thread 1 Temp: 0.61 Retry: 0/10 Idle
Thread 2 Temp: 0.72 Retry: 0/10 Idle
Thread 3 Temp: 0.83 Retry: 0/10 Idle
Thread 4 Temp: 0.94 Retry: 0/10 Idle
Thread 5 Temp: 1.06 Retry: 0/10 Idle
Thread 6 Temp: 1.17 Retry: 0/10 Idle
Thread 7 Temp: 1.28 Retry: 0/10 Idle
Thread 8 Temp: 1.39 Retry: 0/10 Idle
...

E2E Design Demo



Summary

✓ GenAI is Reshaping Biopharma

- Clinical trials as prime transformation area: faster cycles, smarter execution
- Applied to SDTM mapping and downstream deliverables

✓ Knowledge Lib & LLM combined for E2E AI System

- AI for mapping suggestions, code drafting, reasoning, testing and iteration
- Standard and Deterministic methods for QC, traceability

✓ AI + HI Collaboration in SDTM Conversion

- HITL(Human-in-the-Loop) workflows surface uncertainty, isolate errors, enable rapid expert review
- Close the learning loop: reviewer corrections improve models while maintaining compliance

Acknowledgment

Xiang Guo

Cindy Song

You Li

Tao Zhang

Jinling Li

Xuejie Zhou

Haomin Yang

Qianwang Wang

Jing Wang

Chunpeng Zhao

Shuang Gao

Xiaoyuan Huang

Wei Wang

Raymond Chan

Yaohui Zhu

Thank You

- 55 Cambridge Parkway
Suite 700W
Cambridge, MA 02142

- 1 (877) 828-5568

- BeOneMedicines.com

