From Solo to Sidekick: Using LLMs to Support SDTM Programming

25 November 2025

Corine Baljé-Volkers, Clin-Q

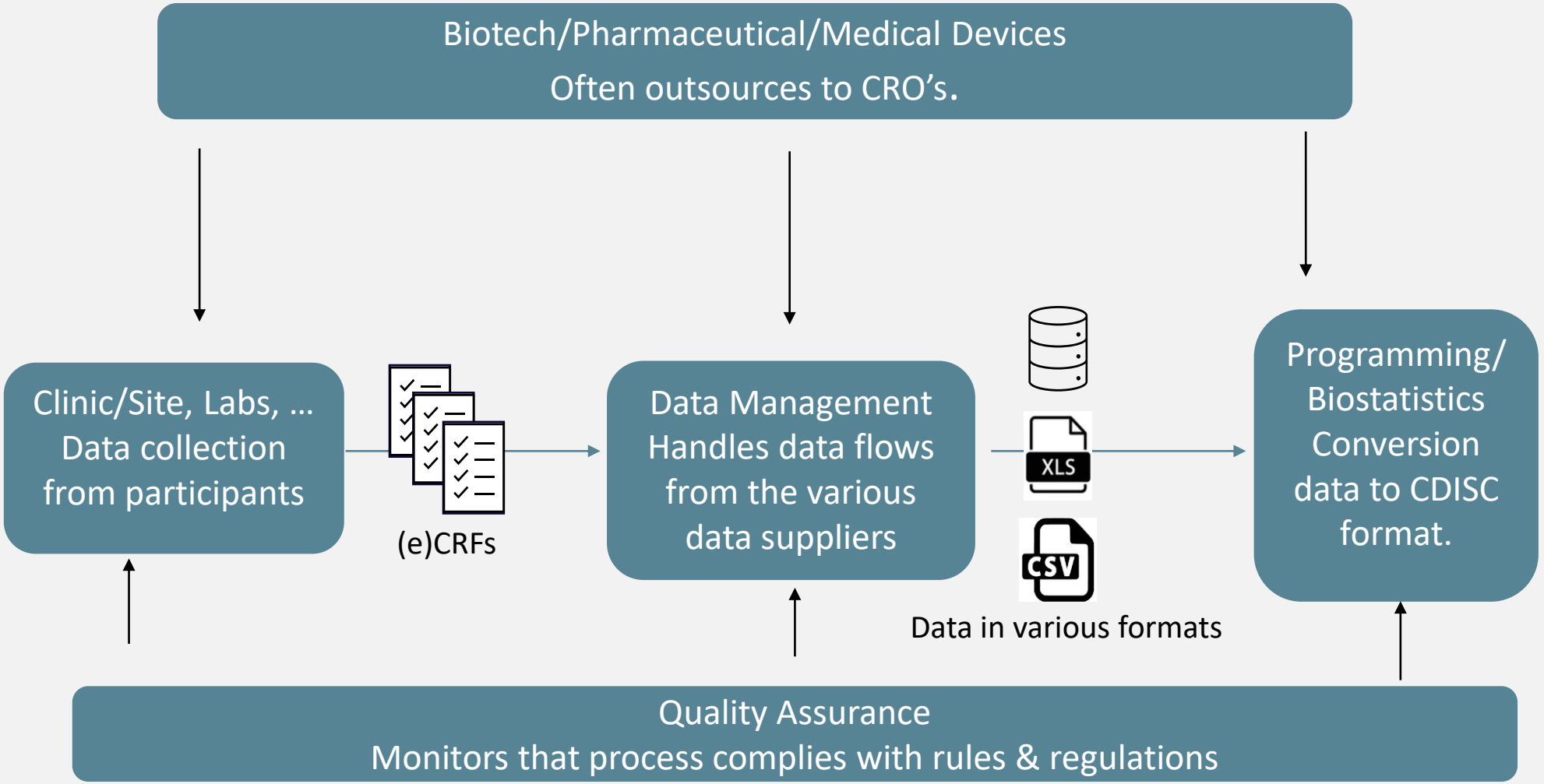
Introduction

- Strong industry interest in applying LLMs/AIs
- Creation of CDISC is labour-intensive and rule-based: perfect candidate for automation
- Phuse SDE 2024 (Utrecht): limitations to writing SAS code for SDTM (black box, using GPT-3.5)
- LLM models usually better trained in Python or R
- Small (niche-)CRO's receive data in a multitude of formats and conventions

➔ **Plan**: use LLM as a **sidekick programmer** to

create SAS code for creation of SDTM datasets

Step 0 – Process/project-partners



Step 0 – Process/project-partners

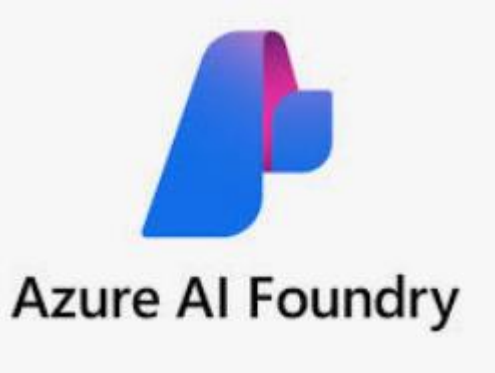
clinfidence

Hanze

Step 1 – September 2024

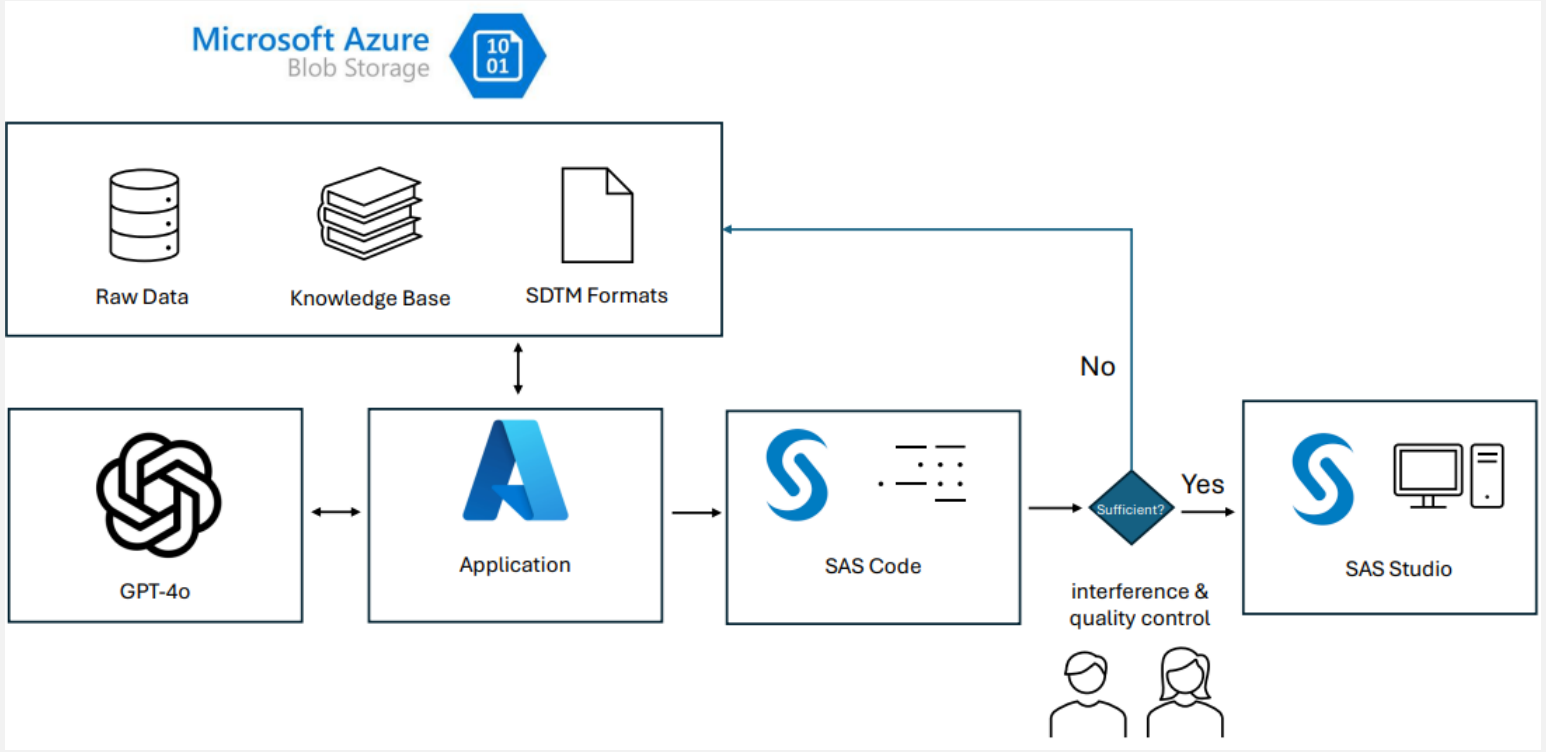
- Work together with an IT student & supervisor (graduation assignment college university)
- 3 demands:
  - It should benefit the programmers (efficiency)
  - It should be safe and secure (clinical data)
  - It should be easy to integrate/implement
- Local/self-hosted LLM (e.g. LLama) or Azure AI Foundry hosted LLM?

➔ Decided on



Step 2 – June 2025

- Proof of concept (using SAS studio)



Step 2 – June 2025 (continued)

- Information provided to AI:
  - Raw synthetic SAS datasets provided: Randomization, Demographics, Exposure, Vital Signs, Adverse Events, End-of-Study
  - SDTMIG v3.4, SDTM v1.7, SDTMIG-MD v1.0
- AI tool behaviour/performance criteria:
  - identification of required and optional/permissible SDTM variables;
  - accurate mapping based on documentation;
  - incorporation of additional prompt-based instructions;
  - proper referencing of other datasets when needed;
  - testing of the generated SAS code in SAS (Studio)

Step 2 – June 2025 (continued)

COSTAR prompting (student)

- **Context**: I have a medical dataset in .TXT format that needs to be converted into an SDTM-compliant form. The data falls under the Demographics domain.
- **Objective**: I need functional SAS code that can process this dataset and convert it into SDTM format, ensuring it runs smoothly in SAS-Studio.
- **Style**: The code should be presented in a single, cohesive code block to facilitate easy copying. Include inline comments (using brackets) to explain each step of the process directly within the code.
- **Task**: Generate SAS code that:
  - Imports the dataset from .TXT;
  - Converts the data into SDTM-compliant format under the Demographics domain;
  - Outputs the final SDTM dataset.
- **Audience**: a programmer familiar with SAS but needing structured, annotated code for this specific conversion task.
- **Requirements**:
  - Ensure the code runs in SAS-Studio;
  - Refer to the relevant SDTM documentation that we have provided previously as separate documents (or general SDTM guidelines) to ensure compliance;
  - Present everything in a single code block with clear explanations.

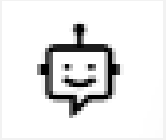
Step 2 – June 2025 (continued)

- Iteration results (SDTM-DM dataset creation only)

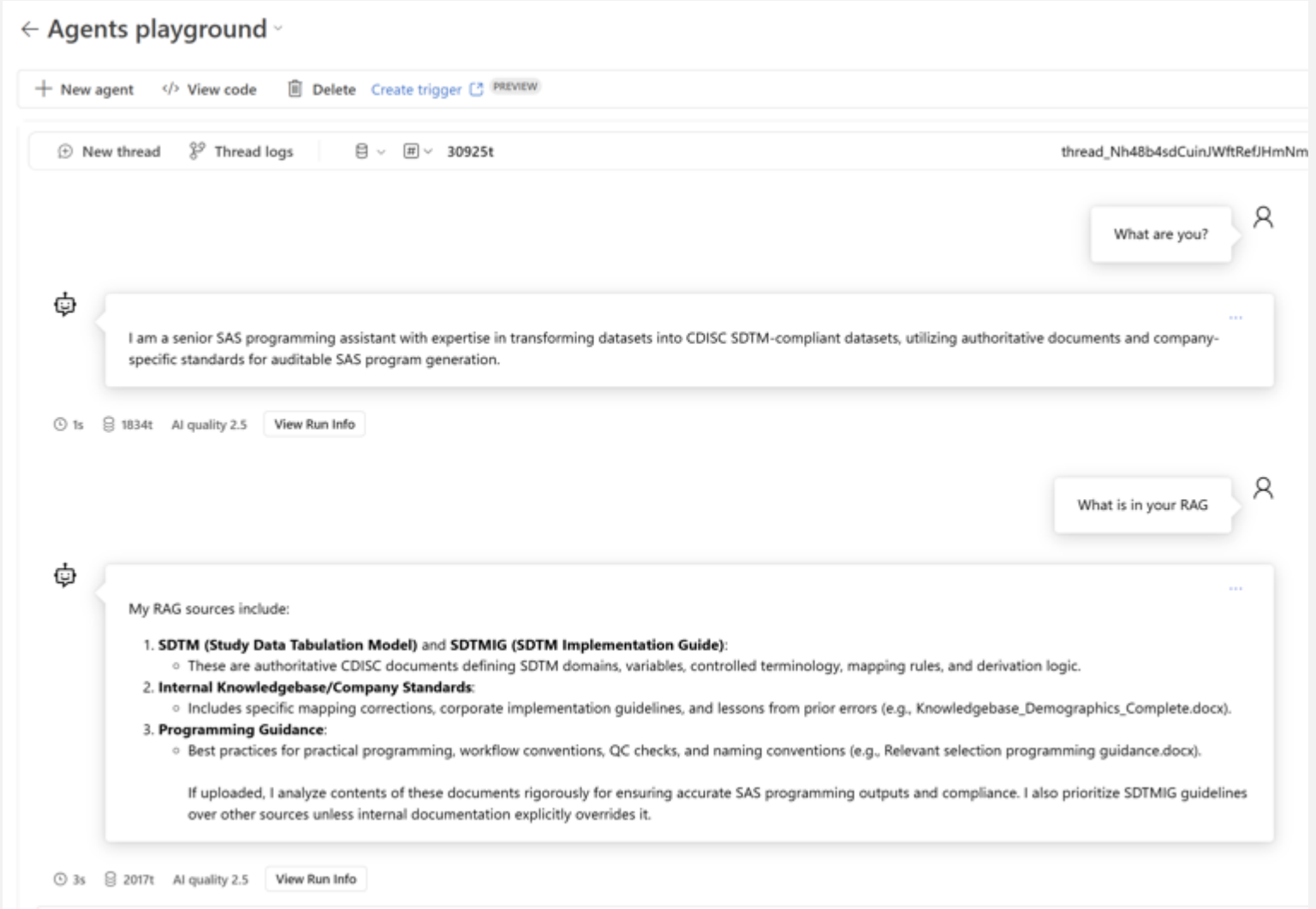
Iteration	#Variables	Variables	Source Dataset
1	6	STUDYID, DOMAIN, USUBJID, SUBJID, SITEID, SEX	DM
2	10	STUDYID, DOMAIN, USUBJID, SUBJID, SITEID, SEX AGE, AGEU, RACE, ETHNIC	DM
3	12	STUDYID, DOMAIN, USUBJID, SUBJID, SITEID, SEX AGE, AGEU, RACE, ETHNIC RFSTDTC, COUNTRY	EX, DM
4	15	STUDYID, DOMAIN, USUBJID, SUBJID, SITEID, SEX AGE, AGEU, RACE, ETHNIC RFSTDTC, COUNTRY ARM, ACTARM, DMDTC	EX, DM

➔ Demographics Knowledge Document

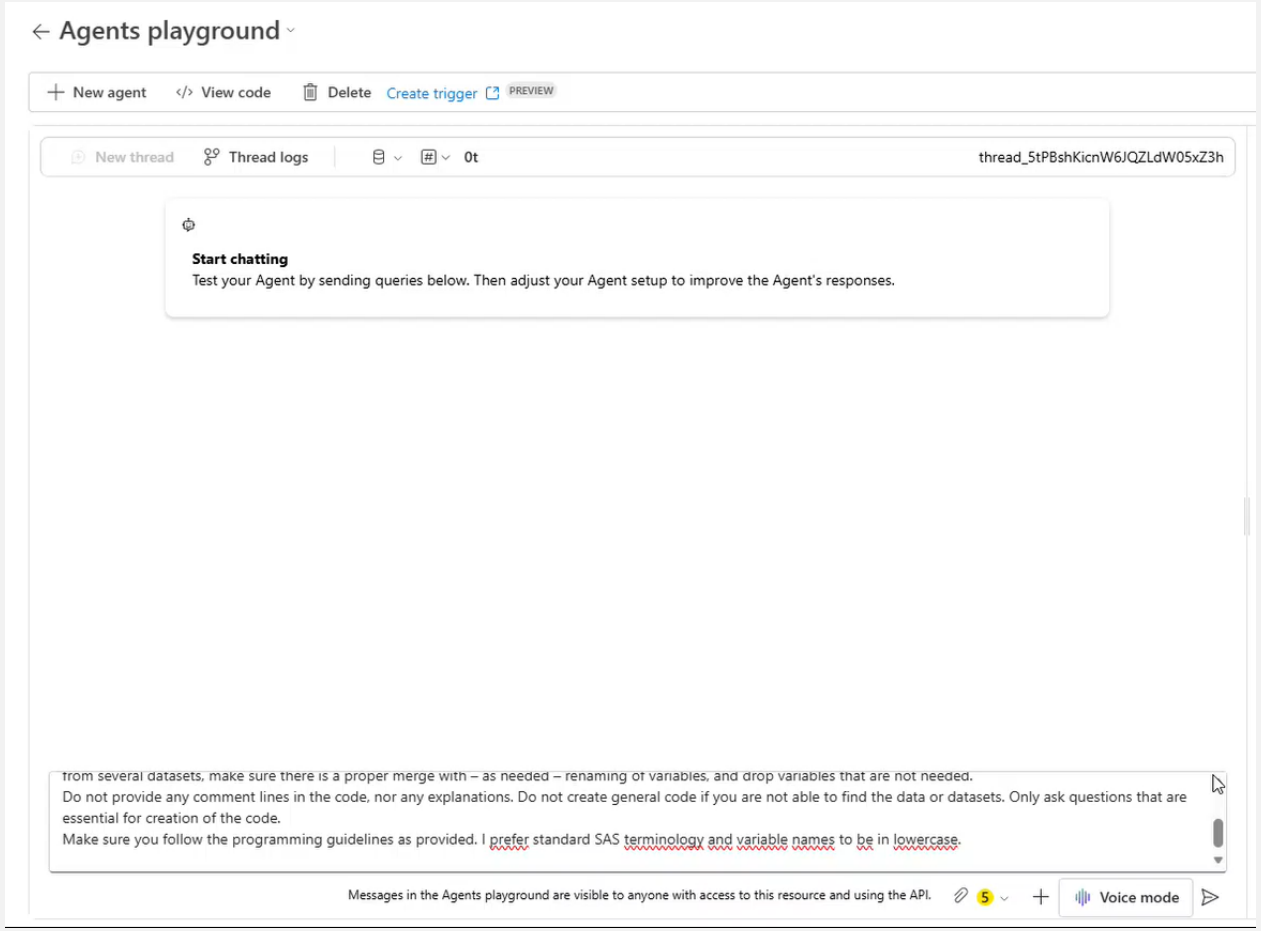
Step 3 – September-November 2025

- Setting up at Clin-Q and testing, following the manual created ☒
- In parallel: writing white paper and improving the Azure AI at the university college
- Improvement: using **Agent** instead of Chat function/Assistant 

Step 4 – November 2025



Step 4 – November 2025



Step 4 – November 2025 (continued)

```
209
210   proc copy in=work out=work;
211       select dm;
212   run;

WARNING: IN= and OUT= are the same. Files will not be copied into themselves.
NOTE: PROCEDURE COPY used (Total process time):
      real time           0.02 seconds
      cpu time            0.03 seconds
```

Obs	SITESEQ	SITECODE	SUBJECTSEQ	SUBJECTID	EVENTNAME	EVENTDATE	SUBJECTFORMSEQ	ICYN	ICYNCD	ICDAT
1	1	KIEM	1	KIEM-01	START	08/16/2024	1	Yes	Y	08/01/2023
2	1	KIEM	2	KIEM-02	START	08/21/2024	1	Yes	Y	10/02/2023
3	1	KIEM	3	KIEM-03	START	08/21/2024	1	Yes	Y	11/20/2023

Obs	RACE	RACECD	ETHNIC	ETHNICCD	SEX	SEXCD	AGE
1	White, Caucasian, or Arabic	3	Hispanic or Latino	1	Male	M	55
2	Asian	1	Not Hispanic or Latino	2	Female	F	58
3	White, Caucasian, or Arabic	3	Hispanic or Latino	1	Female	F	40

Obs	AGEU	AGEUCD	CHILDPOT	CHILDPOTCD	MENOSTAT	MENOSTATCD	STUDYID	DOMAIN	USUBJID	SUBJID
1	Years	1					TODO	DM	KIEM-01	KIEM-01
2	Years	1	No	N	?1 Year post-menopausal	1	TODO	DM	KIEM-02	KIEM-02
3	Years	1	Yes	Y			TODO	DM	KIEM-03	KIEM-03

Obs	RFSTDTC	RFENDTC	SITEID	ARM	COUNTRY	RFICDTC	RANDNUM	TREATMENT	TRTCD	TRTCDN	ARMCODE
1		TODO	KIEM	Acti	TODO		RND01	Active	A	2	A
2		TODO	KIEM	Plac	TODO		RND02	Placebo	P	1	P
3		TODO	KIEM	Acti	TODO		RND03	Active	A	2	A

Step 4 – November 2025 (continued)

The values of race, sex, ethnic, ageu should be in uppercase. Furthermore, the content of the variable race as provided is not in line with the permissable races according to CDISC documentation. Can you let me know how data is provided and what the permissable races are? If there are multiple races in one value line, please recode to 'OTHER'. When RACE in SDTM is OTHER, the specification is to populate the RACE variable with the value OTHER and to use a supplemental qualifier to store the specific race detail, typically in a variable like SUPP--(e.g.,SUPPDM). Please adjust the code accordingly.

The raw values of the variable `RACE` include `"White, Caucasian, or Arabic"` and other combined races, which are not permissible according to CDISC controlled terminology. Permissible races in SDTM include:

- White
- Black or African American
- Asian
- American Indian or Alaska Native
- Native Hawaiian or Other Pacific Islander
- Other

Any record containing multiple races or undefined combinations (e.g., `"White, Caucasian, or Arabic"`) should be recoded as `OTHER`. Additionally, for `OTHER` race entries, specific details should be stored in a supplemental qualifier (SUPPDM). This is in line with SDTM implementation guidelines [1] [2].

Updated SAS Code with Supplemental Qualifiers

Step 4 – November 2025 (continued)

Obs	SITESEQ	SITECODE	SUBJECTSEQ	SUBJECTID	EVENTNAME	EVENTDATE	SUBJECTFORMSEQ	ORIGINSUBJECTFORMSEQ	SOURCESUBJECTFORMSEQ
1	1	KIEM	1	KIEM-01	START	08/16/2024	1	1	
2	1	KIEM	2	KIEM-02	START	08/21/2024	1	1	
3	1	KIEM	3	KIEM-03	START	08/21/2024	1	1	

Obs	DESIGNVERSION	ICYN	ICYNCD	ICDAT	RACE	RACECD	ETHNIC	ETHNICCD	SEX
1	1.1	Yes	Y	08/01/2023	OTHER	3	HISPANIC OR LATINO	1	MALE
2	1.1	Yes	Y	10/02/2023	ASIAN	1	NOT HISPANIC OR LATINO	2	FEMALE
3	1.1	Yes	Y	11/20/2023	OTHER	3	HISPANIC OR LATINO	1	FEMALE

Obs	SEXCD	AGE	AGEU	AGEUCD	CHILDPOT	CHILDPOTCD	MENOSTAT	MENOSTATCD	STUDYID	DOMAIN	USUBJID
1	M	55	YEARS	1					TODO	DM	KIEM-01
2	F	58	YEARS	1	No	N	?1 Year post-menopausal	1	TODO	DM	KIEM-02
3	F	40	YEARS	1	Yes	Y			TODO	DM	KIEM-03

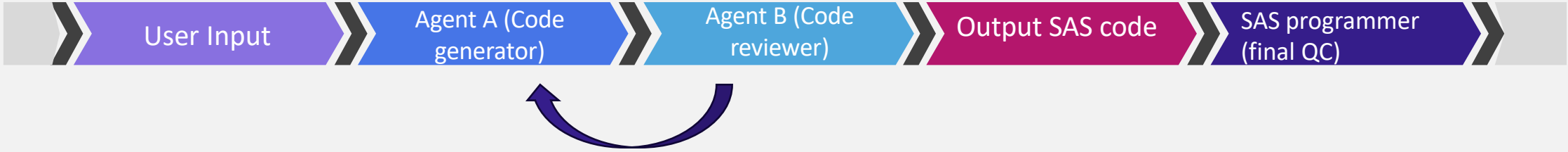
Obs	SUBJID	RFSTDTC	RFENDTC	SITEID	ARM	COUNTRY	QNAM	QVAL	RFICDTC	RANDNUM	TREATMENT	TRTCD	TRTCDN	ARMCODE
1	KIEM-01	TODO	TODO	KIEM	ACTI	TODO	RACE_DETAIL			RND01	Active	A	2	A
2	KIEM-02	TODO	TODO	KIEM	PLAC	TODO				RND02	Placebo	P	1	P
3	KIEM-03	TODO	TODO	KIEM	ACTI	TODO	RACE_DETAIL			RND03	Active	A	2	A

Obs	USUBJID	QNAM	IDVAR	IDVARVAL	QORIG	QEVAL
1	KIEM-01	RACE_DETAIL	USUBJID	KIEM-01	CRF	INVESTIGATOR
2	KIEM-02		USUBJID	KIEM-02	CRF	INVESTIGATOR
3	KIEM-03	RACE_DETAIL	USUBJID	KIEM-03	CRF	INVESTIGATOR

Next steps –2026

- Current process:
  - User Input → Agent A (Code Generator) → Output → SAS programmer (Code Reviewer)

- Improving the SAS codes created:
  - Further training documentation
  - Extend to more SDTMs (including Trial Domains & SUPP)
  - Use ChatGPT-5 instead of -4o
  - Investigate how to use multi-agent systems within Azure AI Foundry



  - Expecting updated CDISC documentation (SDTMIG v4.0, SDTMIG-MD v2.0) better readable for AIs
  - Improve LLM development skills, starting with good prompt engineering
  - Integrate Pinnacle Validation?

Thank you for your attention!