

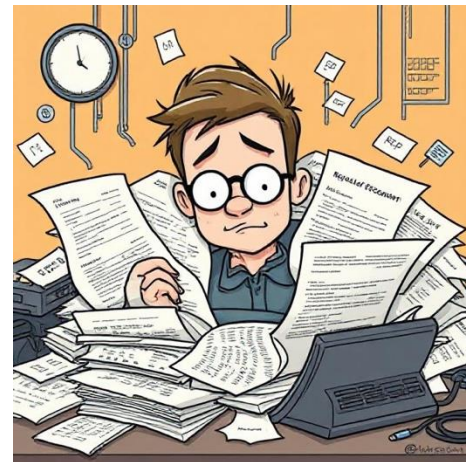
Natural Language to Code: AI-Powered Automation for CDISC-Compliant Datasets

Vina Ro, Eli Lilly and Company

The Challenges of creating CDISC Datasets

Programming CDISC datasets from scratch could lead to these problems...

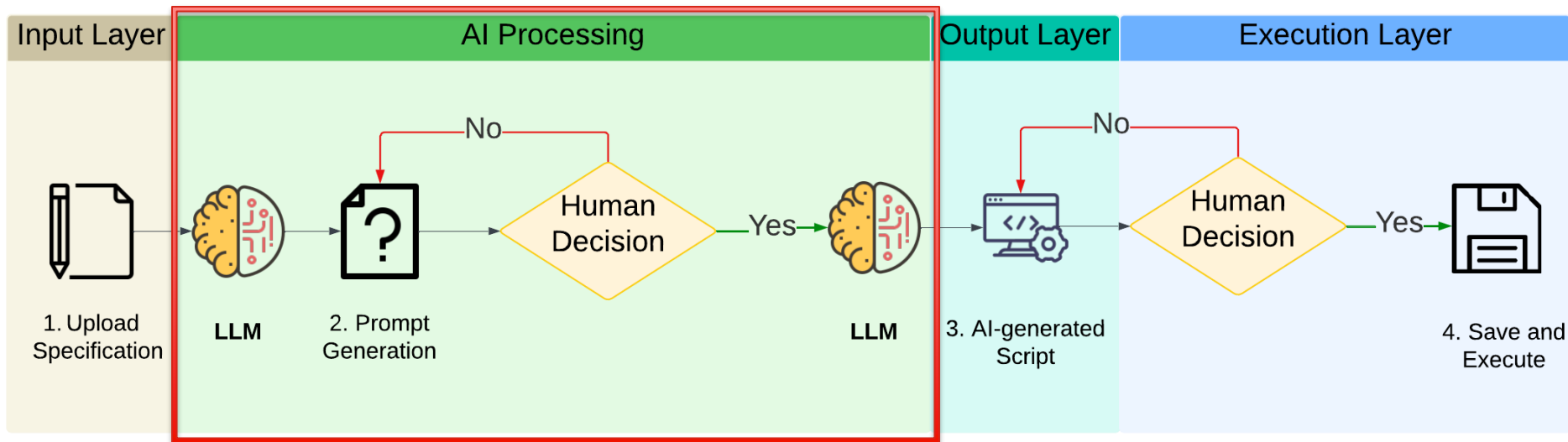
- ◆  Manual and time-consuming
- ◆  Highly dependent on expertise
- ◆  Redundant effort across studies
- ◆  Inconsistencies and errors



 *What if we could automate this process using LLMs while ensuring accuracy and consistency?*

Application Overview

◆ Study Data Tabulation Model (SDTM)



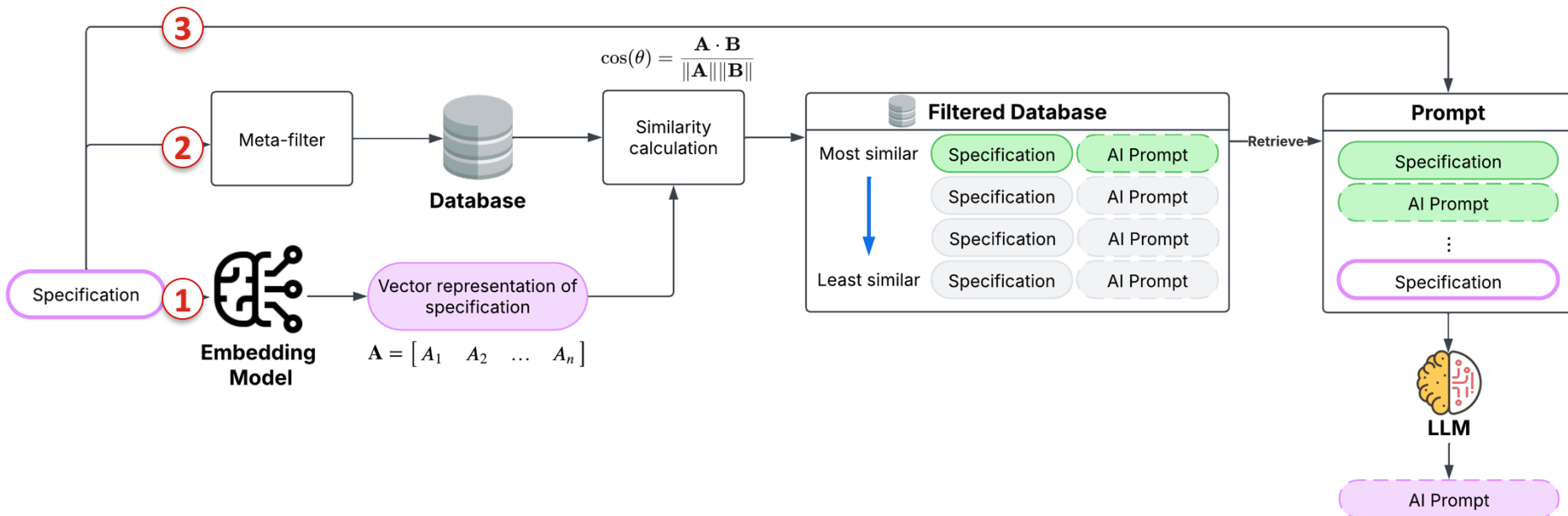
Step 1: Database Creation

- ◆ Records are gathered across all SDTM domains across all therapeutic areas
- ◆ Example entry:

Dataset	Variable	Specification	AI Prompt
DM	DTHDTC	Pull SUBJID, DTHDAT from df_disposition. Convert to ISO8601 format. Merge onto demographics data by SUBJID. Can be null (if no death for a subject).	1. Pull SUBJID, DTHDAT from df_disposition where DTHDAT is not null. Convert DTHDAT into a date object. Name result as df_disposition_subset. 2. Do a left join with df_demographics on the left and df_disposition_subset by SUBJID. 3. Create DTHDTC column in df_demographics and assign values as: DTHDAT formatted to ISO8601 with format_ISO8601().

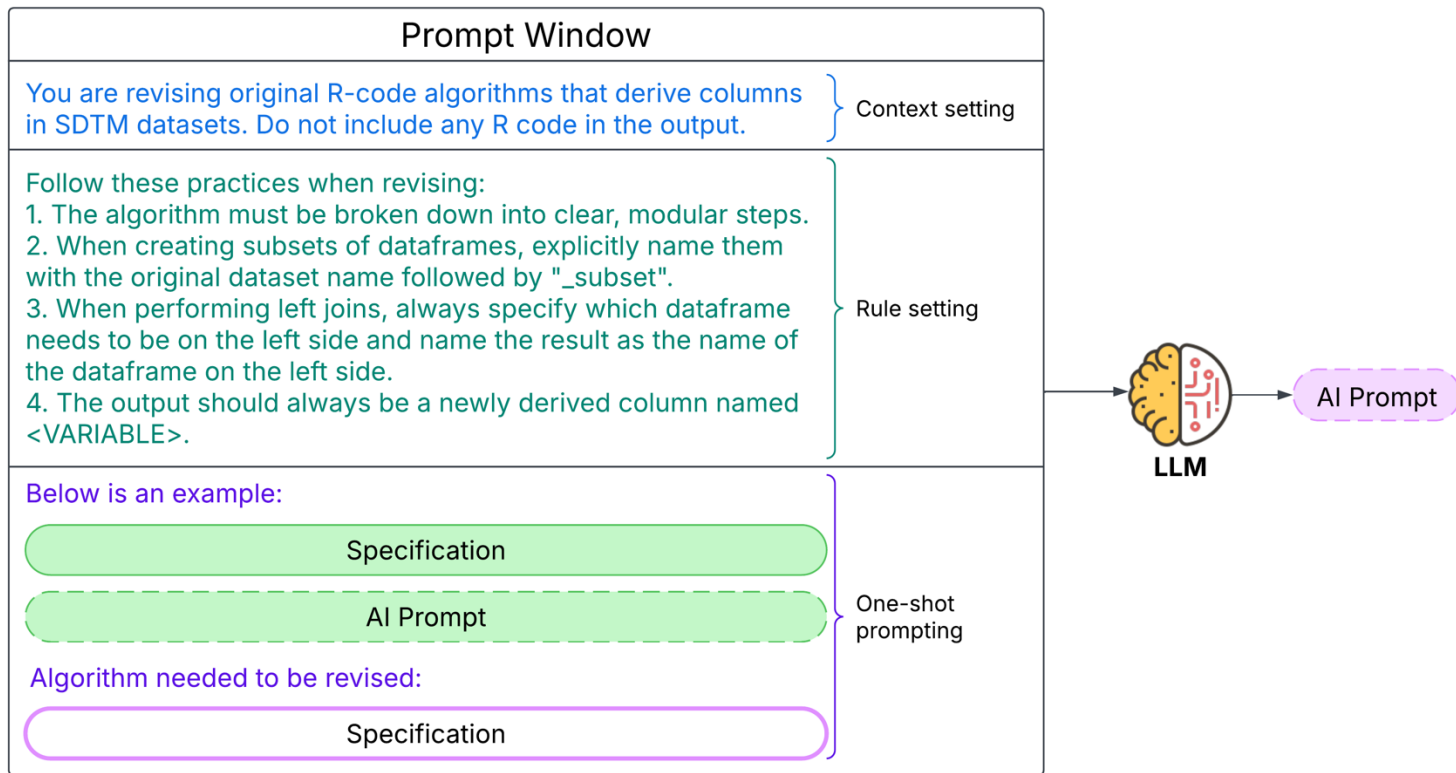
Step 2: Specification-Aligned Matching

- ◆ A method developed by Lilly, inspired by the RAG framework



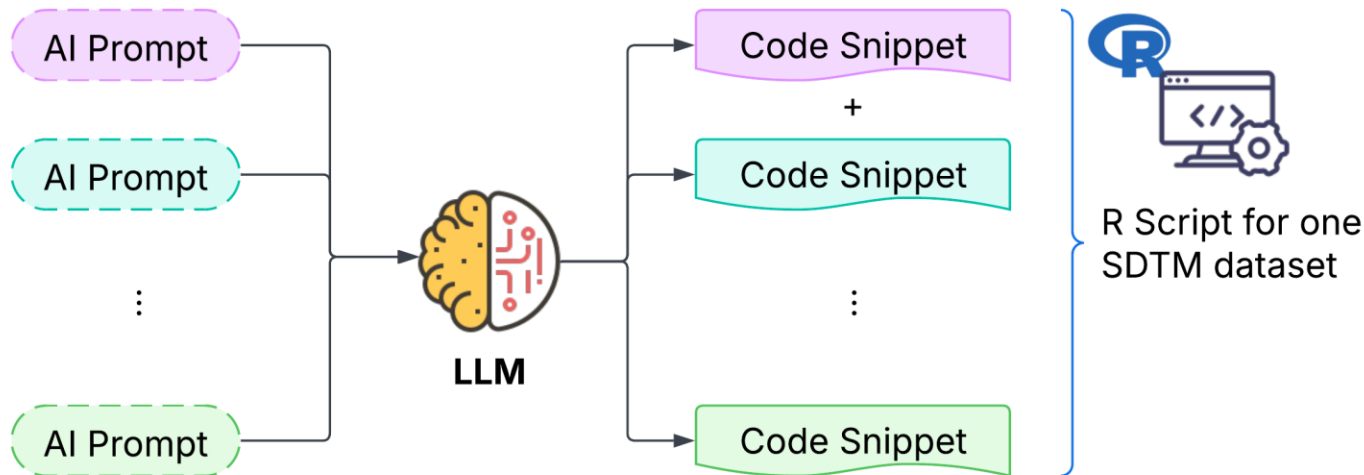
Step 3: Prompt Engineering

◆ Prompt for getting corresponding AI prompt from the input specification



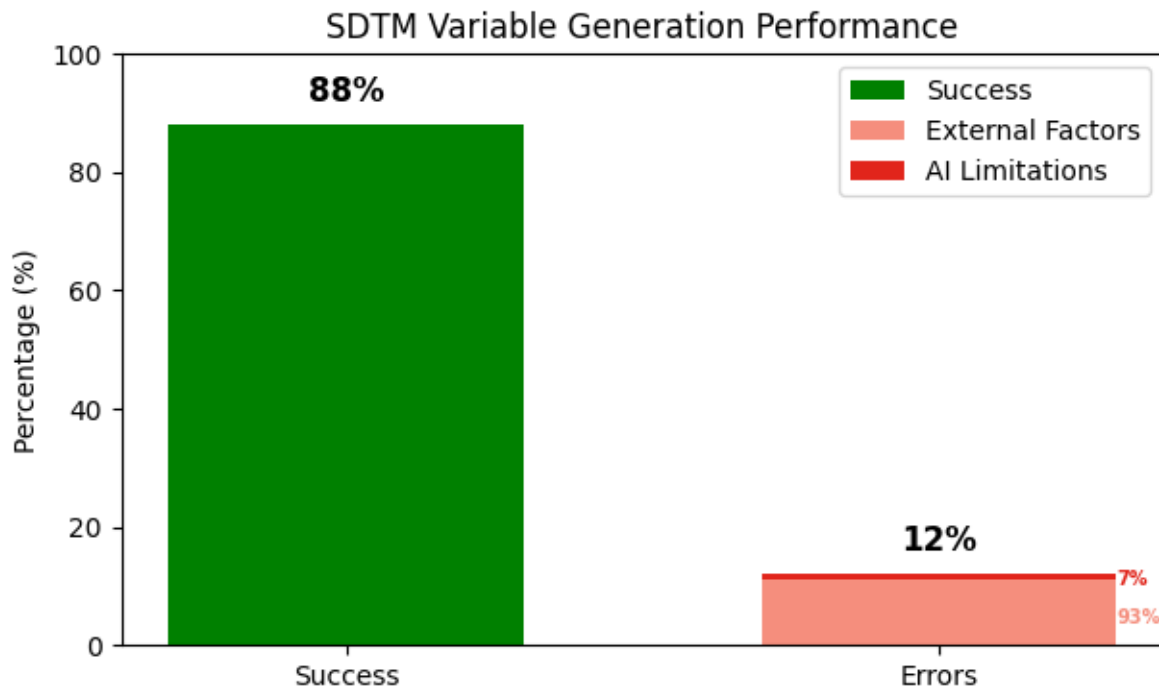
Step 4: Code Generation

◆ Variable-level operation



Results: Accuracy

- ◆ Tested on all SDTM datasets in an ongoing study



Conclusion and Future Direction



Key Highlights

- ◆ Leverages SAM method and LLMs to automate the creation of SDTM datasets with high accuracy
- ◆ Expansion to generate ADaM datasets and Tables, Figures, and Listings (TFLs)



Limitations

- ◆ Performance relies heavily on the quality of the input data
- ◆ Periodic maintenance and refinements are needed for the database

Acknowledgements



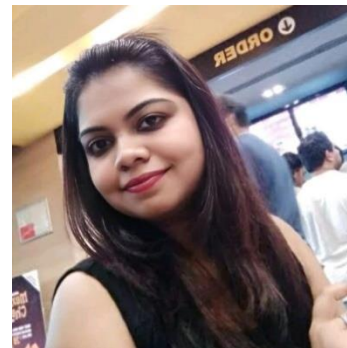
Sathish
Sundaram



Ravi Gupta



Ian Sturdy



Soma Bhadra