Maximizing LLM Potential: How Data Structure and Context Drive Quality Insights

Emily Yates
AD, Data Management and Analytics

GenAI in Pharma
August 19, 2025

Formation Bio

The democratization & open-source challenge with AI in pharma

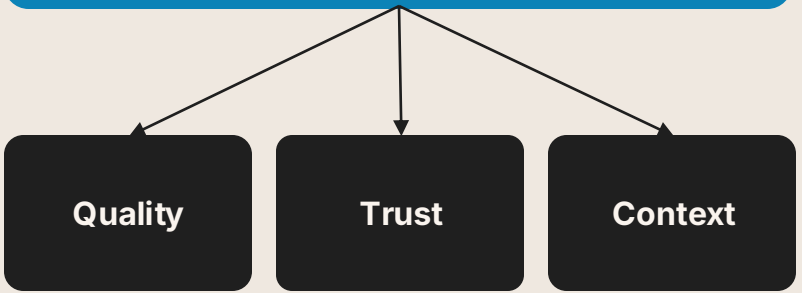
THE PROMISE

Anyone can make a chatbot

General Data Analytics

THE CHALLENGE

How can we make good AI solutions for pharma?



Clinical Data Analysis = Data + Clinical Context

LLM GENERATED R CODE

AI Assistant

Ask questions about the AE and DM datasets. The AI can help you filter, join, and visualize the data.

Show me adverse events by treatment group

✓ Code generated and executed. Check the tabs for results.

USER ENTERED QUERY

show me adverse events by treatment group

Send

Tip: Press Ctrl+Enter to send

Results | DM Dataset | AE Dataset | Generated Code

Data Output

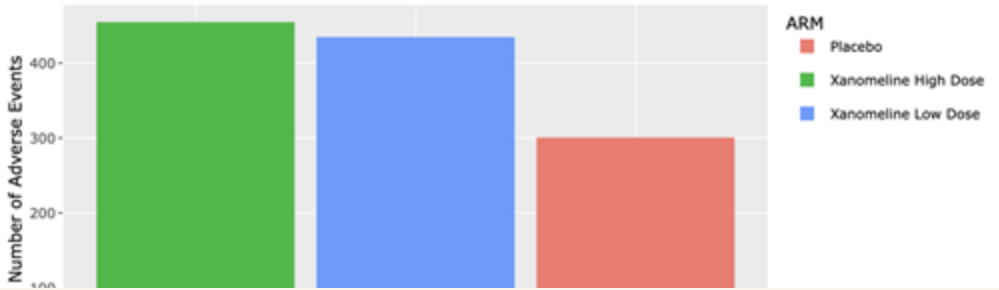
Show 10 entries Search:

	ARM	total_ae_count
1	Xanomeline High Dose	455
2	Xanomeline Low Dose	435
3	Placebo	301

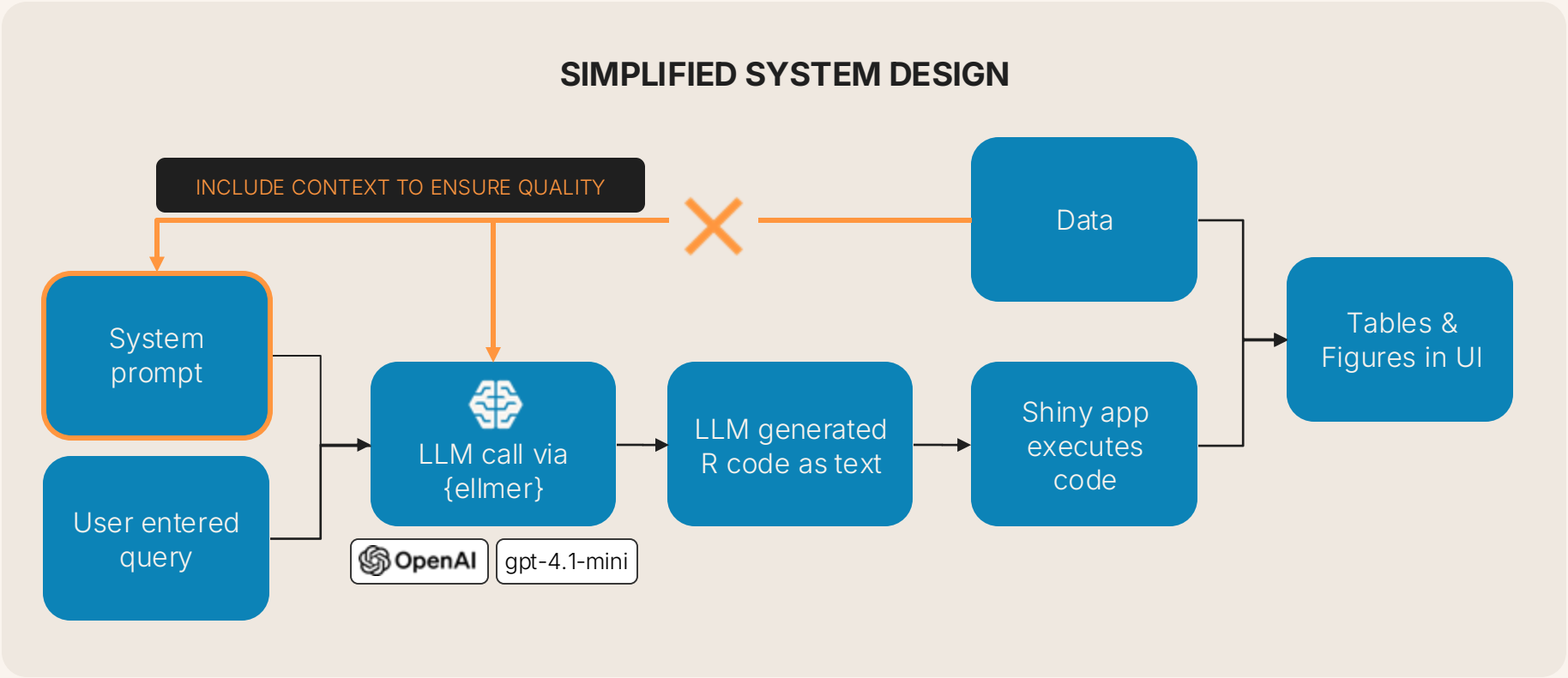
Showing 1 to 3 of 3 entries Previous 1 Next

Visualization

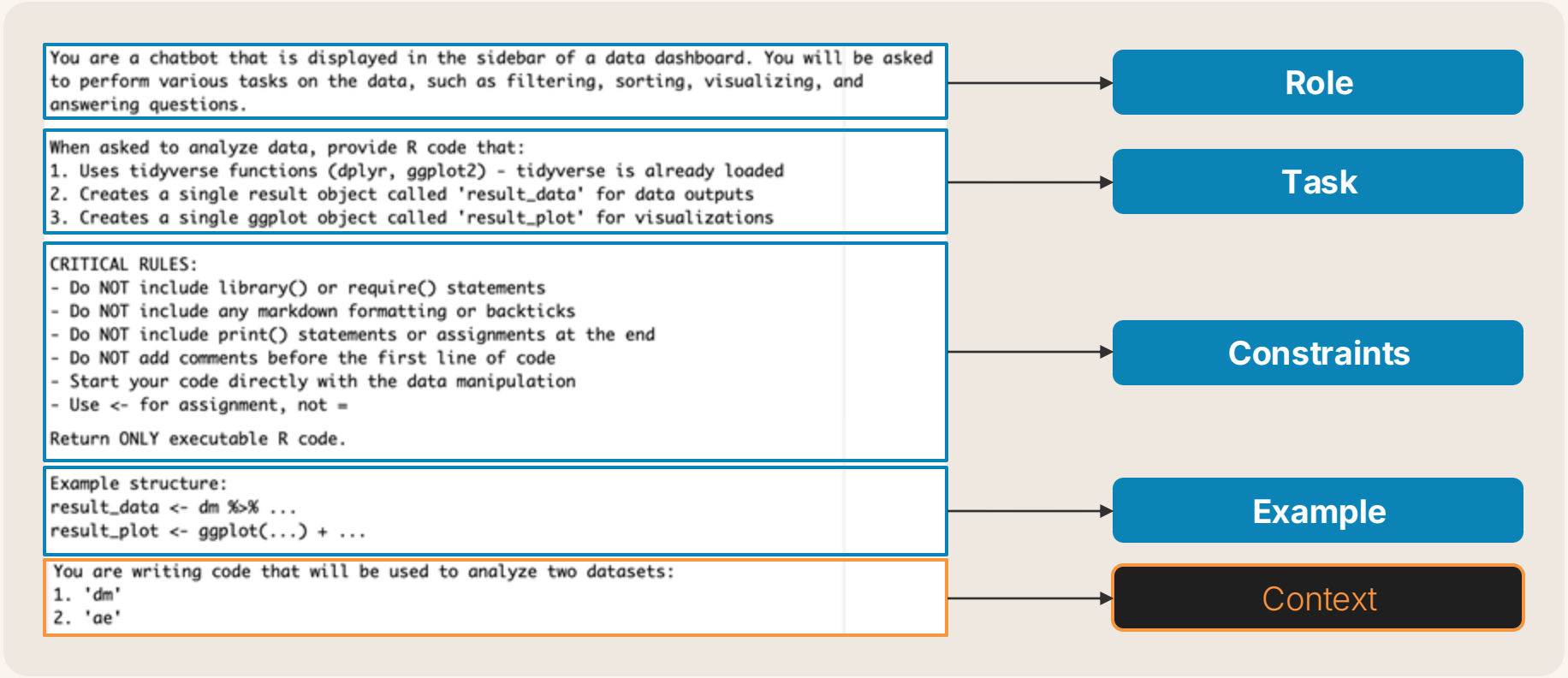
TABLES & FIGURES IN UI



The system prompt is critical since the LLM does not access the data



Iterate on the system prompt's context to increase quality of the output



We will focus on two types of context in the system prompt:

DATA CONTEXT

Code is **technically correct**

Field-level metadata

Schema relationships

Data dictionaries

Missing data pattern

DOMAIN CONTEXT

Analysis is **clinically meaningful**

Semantic mapping

Domain ontologies

Subject matter expertise

Concept harmonization

LLMs will guess without proper context in the system prompt

Iterate: Add more data context to the system prompt

Adding field-level metadata to the system prompt improves LLM output quality by enabling context-aware decision making

Field-level metadata:

- Data type & format
- Valid values & ranges
- Cdisc labels & codelists

BEFORE

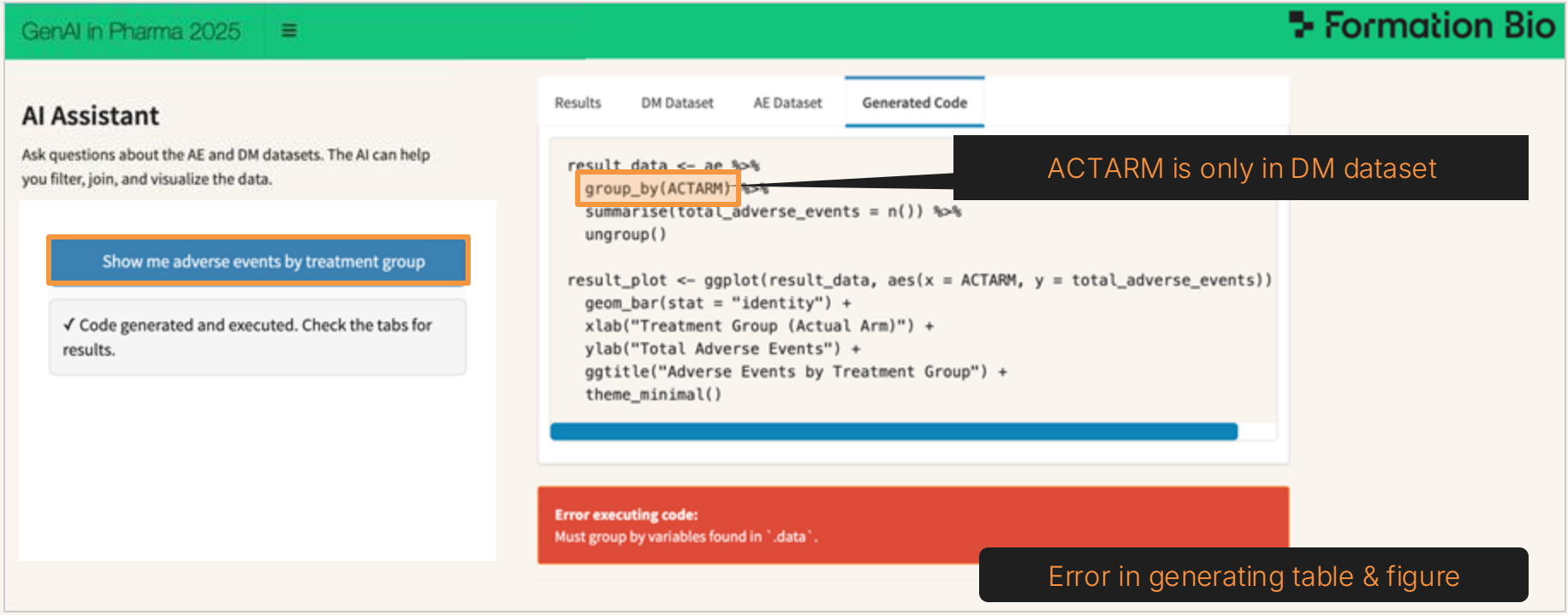
```
You are writing code that will be used to analyze two datasets:
1. 'dm'
2. 'ae'
```

AFTER

```
### 'ae' - Adverse Events
#### Columns:
- **STUDYID** - Study Identifier: Unique identifier for a study.
- **DOMAIN** - Domain Abbreviation: Two-character abbreviation for the domain.
- **USUBJID** - Unique Subject Identifier: Identifier used to uniquely identify a subject across all
studies for all applications or submissions involving the product.
```

```
- **AESEV** - Severity/Intensity: The severity or intensity of the event. Examples: "MILD", "MODERATE",
"SEVERE".
- **AESER** - Serious Event: Is this a serious event? Valid values are "Y" and "N".
```

LLMs need a lot of context to answer even simple questions



Iterate: Add schema relationship to system prompt

Adding schema relationships to the system prompt improves LLM output quality by enabling structurally correct data operations

Schema relationships:

- Table cardinality
- Primary/foreign key
- Aggregation guidelines

BEFORE

```
You are writing code that will be used to analyze two datasets:
1. 'dm'
2. 'ae'
```

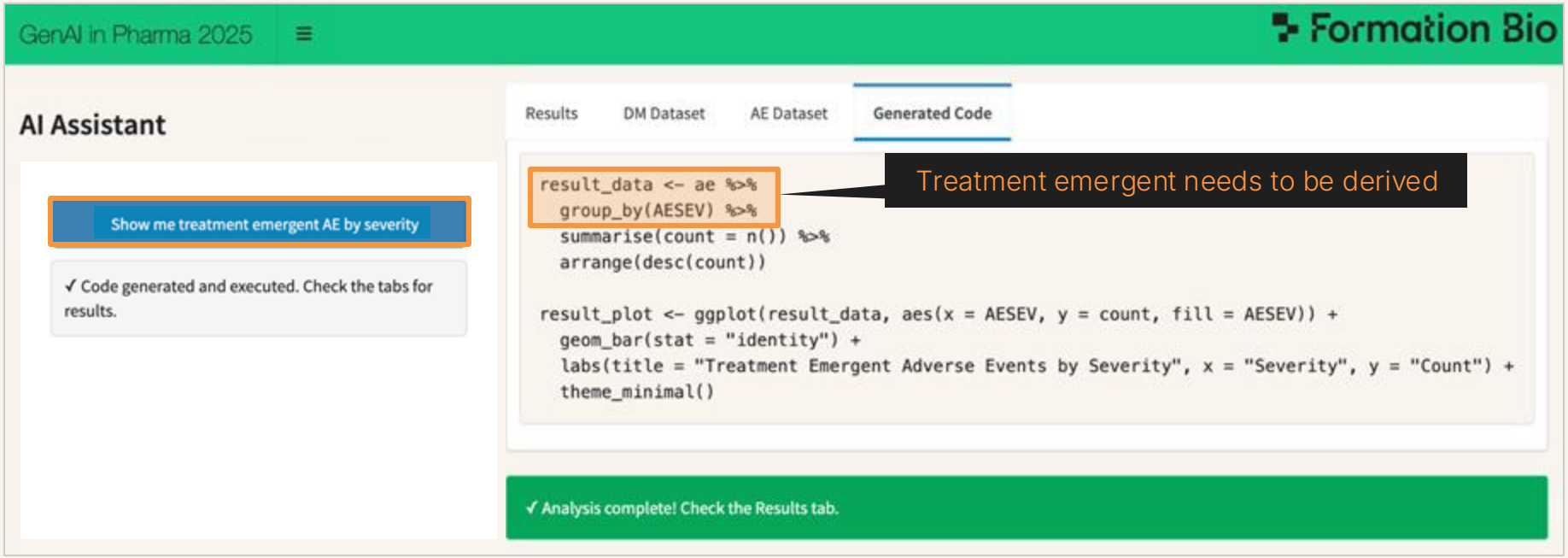
```
### 'ae' - Adverse Events
#### Columns:
- **STUDYID** - Study Identifier: Unique identifier for a study.
- **DOMAIN** - Domain Abbreviation: Two-character abbreviation for the domain.
- **USUBJID** - Unique Subject Identifier: Identifier used to uniquely identify a subject across all
studies for all applications or submissions involving the product.
```

AFTER

```
## Available Datasets
1. **'dm' - Demographics** - One record per subject
   - Key joining column: `USUBJID`

2. **'ae' - Adverse Events** - One record per adverse event per subject
   - Key joining column: `USUBJID`
```

LLM struggles to derive fields without sufficient domain context in the system prompt



Iterate: Add domain context to the system prompt

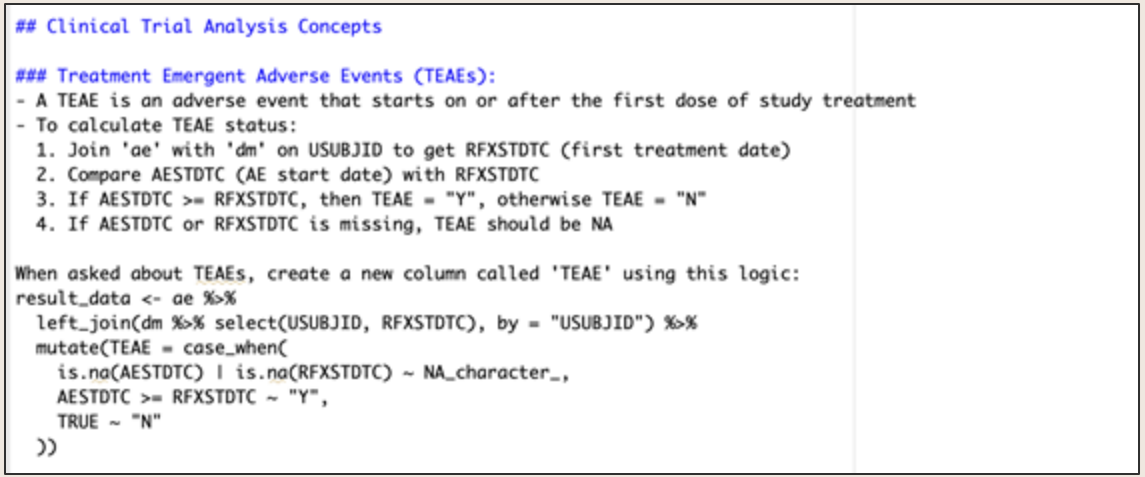
Adding domain context to the system prompt improves LLM output quality by enabling clinically meaningful analysis

Domain context:

- Clinical SME knowledge
- Semantic data mapping
- Derivation logic

BEFORE: NO CLINICAL CONTEXT

AFTER

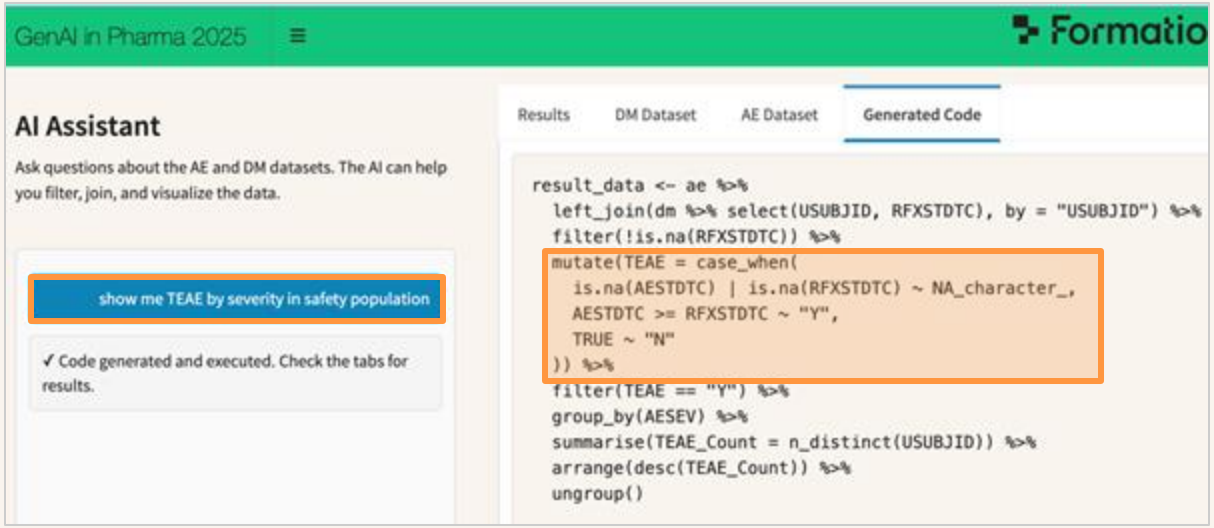


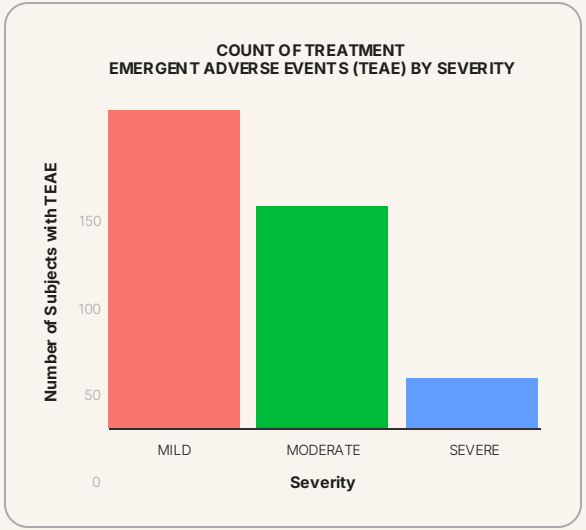
```
## Clinical Trial Analysis Concepts

### Treatment Emergent Adverse Events (TEAEs):
- A TEAE is an adverse event that starts on or after the first dose of study treatment
- To calculate TEAE status:
  1. Join 'ae' with 'dm' on USUBJID to get RFXSTDTC (first treatment date)
  2. Compare AESTDTC (AE start date) with RFXSTDTC
  3. If AESTDTC >= RFXSTDTC, then TEAE = "Y", otherwise TEAE = "N"
  4. If AESTDTC or RFXSTDTC is missing, TEAE should be NA

When asked about TEAEs, create a new column called 'TEAE' using this logic:
result_data <- ae %>%
  left_join(dm %>% select(USUBJID, RFXSTDTC), by = "USUBJID") %>%
  mutate(TEAE = case_when(
    is.na(AESTDTC) | is.na(RFXSTDTC) ~ NA_character_,
    AESTDTC >= RFXSTDTC ~ "Y",
    TRUE ~ "N"
  ))
```

Domain context in the system prompt allows for clinically informed analysis of complex queries

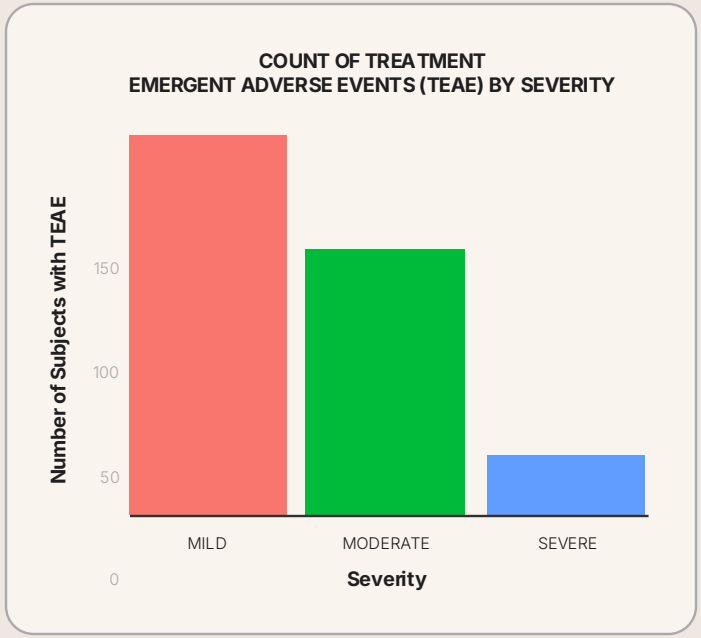




Iterate: Test flexibility with natural language variations of the query

Domain context allows for more consistent results from inconsistent & imperfect human queries

- show me treatment emergent adverse event by severity
- Can you summarize on-treatment AEs by mild/moderate/severe
- What's the severity distribution of treatment-emergent events
- What's the severity of on-treatment side effects
- Show me TEAE by severity
- Generate TEAE summary by severity grade



LLMs need data + domain context to deliver quality output

KEY TAKEAWAYS

Clinical Data Analysis = data + clinical knowledge

Fewer LLM guesses = Higher quality

Iterate and test your system prompts!

+ AI!

Questions?

eyates@formation.bio

Resources

- Ellmer
- Shinychat
- chatlas
- Claude code
- Guide to prompt design
- LLM powered shiny apps with ellmer and chatlas - Carson Sievert
- Shiny & LLMs: Landscape and Applications in Pharma - Phil Bowsher