

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

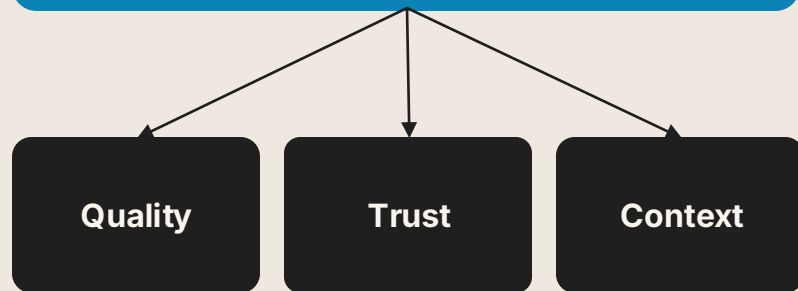
Anyone can make a chatbot



General Data Analytics

THE CHALLENGE

How can we make good AI solutions for pharma?



Clinical Data Analysis = Data + Clinical Context

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:

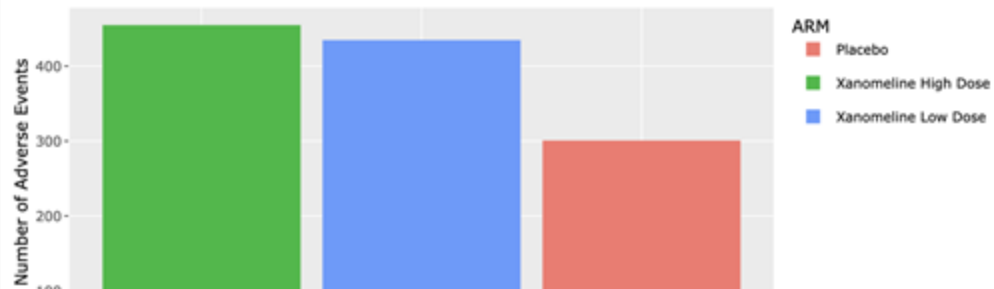
	ARM	total_ae_count
1	Xanomeline High Dose	455
2	Xanomeline Low Dose	435
3	Placebo	301

Showing 1 to 3 of 3 entries

Previous 1 Next

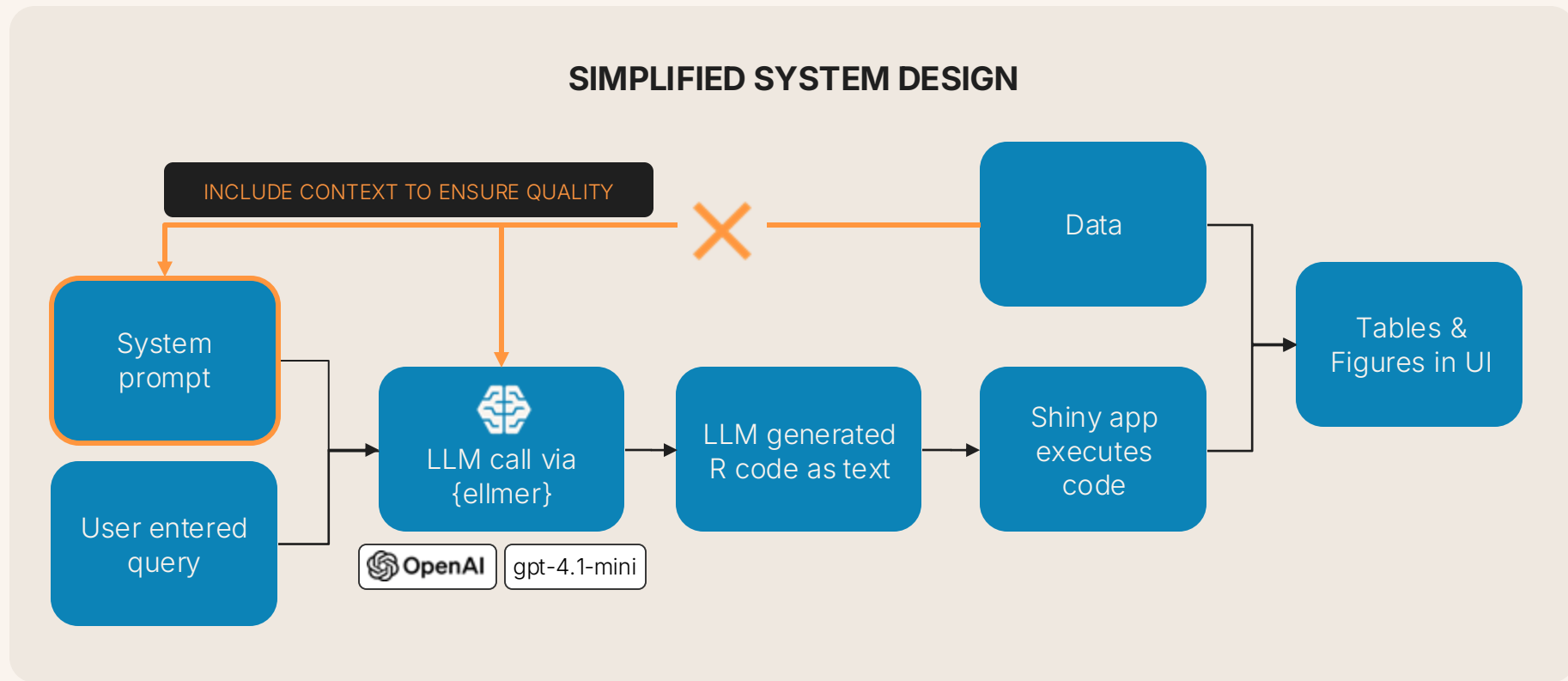
Visualization

Total Adverse Events by Treatment Group



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

You are a chatbot that is displayed in the sidebar of a data dashboard. You will be asked to perform various tasks on the data, such as filtering, sorting, visualizing, and answering questions.

Role

When asked to analyze data, provide R code that:

1. Uses tidyverse functions (dplyr, ggplot2) - tidyverse is already loaded
2. Creates a single result object called 'result_data' for data outputs
3. Creates a single ggplot object called 'result_plot' for visualizations

Task

CRITICAL RULES:

- Do NOT include library() or require() statements
- Do NOT include any markdown formatting or backticks
- Do NOT include print() statements or assignments at the end
- Do NOT add comments before the first line of code
- Start your code directly with the data manipulation
- Use <- for assignment, not =

Constraints

Return ONLY executable R code.

Example structure:

```
result_data <- dm %>% ...  
result_plot <- ggplot(...) + ...
```

Example

You are writing code that will be used to analyze two datasets:

1. 'dm'
2. 'ae'

Context

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

GenAI in Pharma 2025

≡

Formation Bio

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.

ResultsDM DatasetAE DatasetGenerated Code

```
result_data <- ae %>%
  group_by(treatment_group) %>%
  summarize(adverse_event_count = n())

result_plot <- ggplot(result_data, aes(x = treatment_group, y = adverse_event_count)) +
  geom_bar(stat = "identity") +
  labs(title = "Adverse Events by Treatment Group", x = "Treatment Group", y = "Number of Adverse Events") +
  theme_minimal()
```

THESE COLUMNS DON'T EXIST IN AE

Error executing code:
Must group by variables found in '.data'.

ERROR IN GENERATING TABLE & FIGURE

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

GenAI in Pharma 2025

Formation Bio

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.

ResultsDM DatasetAE DatasetGenerated Code

```
result_data <- ae %>%  
  group_by(ACTARM) %>%  
  summarise(total_adverse_events = n()) %>%  
  ungroup()  
  
result_plot <- ggplot(result_data, aes(x = ACTARM, y = total_adverse_events))  
  geom_bar(stat = "identity") +  
  xlab("Treatment Group (Actual Arm)") +  
  ylab("Total Adverse Events") +  
  ggtitle("Adverse Events by Treatment Group") +  
  theme_minimal()
```

Error executing code:
Must group by variables found in '.data'.

ACTARM is only in DM dataset

Error in generating table & figure

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

GenAI in Pharma 2025

Formation Bio

AI Assistant

Show me treatment emergent AE by severity

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

ResultsDM DatasetAE DatasetGenerated Code

```
result_data <- ae %>%  
  group_by(AESEV) %>%  
  summarise(count = n()) %>%  
  arrange(desc(count))  
  
result_plot <- ggplot(result_data, aes(x = AESEV, y = count, fill = AESEV)) +  
  geom_bar(stat = "identity") +  
  labs(title = "Treatment Emergent Adverse Events by Severity", x = "Severity", y = "Count") +  
  theme_minimal()
```

✓ Analysis complete! Check the Results tab.

Treatment emergent needs to be derived

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

GenAI in Pharma 2025

Formatio

AI Assistant

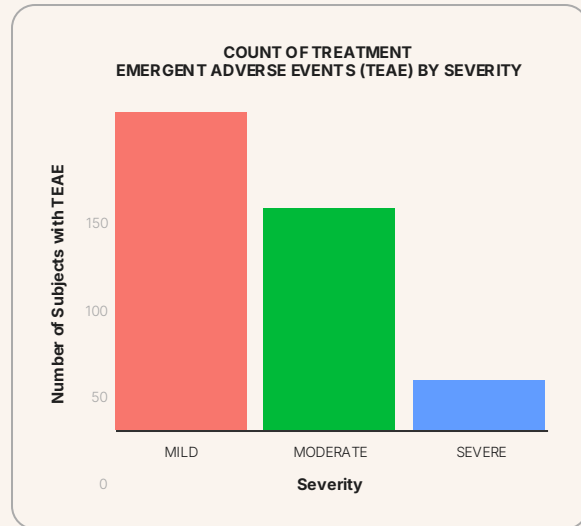
Ask questions about the AE and DM datasets. The AI can help you filter, join, and visualize the data.

show me TEAE by severity in safety population

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

ResultsDM DatasetAE DatasetGenerated Code

```
result_data <- ae %>%  
  left_join(dm %>% select(USUBJID, RFXSTDTC), by = "USUBJID") %>%  
  filter(!is.na(RFXSTDTC)) %>%  
  mutate(TEAE = case_when(  
    is.na(AESTDTC) | is.na(RFXSTDTC) ~ NA_character_,  
    AESTDTC >= RFXSTDTC ~ "Y",  
    TRUE ~ "N"  
  )) %>%  
  filter(TEAE == "Y") %>%  
  group_by(AESEV) %>%  
  summarise(TEAE_Count = n_distinct(USUBJID)) %>%  
  arrange(desc(TEAE_Count)) %>%  
  ungroup()
```



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

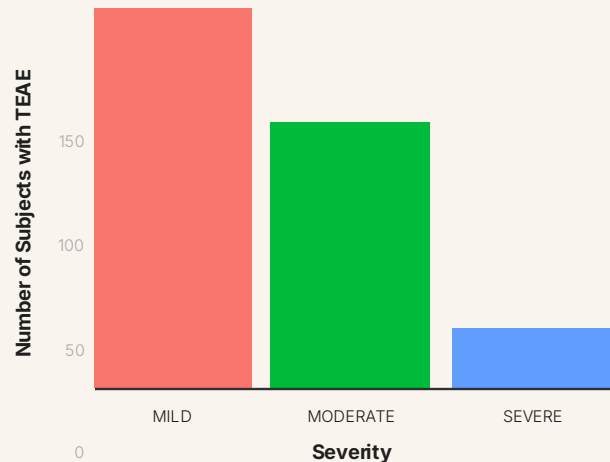
What's the severity distribution of treatment-emergent events

What's the severity of on-treatment side effects

Show me TEAE by severity

Generate TEAE summary by severity grade

COUNT OF TREATMENT
EMERGENT ADVERSE EVENTS (TEAE) BY SEVERITY



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