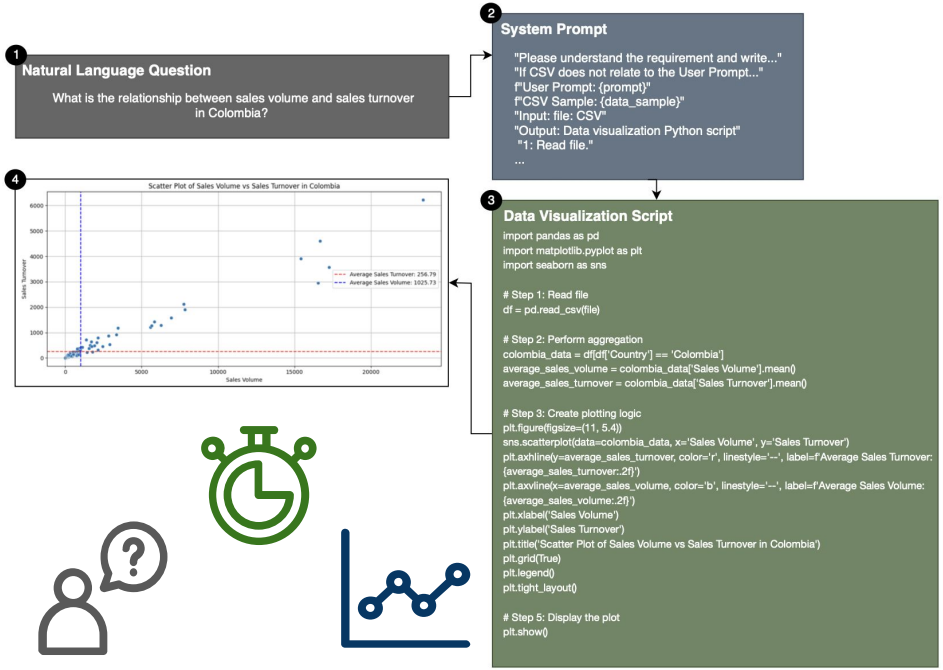
DataViz: A Business Visualization System using LLMs

Maja Binkowska
@VU Amsterdam
binkowskamajalena@gmail.com

1st supervisor: Alexandru Iosup
2nd supervisor: Xiaoyu Chu
2nd reader: Jiahuan Pei



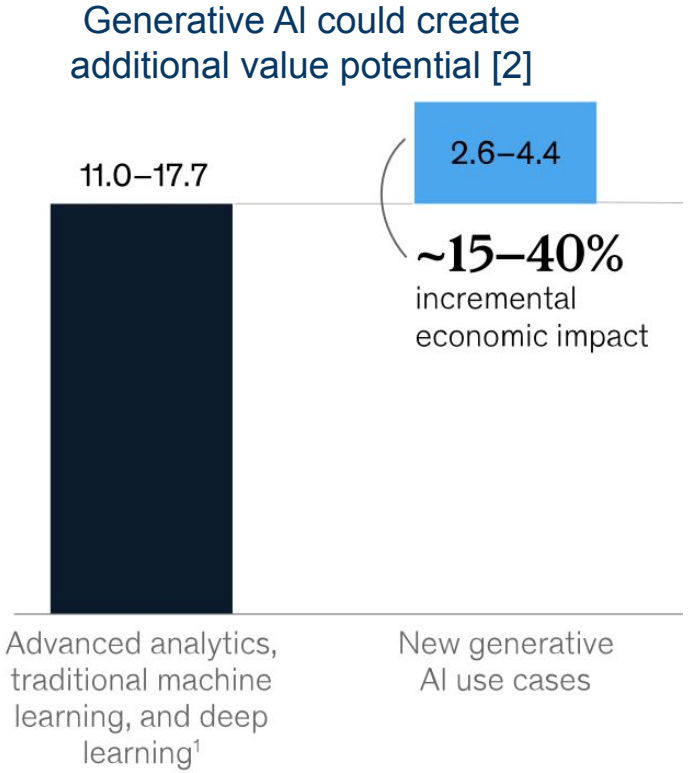
Context

Need for handling rapidly growing data

- Unilever's digital infrastructure handles **~240 TB of data** a week, for **~3 billion transactions** [1]. (Feburary 2025)

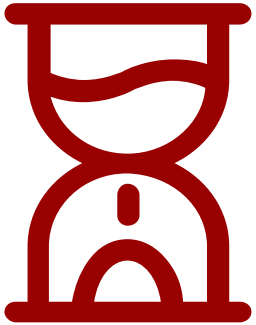
Generative-AI has highest profit potential for business automation.

- Could add **$2.6 – $4.4 trillion annually** in value, (The McKinsey Global Institute)
- **~15-40% economic impact**.



[1] McCrystal, S. (2025, February). How Unilever's digital transformation is driving operational excellence. Unilever News.
[2] Chui, M., Hazan, E., Roberts, R., Singla, A., & Smaje, K. (2023). *The economic potential of generative AI*. McKinsey & Company.

Problem Statement



Time-consuming to extract, aggregate and generate meaningful data visualizations

Need for **tool-specific knowledge** for deployment of BI tools.

Reliance of non-technical users with limited data analytics skills on the dashboard developers.

Research Questions Overview

RQ1 What are the state-of-the-art LLMs applications in data analytics and visualization?

RQ2 How to design a business data visualization system driven by LLMs?

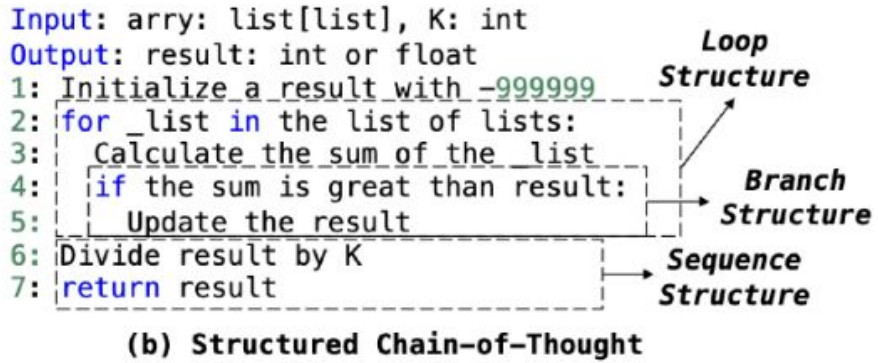
RQ3 How to implement a business data visualization system driven by LLMs?

RQ4 How to evaluate a business data visualization system driven by LLMs?

Literature review on LLMs applications in data analytics and visualization

Reference	Description	Model	Evaluation	Datasets
Prompt4Vis (2024) [29]	In-context learning using multi-objective example mining and schema filtering for simplified database search.	GPT-3.5-Turbo	Compared against ncNet, Seq2Vis, Transformer, RGVisNet using Vis Acc, Axis Acc, Data Acc, Overall Acc.	NVBench
Chat2VIS (2023) [35]	Streamlit-based web app using PE, description and code primers. Supports free-form NL queries and CSV uploads up to 30MB.	ChatGPT, Codex and GPT-3	6 case studies compared against NL4DV and ADVISor; robustness tested with ambiguous prompts.	nvBench, IMDb, Colleges, Energy, Products
ChartGPT (2023) [71]	Six-step reasoning pipeline powered by a fine-tuned LLM including multi-view interface, enabling users to inspect and modify intermediate outputs.	Fine-tuned Flan-T5-XL	Evaluated against NL4DV and ncNet using consistency, similarity (ROUGE-L, BLEU), and user study.	Custom dataset based on NL4DV

F2.2 LLM-based visualization systems often not stress large-scale data and are limited by the token context window.

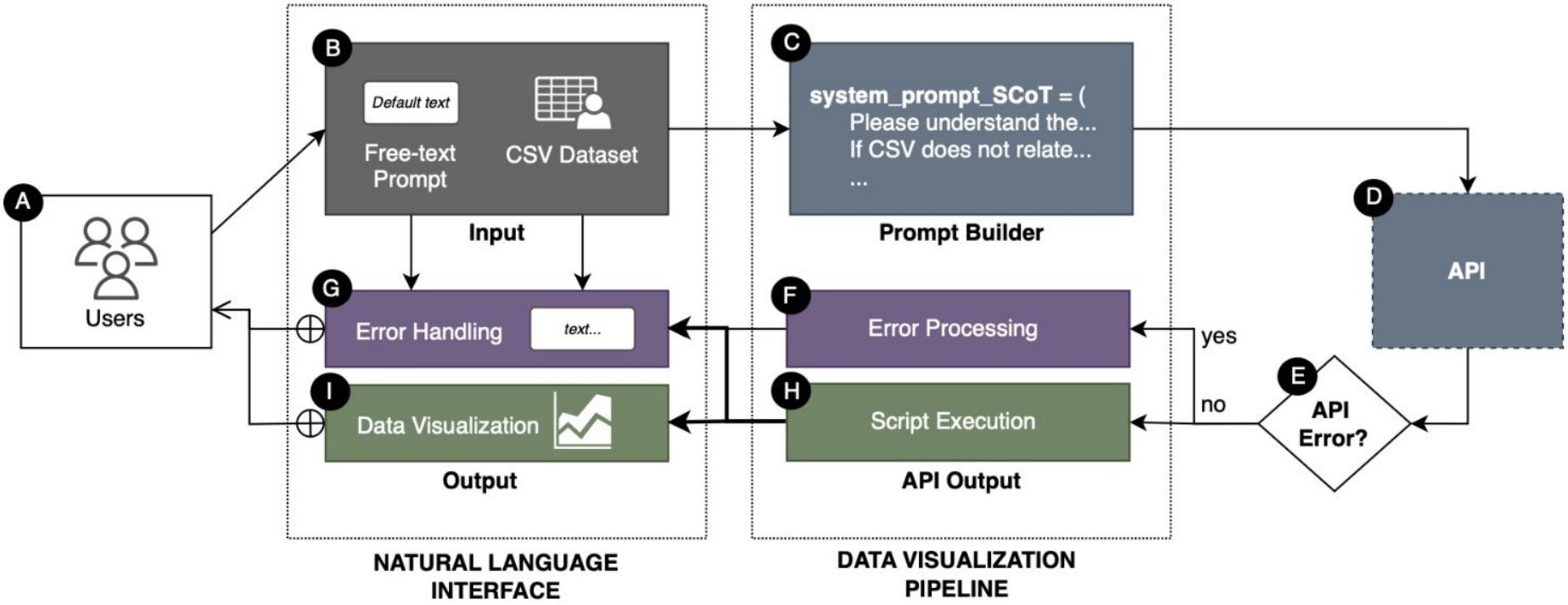


(b) Structured Chain-of-Thought

F2.3 Prompt engineering methods, e.g. SCoT show improved code generation accuracy. [3]

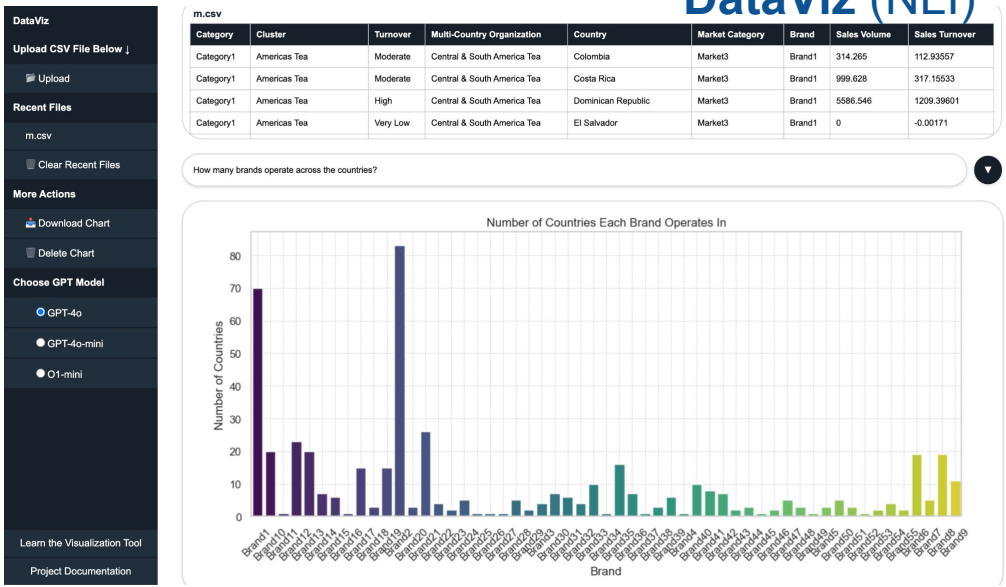
[3] Li, J., Li, G., Li, Y., & Jin, Z. (2025). Structured chain-of-thought prompting for code generation. ACM Transactions on Software Engineering and Methodology, 34(2), 1–23.

Design Overview of DataViz



Design Modules of DataViz

DataViz (NLI)



Category	Cluster	Turnover	Multi-Country Organization	Country	Market Category	Brand	Sales Volume	Sales Turnover
Category1	Americas Tea	Moderate	Central & South America Tea	Colombia	Market3	Brand1	314.265	112.93557
Category1	Americas Tea	Moderate	Central & South America Tea	Costa Rica	Market3	Brand1	999.628	317.15533
Category1	Americas Tea	High	Central & South America Tea	Dominican Republic	Market3	Brand1	5586.546	1209.39601
Category1	Americas Tea	Very Low	Central & South America Tea	El Salvador	Market3	Brand1	0	-0.00171

SCoT (PE) [3]

System Prompt

a **Code Generation Intention**

Please understand the requirement and write data visualisation script based on User Prompt and CSV Sample...

b **Prompt-encoded Error Handling**

If CSV does not relate to User Prompt, return an error explanation...

c **User Prompt**

Plot a bar chart showing relation of sales volume to...

d **CSV Sample**

Category,Cluster,Turnover,Multi-Country Organization,Country,Market Category,...
Category1,Americas Tea,Moderate,Central & South America Tea,Colombia,Market3,...
Category1,Americas Tea,Moderate,Central & South America Tea,Costa Rica,Market3,...

e **SCoT Prompting**

"Input: file: CSV"
"Output: Data visualization Python script using Pandas, Matplotlib..."
"1: Read file."
"2. Perform aggregation using the exact column names from..."
"3: Create plotting logic using Seaborn, ..."
...

LLM-facilitated Error (Error Handling)

Error: The CSV Sample does not contain data related sales turnover instead.

[3] Based on Li, J., Li, G., Li, Y., & Jin, Z. (2025). Structured chain-of-thought prompting for code generation. ACM Transactions on Software Engineering and Methodology, 34(2), 1–23.

Design Modules of DataViz

1 Natural Language Question

What is the relationship between sales volume and sales turnover in Colombia?

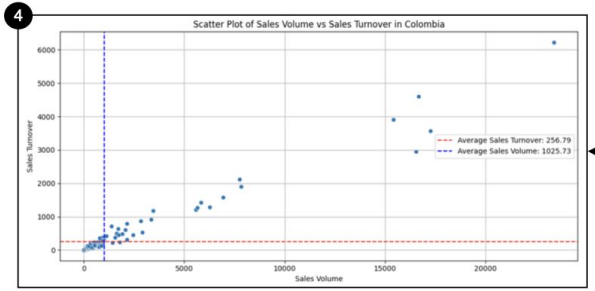
2 System Prompt

```
"Please understand the requirement and write..."
"If CSV does not relate to the User Prompt..."
f"User Prompt: {prompt}"
f"CSV Sample: {data_sample}"
"Input: file: CSV"
"Output: Data visualization Python script"
 "1: Read file."
...
```

3 Data Visualization Script

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Step 1: Read file
df = pd.read_csv(file)

# Step 2: Perform aggregation
colombia_data = df[df['Country'] == 'Colombia']
average_sales_volume = colombia_data['Sales Volume'].mean()
average_sales_turnover = colombia_data['Sales Turnover'].mean()

# Step 3: Create plotting logic
plt.figure(figsize=(11, 5.4))
sns.scatterplot(data=colombia_data, x='Sales Volume', y='Sales Turnover')
plt.axhline(y=average_sales_turnover, color='r', linestyle='--', label=f'Average Sales Turnover:
{average_sales_turnover:.2f}')
plt.axvline(x=average_sales_volume, color='b', linestyle='--', label=f'Average Sales Volume:
{average_sales_volume:.2f}')
plt.xlabel('Sales Volume')
plt.ylabel('Sales Turnover')
plt.title('Scatter Plot of Sales Volume vs Sales Turnover in Colombia')
plt.grid(True)
plt.legend()
plt.tight_layout()

# Step 5: Display the plot
plt.show()
```

4

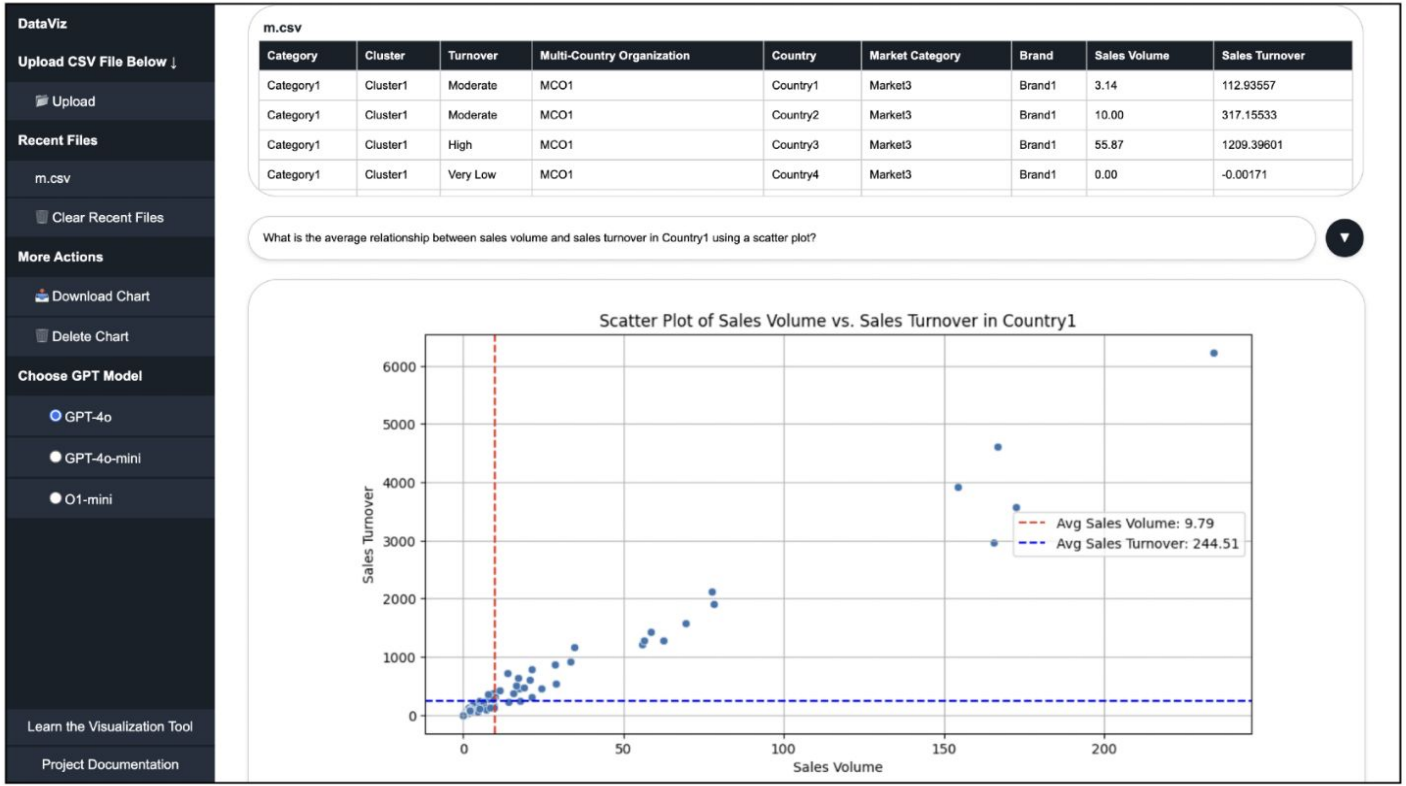


Data Flow

(Data Visualization Pipeline)

[3] Based on Li, J., Li, G., Li, Y., & Jin, Z. (2025). Structured chain-of-thought prompting for code generation. ACM Transactions on Software Engineering and Methodology, 34(2), 1–23.

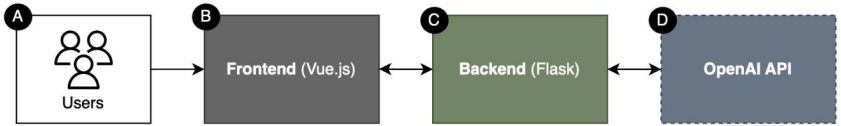
Implementation of DataViz



Category	Cluster	Turnover	Multi-Country Organization	Country	Market Category	Brand	Sales Volume	Sales Turnover
Category1	Cluster1	Moderate	MCO1	Country1	Market3	Brand1	3.14	112.93557
Category1	Cluster1	Moderate	MCO1	Country2	Market3	Brand1	10.00	317.15533
Category1	Cluster1	High	MCO1	Country3	Market3	Brand1	55.87	1209.39601
Category1	Cluster1	Very Low	MCO1	Country4	Market3	Brand1	0.00	-0.00171

F4.1 A fully-automated, visualization-first, LLM-driven web application can be implemented usin a lightweight architecture

F4.2 Resource-efficient prompting enables generation of secure, low-cost visualizations with consistent API latency and low token usage.



Evaluation of DataViz

Category	Cluster	Turnover	Multi-Country Organization	Country	Market Category	Brand	Sales Volume	Sales Turnover
Category1	Cluster1	Moderate	MCO1	Country1	Market3	Brand1	3.14	112.93557
Category1	Cluster1	Moderate	MCO1	Country2	Market3	Brand1	10.00	317.15533
Category1	Cluster1	High	MCO1	Country3	Market3	Brand1	55.87	1209.39601

Aggregation	Data Types	Plot Type	Question
Average	2Q	Scatter Plot	What is the average relationship between sales volume and sales turnover in Country1 using a scatter plot?
Extreme	1C+1Q	Bar Chart	Which country has the highest sales turnover in MCO1 Multi-Country Organization?
Count	1–2C	Bar Chart	How many brands operate across the countries?
Sum	1O+1Q	Line Chart	What is the total sales volume by turnover level for each market using a line chart?

Proposed PE [4]

Dataset Sizes

Dataset Size	Label
10KB	*XXS*
100KB	*XS*
1MB	*S*
10MB	*M*
100MB	*L*
1GB	*XL*
10GB	*XXL*

Evaluation Metrics:

Accuracy

- Human Evaluation,
- VER;

Performance

- Script Execution Latency,
- Token Usage

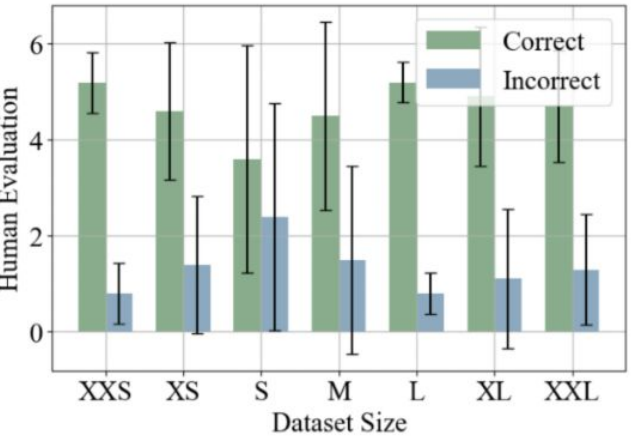
Usability

- User Study

[4] Based on Liu, C., Han, Y., Jiang, R., & Yuan, X. (2021). ADVISor: Automatic visualization answer for natural-language question on tabular data. In Proceedings of the 2021 IEEE 14th Pacific Visualization Symposium (PacificVis) (pp. 11–20).

Evaluation of DataViz

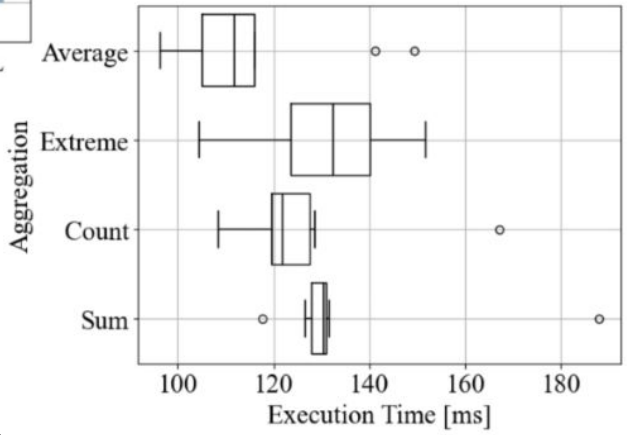
Accuracy (*Average*)



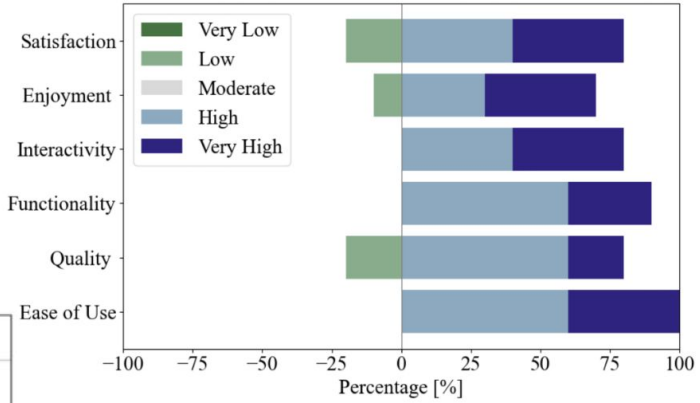
F5.2 Mostly correct, more variability in error bars for *Average* and Extreme agg.
3.7% VER

Performance (*XXS*)

F5.4 Avg script execution latency of **11.29s** with low and stable token usage.



Usability (Task A) [5]



F5.5 80% *very likely* to use LLM-driven data analytics.

[5] Based on Bialkova, S. (2024). The rise of AI user applications. Springer.

Research Contributions Overview

C1 (Conceptual) We conduct a systematic literature review on the state-of-the-art LLMs applications for data analytics and visualization. (**RQ1**)

C2 (Conceptual) We design a business data visualization system (DataViz) that automatically performs data analytics using LLMs. (**RQ2**)

C3 (Technical) We implement and evaluate the system for real-world business data visualization. (**RQ3&4**)

C4 (Open Science) We follow the FAIR principles and will release the data and code as a contribution to open science.

Take-Home Message

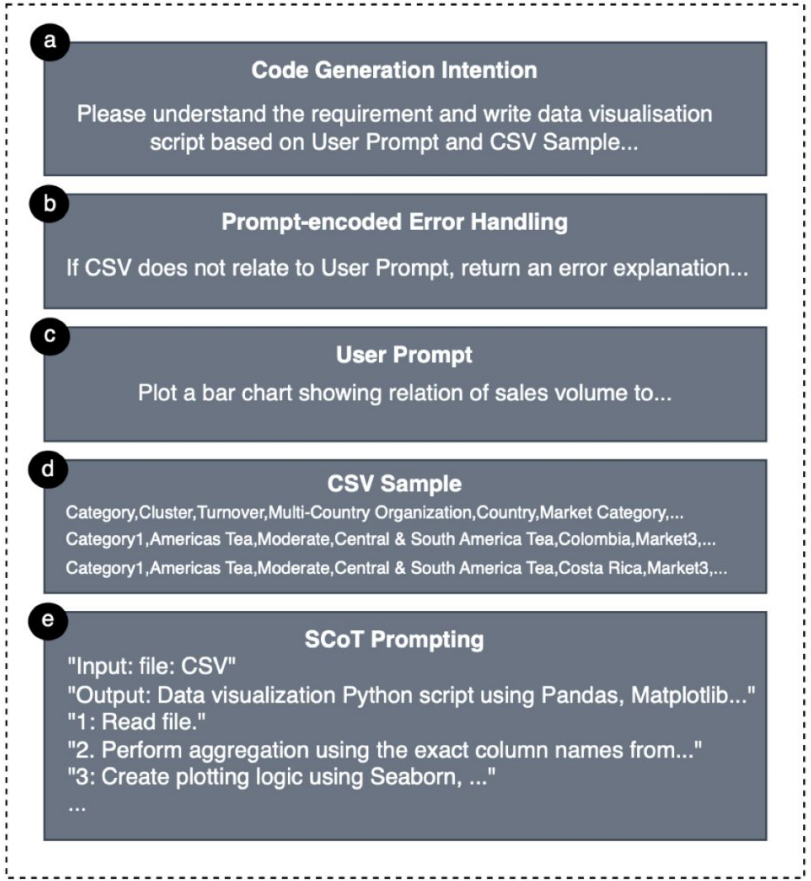
F5.1 DataViz ensures that LLM-driven fully automated visual analytics systems can be accurate, efficient for large-scale datasets and useful in supporting enterprise data.

F5.2, 5.4 We evaluate that DataViz can produce accurate and timely data analytics for varying-scale datasets with 3.7% of visualization error rate and an average execution latency of 11.29 seconds across 10KB-10GB files.

L4.1 Data visualization accuracy can be limited by incorrect understanding of the file structure and / or content by the LLM.

EXTRA SLIDES

System Prompt



1

Natural Language Question

What is the relationship between sales volume and sales turnover in Colombia?

2

System Prompt

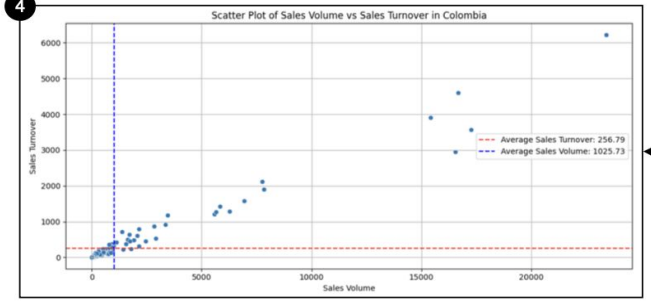
"Please understand the requirement and write..."
"If CSV does not relate to the User Prompt..."
f"User Prompt: {prompt}"
f"CSV Sample: {data_sample}"
"Input: file: CSV"
"Output: Data visualization Python script"
"1: Read file."
...

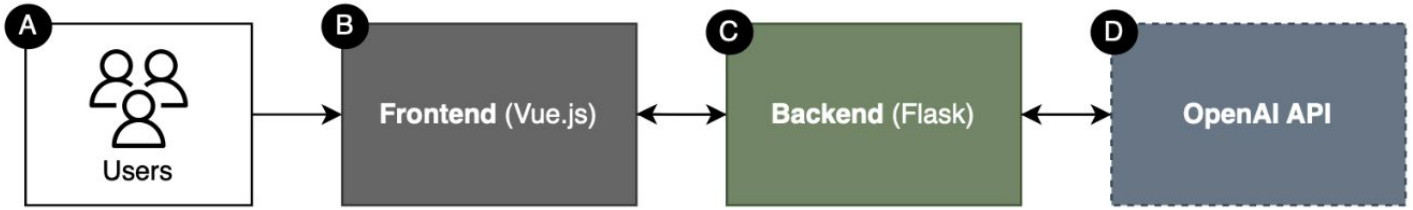
3

Data Visualization Script

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Step 1: Read file
df = pd.read_csv(file)

# Step 2: Perform aggregation
colombia_data = df[df['Country'] == 'Colombia']
average_sales_volume = colombia_data['Sales Volume'].mean()
average_sales_turnover = colombia_data['Sales Turnover'].mean()

# Step 3: Create plotting logic
plt.figure(figsize=(11, 5.4))
sns.scatterplot(data=colombia_data, x='Sales Volume', y='Sales Turnover')
plt.axhline(y=average_sales_turnover, color='r', linestyle='--', label=f'Average Sales Turnover:
{average_sales_turnover:.2f}')
plt.axvline(x=average_sales_volume, color='b', linestyle='--', label=f'Average Sales Volume:
{average_sales_volume:.2f}')
plt.xlabel('Sales Volume')
plt.ylabel('Sales Turnover')
plt.title('Scatter Plot of Sales Volume vs Sales Turnover in Colombia')
plt.grid(True)
plt.legend()
plt.tight_layout()

# Step 5: Display the plot
plt.show()
```

4

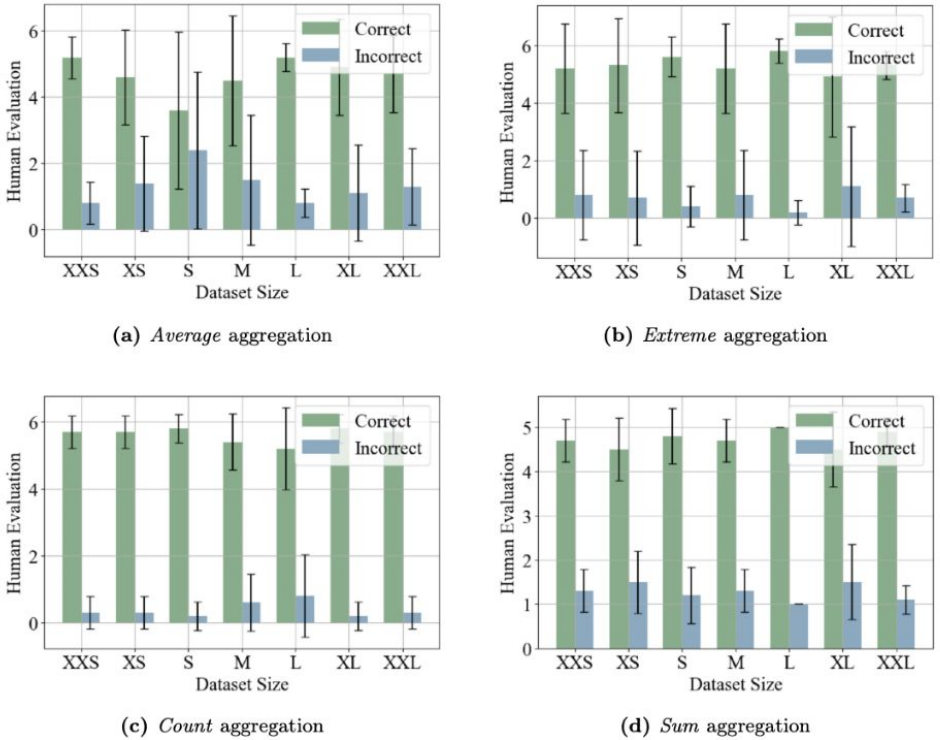




Package	Version	Brief Description
`flask`	3.0.3	Web development framework.
`openai`	1.54.3	Python SDK for OpenAI integration.
`python-dotenv`	1.0.1	Secure usage of environment variables.
`pandas`	2.2.3	Python data processing library.
`matplotlib`	3.9.2	Python plotting library.
`seaborn`	0.13.2	Python plotting library.

Table 4.1: Data dependencies `DataViz` implementation.

Accuracy Evaluation



(a) *Average* aggregation

(b) *Extreme* aggregation

(c) *Count* aggregation

(d) *Sum* aggregation

Figure 5.2: Human Evaluation results for different aggregation strategies.

Accuracy Evaluation

Aggregation	VER [%]
Average	2.7
Extreme	5.4
Count	2.7
Sum	4.1

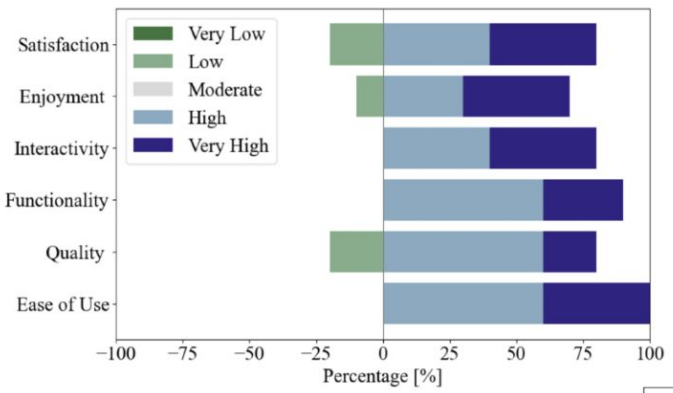
Table 5.7: Results from visualization error rate (VER) metric, based on [14].

Performance Evaluation

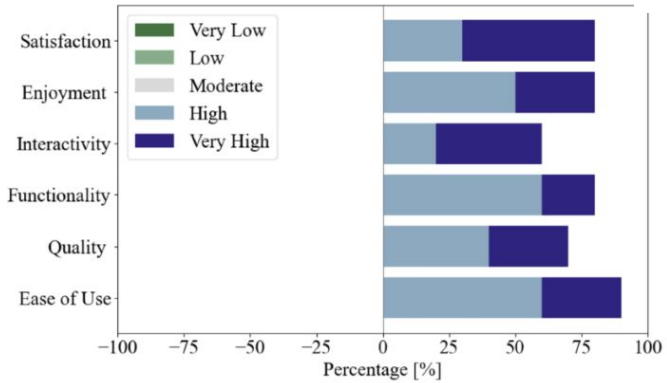
Aggregation	Input Tokens	Ouput Tokens
Average	260	166
Extreme	257	200
Count	257	163
Sum	266	170

Table 5.8: Token Usage per Aggregation.

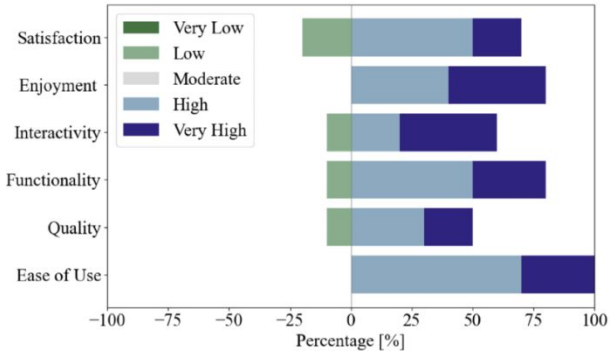
Usability Evaluation



(a) Task A

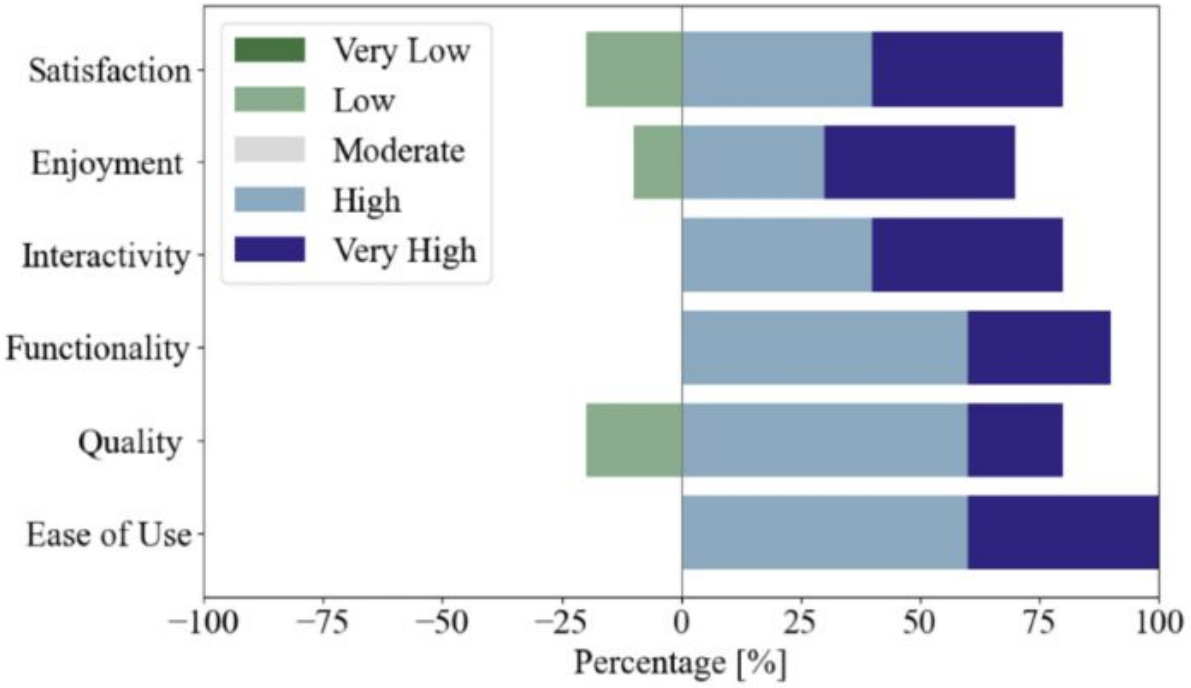


(b) Task B



(c) Task C

Usability Evaluation



(a) Task A

Usability Evaluation

Parameter	Description
Ease of Use	The degree to which the user expects the tech system to be free of effort.
Quality	Information, technical, and overall service success.
Functionality	Reflects a correct technical functioning.
Interactivity	Extent to which users can manipulate technology and/or control device; emphasizes the role of interaction between the user and the system.
Enjoyment	Core affect, typically arising from connection or sensory pleasure, interchanged with happiness.
Satisfaction	Occurs when customers find the products or services meet or exceed their positive expectations.

Table 5.6: Core Parameters in Chatbot Efficiency Evaluation, reproduced from [4].