UNIL | Université de Lausanne
HEC Lausanne

MASTER THESIS

LEVERAGING VISUALIZATION-ORIENTED NATURAL LANGUAGE INTERFACES FOR ENHANCED DATA INTERACTION AND VISUALIZATION IN THE FINANCIAL SECTOR

Author: Pierre Devillers
Supervisors: Michalis Vlachos, Alexandre Batista
Academic Expert: Yash Raj Shrestha

A BRIEF SUMMARY

1. **Problem & Literature Review**
   Quick overview of the problematic and its consideration in academic literature

2. **In Practice: Crafting VizExplorer**
   Development of an intuitive application for data interaction and visualization

3. **Benchmark and Performance measures**
   Evaluating system accuracy and efficiency through benchmarking and performance metrics

4. **Actionable Insights and Limitations**
   Realizing the potential of easy access to data analysis in the financial sector

5. **What's Next?**
   Future steps of development for an accrued efficiency and quality of the solution

THE PROBLEM

In a world with growing data volume and complexity

How do we provide appropriate data for everyone’s decisions

Facilitate insight accessibility and business users’ empowerment

Remove the need of human interaction to custom reporting

GROWING **VOLUME & COMPLEXITY**

Navigating the expanding data landscape and its growing complexities

Diversification of sources & types

The diversification of data sources and types, from structured databases to unstructured text, has transformed how organizations collect and manage information.

Growing volume

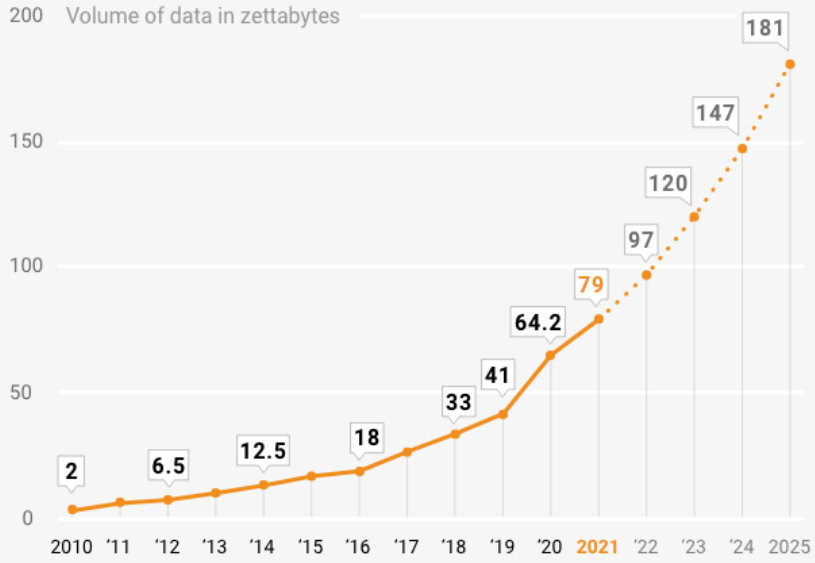
The volume of data generated globally continues to grow exponentially, making traditional methods of analysis increasingly insufficient.

Intensifying complexity

As data-driven decisions become more critical, analyses are becoming more complex, requiring tools that can handle intricate relationships and advanced computations.

Volume of data created, captured, copied and consumed worldwide

The volume of data generated, consumed, copied, and stored is projected to exceed 180 zettabytes by 2025



FINANCIAL **SECTOR FOCUS**

Mastering the evolving demands and resources of the modern financial ecosystem

EXTENSIVE & INTRICATE
FINANCIAL DATA

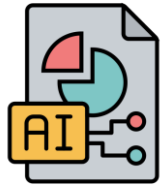
Banks deal with highly detailed and complex datasets, from client portfolios to global market activities, demanding sophisticated handling.

QUICK & EFFICIENT
DATA ANALYSIS

Fast and accurate insights are crucial to stay competitive and ensure operational efficiency in financial markets.

STRINGENT
REGULATORY CONSTRAINTS

The industry operates under strict rules that require absolute precision and accountability in data management.



INCREASED ADOPTION OF
AI & ML TECHNIQUES

Emerging technologies like AI and machine learning are transforming the way institutions uncover trends and make decisions.

RESEARCH **MOTIVATION**

Fuelling analytical excellence and transforming data-driven insights

Practical Challenges

Non-technical professionals often struggle to access insights due to complexity of traditional data tools.

Significant workload

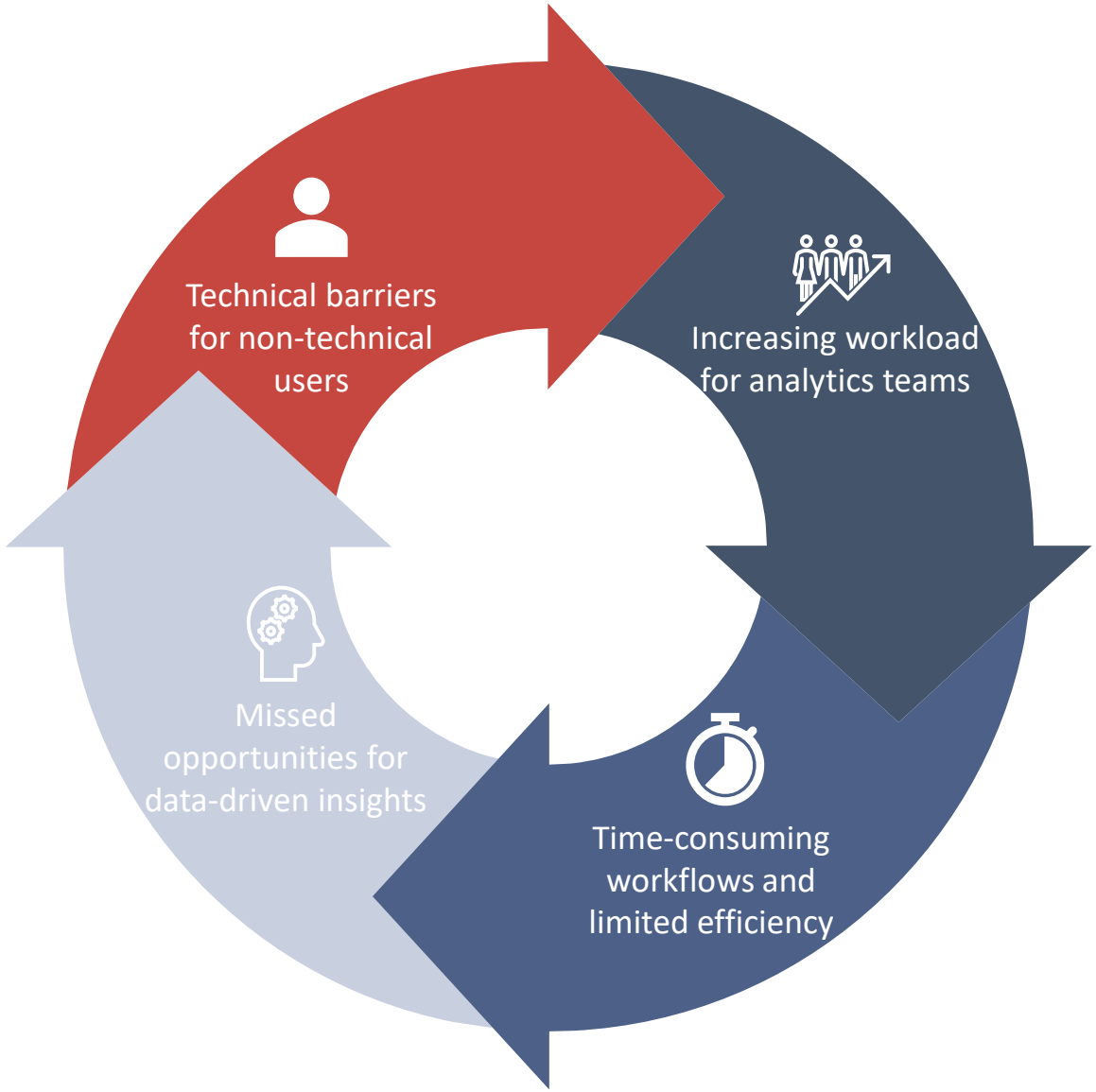
Data analytics teams are burdened with routine queries, limiting their ability to focus on strategic tasks.

Sub-optimized processes

Current processes for data querying and visualization are tedious and can delay decision-making.

Loss of earnings

Organizations risk losing competitive advantages by failing to fully leverage their data potential.



RESEARCH **QUESTIONS**

Evaluating accuracy and usability in relation to technical and methodological requirements

Designing Effective User Interaction

What are the technical and methodological requirements for developing an effective V-NLI?

- Identify key technical requirements, including NLP accuracy, database compatibility, and user interface design
- Explore the methodological challenges of integrating natural language query processing with data visualization tools

Evaluating System Performance

How does the developed system perform in terms of accuracy and usability?

- Evaluate system's ability to accurately translate user queries into SQL and generate meaningful visualizations
- Analyze system performance through benchmarking, focusing on execution accuracy and ranking it in standardized metrics.

Impacts & Challenges

of V-NLIs in financial institutions

What are the main advantages and potential limitations for integrating V-NLIs in financial institutions?

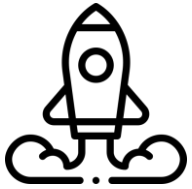
- Highlights the benefits of V-NLIs such as improved data accessibility, operational efficiency, and decision-making support.
- Analyze potential challenges, including scalability, handling complex and ambiguous queries, and real-time processing constraints.

- ACADEMIC RESEARCH -

Foundations of the Study

LITERATURE **REVIEW**

Examining the foundations that shape innovative solutions for data interaction and analysis.

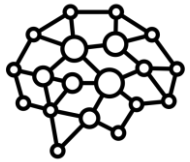


Early Foundations & Systems

- Rule-based natural language understanding for initial conversational systems.
- Semantic parsing for interpreting user intent.

Data Visualization & Natural Language Processing

- Emergence of tools combining natural language interaction with dynamic visualizations.
- Iterative feedback loops to refine queries.



Modern AI techniques & Text-to-SQL tasks

- Transformers models for deep contextual understanding for NLP tasks.
- Text-to-SQL advancements and automated database interaction for complex data queries.

The literature review reveals a growing need to bridge the gap between natural language understanding, interactive visualization, and advanced query automation to make a complex data accessible and actionable.

TIMELINE FOR **REFERENCED RESEARCH**

Tracing the evolution for key innovations in V-NLIs

1966 — **ELIZA**

First program to simulate human-like conversation using simple rule-based NLP

1976 — **SOPHIE**

Incorporating semantic understanding for educational purposes

1992 — **Core Language Engine**

Detailed framework for designing NLP tools in various fields of application

1999 — **Hearst's "User Interfaces And Visualization"**

Highlights the importance of intuitive interfaces and visual representations in enhancing data accessibility

2013 — **WORD2VEC**

Introducing vector embeddings to capture semantic relationships between words

2016 — **EVIZA**

Enabling users to interact with visualization through natural language queries

2022 — **GPT/BERT**

Transformer architectures, enabling deep contextual understanding and high accuracy in text-to-SQL tasks

2022 — **NL2SQL**

Enable Natural Language queries to be translated into SQL, bridging the gap between user intent and database interaction

2024 — **BIRD**

Standardized benchmark to evaluate performance for Text-to-SQL systems

KEY **FINDINGS**

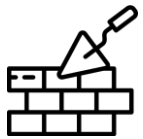
Insights shaping the development direction

1

RULE-BASED NLP AND SEMANTIC GRAMMAR

GROUNDWORK FOR RESEARCH

Syntactic and Semantic understanding of data by the tool with limited handling of ambiguity. Established the foundation for contemporary methods.



2

INTERACTIVE VISUALIZATION TECHNIQUES

ENHANCED INSIGHTS

Shift of the focus from simply generating data to enhancing the user's understanding and decision-making through visualization.

3

EMERGENCE OF AI-DRIVEN MODELS FOR COMPLEX CASES

DISRUPTIVE ADVANCEMENTS

Recent research brought revolutionary breakthroughs in query generation, SQL accuracy, and visualization techniques while highlighting challenges

- IN PRACTICE -

Developing the Solution

SNAPSHOT OF **VIZ EXPLORER**

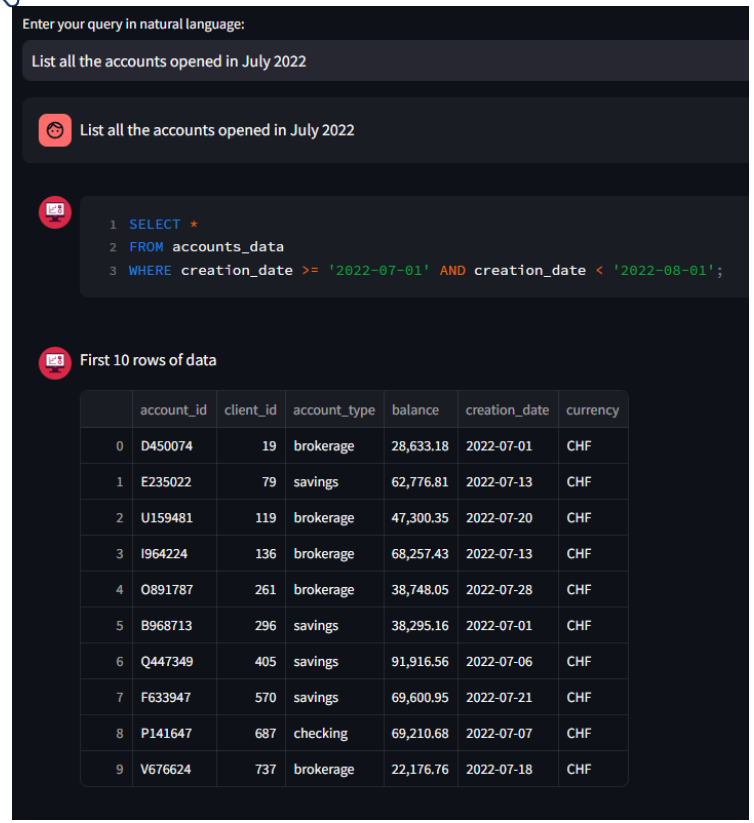
Link to VizExplorer

Data Interaction

“List all the accounts opened in July 2022”

Intuitive interface design

User-defined filtering for precise data retrieval



	account_id	client_id	account_type	balance	creation_date	currency
0	D450074	19	brokerage	28,633.18	2022-07-01	CHF
1	E235022	79	savings	62,776.81	2022-07-13	CHF
2	U159481	119	brokerage	47,300.35	2022-07-20	CHF
3	I964224	136	brokerage	68,257.43	2022-07-13	CHF
4	O891787	261	brokerage	38,748.05	2022-07-28	CHF
5	B968713	296	savings	38,295.16	2022-07-01	CHF
6	Q447349	405	savings	91,916.56	2022-07-06	CHF
7	F633947	570	savings	69,600.95	2022-07-21	CHF
8	P141647	687	checking	69,210.68	2022-07-07	CHF
9	V676624	737	brokerage	22,176.76	2022-07-18	CHF

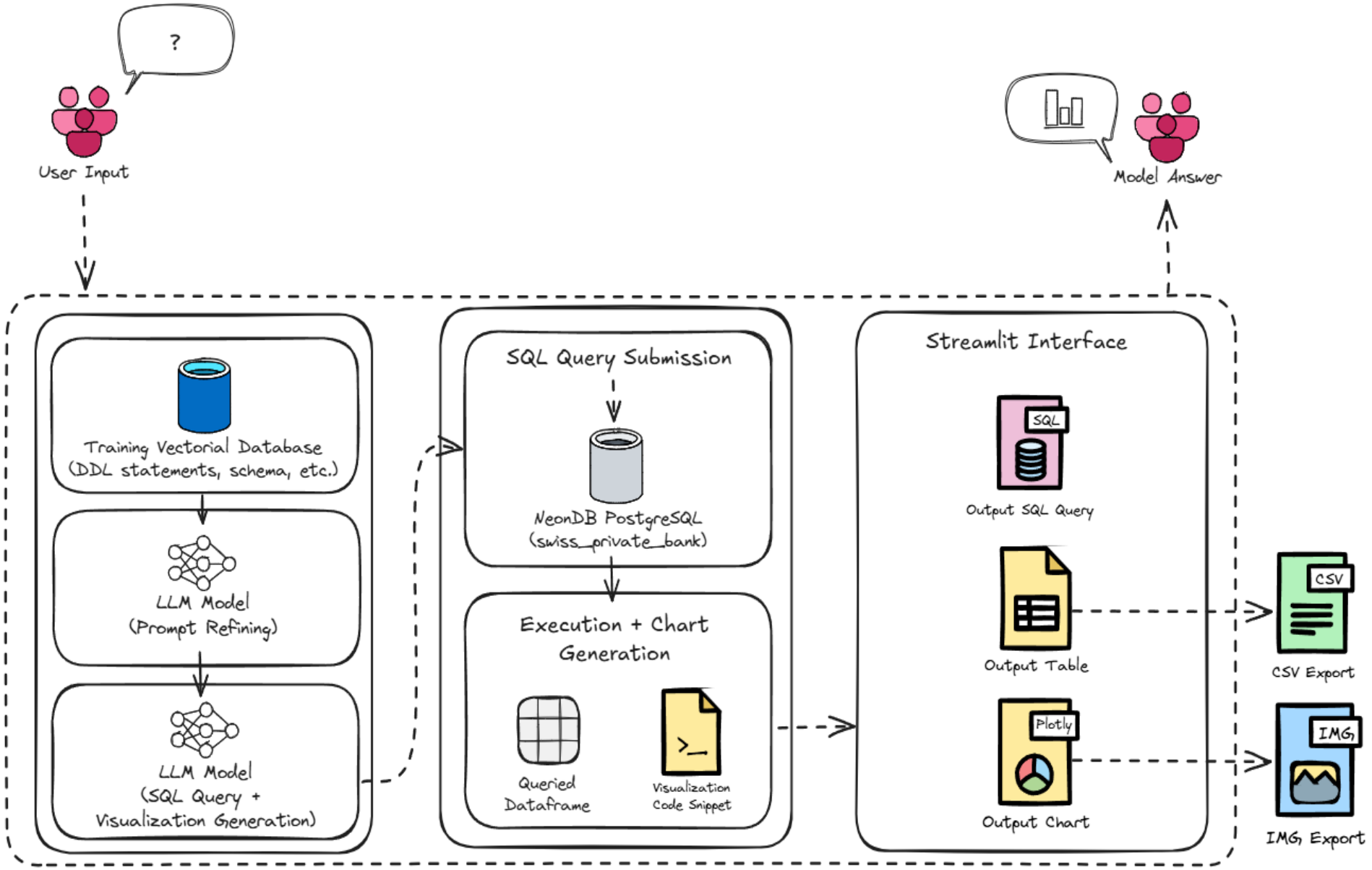
Data Visualization

“What is the performance of Apple over the last year?”

Interactive visualizations: export, detailed zoom, diverse chart types (bar chart, line chart, box plots, etc.)

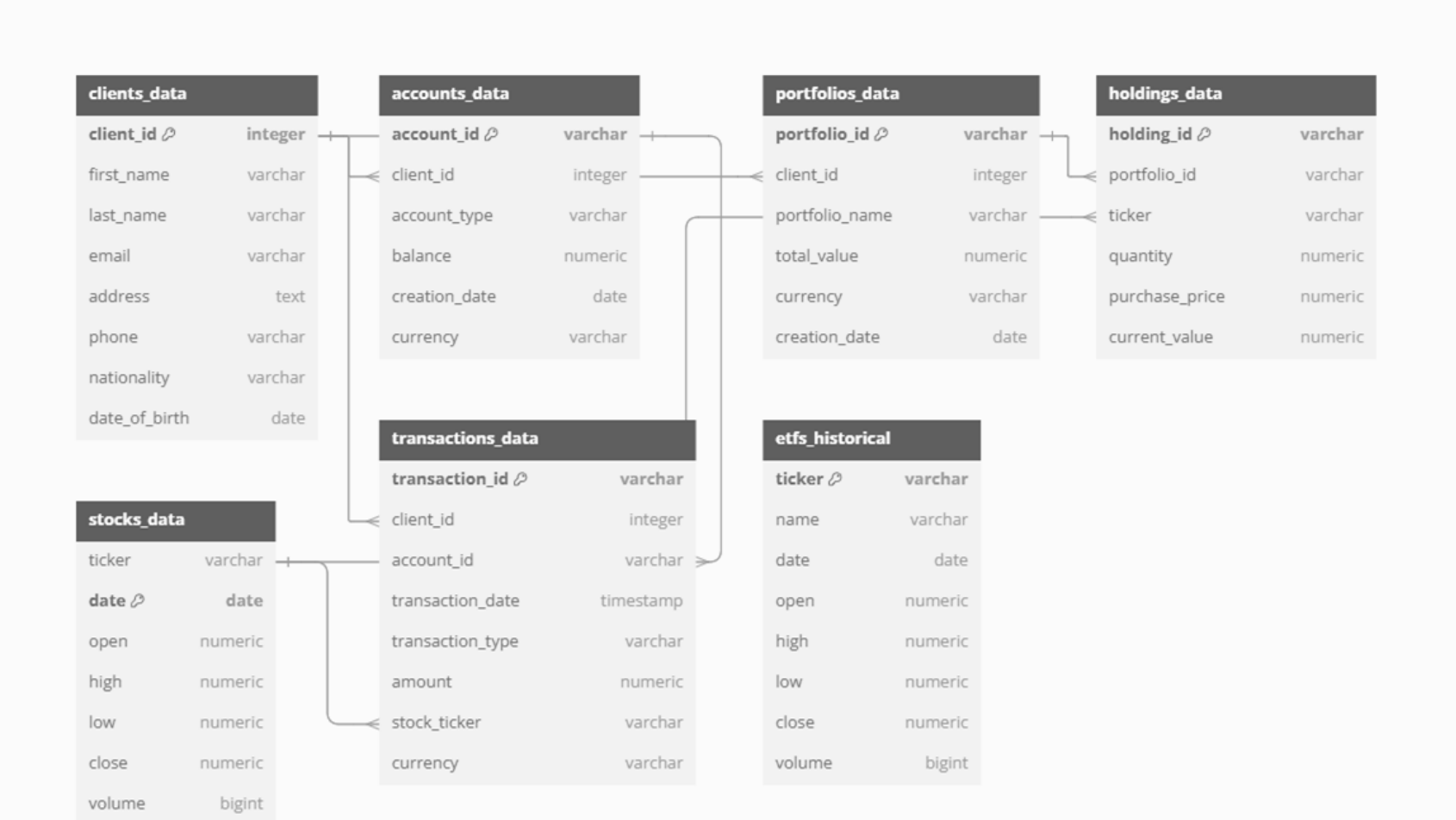
Simple contextual understanding (‘Apple’ → ticker = ‘AAPL’)

GLOBAL **CONCEPTUAL VIEW**



ENTITY RELATIONAL **DIAGRAM**

- **Comprehensive Representation of Financial Activities**
- **Support for Complex Relationships and Queries**
- **Coverage of Financial Instruments and Market Data**



- EVALUATING RESULTS -

Measuring Performance

INTRODUCING **BIRD BENCHMARK**

- (**BI**g Bench for La**R**ge-scale **D**atabase Grounded Text-to-SQL Evaluation)
  - Widely recognized **standardized benchmark** to evaluate performance for **Text-to-SQL systems**
  - Emphasizes **real-world** complexity:
    - **Diverse** database schema
    - Queries of **varying complexity**
    - Robustness to **ambiguous** or **incomplete** output
  - Tests **accuracy** and **execution performance**

User Input

↓

Model

↓

SQL Query

↓

Execution

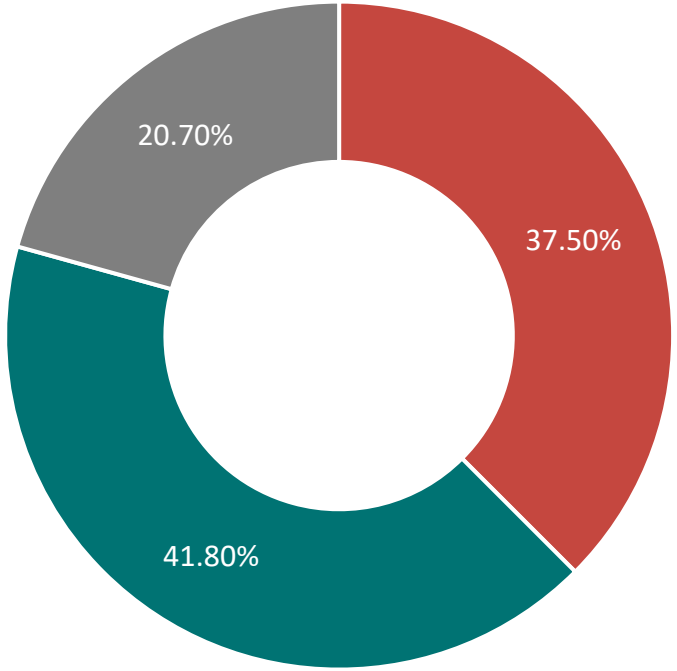
↓

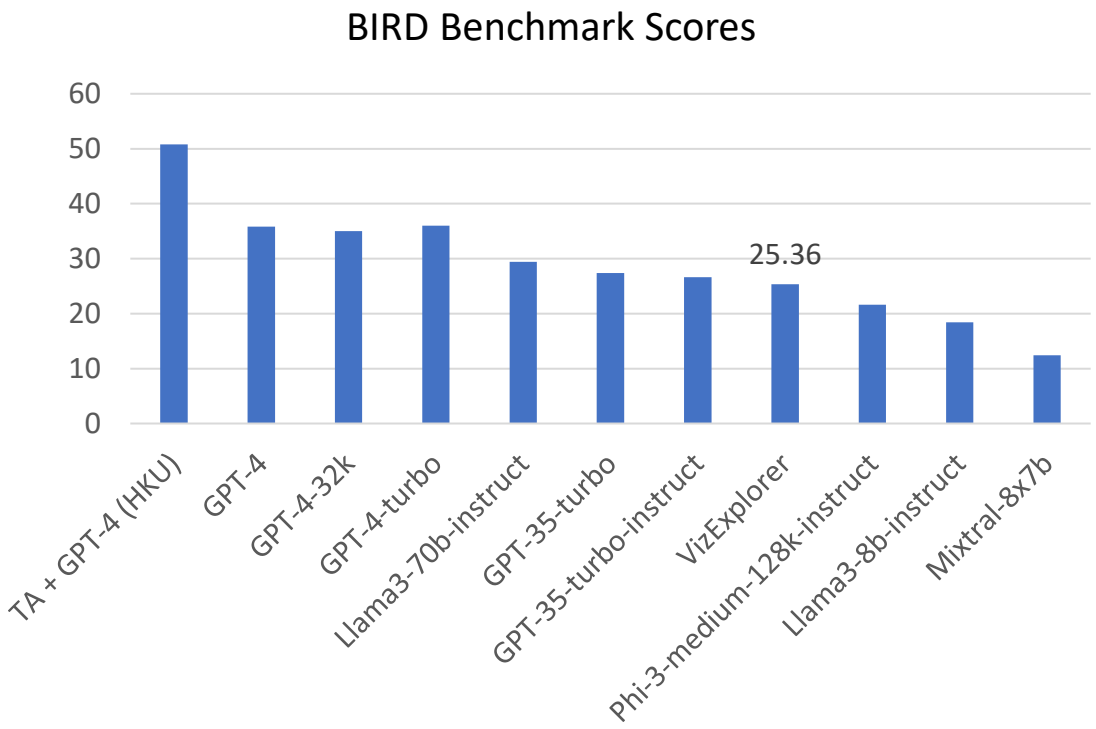
Evaluation Metrics

RESULTS AT A GLANCE

Assessing model's success rate and benchmark performance, showcasing strength in simple cases and identifying areas for improvement

The model demonstrates a **42% success rate** for valid queries, showcasing strong capabilities for straightforward cases. Its competitive ranking – **8th** on the BIRD mini-dev leaderboard with a **25% Execution Accuracy** – highlights promising potential while raising the need for improved handling of complex queries.





DATASET SIZE COMPARAISON

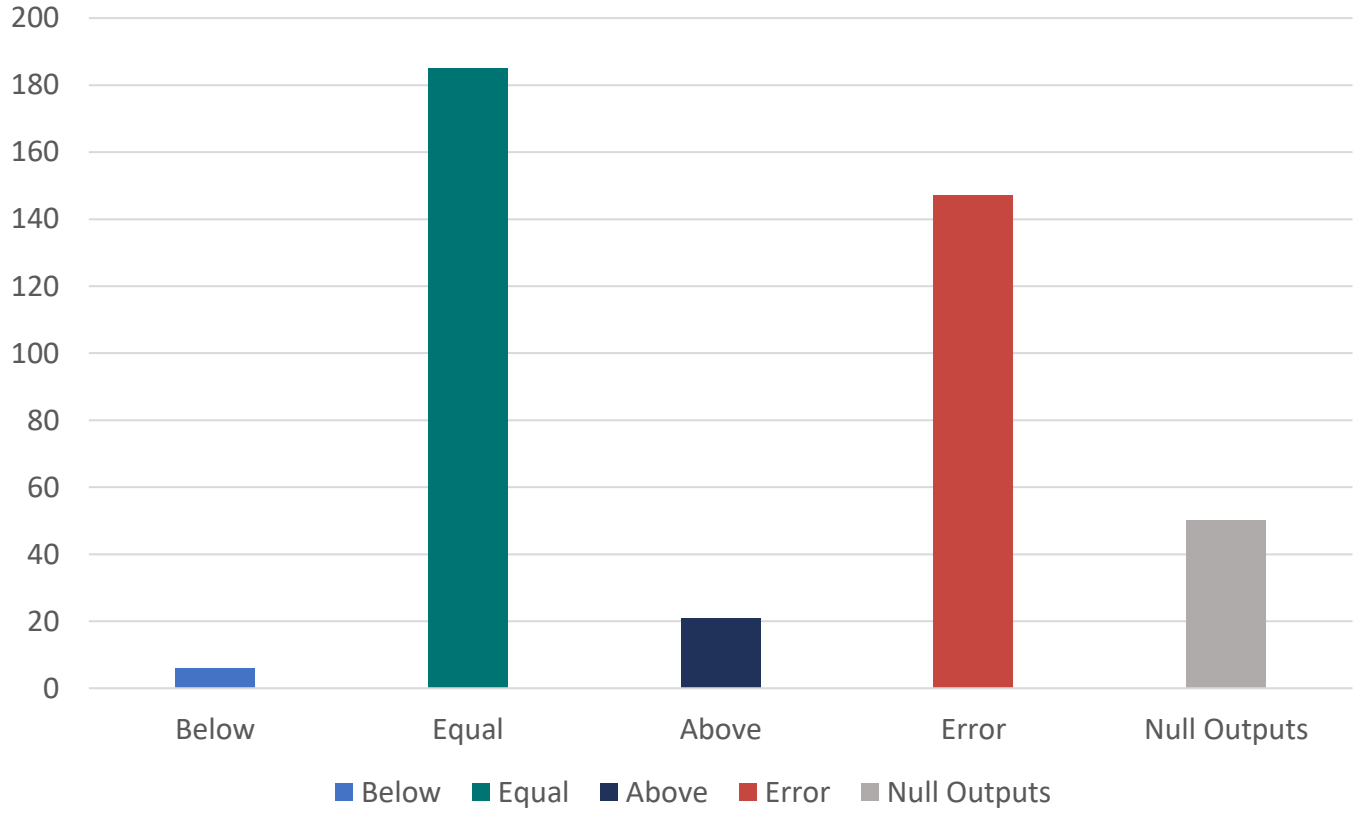
Evaluating model's ability to align query results with reference data, highlighting reliability and areas for refinement

Most queries aligned with the reference dataset size, yielding accurate outputs. However, inconsistencies led to variations in the number of rows returned for certain queries, underscoring the challenges in achieving precise filtering criteria for complex queries.

185 queries which **matched** the **reference dataset size**

147 queries which **failed** to execute due to **syntax errors**

40 queries returning **different number** of rows due to **inconsistencies in filters**



SIMILARITY SCORE ANALYSIS

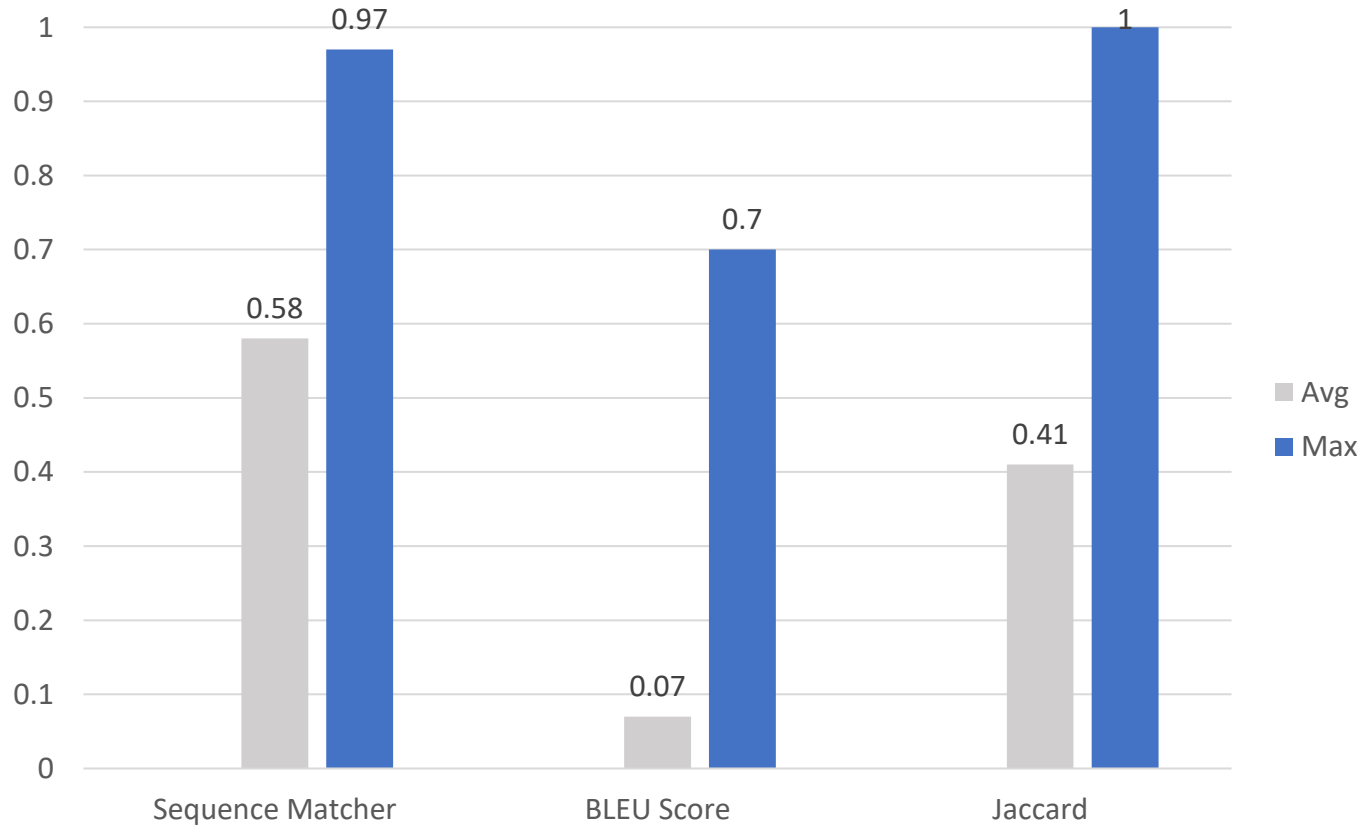
Measuring precision through different scores to respectively evaluate syntactic similarity, token-level accuracy, and query generation consistency

The similarity score analysis highlights the model's effectiveness in handling straightforward queries while revealing challenges in maintaining precision and consistency for more complex SQL structures.

0.58 **SQL syntax similarity score**
Reasonable performance on **simple queries** with improvement needed for more complex cases

0.07 **BLEU syntactic similarity score**
Reflects **challenges** in catching **syntactic correctness** and **semantic understanding** for complex SQL generation

0.41 **Dataset Jaccard similarity score**
Binary performance where generated output either **fully capture the logic**, or **fails completely**



- EXPLORING INSIGHTS -

Implications & Takeaways

WHY DOES IT MATTER?

Unlocking the potential for accessible data analysis with V-NLIs

Data accessibility for non-technical users

User-friendly interface for easy data interaction

- - -

Simplifying complex data into intuitive visual formats

1

Efficiency improvement for analytics teams

Improving our tools to boost reactivity & accuracy

- - -

Reducing manual tasks enabling focus on data-driven decisions

2

Scalability of larger financial datasets

Ensuring compatibility with varied data formats

- - -

Facilitating smooth implementation across diverse teams with different workflows

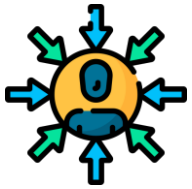
3

WHAT'S NEXT?

Future steps and strategic initiatives for an improved tool in terms of output quality and usability

Improve NLP's contextual understanding

- Dataset expansion to include diverse, multi-dimensional tables with real-world relevance.
- Accrued variety and complexity of the query set to ensure robustness in addressing real-world financial questions.



User-Centric Approach & Interactive Interface

- Involve end-users from various roles in the banking industry in the evaluation process to provide critical feedback for refining the tool.
- Iterative development process to enhance the tool's accuracy and overall performance.

Technical Features & Professional Implementation

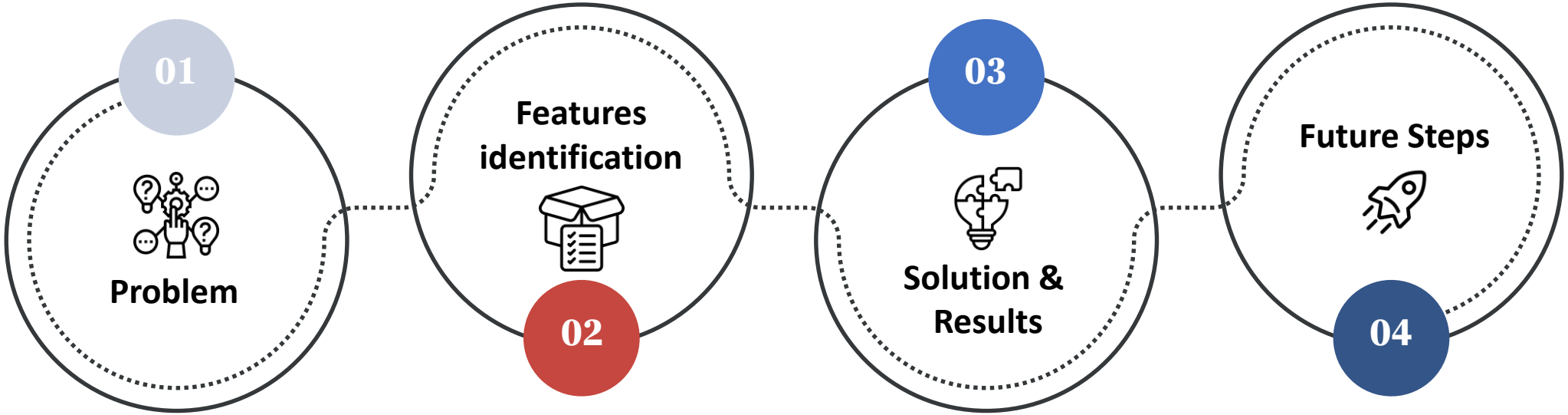
- Improve the User Interface with features such as follow-up prompts, feedback system, visualization customization, etc.
- Adapt the model to real-world environment (datasets size, regulatory requirements, data privacy regulations, industry standards, etc.)

By focusing on these key areas, we can ensure the V-NLI tool to meet the complex needs of the financial industry, delivering robust, user-friendly, and scalable solutions.

- TO CONCLUDE -

Looking back & Moving forward

FROM CHALLENGES TO SOLUTIONS



Data Complexity and technical barriers in the financial sector hinder accessibility and efficiency

Literature highlighted essential features (NLIs, interactive visualizations) to bridge the gap between users and complex data

Development of VizExplorer, a V-NLI combining NLP, SQL generation, and interactive visualizations which demonstrated feasibility

Focus on improving the handling of complex queries, enhancing efficiency, and adapting the tool for domain-specific application

This study represents a step towards making data insights accessible for everyone, anywhere – empowering users through technology.

- THANK YOU -

Q&A Session