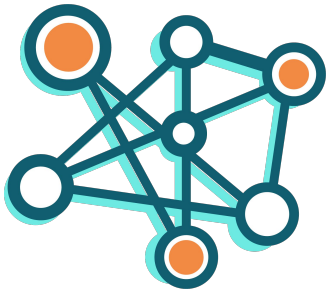
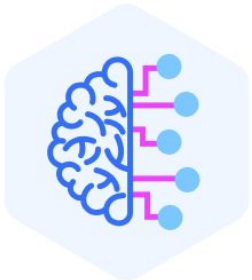
LLM POWERED QUERY ANSWERING
Bridging the Gap between Logic and Flexibility

Panin, Vladimir (12238682)
M.Sc. Data Science
Supervision: Prof. Dr.rer.soc.oec.
Dr.techn. Emanuel Sallinger &
Dr. Eleonora Laurenza

PROBLEM DEFINITION

- Knowledge Representation and Reasoning (KRR) models
  - Combination of data **D** and a set of rules **Σ** in a KG
  - Strong reasoning methods
  - Suffer from rigidity when dealing with imprecise data
  - Deterministic result
- Large Language Model (LLM)
  - Trained on massive amounts of data
  - Semantic capability; fault tolerance in prompt
  - Non - Deterministic result





What are the names of all the dogs in the table?

name	category	id
bill	chien	1
diego	chat	2
chris	dog	3
juan	perro	4

PROBLEM DEFINITION

,,Starting from that point, I propose to use a LLM to develop a query pipeline resolving the issue of residual noise in a database leading to my thesis **LLM-Powered Query Answering**. The query pipeline accepts a user-provided predicate calculus expression as input and automatically identifies noise within the database. Then a translation based approach is going to be used to modify the user-input query to fit the residual noise in the data.”

State of the Art (1)

- **Neurosymbolic Reasoning**: The SOTA involves neurosymbolic methods, combining LLMs with KRR (like the **soft chase** algorithm). In this case, bindings are additionally generated and verified using a LLM based on a NL fact.
- **Integration of LLMs with foundational reasoning skills**: Several techniques to integrate LLMs with structured data (KGs, Databases): Fine-tuning on structured data, embedding ontologies via text encoders, and using Chain-of-Thought Reasoning along few-shot demonstrations among many.

State of the Art (2)

- **Performance Limitations**: While improvements with foundational reasoning skills exist, limitations persist, particularly with larger datasets and complex queries. Symbolic reasoning systems still often outperform LLMs for certain tasks, especially when dealing with complex logical relationships.
- **Research Gap**:
  - Integration of LLM into query pipeline for databases with residual noise
  - LLM generated binding mechanism (**soft binding**) for SQL queries

RESEARCH QUESTIONS

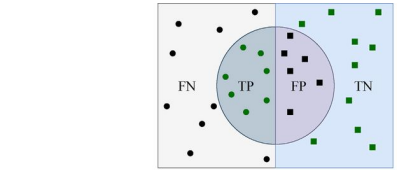
- How can LLMs be integrated into the query processing pipeline to soften the binding of variables in query answering, especially when data contains semantic mismatches or imprecision?
- How do LLM-based softening techniques perform in terms of relevant KPIs compared to approaches that rely on strict binding?
- What is the computational and token cost of the LLM-integrated pipeline compared to a approach with strict binding and how does this impact the query execution time?

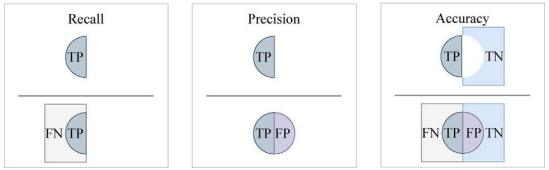
METHODOLOGY & APPROACH

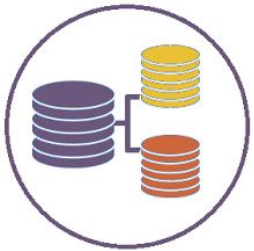
- **LLM Selection** Select a Large Language Model (LLM) appropriate for the task considering its capabilities
- **Test Set** Create a robust test set including residual noise; definition of predicate calculus expression along with ground truth for evaluation/experimentation
- **Soft Binding Approach Exploration** Exploration of an LLM to achieve soft binding; translation-based approach; modification of query using *WHERE* or *CASE* statements
- **Pipeline Implementation** performance and practicability considerations
- **Evaluation and Comparison** Evaluate pipeline performance using the test set and compare it to logical reasoning techniques. Assess effectiveness and computational cost

EXPECTED RESULT

- Crafted Evaluation and Test dataset
- LLM-Integrated query processing pipeline
- Accuracy metrics for evaluated pipeline
    - Precision, Recall, F1 Score and False Positives
- Metrics on query **execution time** and resource **usage**
- Master thesis document







Thank you for your Attention !

Any questions?