



UNIVERSITÄT
LEIPZIG

TEMIR Research Seminar

Text2SQL: Exploring Relational Databases with Natural Language User Interfaces

Leipzig, 27/06/2024

Julian Thilo, supervised by Tim Gollub



WEBIS

TEMIR

The Text2SQL task

„WHO
CONTRIBUTED
TO ECIR'23?“

```
SELECT DISTINCT a.name  
FROM authors a  
JOIN paper_authors pa ON (...)  
JOIN papers p ON (...)  
JOIN conferences c ON (...)  
WHERE c.name = "ECIR'23";
```

The Text2SQL task

„**WHO
CONTRIBUTED
TO ECIR'23?**“

a.name
Adam Jatowt
Arjen P. de Vries
Ebrahim Bagheri
...

The Text2SQL task

„WHO CONTRIBUTED TO ECIR‘23?“

Contributors to ECIR‘23 include Adam Jatowt, Arjen P. de Vries, and Ebrahim Bagheri among others.



TASK BACKGROUND

HISTORY OF TEXT2SQL RESEARCH

- Text2SQL originally based on **careful phrasing** of questions and **hand-crafted rules**
- Trained **neural networks** are much more capable but still **limited**



TASK BACKGROUND

HISTORY OF TEXT2SQL RESEARCH

- Text2SQL originally based on **careful phrasing** of questions and **hand-crafted rules**
- Trained **neural networks** are much more capable but still **limited**
- Many **evaluation datasets** were created for these earlier systems, e.g., **WikiSQL** [Zhong et al., 2017], **SPIDER** [Yu et al., 2018]
- Both feature rather **basic queries** with SPIDER introducing most complex challenges at the time



CURRENT STATE OF THE ART

LLM-BASED APPROACHES

- **Impressive performance** on older datasets
- **New benchmark** datasets with more complexity, e.g., **BIRD** [Li et al., 2023]
- **Room for improvement:** humans still outperform best LLM solution



CURRENT STATE OF THE ART

LLM-BASED APPROACHES

- **Impressive performance** on older datasets
- **New benchmark** datasets with more complexity, e.g., **BIRD** [Li et al., 2023]
- **Room for improvement:** humans still outperform best LLM solution
- LLMs combined with **retrieval-augmented generation** layer
- **Task-specific models** outperform general-purpose models
- Little research on embeddings into **actual database** settings



CURRENT STATE OF THE ART

LLM-BASED APPROACHES

- **Impressive performance** on older datasets
- **New benchmark** datasets with more complexity, e.g., **BIRD** [Li et al., 2023]
- **Room for improvement:** humans still outperform best LLM solution
- LLMs combined with **retrieval-augmented generation** layer
- **Task-specific models** outperform general-purpose models
- Little research on embeddings into **actual database** settings

Are LLM-based Text2SQL models useful in real, complex applications?



OUR GOAL: TESTING THE LIMIT

IR Anthology

- Scholarly **search engine** for Information Retrieval journals and conferences
- **Inaccessible** meta information
- Missing **analytical** features

OUR GOAL: TESTING THE LIMIT



IR Anthology

- Scholarly **search engine** for Information Retrieval journals and conferences
- **Inaccessible** meta information
- Missing **analytical** features
- Use and evaluate two Text2SQL models for **analytical questions**
- Gather real, **complex questions** from the community
- **Optimize database** and schema retrieval to cover more questions

REALISTIC QUESTIONS



- Collection of natural language queries from IR researchers
- Minimal guidance to encourage complex, analytical questions

IR Anthology Analytics

What would you like to know about the IR research community?

Examples:

"Who is new to the community?"

"Which topics were hot in the 1980s?"

"At which conference were LLMs mentioned for the first time?"

Your answer

Submit

Clear form

REFINEMENT PROCESS



Form response

Placeholders

Multiple questions in one

REFINEMENT PROCESS



Form response

Placeholders

Multiple questions in one



Refined question

Single prompt

Specific topics

REFINEMENT PROCESS



Form response

Placeholders
Multiple questions in one



Refined question

Single prompt
Specific topics



Expected answer

Annotation for each question
NL responses?

EVALUATION DATASET

EXPECTED ANSWERS

- Default option: One or more **SQL statements** per question
- Alternative: **information** that needs to be included in response
- Track used SQL features or other **indicator of complexity**

EVALUATION DATASET

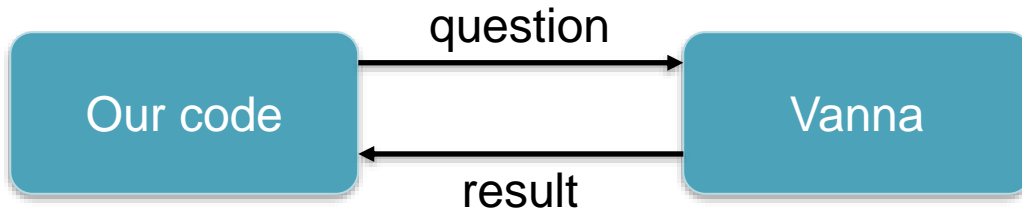
EXPECTED ANSWERS

- Default option: One or more **SQL statements** per question
- Alternative: **information** that needs to be included in response
- Track used SQL features or other **indicator of complexity**

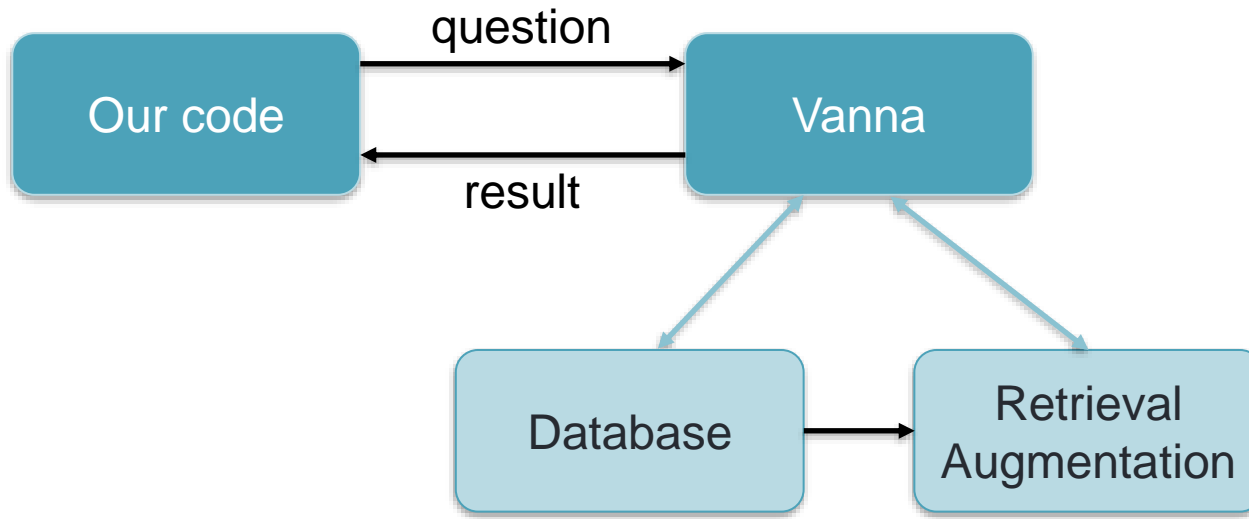
EVALUATION METRICS

- **Execution accuracy:** correct database response [Katsogiannis-Meimarakis and Koutrika, 2023]
- **Component matching:** correct parts of SQL [Katsogiannis-Meimarakis and Koutrika, 2023]
- **Keyword comparison:** correct information included

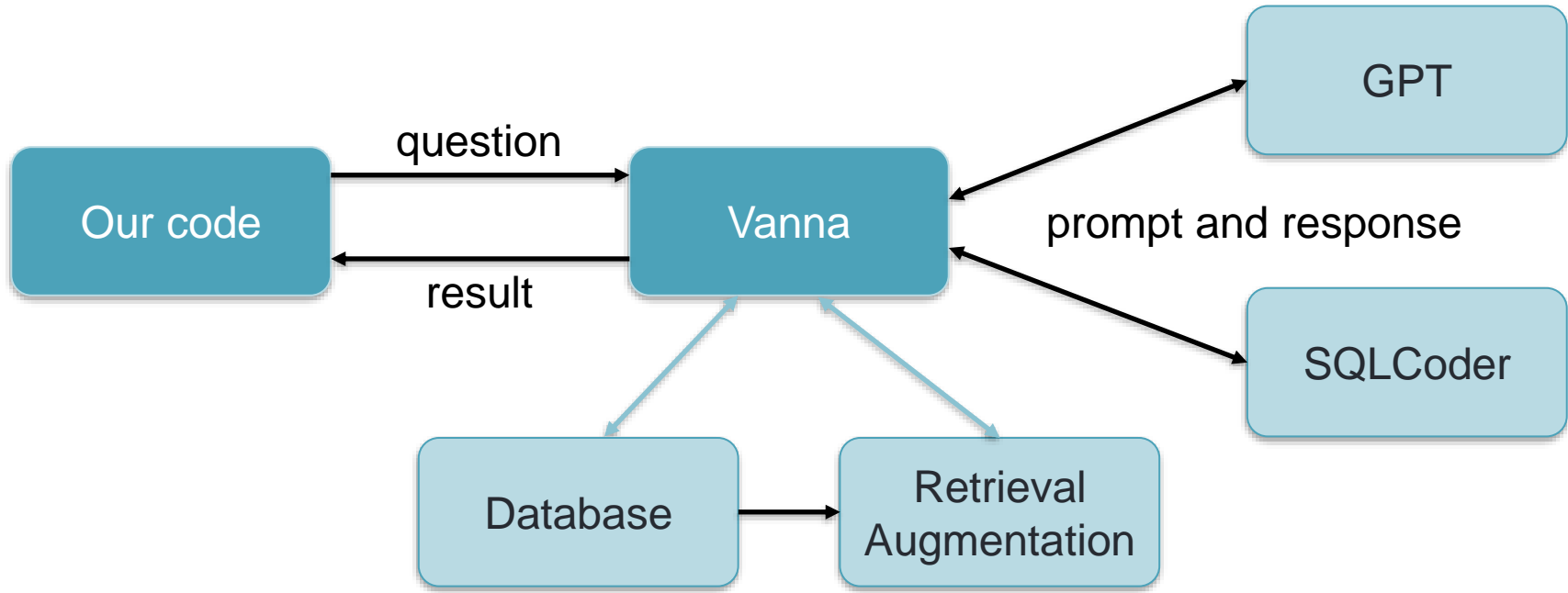
EVALUATION SETUP

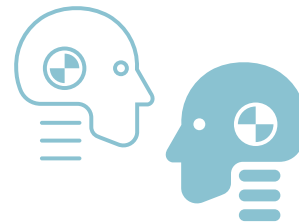


EVALUATION SETUP



EVALUATION SETUP

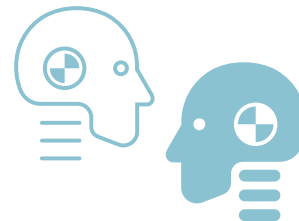




TWO CANDIDATES FOR EVALUATION

DEFOG'S SQLCODER

- Open-source model **fine-tuned** for Text2SQL task from Meta's Llama
- Great performance on authors' **own benchmark** framework
- No literature



TWO CANDIDATES FOR EVALUATION

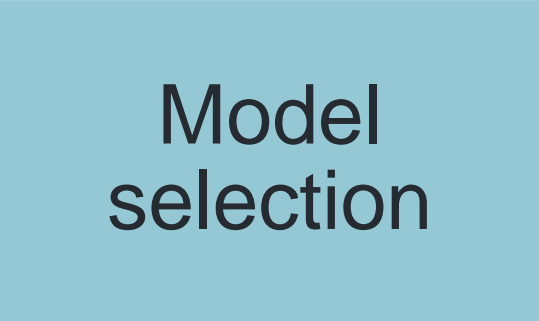
DEFOG'S SQLCODER

- Open-source model **fine-tuned** for Text2SQL task from Meta's Llama
- Great performance on authors' **own benchmark** framework
- No literature

OPENAI'S GPT

- Research has shown promising **zero-shot capability** [Liu et al., 2023]
- **General-purpose** model might be better at inferring from complex questions

CONSIDERATIONS



Model
selection

CONSIDERATIONS

Model
selection

Result
interpretation

CONSIDERATIONS

Model
selection

Result
interpretation

Database
design



SUMMARY

WHAT WE'RE ATTEMPTING

- Evaluate well-versed Text2SQL models in **realistic context**
- Highlight **strengths and limits**
- Provide community with **low-barrier** analytical tooling

SUMMARY



WHAT WE'RE ATTEMPTING

- Evaluate well-versed Text2SQL models in **realistic context**
- Highlight **strengths and limits**
- Provide community with **low-barrier** analytical tooling

WHAT COULD BE NEXT

- Natural language implementation with **explained results**
- Query completion
- Database design based on **schema hallucination**
[Kothyari et al., 2023]