Competence Network for Artificial Intelligence
Kompetenznetzwerk für künstliche Intelligenz
Réseau de compétences en intelligence artificielle
Rete di competenze per l'intelligenza artificiale

Data Science
Competence Center DSCC

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Bundesamt für Statistik BFS
Office fédéral de la statistique OFS
Ufficio federale di statistica UST
Federal Statistical Office FSO

Statbot.Swiss

A Chatbot to query Swiss Open Data

Raphaël de Fondeville, Yara Abu Awad

UNECE Generative AI Project Meeting, 10.06.2024

Agenda

1. Background & Project Goals
2. Challenges and Results
3. Project’s Outputs

Agenda

1. **Background & Project Goals**
2. Challenges and Results
3. Project’s Outputs

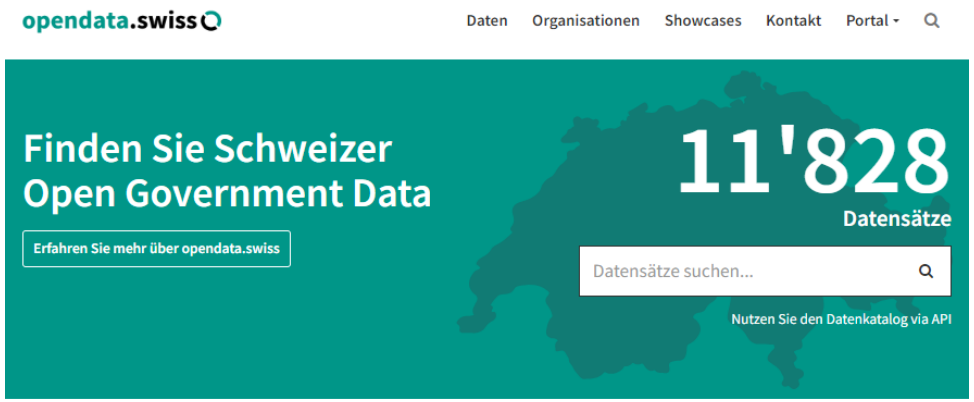
The Digital Switzerland Strategy

> *"Switzerland consistently prioritises digital offerings for the benefit of all people, regardless of gender, age or origin."*
>
> — Digital Switzerland Strategy

- Access to Open Government Data ought to be facilitated to the whole population.
- The Statbot.Swiss project was funded within the framework of the "E-Government Strategy Switzerland **2020**-2023" of Digital Public Services Switzerland.

Open Government Data

- OpenData.swiss is the Swiss public administration's central portal for open government data.
- It (currently) lists 11828 datasets covering 14 categories, e.g., health, environment and economy



https://opendata.swiss
(Access on 06.06.2024)

StatBot.swiss: Principle

A chatbot to automatically query databases

The opendata.swiss portal refers to all open data held by the Swiss authorities.

Objective: easy access to the datasets of the Open Government Data Platform.

Current practice: Datasets must be found on the web portal and downloaded.

➡ Requires a high level of skill in data manipulation.

New solution: Academic partnership to develop a chatbot that automatically queries data sets and provides a simple and structured response.

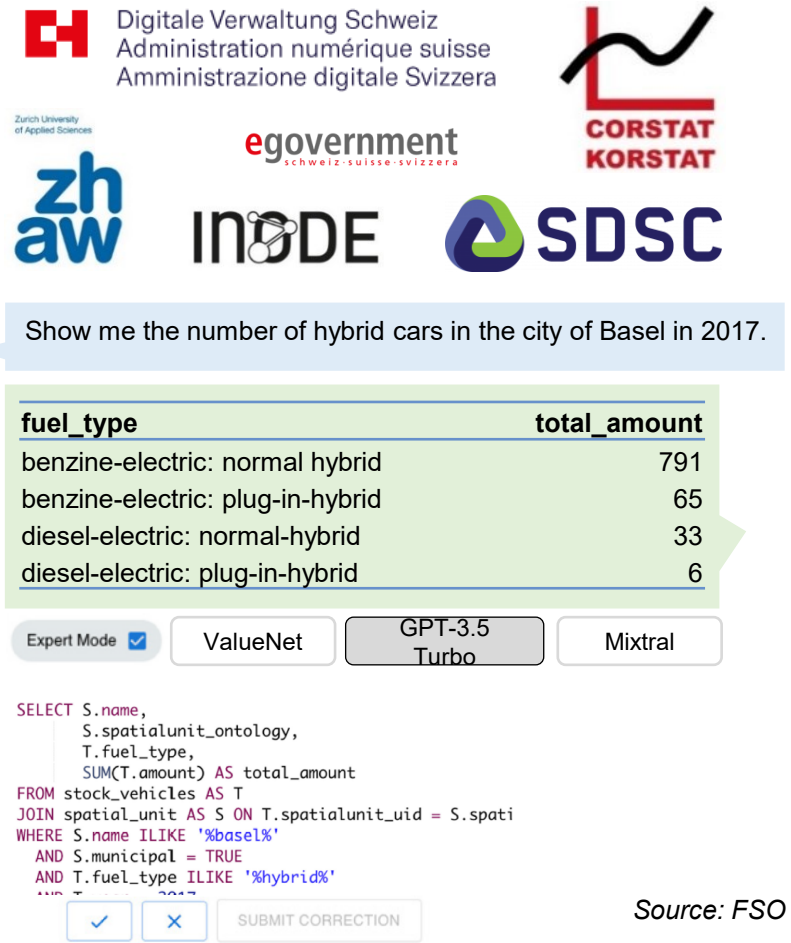
Source: FSO

StatBot.swiss: Implementation

Overall objectives

- Creation of a statistical bot for public statistics based on DCAT (Data Catalog Vocabulary) and in the context of OGD (Open Government Data).

- Evaluate whether it is possible to translate Natural Language into SQL (Structured query language) requests.



Source: FSO

Agenda

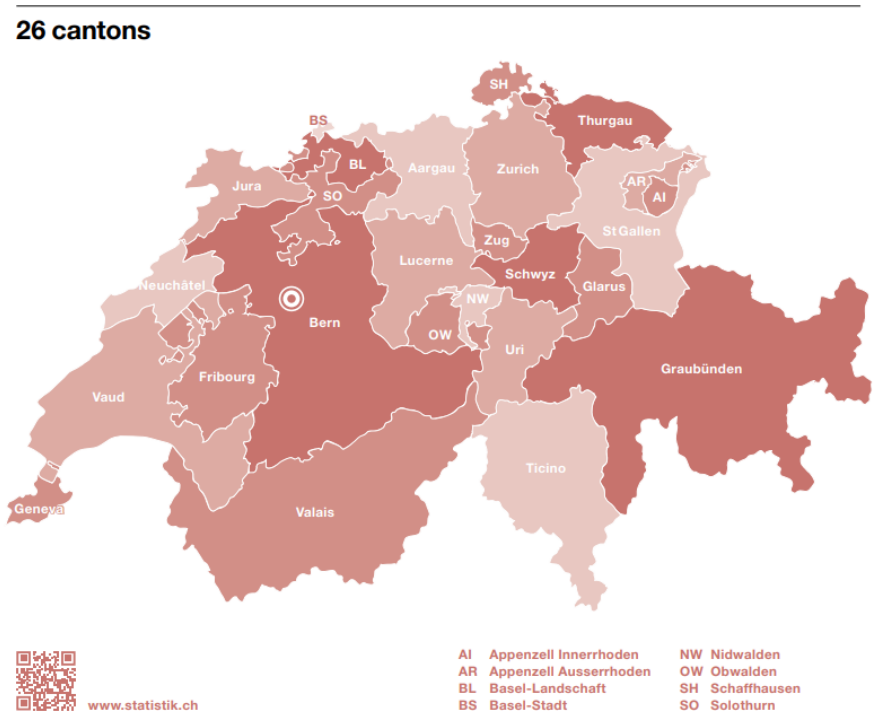
1. Background & Project Goals
2. **Challenges and Results**
3. Project’s Outputs

Challenge 1: Multilingualism

Within Switzerland, state power is shared between **the federal government**, the **cantons** and the **communes**.

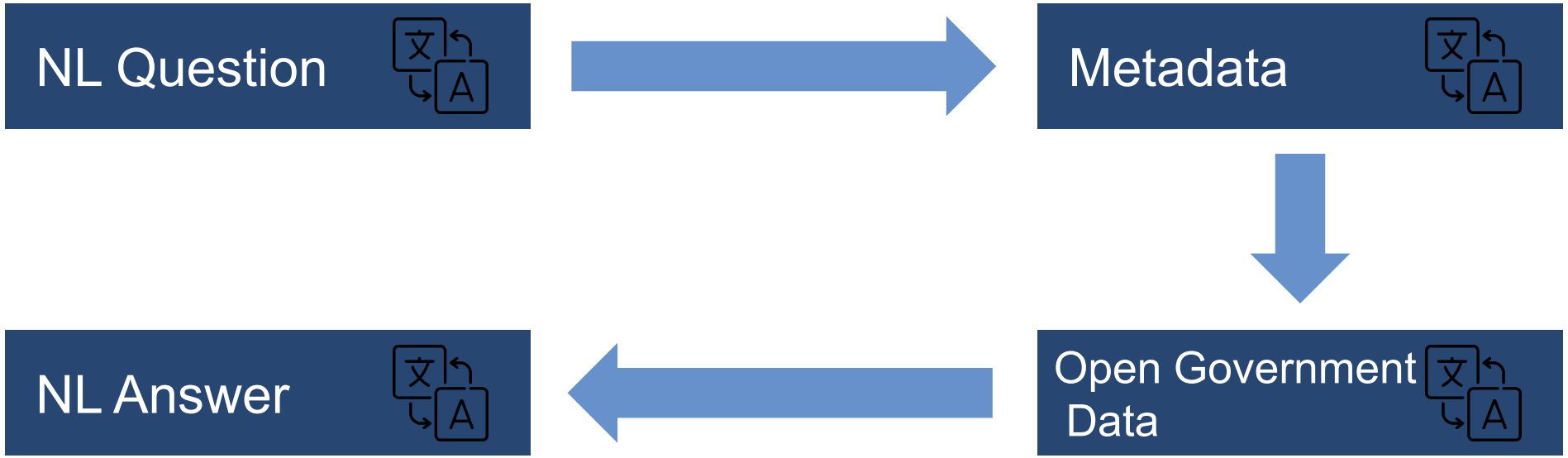
There are four official languages: German, French, Italian and Romansh.

This means that Swiss statistical data faces hurdles in harmonization and standardization, posing challenges in multilingual contexts.



Source: The Swiss Federal Chancellery

Challenge 1: Multilingualism



The chatBot should handle multilingualism at all steps!

Challenge 2: Competences

Interdisciplinary project requiring expertise in multiple fields to create an innovative solution to a complex problem.

Public - Academia Consortium

- NSO: The Data Science Competence Center of the FSO (lead)
- Cantonal Statistical offices: CORSTAT and the canton of Zurich (interoperability).
- Universities: the Swiss Data Science Center (data management) and the Zurich University of Applied Sciences (NL-to-SQL).

egovernment schweiz·suisse·svizzera

Digitale Verwaltung Schweiz
Administration numérique suisse
Amministrazione digitale Svizzera

CORSTAT KORSTAT

Zurich University of Applied Sciences
zh aw

INODE SDSC

Challenge 2: Competences

Quick (non-exhaustive) history of LLMs

- 2017: Scientific paper introducing Transformers.
- 2018: Introduction of the first LLM by OpenAI, namely the Generative Pre-trained Transformers (GPT-1).
- 2020: GPT-3 made available through API calls.
- **2022**: Release of ChatGPT, interactive agent.
- 2022: First source-available models (,e.g, Llama by Meta)
- 2023: Miytral by Mistral AI.

Recall that the Statbot project was initiated in 2020!

Challenge 2: Competences

INODE – Intelligent Open Data Exploration

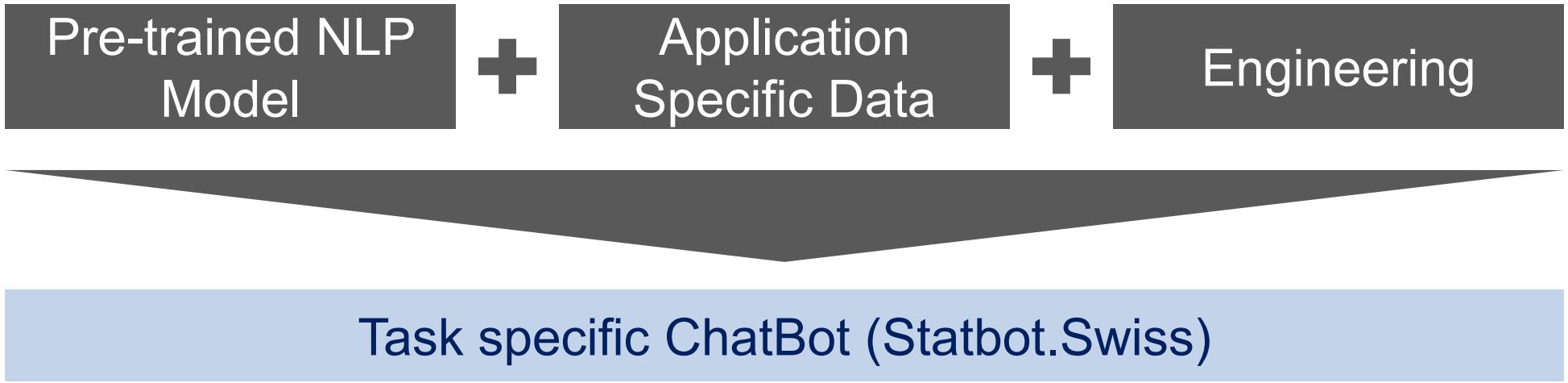
- Inode is a H2020 European research project relying on a consortium of university whose leading house is the ZHAW in Zurich.

> *A classic unified, comprehensive platform that provides extensive access to **open datasets** through **natural language queries** in the fields of Cancer Biomarker Research, Research and Innovation Policy Making and Astrophysics; for a wide range of users from larger scientific communities to public.*

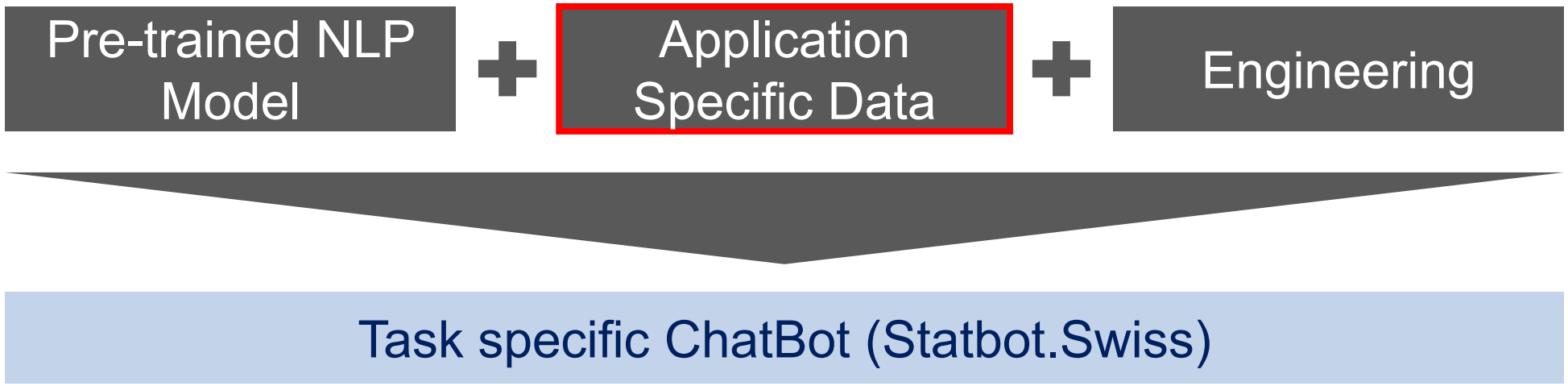
- The INODE platform relies on an NL-to-SQL algorithm “ValueNet” which re-purpose a pre-trained transformers architecture.

Brunner, U., & Stockinger, K. (2021). ValueNet : a natural language-to-SQL system that learns from database information. *Proceedings of the 37th ICDE*, 2177–2182.

Challenge 3: Data



Challenge 3: Data



Challenge 3: Data

- Me: Do you have data?
  - A: Yes plenty! On the XX (Opendata.swiss) website!
- Me: Do you have the **right** data?
  - A: Well, …

What is the number of cars with more than 4 cylinders?

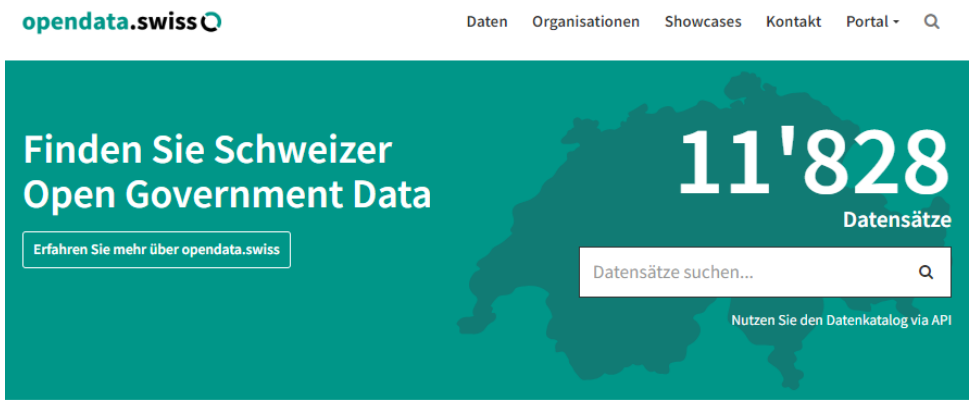
```
SELECT COUNT(*)
FROM cars_data
WHERE cylinders > 4
```

We need pairs of NL-SQL questions and answers…

Challenge 3: Data

- 22 German and 13 English datasets were selected from opendata.swiss.
- Automatic generation of NL-SQL pairs is not reliable.
- Text-to-SQL pairs were generated manually.

Data generation is expensive because “label makers” must be proficient in SQL (which must be thoroughly checked)!



https://opendata.swiss
(Access on 06.06.2024)

Challenge 3: Data

- 455 NL/SQL-pairs covering a total of 35 OGD datasets.
- Different number of pairs per database depending on the dataset complexity.
- Example: 23 pairs for marriage_citizenship, only 5 for greenhouse_gas_emissions_through_consumption.

```
-- Wieviele Angiographiegeräte gab es 2013 in der Schweiz?
SELECT SUM(anzahl_gerate) as anzahl_gerate_in_2013
FROM medizinisch_technische_infrastruktur as T
JOIN spatial_unit as S on T.spatialunit_uid = S.spatialunit_uid
WHERE S.country=TRUE
  AND jahr = '2013'
  AND T.genutzte_infrastruktur LIKE '%Angiographie%';
```

Challenge 3: Data

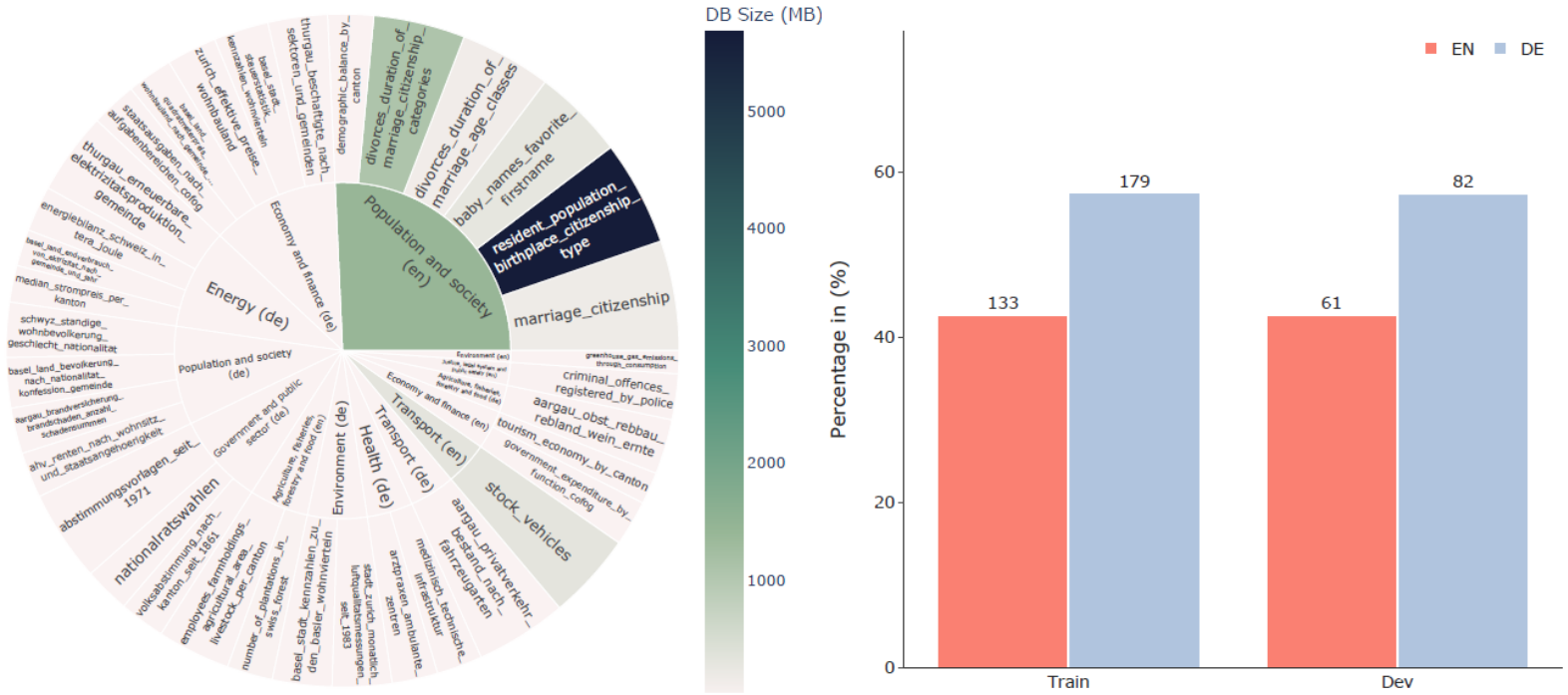


Figure 1: **Dataset distribution:** (a) Left: Knowledge domains, (b) Right: Distribution of natural language/SQL-pairs over the train and development sets. EN = English, DE = German. The numbers on top of the bars denote the number of Text-to-SQL pairs.

F. Nooralahzadeh et al. (2024). StatBot.Swiss: Bilingual Open Data Exploration in Natural Language. *To appear in the proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Challenge 3: Statbot.Swiss Dataset

The statbot.swiss dataset is highly complex and covers more realistic NL questions and more complex SQL syntax than state-of-the art benchmarks.

Dataset	#Examples (#DBs)	#Tables (#Rows)/DB	Language	Function	Granularity	Knowledge	WITH-Queries
WikiSQL(Zhong et al., 2017)	80,654 (26,521)	1 (17)	EN	✗	✗	✗	✗
SPIDER (Yu et al., 2018)	10,181 (200)	5.1 (2K)	EN	✗	✗	✗	✗
KaggleDBQA (Lee et al., 2021)	272 (8)	2.3 (280K)	EN	✗	✗	✓	✗
ScienceBenchmark (Zhang et al., 2024)	5,332 (3)	16.7 (51M)	EN	✗	✗	✓	✗
BIRD (Li et al., 2023)	12,751 (95)	7.3 (549K)	EN	✓	✗	✓	✗
StatBot.Swiss	455 (35)	2 (1.4M)	EN, DE	✓	✓	✓	✓

Table 1: Comparison between StatBot.Swiss and existing state-of-the-art Text-to-SQL datasets. `Function` refers to SQL built-in functions. `Knowledge` stands for the necessity of external knowledge reasoning from the model. `Granularity` refers to the degree of specificity or the level of details. `WITH-Queries` refer to complex sub queries that are broken up into smaller ones. EN = English, DE = German.

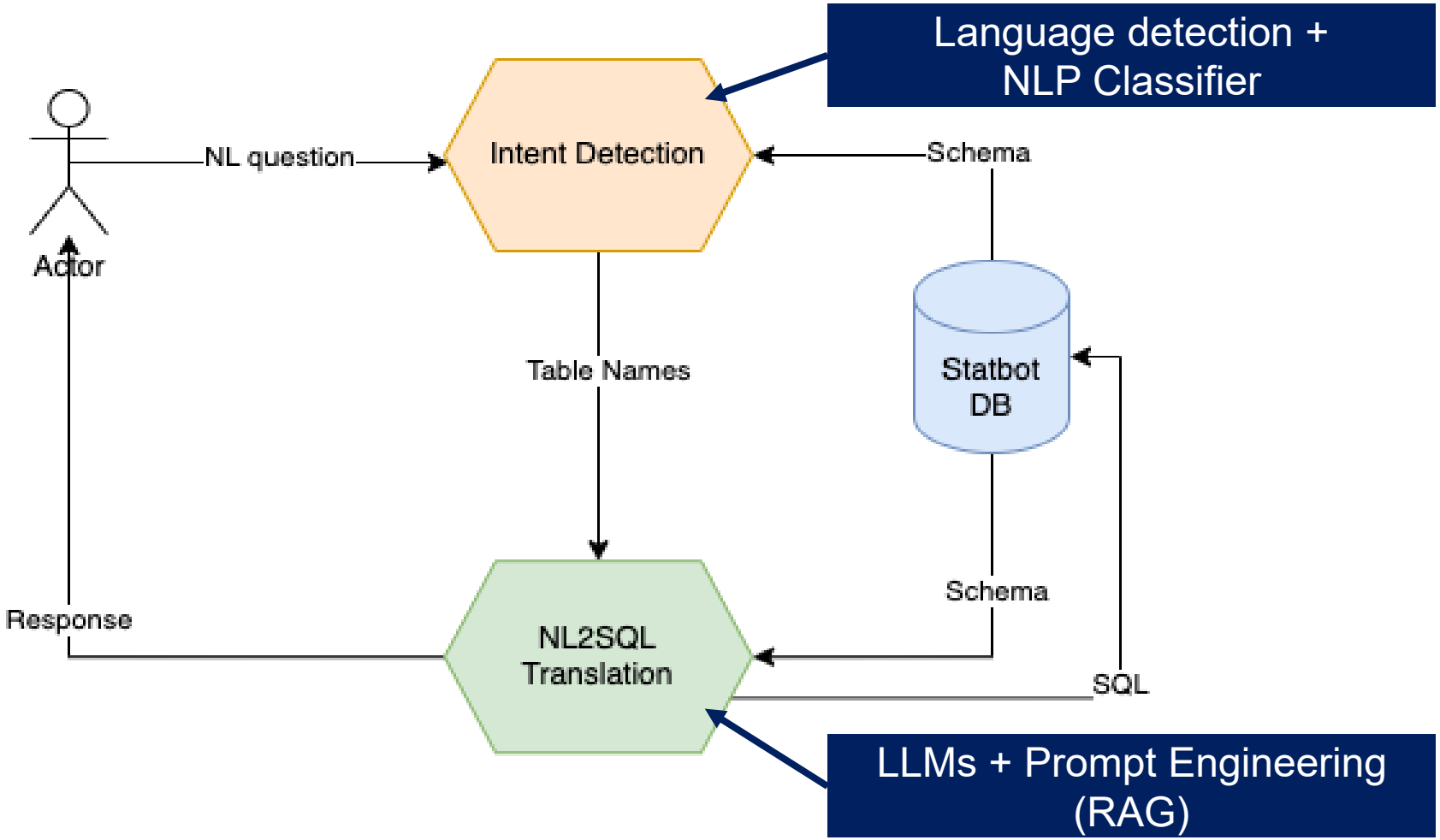
F. Nooralahzadeh et al. (2024). StatBot.Swiss: Bilingual Open Data Exploration in Natural Language. *To appear in the proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Challenge 4: Text-to-SQL Translation

- Original project intent was to use ValueNet (vanilla transformer architecture).
- In 2022, the release of ChatGPT rendered ValueNet's architecture obsolete …

Statbot.swiss had to adapt to this new reality!

Challenge 4: Text-to-SQL Translation



Challenge 5: Performance Evaluation

Q: What were the number of hybrid cars in the city of Basel in 2017?

Region	Total
Basel City	895
Canton Basel-City	2158
Canton Basel-Campaign	1059

Soft

⊃

Region	Total
Basel City	895

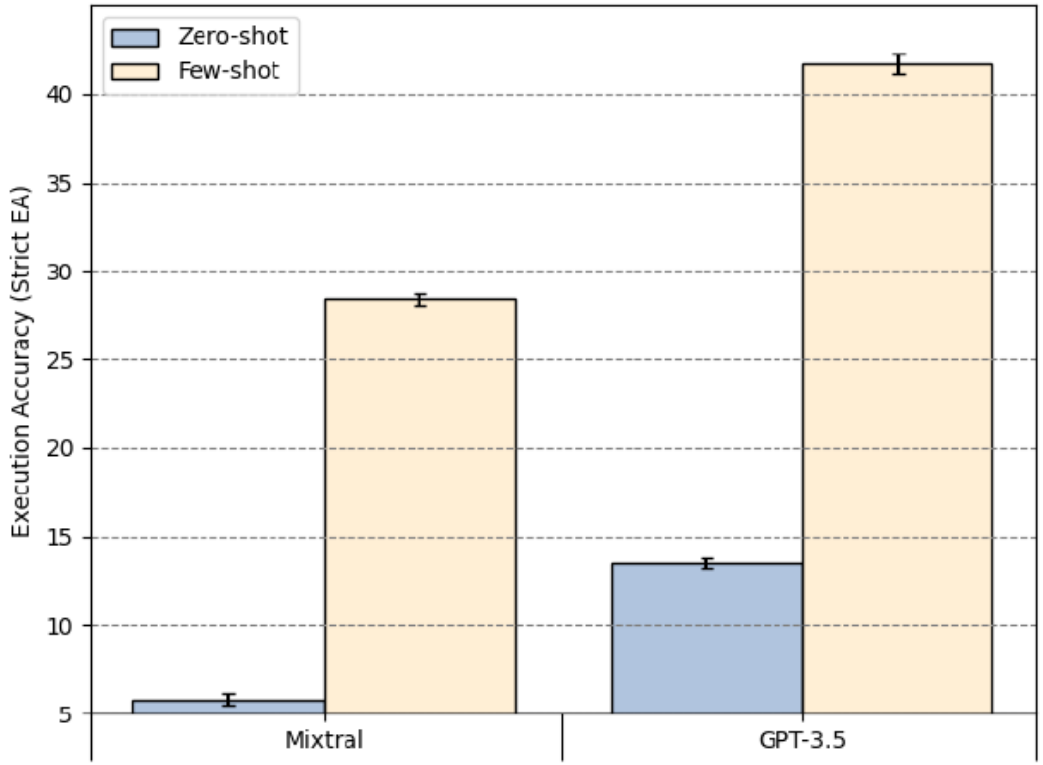
Strict

?

Region	Fuel Type	Total
Basel City	Normal-hybrid	824
Basel City	Plug-in-hybrid	71

Strict underestimate, Soft overestimate performance …

Challenge 5: Performance Evaluation (strict)



F. Nooralahzadeh et al. (2024). StatBot.Swiss: Bilingual Open Data Exploration in Natural Language. *To appear in the proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Challenge 5: Performance Evaluation (strict)

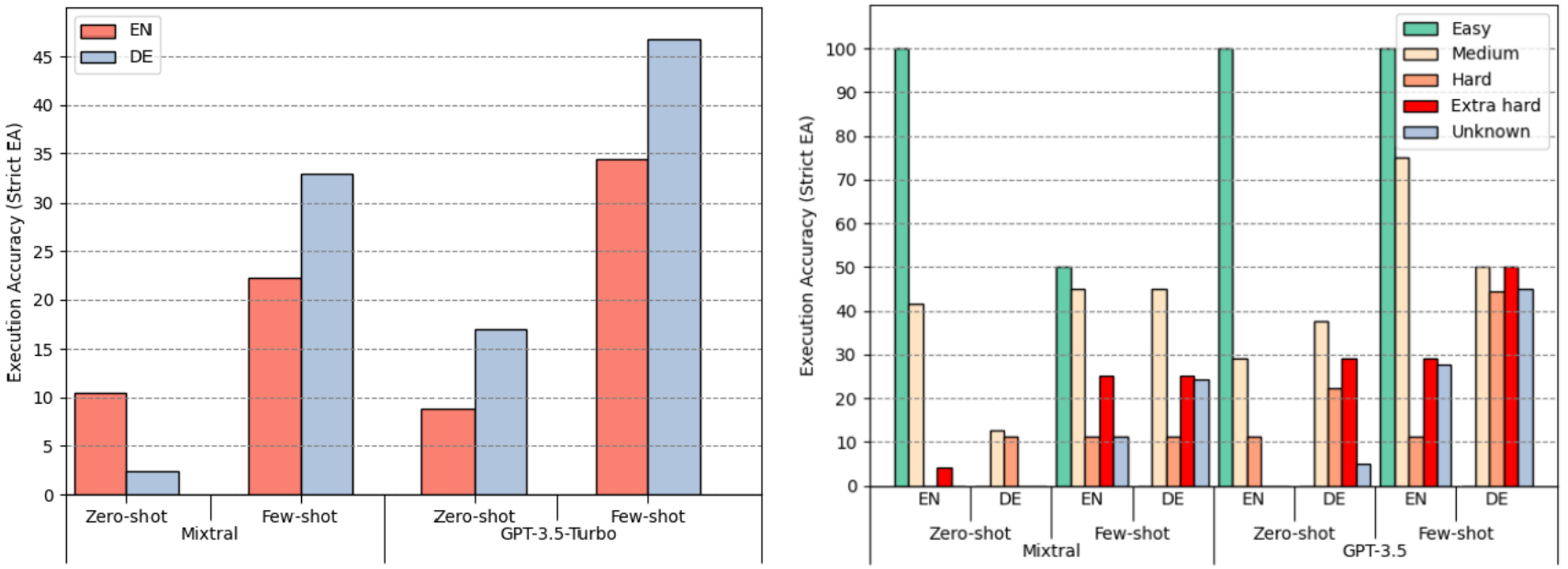


Figure 3: (Left) Strict execution accuracy (EA_{strict}) for each language. (Right) EA_{strict} for each language per query hardness level. All metrics are computed on the development set for zero-shot and few-shot prompting strategies (6-shot in Mixtral, 5-shot in GPT-3.5).

F. Nooralahzadeh et al. (2024). StatBot.Swiss: Bilingual Open Data Exploration in Natural Language. *To appear in the proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Challenges: Summary

- Challenge 1: multilingualism is complex!
- Challenge 2: get the competences where they are!
- Challenge 3: the “right” data is expensive!
- Challenge 4: need to adapt as technology evolve!
- Challenge 5: performance is tricky to evaluate and LLM maturity is not there yet…

Agenda

1. Background & Project Goals
2. Challenges and Results
3. **Project's Outputs**

Project’s outputs - Stay Tuned!

StatBot.Swiss: Bilingual Open Data Exploration in Natural Language

Farhad Nooralahzadeh[1†], **Yi Zhang**[1†], **Ellery Smith**[1], **Sabine Maennel**[2]
Cyril Matthey-Doret[2], **Raphaël de Fondville**[3], **Kurt Stockinger**[1]
[1]Zurich University of Applied Sciences, Switzerland
[2]Swiss Data Science Center, Switzerland
[3]Federal Statistical Office, Switzerland
{farhad.nooralahzadeh, yi.zhang, kurt.stockinger}@zhaw.ch

Abstract

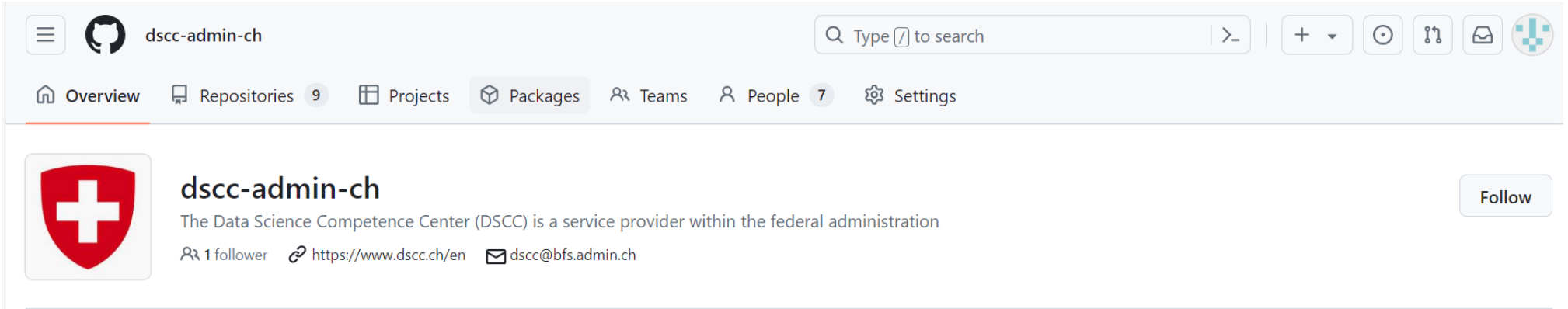
The potential for improvements brought by Large Language Models (LLMs) in Text-to-SQL systems is mostly assessed on monolingual English datasets. However, LLMs' per-

methodological differences in order to know which data are more suitable for the intended usage and more importantly be capable of importing and analyzing the data through statistical software such as a spreadsheet, Python, R or SAS, which all require

To appear in ACL Findings 2024 in August!

Project's outputs - Stay Tuned!

- Project's report will be released shortly by Digital Switzerland (many thanks to my colleague Yara!)



- Code will be available on our GitHub very soon…

Contributions are welcome!

Competence Network for Artificial Intelligence
Kompetenznetzwerk für künstliche Intelligenz
Réseau de compétences en intelligence artificielle
Rete di competenze per l'intelligenza artificiale

Data Science
Competence Center DSCC

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Bundesamt für Statistik BFS
Office fédéral de la statistique OFS
Ufficio federale di statistica UST
Federal Statistical Office FSO

Questions?

www.dscc.ch
www.dscc.ch/fr
www.dscc.ch/it
www.dscc.ch/en

@StatSchweiz / @statsvizzera
@statsuisse / @swissstatistics

@SwissStats

@SwissStatistics

@bfs-ofs

Evaluating Query Complexity

Spider Hardness Metric categorizes SQL queries into 4 levels: easy, medium, hard, extra hard.

Difficulty defined based on the number of SQL components, selections, and conditions.

Easy

What is the number of cars with more than 4 cylinders?

```
SELECT COUNT(*)
FROM cars_data
WHERE cylinders > 4
```

Meidum

For each stadium, how many concerts are there?

```
SELECT T2.name, COUNT(*)
FROM concert AS T1 JOIN stadium AS T2
ON T1.stadium_id = T2.stadium_id
GROUP BY T1.stadium_id
```

Hard

Which countries in Europe have at least 3 car manufacturers?

```
SELECT T1.country_name
FROM countries AS T1 JOIN continents
AS T2 ON T1.continent = T2.cont_id
JOIN car_makers AS T3 ON
T1.country_id = T3.country
WHERE T2.continent = 'Europe'
GROUP BY T1.country_name
HAVING COUNT(*) >= 3
```

Extra Hard

What is the average life expectancy in the countries where English is not the official language?

```
SELECT AVG(life_expectancy)
FROM country
WHERE name NOT IN
   (SELECT T1.name
    FROM country AS T1 JOIN
    country_language AS T2
    ON T1.code = T2.country_code
    WHERE T2.language = "English"
      AND T2.is_official = "T")
```

Source: [1809.08887] Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task (arxiv.org)

Example with hardness "unknown"

Query Types	[db_id] | *Question* | Query
1. GROUP BY > 1 column	[volksabstimmung_nach_kanton_seit_1861] *Welche Kantone haben 2023 gegen das Bundesgesetz über Klimaschutz gestimmt und wieviel Prozent Ja Stimmen gab es dort jeweils?* `SELECT S.name_de AS kanton_gegen_klimaschutzgesetz, T.ja_in_prozent FROM volksabstimmung_nach_kanton_seit_1861 AS T JOIN spatial_unit AS S ON T.spatialunit_uid = S.spatialunit_uid WHERE S.canton = 'TRUE' AND LOWER(T.vorlage) LIKE '%bundesgesetz%klimaschutz%' AND T.jahr = 2023 AND T.ja_in_prozent <= 50 GROUP BY S.name_de, T.vorlage, T.jahr, T.ja_in_prozent ;`

Evaluation of LLMs Results

- Effect of Selection Method: Similarity selection method generally outperforms the Random selection method
- Impact of exemplars on Model Performance:
  - Zero-shot → few-shot: leads to significant improvement
  - Maximal performance is achieved with 5 examples in GPT-3.5, and 6 examples in Mistral using similarity-based selection

Evaluation of LLMs: Setup

- Different strategies were examined for **in-context learning** using two **large language models.**
  - **Zero-shot** (input prompt is limited to a natural language question along with its corresponding database metadata)
  - **Few-shot** (the LLMs' prompts include a small number of natural language and SQL pairs which are inserted between the representation of the database and the target question): Random Selection, Similarity-based Selection
- Experiment was restricted to the **GPT-3.5-Turbo-16k** model, and the **Mixtral-8x7B Instruct** model

Evaluation of LLMs: Database Info

- **Providing prior knowledge** i.e. inclusion of a database's metadata such as table relationships and variable encoding is crucial for enabling effective prompting
- Textual representation of the **database information** using CREATE statement
- Enhance the representation of the **database structure**: integrating metadata column information, such as column name and column description

Lessons Learned

- Handling multilingualism is far from trivial (i.e. question and database in different languages).
- Creating good training data to fine-tune models is crucial but proved to be very time consuming.
- Flexibility is crucial - research around machine learning moves quickly and a change of course may be crucial for best results i.e. during project implementation developments around LLMs rendered previous approaches partially obsolete.

Conclusion

This tool was meant to provide a reliable source of statistical information by making open data more accessible. However, due to the current error rate of state-of-the-art Text-to-SQL solutions based on LLMs, it is currently not suitable as an official source of information because official sources are held to a higher standard. Until this technology becomes more accurate, we would not recommend its use for querying statistical data that requires close to 100% accuracy.

For more information

Link to publication:

StatBot.Swiss: Bilingual Open Data Exploration in Natural Language (arxiv.org)

Link to github:

dscc-admin-ch/statbot.swiss: This repository contains all datasets and evaluations for "StatBot.Swiss: Bilingual Open Data Exploration in Natural Language" paper.