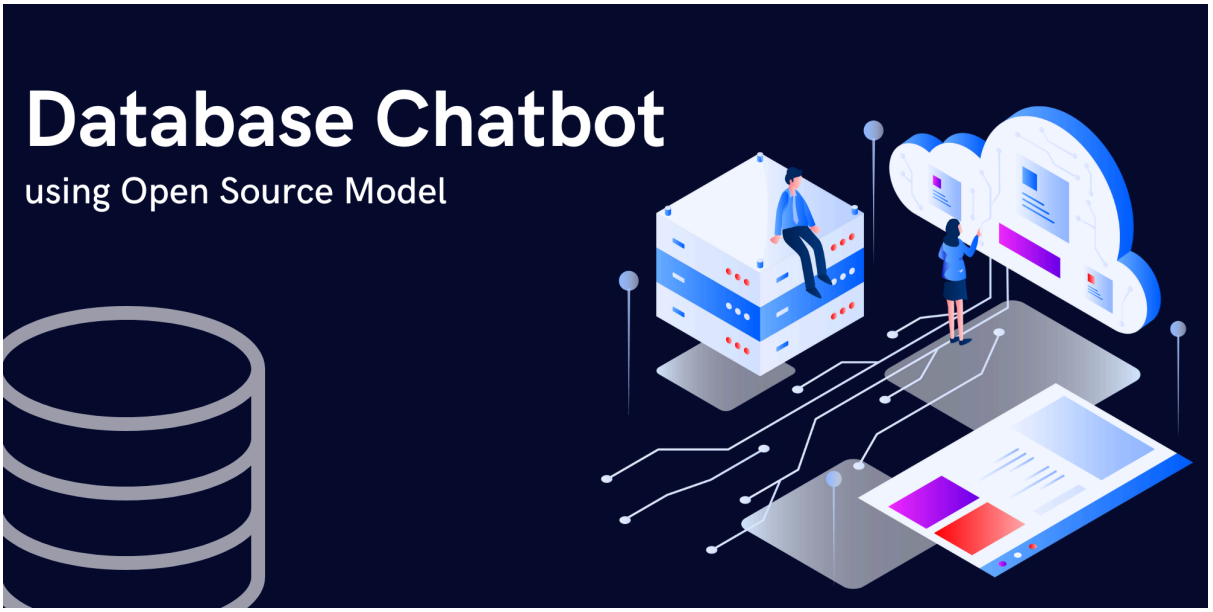
Database Chatbot



Background
Text-to-SQL is a task to translate a user's query spoken in natural language into SQL automatically. It is the project that I'm working on at Microsoft.

If this problem is solved, it's going to be widely useful because the vast majority of data in our lives is stored in relational databases. In fact, healthcare, financial services, and sales all seem to use relational databases exclusively.

Also, writing SQL queries can be prohibitive to non-technical users.

Dataset benchmark
Many models or approaches have been created to do text to SQL.

With multiple approaches and solutions flooding the market, we are left with the problem of evaluation. Which approach is most efficient? Which one more reliably produces accurate answers? Which one adapts to different datasets best? To help answer these questions, the open-source industry and academia put forth several benchmarks, but the three most used today are:

- WikiSQL
- Spider

Overview WikiSQL
Introduced by Salesforce in late 2017, WikiSQL was the first truly large compendium of data built for the text-to-SQL use case. However, it has one major drawback: simplicity.

All of the provided SQL queries are exceedingly simple, with only SELECT, FROM, and WHERE clauses. Furthermore, the tables in the dataset have no linkages to other tables. Although models trained on WikiSQL can still work on new databases, they can only answer simple natural language questions that then translate into simple SQL queries.

Description

- Size: 154.74 MB
- Data points: 87,726 unique question-SQL pairs
- Databases: 24,241 tables from Wikipedia
- Domains: 1

Overview Spider

The Spider dataset aims to cover some of the shortcomings of the WikiSQL dataset. Developed through the efforts of 11 Yale students spending over 1,000 man hours, the Spider dataset introduces two critical elements: complexity and cross-domainality.

- Complexity: The SQL queries go beyond the straightforward SELECT and WHERE clauses that WikiSQL is limited to, covering the more complex GROUP BY, ORDER BY, and HAVING clauses along with nested queries. Furthermore, all databases have multiple tables linked through foreign keys, allowing for complicated queries that join across tables.
- Cross-domainality: With 200 complex databases across a high number of domains, Spider is able to include unseen databases in the test set, allowing us to test the model's generalizability.

Description

- Size: 919.2 MB
- Data points: 10,181 questions and 5,693 unique complex SQL queries
- Databases: 200
- Domains: 138

Model

Based on the explanation of the benchmark dataset above, here are several popular or best models based on the WikiSQL and Spider datasets.

Based on WikiSQL

Leaderboard

LINK LINK

Model	Base model	Num of parameters	Year	Description	LINK
NL2SQL-BERT	BERT	~110M	2019		LINK LINK
TAPEX-Large	BART	~400M	2021	Pretrained-model (weak supervision)	LINK LINK LINK
ReasTAP-Large		~800M	2022	(weak supervision)	LINK LINK LINK

TAPAS-Large	BERT	~110M	2020	Pretrained-mode (weak supervision)	LINK LINK

Model	Execution Accuracy	Exact Match Accuracy
NL2SQL-BERT	89.2	83.7
TAPEX-Large	89.5	
ReasTAP-Large	89.2	
TAPAS-Large	83.6	

Note:

- Strict denotation accuracy is the percentage of predicted queries that when executed produce the same results as the reference query. [LINK]
- **Exact Set Match Accuracy (EM)** The exact set match accuracy (without values) is calculated by comparing the ground-truth SQL query and the predicted SQL query
- **Execution Accuracy (EX)** Execution accuracy (with values) is calculated by comparing the output results of executing the ground-truth SQL query and the predicted SQL query on the database contents shipped with the test set.

Based on Spider Leaderboard
LINK LINK

Model	Base model	Num of parameters	Year	Description	LINK
RESDSQL-3B + NatSQL		~3 billion	2023		LINK LINK
LEVER + Codex			2023	Using openAPI	LINK LINK LINK
RASAT+PICARD		~3 billion	2022		LINK LINK
Graphix-3B + PICARD		~3 billion	2023		LINK LINK LINK
T5-3B + PICARD	T5	~3 billion	2021		LINK LINK LINK

Model	(Dev) Execution Accuracy	(Dev) Exact Match Accuracy	(Test) Execution Accuracy	(Test) Exact Match Accuracy
RESDSQL-3B + NatSQL	80.5	84.1	72.0	79.9
LEVER + Codex	81,9			
RASAT+PICARD	80,5	75,3	75,5	70,9
Graphix-3B + PICARD	81,0	77,1	77,6	74,0
T5-3B + PICARD	79,3	75,5	75,1	71,9