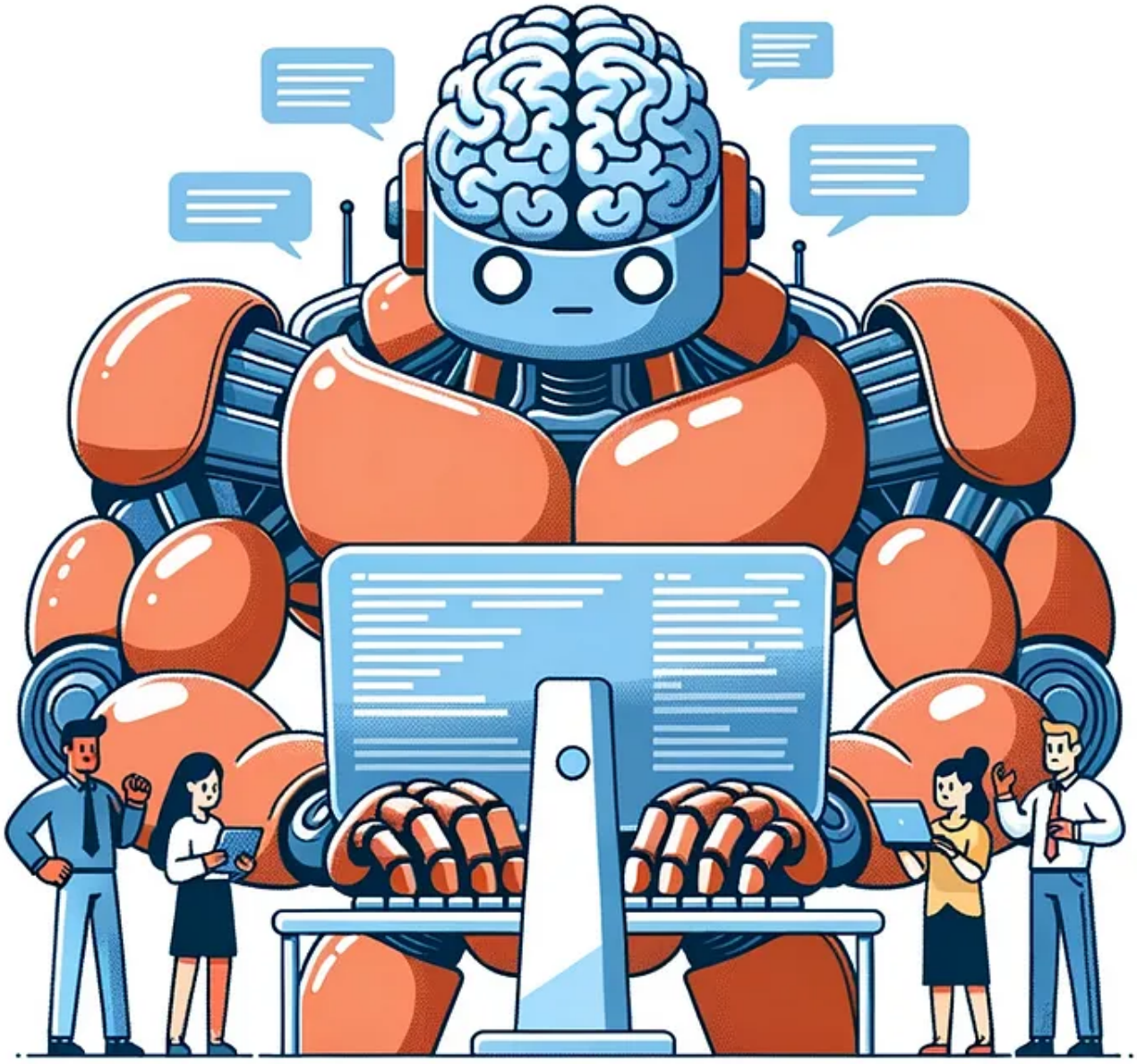
Comparing Open-source LLMs for NL-to-SQL

Mo Pourreza · Follow
Published in Dataherald
4 min read · Oct 31



Generated BY DALLE3

Background

The NL-to-SQL (Natural Language to Structured Query Language) task is a hard problem in the field of natural language processing (NLP). It involves transforming natural language questions into SQL queries, which can then be executed against relational databases to answer the question. This task is a specialized subfield within NLP and is closely related to the broader domain of natural language understanding (NLU) and the interface between natural language and databases (NLIDB).

With the recent advances in the development of Large Language Models (LLMs), such as GPT-4, Llama2, and Falcon the focus in industry and academia excel for NL-

to-SQL has moved to leveraging these LLMs to generate SQL for real world use cases. This would be immensely powerful as it would allow non-technical users to directly find insights from data.

At Dataherald, we have built an open source natural language-to-SQL engine which can be used with different LLMs, though we are using GPT-4–32K in our enterprise deployments. The current state of the art research for NL-to-SQL such as DAIL-SQL, C3, and DIN-SQL also uses closed-source LLMs like GPT-4 and GPT-3.5-turbo. These models are both expensive and raise data privacy concerns for enterprises. Therefore we set out to see how open-source LLMs, like Llama2 and Mistral stack up against OpenAI's models. The following are our results.

Open-source LLMs

In this blog post, we will explore the capabilities of open-source LLMs (Large Language Models) from various families. The information shared here is derived from three recent papers, namely Battle of the Large Language Models, Text-to-SQL Empowered by Large Language Models, and Decomposed In-Context Learning of Text-to-SQL as well as our own internal testing conducted using a Google Colab A100 GPU.

The list of LLMs looked at was:

1. Llama-7B
2. Llama-33B
3. MISTRAL-7B
4. Alpaca-7B
5. Llama-2-CHAT-7B
6. Llama-2-CHAT-13B
7. Vicuna-7B
8. Vicuna-33B
9. BARD-LAMDA
10. BARD-PALM2

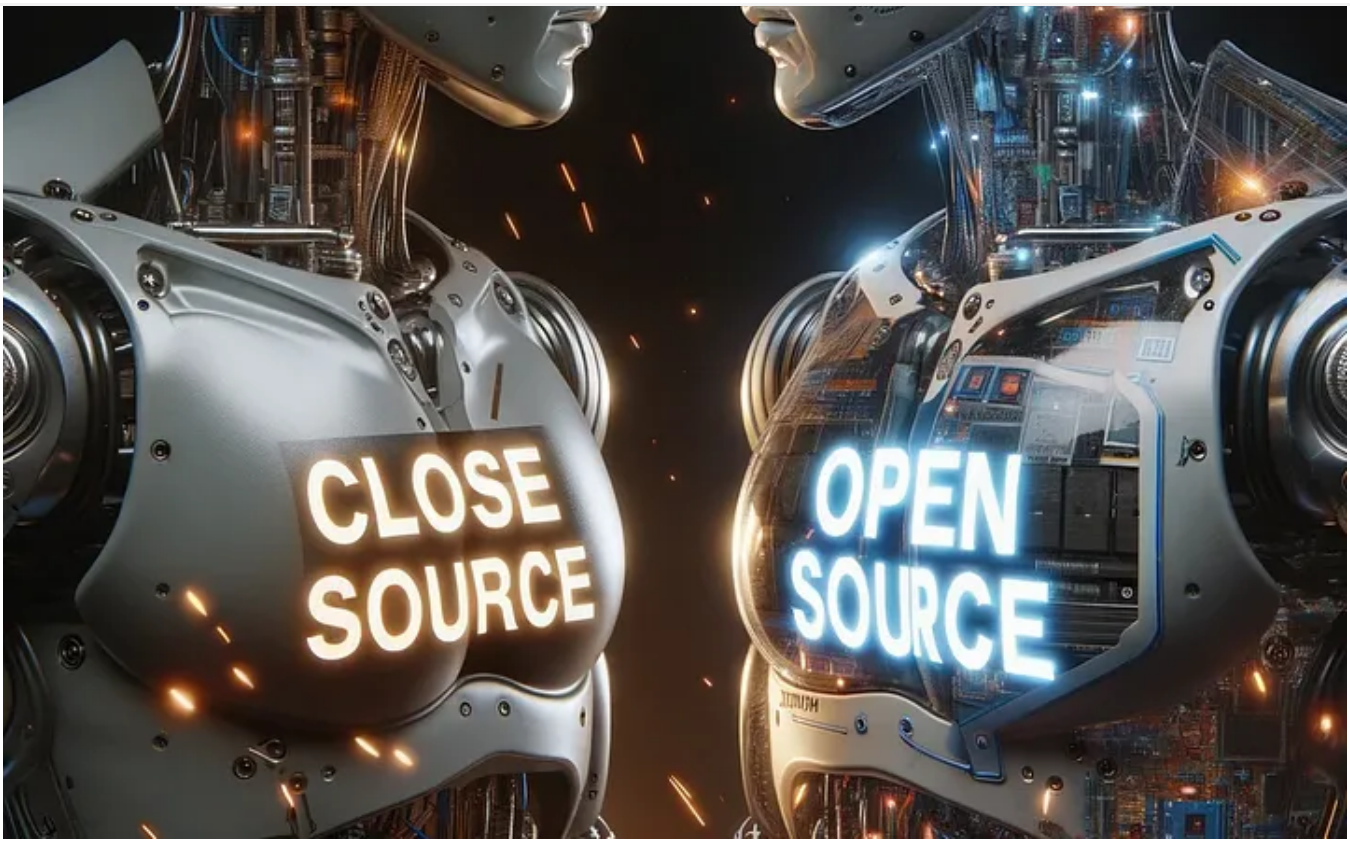
11. GPT-3.5-turbo

12. GPT-4

Some of these models, like Llama, Llama2, and MISTRAL , are pre-trained models similar to GPT-3.5-Turbo, that have undergone supervised fine-tuning and contrastive fine-tuning. The other models have gone through an alignment process, which involves additional instruction tuning, and essentially share the same architecture as the pre-trained ones. In particular, Vicuna, Guanaco, and Alpaca are aligned versions of the Llama model trained on specific datasets.

Zero-shot NL-to-SQL Performance

In this section, we will evaluate the zero-shot performance of open-source LLMs and contrast it with their larger closed-source counterparts. Zero-shot performance refers to the LLM's ability to generate a SQL query solely based on the given question and the corresponding database schema referenced by the question, without any few shot examples in the prompt.



Generated by DALLE3

To ensure a fair comparison, we maintained consistent prompts across all LLMs used in our assessment. The specific prompt employed for reporting zero-shot performance is based on the template proposed by Rajkumar et al., known for its superior performance when compared to other prompt formats.

We evaluated the performance of these models based on execution accuracy, which involves executing both the generated SQL query by the model and the reference SQL query on the database, and then comparing their results. The results are obtained by using the LLMs on the development set of the Spider dataset.

Here are the results:

Performance of LLMs on Dev set of Spider

#	Model	Exec Acc	Type
1	Llama-7B	16.3	Open-source
2	Vicuna-7B	24.0	Open-source
3	Llama2-chat-7B	25.5	Open-source
4	Alpaca-7B	32.1	Open-source
5	Llama2-chat-13B	40.0	Open-source
6	Llama-33B	42.8	Open-source
7	Mistral-7B	43.0	Open-source
8	Vicuna-33B	43.3	Open-source
9	BARD-PaLM2	48.7	Close-source
10	BARD-LAMDA	52.5	Close-source
11	GPT-3.5-turbo	67.2	Close-source
12	GPT-4	72.3	Close-source

Takeaways

The takeaways are clear

1. Closed-source models (GPT models and BARD) significantly outperform their open-source counterparts in NL-to-SQL. It is safe to assume this is due to the higher number of parameters they were trained on.

2. Models aligned with an additional supervised fine-tuning step exhibit a notable performance improvement when compared to their predecessor models. For example, the Alpaca-7B model showcases nearly a 16 percent improvement over its predecessor, Llama-7B. This underscores the potential of fine-tuning to achieve enhanced performance using the same underlying architecture.

3. Newer open source models such as Mistral-7B and Llama2 exhibit superior performances compared to the predecessors and are closing the gap with closed-source models.

Generated by DALLE3

Looking Ahead

For NL-to-SQL workloads while newer open source models are closing the gap with the OpenAI models, there is still a substantial gap when it comes to accuracy out-of the-box. However, it seems that fine-tuning for a specific dataset can greatly improve accuracy even on the same architecture.

At Dataherald, we are planning to add API support to our engine to enable fine-tuning models for NL-to-SQL workloads. We will be sharing results from our fine-tuned models in future posts.

About Dataherald

- Our open-source engine is available on Github.

- You can join our waitlist today for the hosted version.
- Join our Discord to learn more about the project.

Llm OpenAI Text To Sql Agents Evaluation

Follow

Written by Mo Pourreza

101 Followers · Editor for Dataherald

My email: pourreza@ualberta.ca

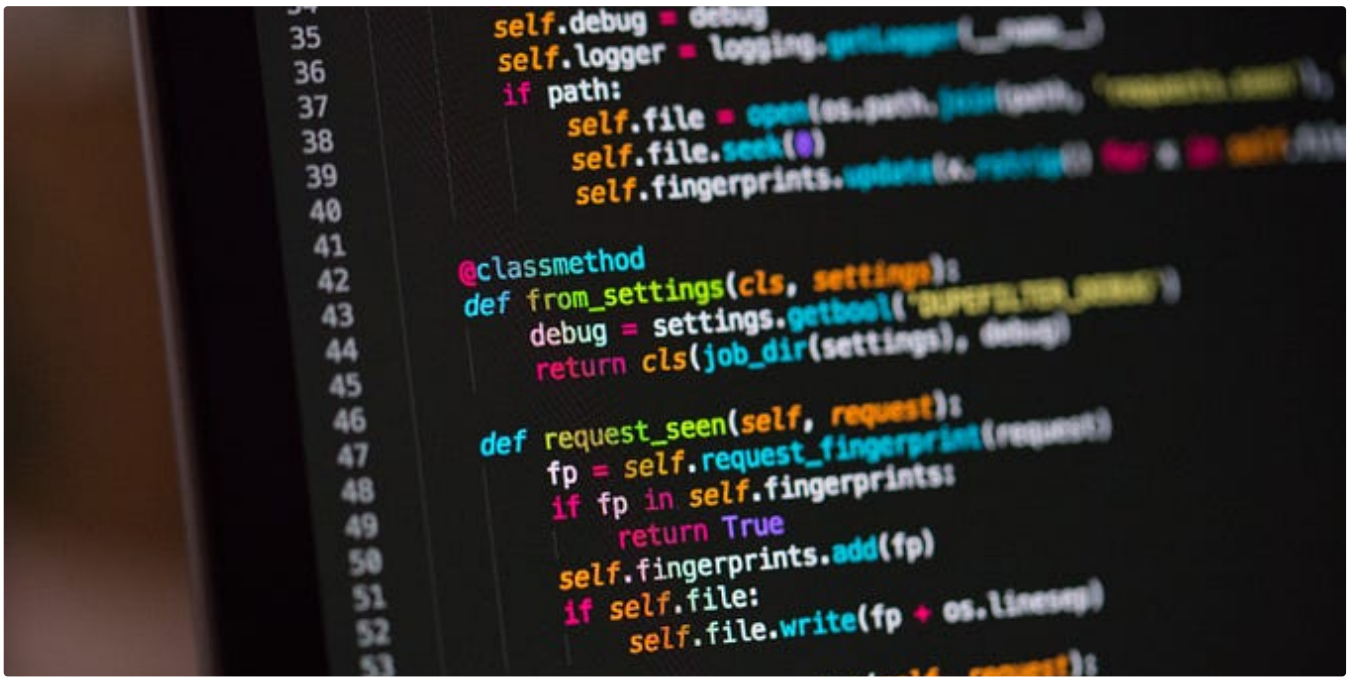
More from Mo Pourreza and Dataherald

Mo Pourreza in Dataherald

How to connect OpenAI's Assistant API to your SQL database

A step-by-step guide

5 min read · Nov 10



Dishen Wang in Dataherald

How to connect LLM to SQL database with LangChain SQLChain

How to Tutorial for using LangChain SQLChain

6 min read · Jun 16

182 6

Dishen Wang in Dataherald

How to connect LLM to SQL database with LlamaIndex

How to Tutorial for using LlamaIndex for text-to-SQL

5 min read · Jun 30

115

Mo Pourreza in Dataherald

Fine-tuning GPT-3.5-Turbo for Natural Language to SQL

Tutorial and comparison with RAG methods

10 min read · Aug 31

Jerry Liu in LlamaIndex Blog

Easily Finetune Llama 2 for Your Text-to-SQL Applications

Llama 2 is a huge milestone in the advancement of open-source LLMs. The biggest model and its finetuned variants sit at the top of the...

7 min read · Aug 17

275 11

Ashhadul Islam in Python in Plain English

Super Quick: LLAMA2 on CPU Machine to Generate SQL Queries from Schema

Using NSQL-Llama-2–7B, to generate SQL queries from free-flowing text instruction.

· 5 min read · Aug 30

34

Lists

Natural Language Processing

997 stories · 483 saves

AI Regulation

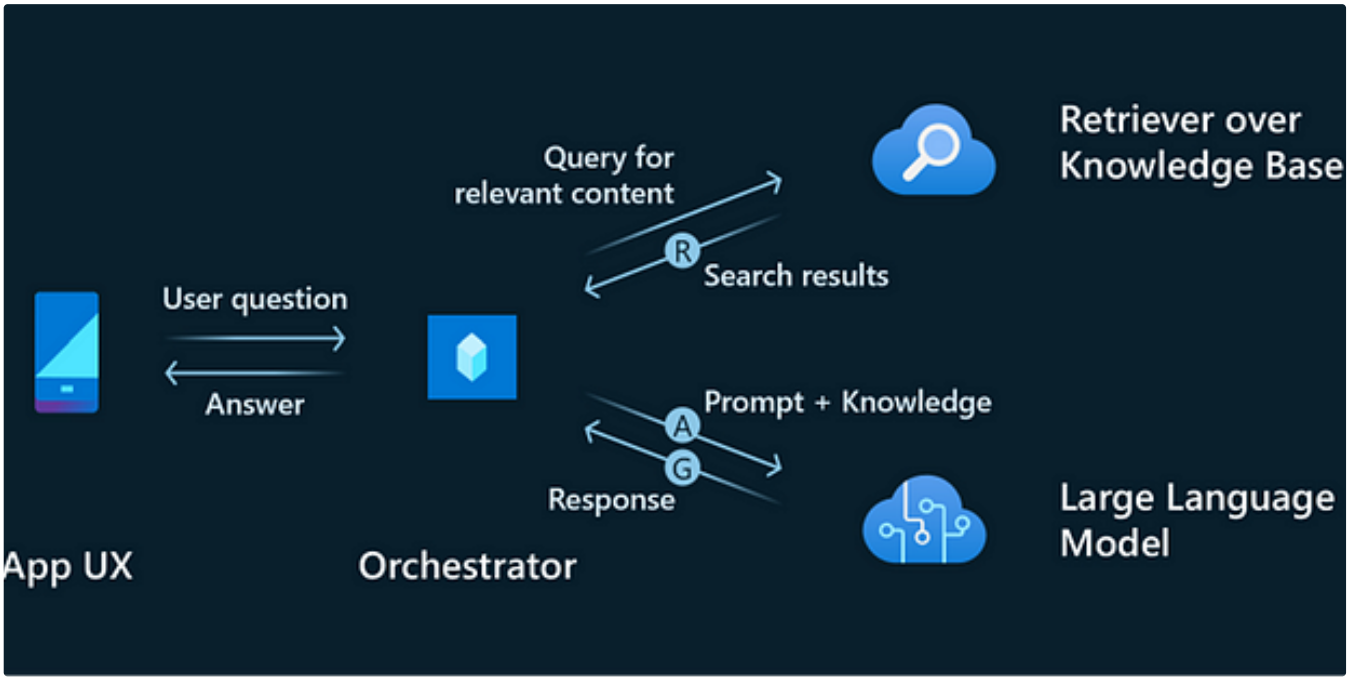
6 stories · 222 saves

data science and AI

38 stories · 9 saves

Coding & Development

11 stories · 329 saves



James Nguyen

Forget RAG: Embrace agent design for a more intelligent grounded ChatGPT!

The Retrieval Augmented Generation (RAG) design pattern has been commonly used to develop a grounded ChatGPT in a specific data domain...

6 min read · Nov 18

584 7

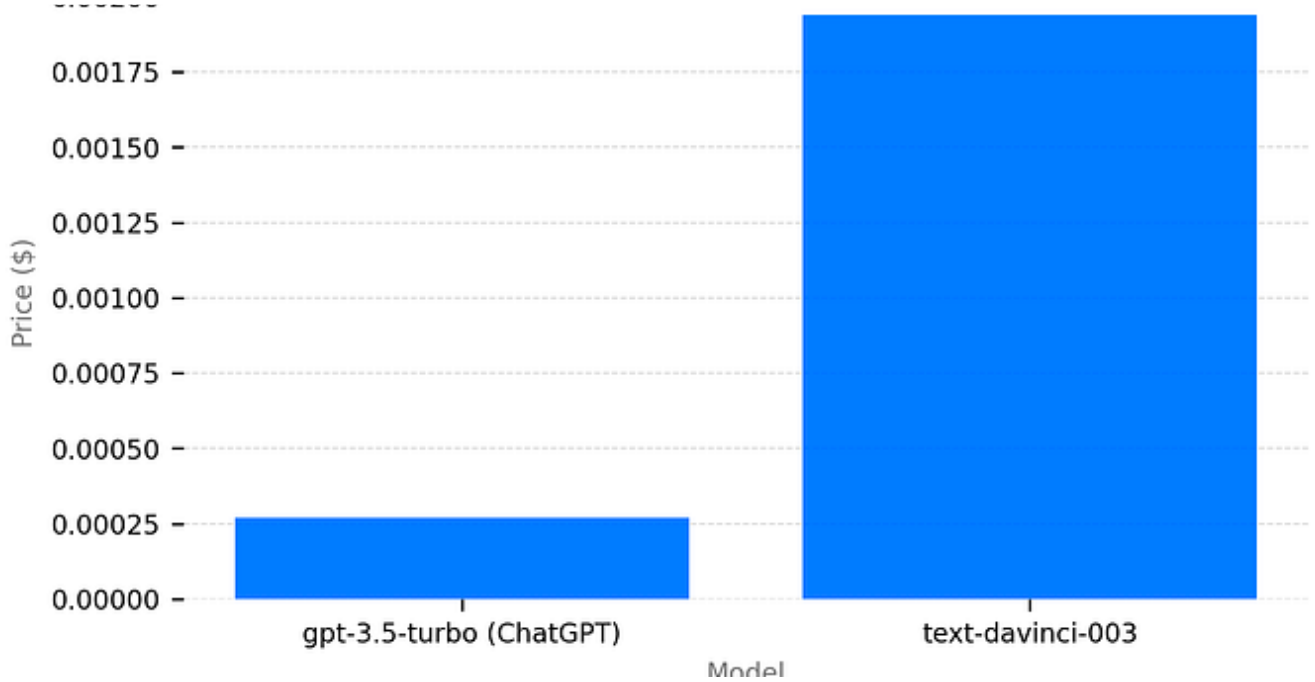
Satwiki De in Towards Data Science

‘Talk’ to Your SQL Database Using LangChain and Azure OpenAI

Explore the power of natural language processing using LLMs for your database queries

10 min read · Sep 28

150 3



Daniel Liden in The Inner Join

Make ChatGPT Stop Chatting and Start Writing SQL

You are a natural language to SQL code translator. Your role is to translate text to SQL. Return only SQL. Do not include...

8 min read · Mar 9

180 2

Madhav Thaker

Build your own RAG with Mistral-7B and LangChain

LLMs have taken the world by storm and rightfully so. They help you with every day tasks like building a coding project, creating a recipe...

14 min read · Nov 14