



AI Geeks

Team members-

Krish Kumar - BT20CSE077

Vaibhav Aggarwal - BT20CSE102

Kaushal Shirode - BT20CSE132

Aditi Soholkar - BT20CSE210



Hackathon registered

Link - <https://lablab.ai/>

AI based hackathon , providing access to GPT-4 and several other cutting edge technical models.



Problem statement

Text to SQL conversion using GPT4 Api

Input - Natural Lang question

Output - SQL query relevant to that question

Significance - Can be used in various enterprises that deal with large and unstructured data , so that even a person with merge knowledge of SQL and manipulate databases.



Insights Collection

- We have analyzed SPIDER benchmark , where we found DIN-SQL has a considerable accuracy of 85.3% on dev dataset (around 20+ databases and 10,000+ NLQ)

Rank	Model	Test
1 Nov 2, 2023	MiniSeek <i>Anonymous</i> Code and paper coming soon	91.2
1 Aug 20, 2023	DAIL-SQL + GPT-4 + Self-Consistency <i>Alibaba Group</i> (Gao and Wang et al.,'2023) code	86.6
2 Aug 9, 2023	DAIL-SQL + GPT-4 <i>Alibaba Group</i> (Gao and Wang et al.,'2023) code	86.2
3 October 17, 2023	DPG-SQL + GPT-4 + Self-Correction <i>Anonymous</i> Code and paper coming soon	85.6
4 Apr 21, 2023	DIN-SQL + GPT-4 <i>University of Alberta</i> (Pourreza et al.,'2023) code	85.3
5 July 5, 2023	Hindsight Chain of Thought with GPT-4 <i>Anonymous</i> Code and paper coming soon	83.9



Why DIN-SQL?

- If we read papers of other approaches most of them are made keeping DIN-SQL as base papers and adding new functionalities in them for ex - self consistency.
- Selecting a paper will lower accuracy will give us room for improvement and experimentation with prompts and code.
- Also the approach of DIN-SQL is quite easy and understandable giving us a good knowledge of problems faced by them so that we can find a solution for it.
- Lot of tools such as evaluations scripts are already made in this repo , giving us the time to focus on main challenge of SQL generation rather that evaluation criterion.



Initial approach

- Understanding DIN-SQL code and its loopholes .
- Trying to modify the prompts written by them , by changing prompting techniques like few shot to zero shot or more profound few-shot etc.
- We can adopt different approaches over the period of time such as - relevant table selection using NL capacity of LLM.
- Correct way and fashion of providing schemas and data matter a lot , we can think of different strategies to do so , which is compatible with context and token size of GPT-4.