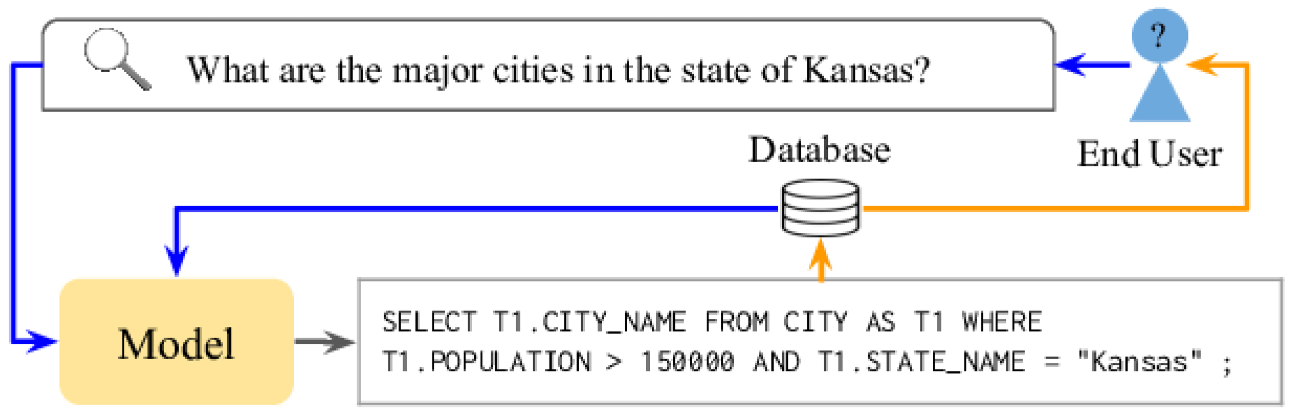
Text-to-SQL: Converting Natural Language Queries into SQL Statements

Youssef Mansour 900212652
Mohamad Abbas 900211252
THE AMERICAN UNIVERSITY IN CAIRO

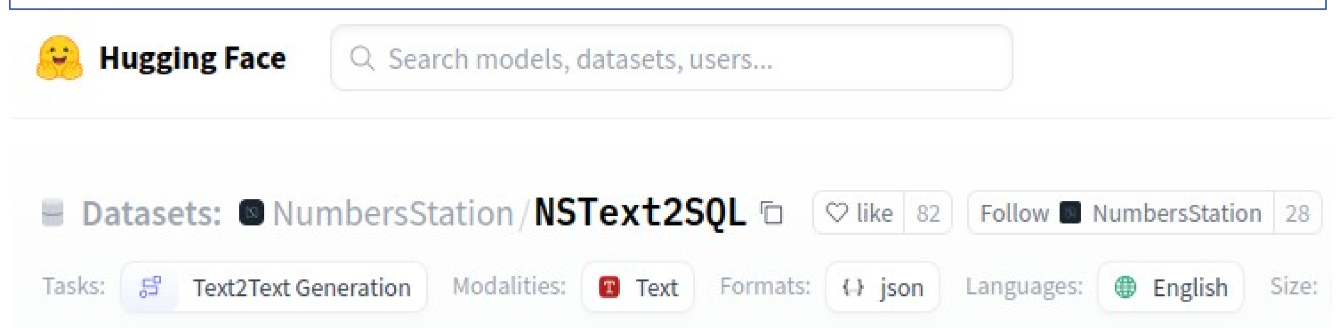
Problem Statement

Given a natural language query (NLQ) on a Relational Database (RDB) with a specific schema, produce a SQL query



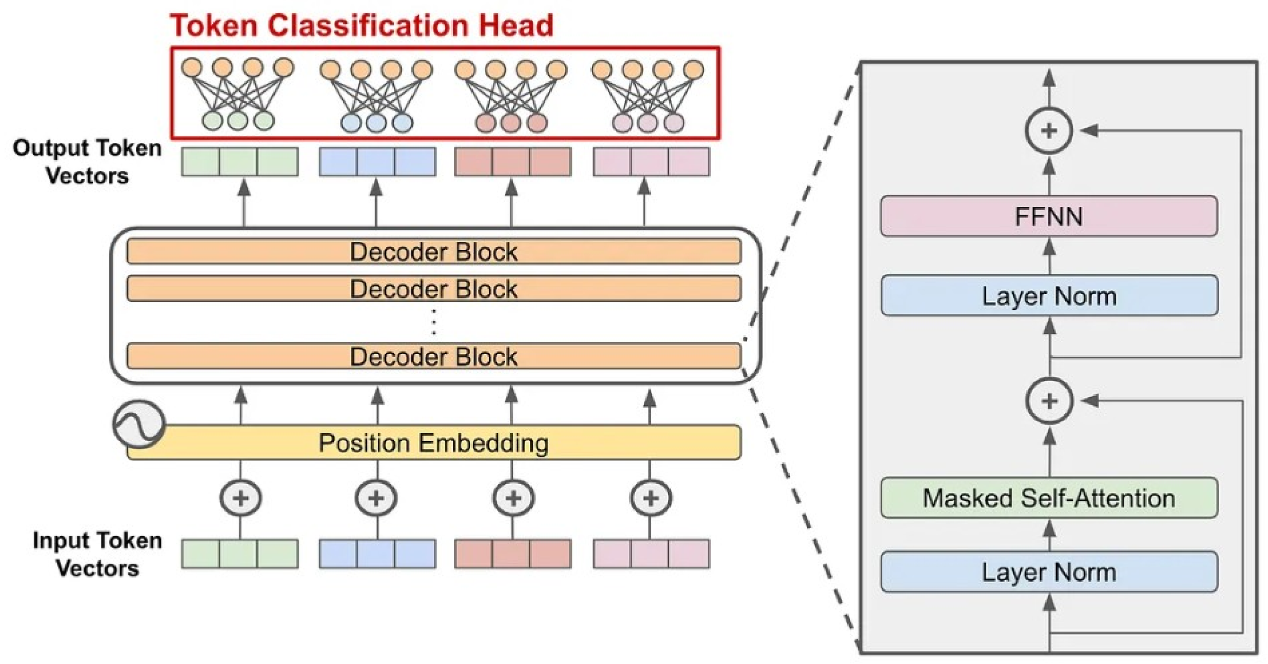
Dataset

NSText2SQL dataset used to train our model. The data is curated from more than 20 different public sources across the web. All of these datasets come with existing text-to-SQL pairs. Applied to the data are cleaning and pre-processing techniques including table schema augmentation, SQL cleaning, and instruction generation using existing LLMs. The resulting dataset contains around 290,000 samples of text-to-SQL pairs.



Original Model

Deepseek-coder-1.3b-instruct This model is part of series models built upon the same framework as the DeepSeek Large Language Model (LLM) outlined by DeepSeek-AI (2024). **It is decoder-only Transformer.**



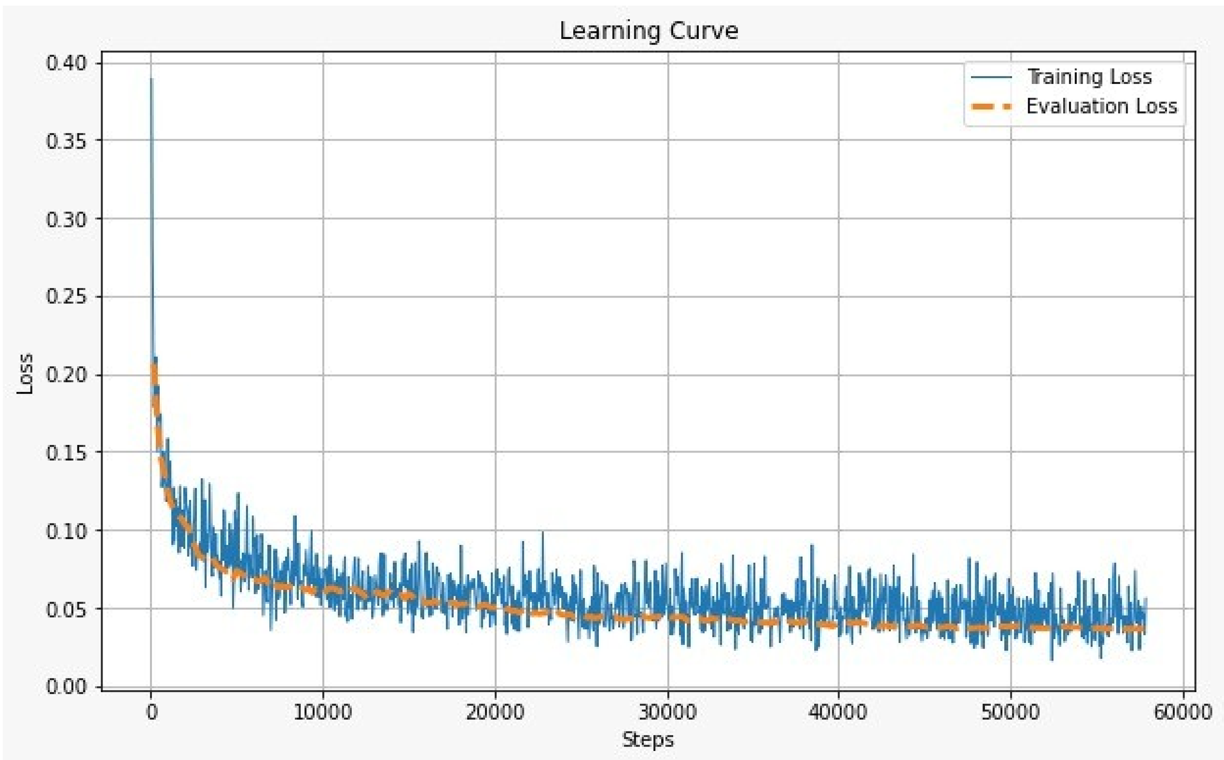
Methods

Implemented **QLoRA** for parameter-efficient fine-tuning. 4-bit Quantization: Cuts memory and computation with 4-bit weights. Double Quantization: Improves accuracy using NF4 with minimal performance drop. LoRA: Fine-tunes key layers, reducing parameters.

Implemented **RAG (Retrieval-Augmented Generation)** To retrieve relevant information from knowledge store based on the user query to help enhance the relevance of model output.

Fine-tuned hyptерparamters of the model like learning rate, epochs, batch size, model_max_length, putting into account training time and hardware resources.

Results



Leaderboard

	LLM name	submitted by	score, %
1	gpt4-0125	SG	49
2	gpt-4o-2024-05-13	NP	45
3	gpt4o-11.09.24	prafigon	43
4	gtp4-1106	NP2	42
5	Claude-3-Opus	NP2	40
6	Claude-3-Haiku	Stephen Randolph	39
7	Claude-3-Sonnet	Stephen Randolph	38
8	Claude-2	Stephen Randolph	36
8	anthropic.claude-v2	str	36
10	llama3.1 70b int8 11.09.24	prafigon	34
10	XYZ-ITALY	keenane	34
12	Claude-2.1	Stephen Randolph	31
13	Gemini-1.0-pro	NP2	29
14	Claude-Instant-1	Stephen Randolph	27
15	ktrv0	adam	25
16	LLama3.1 70B INT8	prafigon	23
17	AskSQL	Youssef and Mohamed	22
18	AskSQL	Youssef & Mohamed	18
19	mixtral-8x7b-32768	Alex Kira	16
20	deepseek-ai/deepseek-coder-1.3b-instruct	Mohamed&Youssef	14

Progress of our score during different phases compared to the plain baseline model score on Text-2-SQL Benchmark

Input/Output Example

Input:
What is the lowest Share, when Rating is greater than 1.3, and when Air Date is May 28, 2008? and the database schema

Output:
SELECT power_output
FROM table_name_88
WHERE wheel_arrangement = 'b-b' AND build_date = '1952'

Conclusion

1. **Hardware Constraints and Optimization**: Hardware capabilities and limitations play a critical role in deep learning, often necessitating strategies to reduce trainable parameters while focusing on essential aspects, such as leveraging techniques like QLoRA.

2. **Transfer Learning Benefits**: Transfer learning is a powerful approach that can significantly reduce training time while delivering excellent results by building on pre-trained models.

3. **Enhanced Inference with Advanced Systems**: Incorporating advanced inference enhancements, such as Retrieval-Augmented Generation (RAG) systems, can lead to substantial performance improvements.

4. **Effective Learning from Limited Data**: Deep learning models can still perform remarkably well when trained on relatively small datasets, even when these datasets are small in comparison to the model's complexity and number of parameters.

5. **Insights from Diverse Benchmarks**: Different benchmarks highlight the strengths and weaknesses of models and datasets in specific areas, making it essential to base decisions on collective insights drawn from multiple benchmarks for a comprehensive evaluation.

References

Benchmark URL: LLM SQL Streamlit App.
Model URL: Model Deepseek.
Dataset URL: NumbersStation/NSText2SQL.