INTERNSHIP REPORT

TEXT TO SQL WITH THE USE OF LLM

A REPORT

Submitted by: TANMAY GUPTA

*Under the Guidance of :*Ms. GEETIKA PANDEY

Deputy Manager-Automation and Digital

Department: Automation and Digital

University: SRM Institute of science and Technology

Submission Date:

T.EN Global Business Services Private Limited

9th Floor-Tower B WTT ,Sec-16

Noida-210301

FEB 2025

\

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to Technip Energies and my mentor **Ms. Geetika Pandey** for their invaluable guidance and support during my internship. I extend my appreciation to the Head of the Department, **Mr. Moshin Khan**, for his encouragement and assistance throughout the internship process. Additionally, I am grateful to **Mr. Ranjodh Singh**, with whom I was aligned during my internship, for his mentorship and insights that enriched my learning experience. Finally, I thank my university and faculty members for providing me with this opportunity.

TANMAY GUPTA

ABSTRACT

This report presents my internship experience focused on developing a Text-to-SQL system using Large Language Models (LLMs). The primary objective was to fine-tune LLMs to accurately translate natural language queries into SQL statements. The internship involved exploring fine-tuning techniques, schema linking, and query optimization strategies to enhance model accuracy. Techniques such as Retrieval-Augmented Generation (RAG) and advanced prompt engineering were employed to handle multi-table queries efficiently. Key takeaways from this internship include improved execution accuracy, domain-specific adaptability, and enhanced query generation capabilities. This experience provided hands-on exposure to AI-driven database querying solutions, contributing to advancements in automated SQL generation.

TABLE OF CONTENTS

INTRODUCTION

Introduction to Text-to-SQL Using LLM

Text-to-SQL is an advanced Natural Language Processing (NLP) task that focuses on translating human language queries into SQL statements. The primary challenge lies in accurately interpreting natural language inputs while maintaining semantic consistency with database schemas. Large Language Models (LLMs), such as T5, SQL-PaLM, and Codex, have significantly improved Text-to-SQL performance by leveraging fine-tuning techniques, prompt engineering, and Retrieval-Augmented Generation (RAG).

Importance of Text-to-SQL

Traditional database querying requires users to have proficiency in SQL, which limits accessibility for non-technical users. Text-to-SQL bridges this gap by enabling users to retrieve data using conversational queries. Fine-tuned LLMs enhance query accuracy by improving schema linking, understanding multi-table relationships, and generating syntactically correct SQL queries.

Challenges in Text-to-SQL

1. **Schema Complexity** – Large-scale databases contain intricate table relationships, making schema linking difficult.
2. **Ambiguity in Natural Language** – Different users may phrase the same query in multiple ways, requiring contextual adaptation.
3. **Execution Accuracy** – Generated SQL queries must return the correct results, necessitating execution-based refinements.
4. **Handling Multi-Table Queries** – Many real-world queries involve joins and aggregations across multiple tables, requiring sophisticated query generation techniques.

Advancements in LLMs for Text-to-SQL

1. **Fine-Tuning Techniques:** Models like SQL-PaLM and MedT5SQL improve accuracy through domain-specific training.
2. **Retrieval-Augmented Generation (RAG):** Enhances query precision by dynamically retrieving relevant schema information.
3. **Schema Linking Approaches:** Techniques such as relation-aware transformers and grammar-based decoding help models map query terms to the correct database structures.

COMPANY OVERVIEW

Technip Energies is a leading engineering and technology company dedicated to accelerating the energy transition. With over 65 years of experience, the company specializes in designing and delivering advanced solutions for the energy sector, including liquefied natural gas (LNG), hydrogen, sustainable chemistry, and carbon capture. Operating in 34 countries, Technip Energies employs more than 17,000 professionals worldwide.

In India, Technip Energies has a significant presence through Technip Energies India Limited, headquartered in Noida, Uttar Pradesh. The Noida office serves as a central hub for the company's operations in the region, providing comprehensive engineering, procurement, and construction services. The team in Noida collaborates closely with clients to deliver innovative and sustainable solutions tailored to the local market's needs.
Technip Energies India Limited is actively involved in various projects that contribute to the country's energy infrastructure development. The company's expertise encompasses a wide range of services, including process design, project management, and construction supervision, ensuring the successful execution of complex energy projects. By leveraging its global experience and local knowledge, Technip Energies India Limited plays a pivotal role in supporting India's energy transition and sustainable growth.
As part of its commitment to innovation and sustainability, Technip Energies continues to expand its footprint in India. The company's focus on research and development, combined with its dedication to excellence, positions it as a key player in the Indian energy sector, driving progress toward a more sustainable and efficient energy landscape.

INTERNSHIP DETAIL

Department & Supervisor: Automation and Digital
Supervisor: Ms. Geetika Pandey
Roles & Responsibilities:

- Researched and implemented fine-tuning strategies for Text-to-SQL models.
- Worked with datasets like Spider and BIRD-SQL to train models.
- Developed schema linking and prompt engineering techniques for enhancing query accuracy.
- Integrated Retrieval-Augmented Generation (RAG) to improve multi-table SQL generation.

- Evaluated execution accuracy and optimized query refinement strategies.

Tools/Technologies Used:

- Python, SQL, TensorFlow, PyTorch
- Large Language Models (T5, Codex, SQL-PaLM)
- Retrieval-Augmented Generation (RAG) Frameworks
- Database Management Systems (MySQL, PostgreSQL)
- Vscode,Google Colab,JupiterNotebook,Anaconda

LEARNING & SKILLS GAINED

Technical Skills:

- Fine-tuning LLMs for Text-to-SQL tasks
- Implementing Retrieval-Augmented Generation (RAG) for multi-table queries
- SQL optimization and schema linking techniques
- Query decomposition for handling complex SQL structures
- Utilizing table serialization techniques to improve LLM query interpretation

Soft Skills:

- Effective communication of AI-driven solutions
- Collaboration in cross-functional teams
- Problem-solving in real-world AI applications
- Analytical thinking for debugging and refining SQL queries

Industry Knowledge:

- AI-driven automation in database querying
- Energy sector digital transformation initiatives
- Application of LLMs in structured data retrieval
- Enhancements in NLP models for SQL generation

CHALLENGES & SOLUTIONS

Challenges Faced:

- Implementing free APIs in the project while ensuring efficiency and reliability.
- Processing large datasets, which led to long execution times on a standard laptop.
- Adapting LLMs to generate accurate multi-table SQL queries.
- Handling large and complex database schemas efficiently.
- Managing ambiguous natural language queries that required contextual adaptation.

Solutions Implemented:

- Explored cost-effective API solutions and optimized API calls to minimize processing time.

- Used cloud-based resources for large dataset processing to improve efficiency.
- Fine-tuned models on domain-specific datasets for better schema linking.
- Implemented RAG to enhance retrieval precision and execution accuracy.
- Applied table serialization techniques and query decomposition for handling complex SQL structures.
- Used **Groq API** for free, reducing reliance on costly API services.
- Downloaded and integrated **Ollama**, a free tool for enhancing model performance without additional costs.

CONCLUSION AND RECOMMEDATIONS

Conclusion

The internship at Technip Energies Noida provided an in-depth experience in Text-to-SQL using Large Language Models. Through hands-on involvement in fine-tuning models, optimizing query accuracy, and integrating Retrieval-Augmented Generation (RAG), I gained valuable insights into AI-driven database solutions. The challenges faced, such as handling large datasets and improving schema linking, allowed me to explore innovative problem-solving techniques. This experience has strengthened my technical expertise and analytical skills, making it a significant step in my career development.

Recommendations

1. **Enhance Model Optimization:** Future interns should focus on refining fine-tuning techniques to reduce computational overhead while maintaining accuracy.
2. **Improve Execution Accuracy:** Incorporating real-time execution feedback can help refine SQL query generation.
3. **Leverage More Efficient APIs:** Exploring lightweight and cost-effective APIs, like Groq API and Ollama, can further streamline performance.

4. **Utilize Cloud-Based Solutions:** Given the limitations of local hardware, using cloud-based processing can improve efficiency when dealing with large datasets.
5. **Expand Schema Linking Strategies:** Improving schema understanding through advanced prompt engineering and relation-aware transformers can enhance SQL generation.
6. **Explore Multi-Model Integration:** Combining different LLMs can provide more robust and contextually aware query results.

REFRENCES

1. Large Language Model Enhanced Text-to-SQL Generation: A Surveyhttps://arxiv.org/html/2410.06011v1?utm_source=chatgpt.com
2. SQL-PaLM: Improved Large Language Model Adaptation for Text-to-SQL (extended)https://arxiv.org/abs/2306.00739?utm_source=chatgpt.com
3. Extreme RAG for Table To Answers with Large Language Modelshttps://arxiv.org/html/2405.03963v2
4. Enhancing LLM Fine-tuning for Text-to-SQLs by SQL Quality Measurementhttps://arxiv.org/html/2410.01869v1
5. BIRD-SQLhttps://bird-bench.github.io/

APPENDICES

1. NLP REPORT:https://d.docs.live.net/9e195d081c31509d/Desktop/NLP%20REPORT.docx
2. NLP PPThttps://d.docs.live.net/9e195d081c31509d/Desktop/Natural-Language-Processing.pptx
3. AI AGENT:https://docs.google.com/document/d/1DxMvQLcoqqLk7_jbx-XEX2FxG9KRwXtZ5c0eSMMyrO4/edit?usp=sharing
4. VANNA AIhttps://docs.google.com/document/d/1JbMJ3pGNIj_XE2zfuluv3FBeHxpVo7gNIQdyRDBuqQo/edit?usp=sharing
5. DEEP SEEKhttps://docs.google.com/document/d/1H2IB8nBIq6FLiMqbpFVUhgEaKe4KrxZJlynKyfRWNPY/edit?usp=sharing
6. TEXT-TO-SQLGOOGLE COLABhttps://colab.research.google.com/drive/1DKrigj1EGi6KaZ36fvOJZAYdvr3IZZfO?usp=sharing
7. LANGCHAIN GOOGLECOLABhttps://colab.research.google.com/drive/1IumB-RQ_nV5mq-I9TchmJ1dmkgondQKS?usp=sharing

8. AIAGENT(BOTPRESS)https://cdn.botpress.cloud/webchat/v2.2/shareable.html?configUrl=https://files.bpcontent.cloud/2025/01/21/17/20250121170656-QGE3T6NW.json
9. GENAI APPUSING DEEPSEEK AND LANGCHAINhttps://github.com/t07g/End-To-End-Gen-AI-App-Using-DeepSeek-R1-With-Langchain-And-Ollama.git

Signature:Tanmay Gupta
Date:16 February 2025