Talk to your data: Enhancing Business Intelligence and Inventory Management with LLM-Driven Semantic Parsing and Text-to-SQL for Database Querying

Jerry Zhu
Faculty of Mathematics
University of Waterloo
Ontario, Canada
j25zhu@uwaterloo.ca

Saad Ahmed Bazaz
Dept. of Computer Science
National University of Computer and Emerging Sciences
Islamabad, Pakistan
i180621@nu.edu.pk

Srimonti Dutta
School of Agro and Rural Technology
Indian Institute of Technology Guwahati
Guwahati, India
srimonti@iitg.ac.in

Bhavaraju Anuraag
Dept. of Electronics and Communication Engineering
Vellore Institute of Technology
India
bhavarajuanuraag@gmail.com

Imran Haider
Dept. of Industrial and Systems Engineering
Indian Institute of Technology Kharagpur
Kharagpur, India
ihalig14@iitkgp.ac.in

Srijita Bandopadhyay
Dept. of Electronics and Communication Engineering
University of Engineering and Management
Kolkata, India
srijitabando30@gmail.com

***Abstract*—This research paper explores the transformative potential of Large Language Models (LLMs) in the context of business intelligence and inventory management through semantic parsing and Text-to-SQL methodologies. The study conducts extensive evaluations of various models, including DIN-SQL, DSP, NSQL, GPT, CoPilot, and LLaMa, shedding light on their capabilities and contributions to this domain. Here, results from two analyses have been presented. The first compares state-of-the-art LLM models using metrics like cosine similarity and cost considerations. CoPilot stands out for its cost-effectiveness and accuracy. In contrast, open-source models exhibit varying performance and infrastructure requirements. The analysis also highlights the significance of prompting in model performance. The second analysis focuses on improving GPT's accuracy through prompt engineering like few-shot, exploring frameworks like DIN-SQL NSQL, and DSP. DIN-SQL demonstrates a substantial accuracy boost, while NSQL shows favourable potential in certain scenarios. This study showcases the potential of LLM-driven models to revolutionize business intelligence and inventory management. DIN-SQL emerges as a standout performer, promising a paradigm shift in inventory management practices. GPT exhibits versatile capabilities through fine tuning beyond traditional programming tasks, and CoPilot offers a cost-effective alternative. This research also emphasizes the importance of cost-effectiveness in real-world implementations, with LLaMa and CoPilot emerging as practical choices. NSQL provides a budget-friendly, semi-accurate solution, particularly suited for semantic parsing in growing companies. In summary, this research provides a comprehensive overview of the advancements facilitated by LLM-driven semantic parsing and Text-to-SQL capabilities. These insights serve as a foundation for further innovation, promising unparalleled efficiency and competitiveness across industries in the evolving AI landscape.**

***Index Terms*—Semantic parsing, Text to SQL, Structured Query Language, Large Language Models, Natural Language Processing, Business Intelligence**

I. Introduction

In today's data-driven world, businesses rely heavily on advanced tools for accessing and managing data from complex databases. Accurate and timely data extraction is crucial for informed decision-making, operational efficiency, and staying competitive. However, there's often a gap between how databases are organized and the natural language queries that business users prefer. Can modern technology simplify this process, allowing decision-makers to ask questions from their data without needing to write complex SQL queries?

Semantic parsing and text-to-SQL translation are emerging solutions that bridge the gap between human users and databases. These approaches enable users to express queries in plain language, which is then converted into structured SQL queries for execution. This eliminates the need for extensive knowledge of database structures and query languages, making data more accessible to a wider audience. Recent advances in Natural Language Processing (NLP) have led to Large Language Models (LLMs) like OpenAI's GPT series. These models are exceptionally skilled at understanding and generating human-like text, including complex linguistic nuances and logically consistent responses. As businesses increasingly seek innovative ways to leverage data for strategic decisions, the integration of LLMs into semantic parsing and text-to-SQL frameworks presents exciting possibilities.

Our research paper explores the impact of LLM-driven semantic parsing and text-to-SQL techniques on business intelligence. We aim to demonstrate how LLMs can optimize querying, and evaluate costs of running them. In addition to theory, we will present empirical evidence of the benefits LLMs bring to accuracy, efficiency, and user-friendliness in business database querying. By leveraging benchmark datasets and performance metrics, we will showcase the improvements these models offer.

The paper is structured as follows: Section II discusses the existing research around these topics, Section III discusses the solution methodology, followed by the results and discussions in Section IV. Finally, the study is concluded in Section V.

II. Literature review

Previous research has made efforts to examine the relationship between Language Learning Models (LLMs) and semantic parsing, specifically within the domain of business intelligence.

[1] presented BIRD, a pioneering big Benchmark for Large-Scale Database grounded text-to-SQL parsing. Unlike existing benchmarks, BIRD addresses the real-world gap by encompassing 12,751 text-to-SQL examples across 95 expansive databases spanning various domains. The emphasis on database values reveals challenges in handling noisy content, external knowledge integration, and SQL efficiency. This work introduces the Valid Efficiency Score (VES) metric, evaluating query efficiency within the context of substantial and intricate databases. Despite leveraging state-of-the-art models like ChatGPT, BIRD demonstrates the models' struggle to generalize, with execution accuracies well below human performance. BIRD not only bridges research and applications but also provides insights for advancing text-to-SQL solutions within realistic settings. Notably, the Spider benchmark highlights the performance gap that BIRD aims to address. In another study, [2] introduced SQL-PaLM, an advanced Large Language Model (LLM) adaptation for Text-to-SQL tasks. Leveraging PaLM-2, they achieve state-of-the-art performance in both few-shot prompting and fine-tuning scenarios. Their approach, executed through an execution-based self-consistency prompt, elevates accuracy by up to 4.7% compared to existing methods. Remarkably, SQL-PaLM surpasses previous SoTA fine-tuned models by 3.8%, even outperforming the latest in-context learning SoTA. The model showcases its potential through its generation of complex SQL outputs and remarkable understanding of the SQL language. SQL-PaLM demonstrates its superiority in diverse scenarios, solidifying its status as a benchmark in Text-to-SQL tasks. The achieved results on Spider further underline its efficacy. [3] presented a novel approach, DIN-SQL, that strategically decomposes intricate text-to-SQL tasks into manageable sub-tasks. By feeding solutions of these sub-tasks into Large Language Models (LLMs), they significantly enhance LLM performance in complex reasoning processes. Their technique effectively narrows the performance gap between fine-tuned models and prompting approaches, pushing LLM accuracy to match or surpass fine-tuned SoTA models. Evaluated on the challenging Spider dataset, DIN-SQL achieves an execution accuracy of 85.3%, outperforming existing methods by a substantial margin. Notably, DIN-SQL stands out on the Spider leaderboard for execution accuracy without requiring database cells, setting new ground for enhanced text-to-SQL performance. This study highlights the potential of strategically decomposing tasks to enhance prompting approaches for intricate NLP challenges like Text-to-SQL.

Research Gaps: The present study aims to make a valuable contribution to the ongoing discussion by conducting an extensive empirical inquiry into the effects of LLMs on business intelligence systems. This research endeavor aims to provide valuable insights into the practical effectiveness of LLMs, the obstacles they present, and the potential enhancements they bring about.

For example, new frameworks like open source LLM LLaMa and Alpaca and frameworks such as DSP (Demonstrate-Search-Predict) shown in [4] and NSQL in [1] are leading the newest phase of fine-tuned language models. In particular, NSQL-350M is an autoregressor trained on one million training data points of SQL queries. There are research gaps considering how these models perform against the standard text-to-SQL methods.

In the following sections, a comparison between state-of-the-art (SoTA) LLMs in the current market is presented. The subsequent analysis and synthesis of pertinent studies are conducted to unravel the dynamic and changing landscape of advancements in business intelligence driven by LLM. In this analysis, our objective is to provide a thorough overview of the current state-of-the-art in the field which is relevant to businesses. Subsequently, we aim to contribute to the ongoing discourse by presenting our empirical evaluations and findings in this domain.

III. Solution Methodology

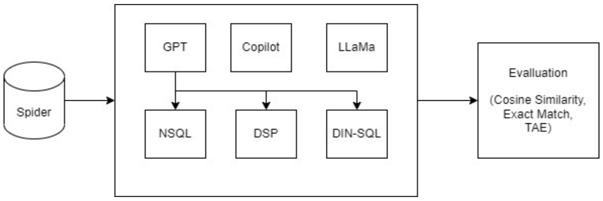

Fig. 1: Workflow Diagram

A. Dataset

In this paper, the dataset we will be using is the SoTA Spider dataset, generated by Yale's semantic parsing competition. It is a complex semantic parsing dataset, annotated with over 10,000 question-and-answer pairs of various SQL queries on over 100 databases. For the data generation of our validation data, we sampled 100 values from the spider dataset, trying to keep a subset of each type of data in the spider dataset. We worked with a smaller more advanced set of test data in order to get a better idea of how to fine tune our models to improve accuracy.

B. Data Comparison

Consider a business intelligence database. This schema will contain a certain number of columns with text or number data. An example of data on inventory is taken from Kaggle. This schema is very similar to the Spider dataset, so any results generated by Spider (for a more generalized use case) can easily be extrapolated for the use case of business intelligence. In particular, this gives us a new way to generate data: any existing datasets used for database querying can be annotated for semantic parsing, which is something we have shown in the results.

C. Breakthrough Methodologies

We will be exploring two different types of improvements to semantic parsing. The first compares regular SoTA models to find the one with the highest cost-benefit analysis, using accuracy and cost analysis as metrics. The second improvement is looking at the evaluation accuracy of recent advancements in the field as part of the literature review.

D. Performance Evaluation

In this paper, two types of evaluation metrics are chosen. For comparing various LLM models, we will use cosine similarity to check the ”distance” between the vectorized prediction and results. We focused on this metric compared to other similar NLP text distance metrics, as it has the closest resemblance to a semantic similarity output. In particular, given two generated vector matrices of the two texts in question, we can calculate the similarity as:

$$\cos(\mathbf{A},\mathbf{B}) = \frac{\mathbf{A}\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n}\mathbf{A}_i\mathbf{B}_i}{\sqrt{\sum_{i=1}^{n}(\mathbf{A}_i)^2}\sqrt{\sum_{i=1}^{n}(\mathbf{B}_i)^2}} \tag{1}$$

For comparing the improvements of methods on standard GPT, we will be using exact match accuracy. Spider also has its own metric, called Test Suite Evaluation Accuracy (TAE). Exact match accuracy is exactly as its name suggests; it takes the number of correct predictions divided by the number of total predictions. Test Suite evaluation accuracy builds off of EMA and creates a stricter lower bound by rooting out false positives. In particular, it attempts to optimize test suites with strong neighbor query correlation. Formally, according to [2, TAE attempts to optimize S_g for a test suite under group g with:

Goal = minimize $\|S_g\|$ s.t. $\forall q \in N_g, D_{S_g(g;q)} = 1$

This is done by generating several random samples through ”fuzzing” to test the database suite against the proposed target property [2]. However, since this is extremely computationally expensive, we will only run this for models we want to compare against the accuracy of Spider models such as BIRD for DIN-SQL. For SoTA or lower-performing EMA models, we will not be running a full TAE test suite.

IV. Results and Discussions

The experimental setup is computed on Lambda Labs with 1x A100 GPU, 30 vCPUs, 200 GiB RAM, and

512 GiB SSD for optimal performance, but due to infrastructure limitations, Python scripts were created to use the APIs for the LLMs.

We performed two analyses. The first is a direct comparison of state-of-the-art LLM models, performed on a basic prompt, asking the LLM to generate a MySQL query given a text response for a particular query. The second analysis compares several advancements in the text-to-SQL field outlined by Spider, as discussed earlier in section III.

Table 1 highlights state-of-the-art (SoTA) models, including GPT, CoPilot, LLaMa, and Vicana. We compare them with the following metrics; cosine similarity and cost per 1000 tokens. Cosine similarity is a measure used to assess the similarity between two vectors, often employed in fields like natural language processing and recommendation systems. It quantifies the cosine of the angle between the vectors, providing a value between -1 and 1, with higher values indicating greater similarity. In NLP, tokens are the fundamental units of text, typically representing words or characters, and are used for language processing tasks. They enable NLP models to analyze and understand text. Costs are measured in US dollars (as per August 2023).

TABLE I: Performance comparisons of different models

Data generation methods	Cosine Similarity	Cost/1000 tokens
GPT-4	0.79	0.12
Copilot	0.84	0.004
LLaMa	0.44	0
Vicuna	0.38 (timed out)	0

We can extract several conclusions from Table 1. For example, CoPilot almost edges out GPT-4 in terms of accuracy. This makes sense as GPT-4 is a strong model, but gives a reasonable alternative, as CoPilot is hundreds of times cheaper and runs at approximately the same inference speed. On the open source side, we see a significant drop in infra. Running on a T4 GPU, and Vicuna still times out on a Lambda Labs A100 GPU 512 GiB SSD.

Another result not shown by this analysis is the sheer amount of prompting required for these models to run. LLaMa will achieve almost no accuracy unless prompted, while GPT also requires some amount of prompting, see the below graph. However, since CoPilot runs OpenAI Codex, trained on instruction following embeddings, it requires almost no fine tuning to run with high accuracy.

In terms of accuracy, although CoPilot outshines the other models, it has close to no customization. LLaMa and Vicuna can be fine-tuned further, and GPT-4 has many more improvements that can be made. In the second analysis, we will discuss improvements to GPT in order to improve the accuracy using the same base model. We then delve deeper into GPT. Given that GPT can be fine-tuned by prompt-engineering, using DIN-SQL and DSP-based frameworks, we evaluate how simply prompting GPT can produce better results for text-to-SQL queries.

DIN-SQL is a framework that enhances Large Language Models' (LLMs) performance in complex text-to-SQL tasks by breaking them into sub-tasks and improving accuracy by approximately 10%, exceeding the state-of-the-art on datasets like Spider, with an execution accuracy of 85.3%, and outperforming heavily fine-tuned models by at least 5%.

DSP (Directional Stimulus Prompting) is a framework that enhances large language models (LLMs) by providing precise cues in prompts, improving LLM performance through a small model and reinforcement learning. It achieved a remarkable 41.4% improvement on the MultWOZ dataset for ChatGPT and offers publicly available code.

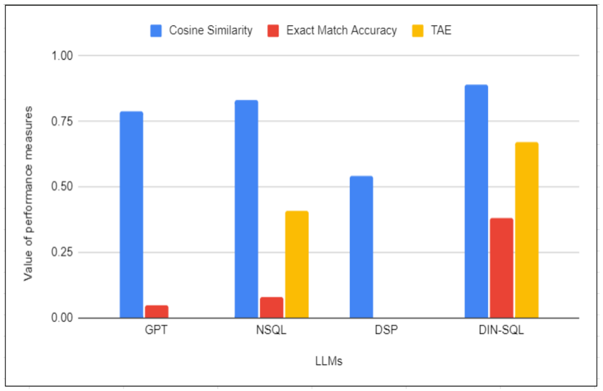

Fig. 2: Graphical Representation of the performance measures.

In Figure 2, we highlight improvements achieved through fine-tuning and prompt engineering, exemplified by Din-SQL and DSP-based frameworks. The paper aims to delve into two directions: standard Language Model Models (LLMs) and prompt engineering. We could not generate results for DSP due to extremely low accuracy, with an exact match accuracy of zero. Additionally, there's no Targeted Adversarial Evaluation (TAE) for DSP because it doesn't make sense to run a computationally expensive TAE when the exact match accuracy is already at zero. TAE is noted for its significant computational cost.

V. Conclusion

In the contemporary landscape of business intelligence and inventory management, the convergence of advanced technologies has paved the way for unprecedented enhancements. This study delved into the realm of semantic parsing and Text-to-SQL methodologies, harnessed by the power of Large Language Models (LLMs), to invigorate the domain. By rigorously evaluating and comparing the performance of various models, including DIN-SQL, DSP, BIRD, GPT, CoPilot, and LLaMa, this research contributes a comprehensive understanding of the capabilities and potential of these technologies. The results obtained from this investigation unequivocally illustrate the potency of these LLM-driven models in transforming the landscape of business intelligence and inventory management. DIN-SQL emerged as a standout performer among the models assessed, showcasing its prowess in the intricate task of semantic parsing and SQL generation for database querying. This resounding success underscores its potential to revolutionize conventional practices, potentially catalyzing a paradigm shift in how businesses manage and leverage their inventory systems.

Furthermore, it is evident that GPT, renowned for its language generation abilities, demonstrated notable proficiency in this context. Its ability to understand and respond to nuanced queries signifies a broader spectrum of applications in the domain beyond traditional programming tasks.

A notable observation from the analysis pertains to cost-effectiveness, a pivotal factor in real-world implementations. LLaMa emerged as an affordable alternative, boasting commendable performance that positions it as a suitable candidate for practical adoption. Additionally, CoPilot presents itself as a close second in terms of cost-efficiency, providing organizations with an array of options to align technological choices with budgetary considerations.

From a business intelligence standpoint, NSQL is a cheap, semi-accurate, and low infra solution. It doesn't have the cost overhead that GPT-4 does, the high infra and prompting annoyance that LLaMa has, or the complexity of running loopbacks with DIN-SQL. NSQL is also fine-tuned towards SQL querying, which makes it perfect for semantic parsing. The only downside is its low EMA score, which can be salvaged with diligent error-checking and annotation, which should be done with any ML model.

CoPilot's performance is also a reasonable alternative due to its good ease-of-use; it could mean that businesses could save a significant budget by having a smaller model that is more fine-tuned toward answering questions with SQL code.

In summary, this research presents a holistic overview of the advancements in business intelligence and inventory management facilitated by LLM-driven semantic parsing and Text-to-SQL capabilities. The findings underscore the potential of these technologies to redefine established norms, empower decision-makers, and streamline operations. It is to be noted, however, to bring this into practicality involves some security practices such as preventing the generated SQL from modifying or destroying vital data, and access to classified data. In the future, LLMs may be part of the Database Management System (DBMS) layer, to assist in natural queries and provide better optimization at a lower-level, and to provide more fine-grained control. As the realm of AI continues to evolve, these insights provide a foundation upon which further innovations can be fostered, with the potential to unlock unparalleled efficiency, accuracy, and competitiveness across industries.

References

[1] J. Li, B. Hui, G. Qu, B. Li, J. Yang, B. Li, B. Wang, B. Qin, R. Cao, R. Geng *et al.*, "Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls," *arXiv preprint arXiv:2305.03111*, 2023.

[2] R. Sun, S. O. Arik, H. Nakhost, H. Dai, R. Sinha, P. Yin, and T. Pfister, "Sql-palm: Improved large language modeladaptation for text-to-sql," *arXiv preprint arXiv:2306.00739*, 2023.

[3] M. Pourreza and D. Rafiei, "Din-sql: Decomposed in-context learning of text-to-sql with self-correction," *arXiv preprint arXiv:2304.11015*, 2023.

[4] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar *et al.*, "Holistic evaluation of language models," *arXiv preprint arXiv:2211.09110*, 2022.