Information Processing and Management 62 (2025) 104136

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Optimizing text-to-SQL conversion techniques through the integration of intelligent agents and large language models

Samuel Ojuri [a], The Anh Han [a], Raymond Chiong [b,c], Alessandro Di Stefano [a,*]

[a] School of Computing, Engineering and Digital Technologies, Teesside University, Middlesbrough, United Kingdom
[b] School of Science and Technology, The University of New England, Parramatta, Australia
[c] School of Information and Physical Sciences, The University of Newcastle, Callaghan, Australia

ARTICLE INFO

Keywords:
Text-to-SQL conversion
Large language models
Intelligent agents
Model fine-tuning
Few-shot learning

ABSTRACT

In many organizations, retrieving valuable information from complex databases has traditionally required specialized technical skills, often leaving non-technical professionals dependent on others for timely insights. This study presents an approach that allows anyone, even without knowledge of query languages, to directly interact with databases by asking questions in everyday language. We achieve this by combining advanced generative language models, such as a high-capacity Generative Pre-trained Transformer (GPT) model, with intelligent software agents that translate natural language queries into precise SQL statements. Our evaluation compares different strategies, including models specifically trained on a particular database domain versus those guided by only a handful of examples. The results show that training a model with tailored examples yields more accurate and reliable database queries than relying solely on minimal guidance for the given use case. This work highlights the practical value of refining model complexity and balancing computational costs to empower business users with easy, direct access to data. By reducing reliance on technical teams, organizations can enable faster, more informed decision-making and foster a more inclusive environment where everyone can uncover data-driven insights on their own.

1. Introduction

As recent advancements in artificial intelligence make it easier to leverage the use of large language models (LLMs) in business solutions, the potential for widespread adoption of this approach is increasingly promising (Grohs, Abb, Elsayed, & Rehse, 2023; Kanbach, Heiduk, Blueher, Schreiter, & Lahmann, 2024; Powers et al., 2023). Most organizations across several industries store their data about business processes and transactions in relational database management systems (Fadloun, Meshoul, Hosseini, & Choutri, 2023). Retrieving business-critical insights from these data stores has the limitation of technical expertise in structured query language (SQL). This work aimed at providing a solution that gives non-technical stakeholders access to easily query databases using natural language as well as retrieve insights about business key-performance-indicators (KPIs) and opportunities in real-time by building a prototype system. The prototype is built on the text-to-SQL task in natural language processing, which involves generating SQL queries automatically from natural language text (Elgohary, Hosseini, & Hassan Awadallah, 2020). In the realm of enterprise analysis, this involves using natural language text to retrieve business insights from a relational database storing enterprise data, such as business transactions instead of having to query the database directly with SQL (Sen et al., 2019). The approach adopted in this work involves optimizing text-to-SQL conversion techniques using a synergized reasoning and acting (ReAct) approach while

* Corresponding author.
E-mail address: a.distefano@tees.ac.uk (A. Di Stefano).

https://doi.org/10.1016/j.ipm.2025.104136
Received 22 September 2024; Received in revised form 25 December 2024; Accepted 6 March 2025
Available online 27 April 2025

providing an intelligent agent with the appropriate tools to use an LLM as a reasoning engine for the specified task (Yao et al., 2023).

The synergy between language-based game theory and artificial intelligence has been explored in recent works, notably by Capraro, Di Paolo, Perc, and Pizziol (2024), who examined the role of intelligent agents in facilitating complex interactions. Their findings highlight the potential of LLMs as reasoning engines in decision-making frameworks, emphasizing their applicability across diverse domains, including business analytics. This paper builds on such insights by focusing on enterprise-specific challenges, particularly in bridging the gap between complex database queries and non-technical users. By leveraging intelligent agents, we aim to transform natural language inputs into actionable SQL queries, thereby democratizing data access and enhancing decision-making capabilities.

2. Research objectives

The primary aim is to bridge the gap between complex database interactions and non-technical stakeholders, thereby democratizing data access within enterprises. The research aims to leverage a novel approach that explores how LLMs can be enhanced to convert natural language queries into SQL queries for business intelligence purposes. This involves retrieving business insights from a relational database management system using an intelligent agent abstraction of the LLM that ensures errors or subtasks required to return the appropriate response in natural language to the user's question are effectively handled. The study aims to evaluate various methods of optimizing LLMs for the specified task and establish benchmarks for comparing optimized systems developed during this research, focusing on their efficiency, accuracy, and usability (Jahan, Laskar, Peng, & Huang, 2024). Specifically, we aim at experimenting with fine-tuning a model for the domain-specific task involved in this work and evaluating the performance of the model in comparison with few shot in-context learning as performance enhancement techniques for pre-trained language models (Mosbach, Pimentel, Ravfogel, Klakow, & Elazar, 2023). Other task optimization methods such as optimized prompt engineering are also explored to achieve the highest level of efficiency and accuracy for the prototype system (Velásquez-Henao, Franco-Cardona, & Cadavid-Higuita, 2023). The study also evaluates the prototype system's impact on business KPIs such as query response time, user engagement, and the quality of insights generated, while highlighting the trade-offs between model sophistication, computational costs, and practical usability, which are crucial considerations in enterprise environments.

3. Related work

This research has been inspired by some significant related work such as recent advancements in text-to-SQL conversion which focused on improving the accuracy and efficiency of converting natural language queries into SQL. In this work, we exploit the use of *intelligent agents* that offer advantages over traditional methods in text-to-SQL conversion for enterprise analysis. In a related research work, a comprehensive survey was provided on LLM-based agents, emphasizing their potential for Artificial General Intelligence (AGI) and their applications in various domains, including business analytics and decision-making processes (Xi et al., 2023). This highlights the suitability of LLMs as a foundation for intelligent agents in complex tasks such as text-to-SQL conversion. Also, the use of LLMs as a system of multiple expert agents to solve the Abstraction and Reasoning Corpus (ARC) Challenge was explored in a similar work, demonstrating the flexibility and effectiveness of LLMs in handling diverse tasks through zero-shot and few-shot learning (Tan & Motani, 2023). This approach aligns with our methodology of using LLMs to convert natural language queries into SQL queries for data retrieval. The ReAct framework has been used to synergize reasoning and acting in language models, allowing for interleaved reasoning traces and task-specific actions (Yao et al., 2023). This methodology improves the interpretability and accuracy of LLM-generated outputs, which is crucial for ensuring the correctness of SQL queries generated from natural language inputs. A previous work reviewed image analysis methods using intelligent agents in decision-making systems, highlighting the advantages of intelligent agents in providing intellectual support for complex tasks (Kotova, Pisarev, & Pisarev, 2023). Their findings support the use of intelligent agents in enhancing the performance of text-to-SQL systems by integrating reasoning capabilities with LLMs.

A comprehensive review of recent advancements in text-to-SQL conversion highlights the potential of integrating LLMs with intelligent agents. Works such as (Sun et al., 2024) and Tan et al. (2024) have explored novel models like SQL-PaLM and TREQS, which demonstrate significant improvements in parsing accuracy and usability. Additionally, Li, Zhang, Li, and Chen (2023) introduced RESDSQL-3B + NatSQL, a model that optimizes semantic parsing for complex SQL queries. These studies provide a solid foundation for evaluating the capabilities of GPT models, as well as emerging open-source alternatives, in handling enterprise-specific text-to-SQL tasks. Building on this foundation, our research situates itself within the broader context of enabling intuitive database interactions for non-technical stakeholders. By leveraging the capabilities of performance enhancement techniques for large language models, this study contributes to the ongoing discourse on enhancing the accessibility and efficiency of AI-powered business analytics tools.

The advancements in parameter-efficient fine-tuning techniques emphasize the importance of optimizing LLMs for specific tasks while minimizing computational costs (Ding et al., 2023). This is particularly relevant for our research, which involves fine-tuning LLMs for domain-specific text-to-SQL conversion tasks. *Large Language Model fine-tuning* involves using task-specific training data and further training the pre-trained model on a smaller, task-specific dataset. This dataset contains examples that are representative of the tasks the model will perform, ultimately, enhancing the LLMs performance. Despite the potential of LLM fine-tuning, its application in various fields and domains, including medicine, engineering, social science, and humanities, requires further exploration (Baldazzi et al., 2023). In comparison to using LLM fine-tuning technique as a means of adapting models to a specified task, the concept

of *few-shot in-context learning* was explored in this work. In related work few-shot learning methodology in medical imaging was explored, demonstrating the potential of few-shot learning to handle data scarcity in specialized domains (Nayem et al., 2023). This supports our comparative analysis of few-shot in-context learning and fine-tuning techniques for enhancing LLM performance in text-to-SQL tasks. In-context learning refers to the ability of a large language model to understand and respond based on the context provided in the input. It involves optimizing a model for a specific task without updating the weights of the model (Brown et al., 2020). It is a learning paradigm that aims to train models with limited data by providing them with relevant context examples within the model prompt at run time. By integrating these recent studies, our research aligns with current advancements in the field and also addresses the critical need for efficient, accurate, and user-friendly text-to-SQL conversion systems.

4. Methodology

4.1. Research design

We evaluated and compared methods of enhancing the latest text-to-SQL conversion techniques, focusing on the integration of intelligent agents and Large Language Models (LLMs) for enterprise analysis. A blend of qualitative and quantitative research paradigms was used such as; exploring new methods of text-to-SQL conversion, review of current state-of-the-art methods, performance evaluation of models, and the comparative analysis of model fine-tuning versus few-shot learning for model optimization. The research design adopted is a mixed-method design, incorporating both exploratory and experimental elements. This allowed the exploration of complex research questions through various lenses, enhancing the robustness and applicability of the research findings.

The requirements for the design of the prototype in this research involve using LLMs for a specified text-to-SQL conversion task. An initial study involved a thorough literature review to understand current state-of-the-art systems for text-to-SQL conversion tasks. This study aims at identifying the top-performing LLMs that can be used in production with minimal errors for the text-to-SQL conversion task, and also achieve efficient results and outputs that are satisfactory to the end users. LLM rating benchmarks were considered to decide the most appropriate models for the use case in this study, considering the sophistication of the tasks involved and the need for reliable performance. A common approach to rating LLM abilities is the use of the Elo rating system for benchmarking LLM performance (Razali, Mustapha, Aziz, & Mostafa, 2023; Wu & Aji, 2023). For example, it has been used in calculating the relative skill levels of players in competitive games such as Chess. It is based on the principle that the difference in the ratings of two players predicts the outcome of a match. This system is particularly effective in contexts where multiple models compete in pairwise battles. In the following we elaborate the method.

The expected probability $E_{A,B}$ of player A with rating R_A winning against player B with rating R_B, is given by the formulas:

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}}$$

$$E_B = \frac{1}{1 + 10^{(R_B - R_A)/400}}$$

In this context, the factor 400 (Elo, 1978) accurately adjusts the sensitivity of the expected results to differences in ratings. A rating advantage of 400 points means 10:1 odds favoring the higher rated model, providing an interpretable metric for comparing performance. For an evenly matched pattern $R_A = R_B$, both E_A and E_B are equal to 0.5, representing a 50 : 50 chance of winning between the two patterns (Boubdir, Kim, Ermiş, Hooker, & Fadaee, 2023).

After each game, players' ratings are updated linearly based on their expected score versus the actual score. If player A's actual score in a game is S_A and the expected score was E_A, then the updated rating R'_A is calculated using:

$$R'_A = R_A + K \cdot (S_A - E_A)$$

In these formulas, K is a factor that determines the maximum possible adjustment per game, which can vary depending on the rules of the rating system in use.

Another benchmark for evaluating large language models is the Massive Multitask Language Understanding (MMLU) which measures a text model's multitask accuracy. Using these benchmarkratings shows the best available proprietary and open source models that were considered suitable for our model enhancement technique experiments. An additional performance optimization strategy adopted for the prototype system is optimized prompt engineering which facilitated bringing together all the various components required for retrieving the appropriate response from the model (Velásquez-Henao et al., 2023). In the context of the prototype system for this research, this involves including the database schema, the database tables description, and the natural language question in each prompt to ground the LLM in the relevant context for optimized response. Without optimized prompt engineering with the relevant components in this use case, the model responds with a default message stating the limitation of the scope of its training data for a domain-specific task (Marvin, Hellen, Jjingo, & Nakatumba-Nabende, 2024). Illustrations of the prompting techniques are given below:

- **Baseline Prompt**: "Generate SQL for: *What are the top 5 customers by revenue in 2004?*"
- **Optimized Prompt for Prototype System**: "Using the schema *[Schema Details]* and table descriptions *[Table Details]*, generate SQL for: *What are the top 5 customers by revenue in 2004?*".

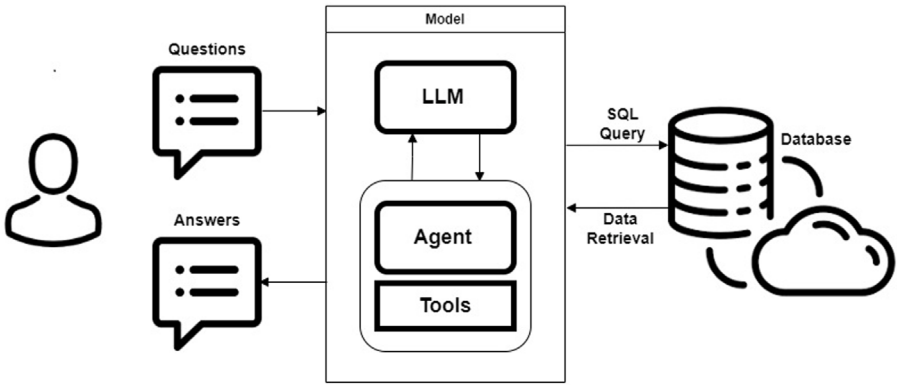

Fig. 1. Solution Workflow: This figure illustrates the overall workflow of the solution, showing how natural language queries are processed through an intelligent agent integrated with an LLM to generate actionable SQL queries. It highlights the synergy between LLMs and agents in simplifying complex database interactions for non-technical users.

In the context of the prototype system, the optimized prompt involved equipping an AI agent with the necessary tools for interacting with the database effectively using an LLM as a reasoning engine as shown in Fig. 1.

The application developed provides an intuitive interface for users to derive business insights directly from their databases by using natural language queries. Leveraging OpenAI Assistant's API as an intelligent agent and OpenAI's GPT-4 model as a reasoning engine, the application executes SQL queries behind the scenes to generate informative responses. The use case for this research is a domain-specific text-to-SQL conversion task which involves providing natural language questions as inputs to the system and retrieving the appropriate natural language answers, analytical chart, or graph as output. The database used for this research is the MySQL classicmodels sample database. The sample database was created from the publicly available dump script and hosted on a Google Cloud server for easy access. The classicmodels database was chosen for this research because it makes available a dump script that can be used to recreate the database and all its database records and relationships which depicts a database management system in a real-world setting. This makes it possible to build a database that the research prototype connects and interacts with to retrieve real-time insights. Popular datasets such as WikiSQL and KaggleDBQA only have question and query pairs that have been tested against specific benchmarks publicly available, but the dump scripts or database files needed to recreate the databases from which the datasets are derived are not publicly available. As such the classicmodels database was chosen to be able to implement a working prototype that demonstrates an end-to-end process of the research requirements including testing against appropriate benchmarks.

4.2. Experimentation with model enhancement techniques

4.2.1. Model fine-tuning setup

To address the domain-specific task of natural language to SQL query translation, we prepared the GPT-3.5-turbo and Meta-Llama-3-8B-Instruct models through a carefully designed fine-tuning process. The methodology involved creating a comprehensive set of example natural language questions paired with their corresponding SQL queries, focusing on the classicmodels database schema. This approach enabled a comparative analysis of model enhancement techniques, including fine-tuning and few-shot learning (Goswami, Prajapati, Saha, & Saha, 2024).

The training data was strategically developed through two primary methods. First, some question-and-query pairs were carefully crafted to represent various types of SQL queries necessary for retrieving information from a relational database. They were curated to encompass diverse SQL query types, including one-to-many, many-to-many, and single entity queries. These pairs were specifically tailored to the table names and fields of the MySQL classicmodels database. Additionally, we supplemented these examples by using the advanced GPT-4 model to synthetically generate high-quality training data specific to the classicmodels database schema (Puri, Spring, Patwary, Shoeybi, & Catanzaro, 2020; Tang, Han, Jiang, & Hu, 2023; Tremblay et al., 2018).

The entire dataset of question-and-query pairs was divided into training and test sets, with the training set used to fine-tune the fine-tuned models for text-to-SQL conversion tasks optimized for the specific database context. The test set had about 200 test samples for testing the various model configurations.

4.2.2. Few-shot in-context learning setup

Another GPT-3.5-turbo and a Meta-Llama-3-8B-Instruct-Turbo model without fine-tuning were set up for model enhancement techniques comparison. This time around examples of the curated question-and-query pairs were included in each input prompt given to the model when asking a new natural language question. As such the optimized prompt for the few-shot in-context learning setup included some examples of questions and their corresponding query, in addition to the new question posed to the model (Wang, Liu, Zhang, Leng, & Lu, 2023). In contrast, in other model configurations used for comparison, the input prompts did not contain these examples of question-and-query pairs. This distinction in setup was intended for a comparative analysis of the model enhancement techniques.

4.2.3. Model evaluation process

The experimentation involved an evaluation of six distinct models across different model enhancement techniques (Chang et al., 2024). The evaluation framework included two primary categories of models: fine-tuned models and few-shot in-context learning models. Specifically, the lineup consisted of a GPT-3.5-turbo and a Meta-Llama-3-8B-Instruct model that were fine-tuned using classicmodels database question-and-query pairs, alongside their counterparts configured with few-shot in-context learning. Additionally, two state-of-the-art (SOTA) models, GPT-4 and Llama-3.3-70B-Instruct-Turbo were included to provide a benchmark for comparison. These models were subsequently tested across three evaluation benchmarks to assess their performance comprehensively, these include:

Valid SQL (VA): This metric assesses whether the SQL queries generated by the model are syntactically correct. It is a binary check, marking a query as valid if it adheres to the SQL syntax rules and invalid otherwise. The steps involved in this evaluation benchmark include compiling the set of natural language questions in the test set that was used to test the models being evaluated. The questions were imputed into the models to generate SQL queries. The SQL queries generated were evaluated as either 'valid' or 'invalid' based on their adherence to SQL syntax rules. This is a binary evaluation – the query is either correct or not, with no partial credit. The number of valid queries for each model was tallied. The percentage of valid SQL queries for each model was calculated by dividing the number of valid queries by the total number of queries tested.

Execution Accuracy (EX): This involves running the generated SQL queries against a database and checking if they return the correct results. The accuracy is measured by comparing the output of the query with the expected result. The steps involved in this evaluation benchmark include compiling the set of natural language questions in the test set that was used to test the models being evaluated. The questions were imputed into the models to generate SQL queries. The outputs from the SQL queries generated were compared with the results returned by the corresponding queries for the input question in the test set. This involved checking if the retrieved data matches the expected data or not. The execution accuracy for each model was determined by calculating the percentage of queries that returned the correct result.

Test-Suite Accuracy (TS): This metric tests the semantic accuracy of the SQL queries. It checks if different variations of a query (that might have different syntax but are meant to achieve the same result) are semantically equivalent. This is usually done by running a suite of tests to ensure that the queries retrieve the correct data regardless of their syntactical differences (Zhong, Yu, & Klein, 2020). The steps involved in this evaluation benchmark include creating a set of distilled databases from the MySQL classicmodels database. These are smaller or modified versions of the original database that maintain its structural integrity but vary in data content. The aim was to cover a wide range of possible query scenarios. The model-generated queries and the gold queries (correct SQL queries from the test dataset) are collated. The model-generated queries and the gold queries are run against the distilled test databases. The results (denotations) of these queries are compared. A model-generated query is considered semantically accurate if its results match those of the gold query across all the test databases. If a model-generated query is semantically accurate, it should return results consistent with the logic of the gold query across these varied databases. For example, if the query is about finding "all customers who have placed more than three orders" the semantic accuracy is about correctly identifying such customers, irrespective of who they are or how many there are in each distilled database.

Each metric provides a different angle of assessment, which evaluates syntactic correctness, functional correctness, and semantic equivalence, respectively. These metrics together give a comprehensive understanding of the model's capabilities in converting natural language to SQL.

5. Results and discussion

5.1. Comparative analysis of LLMs for text-to-SQL conversion

After refining our approach and expanding our evaluation, we tested six different models on a set of 200 domain-specific text-to-SQL queries derived from the MySQL *classicmodels* database. These included two fine-tuned models (GPT-3.5-turbo and Meta-Llama-3-8B-Instruct), their respective few-shot in-context learning configurations, and two state-of-the-art (SOTA) models (GPT-4 and Llama-3.3-70B-Instruct-Turbo) serving as performance benchmarks. The evaluation was conducted using three primary metrics:

- **Valid SQL (VA):** Measures the percentage of syntactically correct SQL queries.
- **Execution Accuracy (EX):** Assesses whether executing the generated SQL against the database yields correct results.
- **Test-Suite Accuracy (TS):** Evaluates the semantic equivalence of the generated SQL queries to the reference queries by testing their performance across multiple distilled test databases.

Table 1 summarizes the performance of all six models on these three metrics.

5.1.1. Valid SQL (VA) evaluation

Across the evaluated models, GPT-4 and Llama-3.3-70B-Instruct-Turbo achieved the highest VA scores, both at 99.00% as shown in Table 1. Their ability to produce syntactically correct SQL queries at this rate underscores the sophistication of these advanced LLMs. Among the remaining models, the fine-tuned GPT-3.5-turbo model reached 97.00%, while its few-shot configuration scored 94.50%. This indicates that fine-tuning GPT-3.5-turbo on domain-specific data improves its grammatical adherence to SQL compared to relying solely on a few-shot prompt structure.

The Meta-Llama-3-8B-Instruct models exhibited lower VA scores: 82.50% when fine-tuned and 69.50% with few-shot learning. These results suggest that while fine-tuning can substantially boost performance even for smaller or less capable models, some models may still lag behind top-tier LLMs in raw syntactic accuracy.

Table 1
Performance evaluation of LLMs on domain-specific Text-to-SQL conversion task.

Model	Valid SQL (VA)	Execution accuracy (EX)	Test-suite accuracy (TS)
Fine-tuned GPT-3.5-turbo	97.00	91.50	79.50
GPT-3.5-turbo with Few-shot	94.50	87.00	74.00
GPT-4	99.00	95.50	82.00
Fine-tuned Meta-Llama-3-8B-Instruct	82.50	79.00	72.50
Meta-Llama-3-8B-Instruct-Turbo with Few-shot	69.50	62.50	56.00
Llama-3.3-70B-Instruct-Turbo	99.00	95.00	81.50

5.1.2. Execution accuracy (EX) evaluation

Execution accuracy assesses how well the generated queries retrieve correct results from the underlying database. GPT-4 again led with an EX score of 95.50%, followed closely by Llama-3.3-70B-Instruct-Turbo at 95.00%. These near-parity performances highlight that both proprietary and cutting-edge open-source LLMs can reliably translate natural language queries into accurate database commands.

Fine-tuned GPT-3.5-turbo achieved 91.50% EX, outperforming the GPT-3.5-turbo with few-shot setup at 87.00%. Similarly, fine-tuning improved the Meta-Llama-3-8B-Instruct model's EX from 62.50% (few-shot) to 79.00%. These results emphasize that tailored training data can substantially enhance a model's ability to not only produce syntactically correct SQL but also retrieve the intended information accurately.

5.1.3. Test-suite accuracy (TS) evaluation

Test-suite accuracy measures the semantic fidelity of the generated queries, whether they produce correct and consistent results across a range of scenario variations. GPT-4 achieved the highest TS score (82.00%), slightly surpassing Llama-3.3-70B-Instruct-Turbo at 81.50%. The close scores suggest that both models excel at understanding the underlying semantics of the queries, maintaining their performance across different test databases.

Fine-tuned GPT-3.5-turbo outperformed its few-shot variant, scoring 79.50% compared to 74.00%. This again reinforces the benefit of domain-specific fine-tuning in capturing deeper semantic nuances. The Meta-Llama-3-8B-Instruct models followed a similar pattern, with fine-tuning elevating TS from 56.00% to 72.50%.

Fig. 2 presents a comparative analysis of the six models' performance across three key metrics: valid SQL syntax, execution accuracy, and test-suite accuracy.

5.2. Discussion of results

The evaluation of tested models provides several key insights. Firstly, both GPT-4 and Llama-3.3-70B-Instruct-Turbo stand out as top performers across all metrics. Their consistently high scores for VA, EX, and TS indicate advanced levels of syntactic correctness, accuracy in execution, and robust semantic understanding. This suggests that the growing sophistication of both proprietary (GPT-4) and high-parameter open-source (Llama-3.3-70B) models can reliably handle domain-specific text-to-SQL tasks at scale.

Secondly, the comparisons between fine-tuned and few-shot versions of GPT-3.5-turbo and Meta-Llama-3-8B-Instruct reveal a clear trend. Fine-tuning significantly boosts performance in all three metrics. By exposing the model to specialized question-query pairs from the target database schema, fine-tuning enhances both the syntactic and semantic accuracy of the generated queries. This result confirms that tailor-made training data and fine-tuning procedures can narrow the performance gap between smaller, less capable models and the top-tier LLMs.

Thirdly, while few-shot in-context learning offers flexibility and rapid adaptation to new tasks without retraining, it generally underperforms compared to fine-tuning. For domain-specific requirements, especially those demanding high precision and consistency, the investment in fine-tuning appears well justified. Given that enterprises often seek maximum reliability in automated database querying, the superior performance of fine-tuned models has practical significance.

The results highlight a variety of choices for organizations. Models like GPT-4 and Llama-3.3-70B-Instruct-Turbo deliver exceptional performance but at potentially higher computational and resource costs. Smaller models, such as GPT-3.5-turbo or Meta-Llama-3-8B-Instruct, can achieve respectable accuracy when fine-tuned, potentially offering a more cost-effective solution. This trade-off between model sophistication, resource investment, and end-user usability is crucial for enterprises seeking to deploy AI-driven analytics tools. Overall, the results confirm that advanced LLMs and intelligent agents can make text-to-SQL conversion more accessible and efficient. As such, organizations can tailor solutions that strike the right balance between performance, cost, and usability by combining appropriate model selection strategies, ranging from using cutting-edge proprietary or open-source models to fine-tuning them on domain-specific data.

Our statistical analysis, using McNemar's test, which was designed to handle paired nominal (correct/incorrect) data, confirms that the fine-tuned configurations of both GPT-3.5-turbo and Meta-Llama-3-8B-Instruct significantly outperform their respective few-shot setups in terms of Execution Accuracy (EX). Specifically, when comparing fine-tuned GPT-3.5-turbo (91.5% EX) to its few-shot counterpart (87.0% EX), the contingency tables were formed based on each model's correct/incorrect answers out of 200 questions. The McNemar's test yielded a chi-square value of approximately 4.26 ($p \approx 0.039$), indicating a statistically significant difference at the 5% level. A similar comparison for fine-tuned Meta-Llama-3-8B-Instruct (79.0% EX) against its few-shot version (62.5% EX) resulted

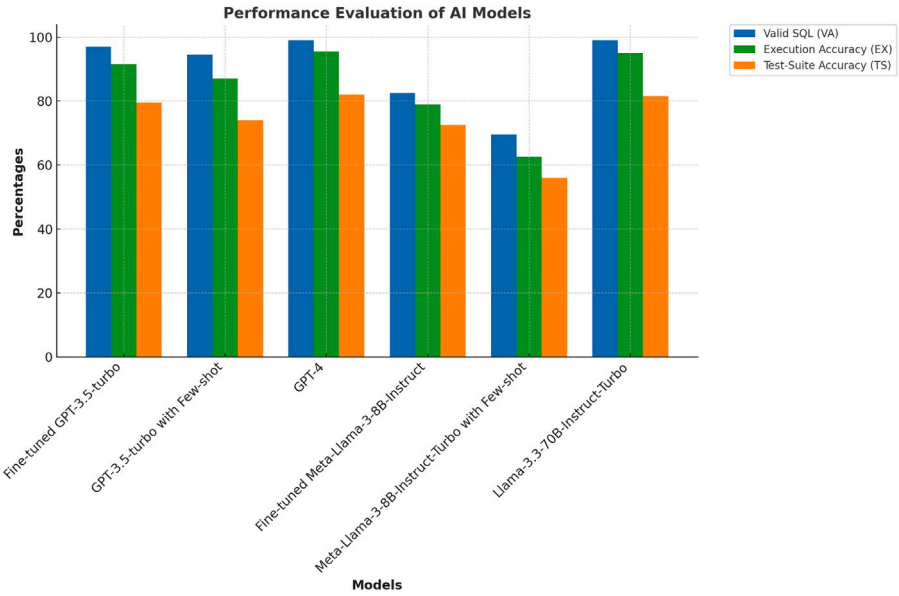

Fig. 2. Performance Evaluation of tested Models on Evaluation Metrics: This figure compares the performance of different LLMs on valid SQL syntax, execution accuracy, and test-suite accuracy. It shows that fine-tuned models generally outperform few-shot setups.

in a chi-square value of approximately 25.33 ($p \approx 4.8 \times 10^{-7}$), again demonstrating a highly significant advantage of fine-tuning. These results demonstrate that the observed performance improvements for fine-tuned models over few-shot approaches represent meaningful gains in accuracy.

5.3. Theoretical and practical implications

This research significantly advances the understanding of integrating LLMs with intelligent agents for text-to-SQL conversion. By comparing different optimization techniques (specifically, fine-tuning versus few-shot learning) the study demonstrates that model fine-tuning yields superior accuracy and usability for domain-specific tasks such as tasks for specific enterprise databases. This finding reinforces the importance of tailored optimization in the development of LLMs for specialized applications. Moreover, this research introduces an innovative approach to converting SQL results into actionable insights, both textually and graphically, which democratizes data access and makes complex data analysis accessible to non-technical stakeholders. This comprehensive end-to-end process showcases the potential of LLMs to enhance traditional business intelligence frameworks by providing more intuitive and effective data interaction capabilities (see Figs. 3–9).

From a practical standpoint, the integration of LLMs for text-to-SQL conversion addresses a critical need in enterprises to democratize data access. Non-technical stakeholders can interact with databases using natural language queries, thereby reducing dependency on technical teams and improving efficiency in data retrieval and decision-making processes. The study also highlights important trade-offs between model sophistication, computational costs, and enterprise usability. For instance, while GPT-4 offers superior performance, it comes with higher computational costs. These insights are crucial for enterprises to make informed decisions about which model to deploy based on their specific needs and resource constraints. Furthermore, the research suggests that fine-tuning models for specific enterprise databases can lead to significant performance improvements, implying that enterprises should invest in customizing local LLMs to their datasets to maximize efficiency and accuracy. These practical implications underline the transformative potential of AI-powered business analytics tools in enhancing decision-making and operational efficiency.

To clarify the research aims and their practical value for non-technical workers, we highlight several scenarios where intuitive text-to-SQL conversion tools can democratize data access: for example, a marketing associate without SQL expertise can quickly assess sales trends, a human resources manager can track employee performance metrics, or a logistics coordinator can promptly review shipment and delivery histories. By allowing such users to interact with databases through natural language queries rather than requiring technical SQL skills, these tools potentially reduce the dependency on specialized data teams. In fact, this approach may exemplify the “inverse-skilled bias”, whereby generative AI technology disproportionately benefits workers with fewer technical

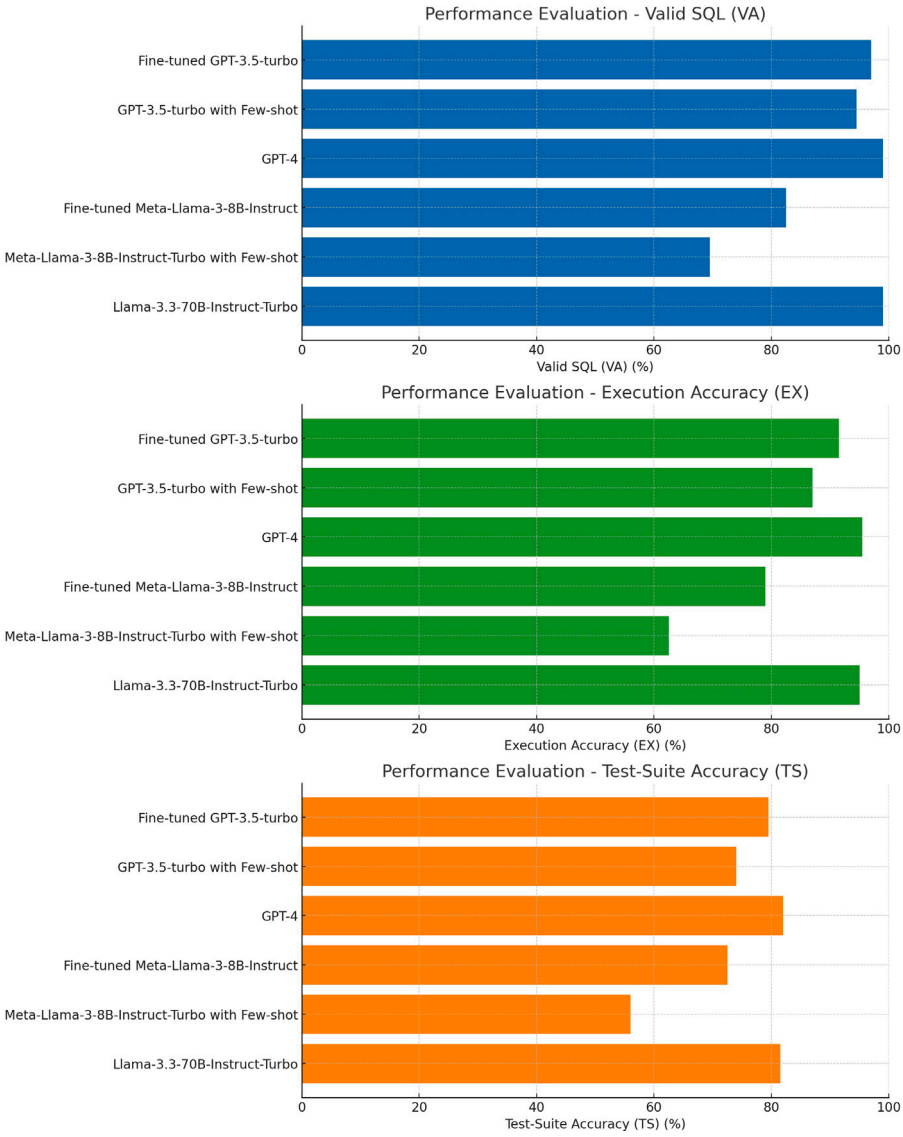

Fig. 3. Performance Evaluation of LLMs on Domain Specific Text-to-SQL Conversion Task: The chart showcases the superior performance of GPT-4 and Llama-3.3-70B-Instruct-Turbo in domain-specific tasks across various evaluation metrics, emphasizing the effectiveness of advanced LLMs for enterprise use.

skills, such as those lacking formal database training, by granting them more direct, accessible pathways to critical data-driven insights, thereby leveling the playing field and empowering a broader range of stakeholders (Capraro et al., 2023).

Although our research demonstrates the potential of LLMs and intelligent agents for text-to-SQL conversion in enterprise analytics, several limitations must be acknowledged. Firstly, our study relies on a relatively small, domain-specific dataset, which may limit the broader applicability of the findings. The chosen database and synthesized question-query pairs, while useful for demonstrating proof-of-concept, may not fully capture the complexity and diversity of queries seen across different industries, languages, or database schemas. Additionally, relying on benchmark models used in this study inherently introduces biases stemming from their pre-training data, potentially skewing performance in unforeseen ways. Lastly, as LLM outputs are probabilistic, model responses could be sensitive to variations in prompts, and subtle differences in prompt engineering might influence accuracy, potentially introducing a source of bias and making exact reproducibility challenging.

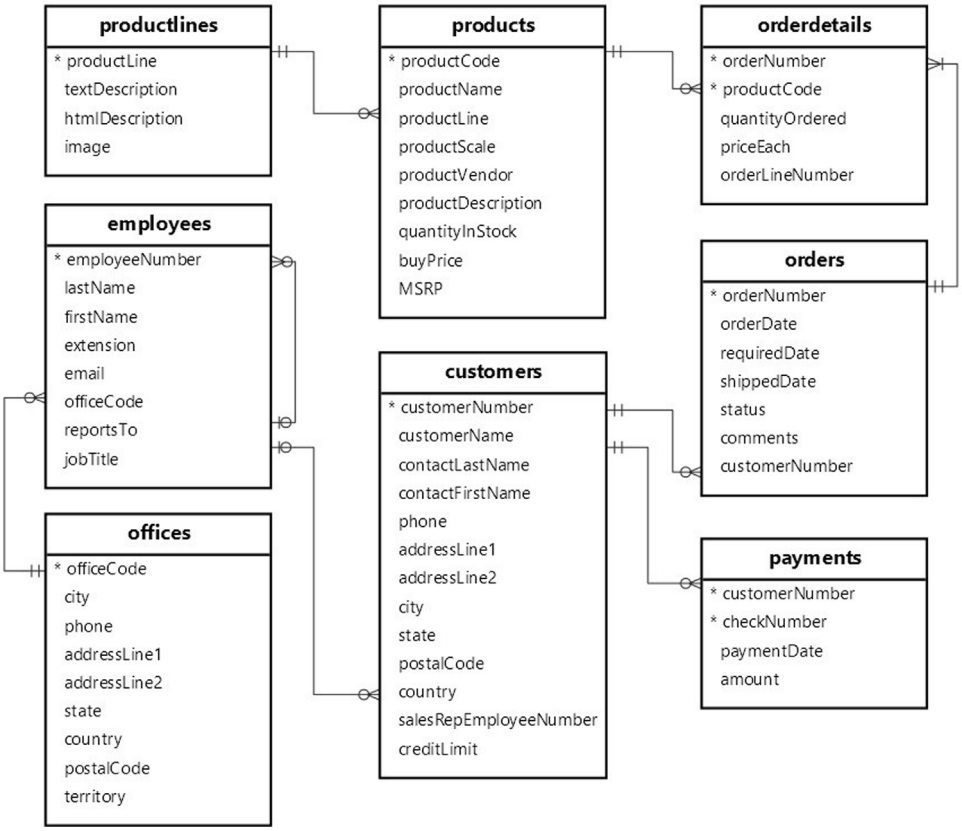

Fig. 4. Classicmodels Database Schema: This schema illustrates the structure of the MySQL classicmodels database used in the study. It provides essential context for understanding how the LLM-generated SQL queries interact with the database.

6. Conclusions and recommendations

6.1. Conclusions

This research offers a comprehensive and rigorous examination of how large language models (LLMs) and intelligent agents can revolutionize text-to-SQL conversion within the realm of business analytics. By expanding the evaluation scope to encompass a significant number of test queries and incorporating a diverse range of models (including GPT-3.5-turbo, Meta-Llama-3-8B-Instruct, GPT-4, and Llama-3.3-70B-Instruct-Turbo) this study demonstrates the remarkable capacity of both proprietary and advanced open-source models to accurately translate natural language questions into executable SQL code. Notably, GPT-4 and Llama-3.3-70B-Instruct-Turbo consistently exhibit superior performance across key metrics, including valid SQL syntax, execution accuracy, and semantic fidelity. Furthermore, the research underscores the significant role of fine-tuning in enhancing domain-specific accuracy, often surpassing the performance achieved through few-shot in-context learning. Specifically, customizing models such as GPT-3.5-turbo or Meta-Llama-3-8B-Instruct to align with a specific enterprise database demonstrably improves their SQL generation capabilities. These findings emphasize the critical importance of selecting the most appropriate model architecture, whether proprietary or open-source, while concurrently employing targeted fine-tuning or innovative prompt engineering techniques to optimize the balance between cost-effectiveness and accuracy. This study also highlights the transformative potential of integrating LLMs into intelligent-agent frameworks. By empowering non-technical users to query databases using everyday language, these systems have the capacity to significantly democratize data access. This automation reduces reliance on specialized data teams, thereby accelerating organizational decision-making processes. By facilitating direct and intuitive interaction with business-critical information through user-friendly interfaces and precise data retrieval mechanisms, these AI-driven tools are poised to reshape the landscape of business analytics, fostering a more inclusive and efficient environment for data-driven insights.

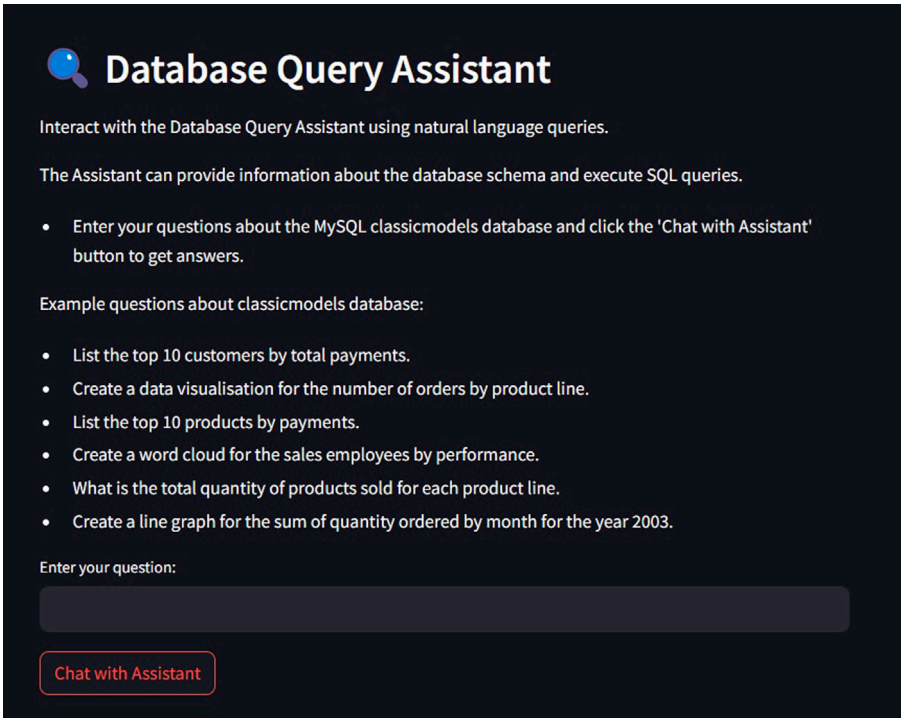

Fig. 5. Prototype's Frontend Interface: This figure displays the intuitive frontend interface designed for non-technical users to input natural language queries. The key message is the accessibility and usability of the system for business stakeholders.

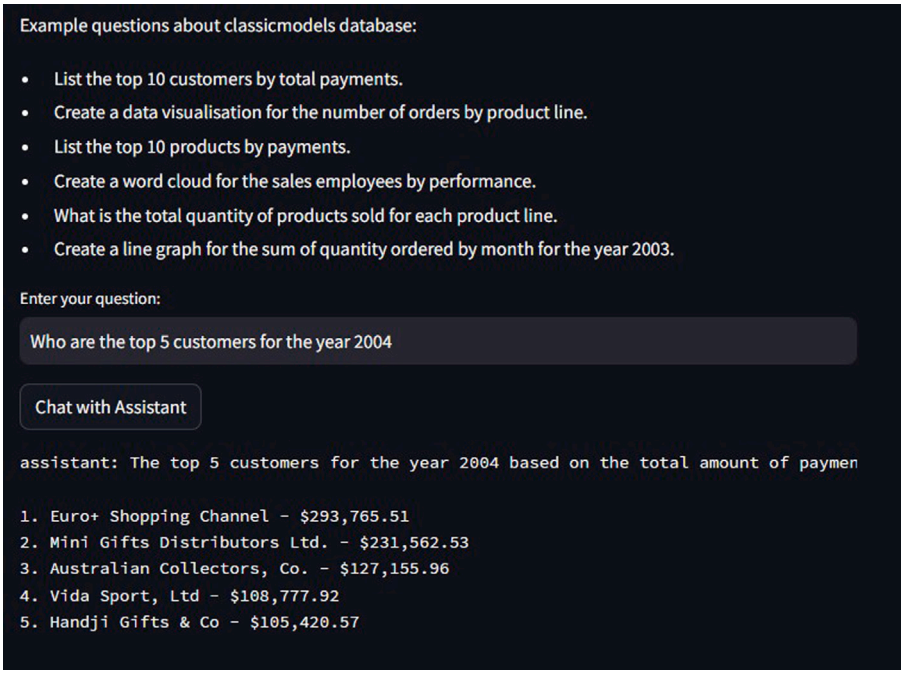

Fig. 6. Prototype's response to a question about top 5 customers in 2004: This example demonstrates the system's ability to generate accurate SQL queries and provide insightful answers, showcasing its practical application in enterprise scenarios.

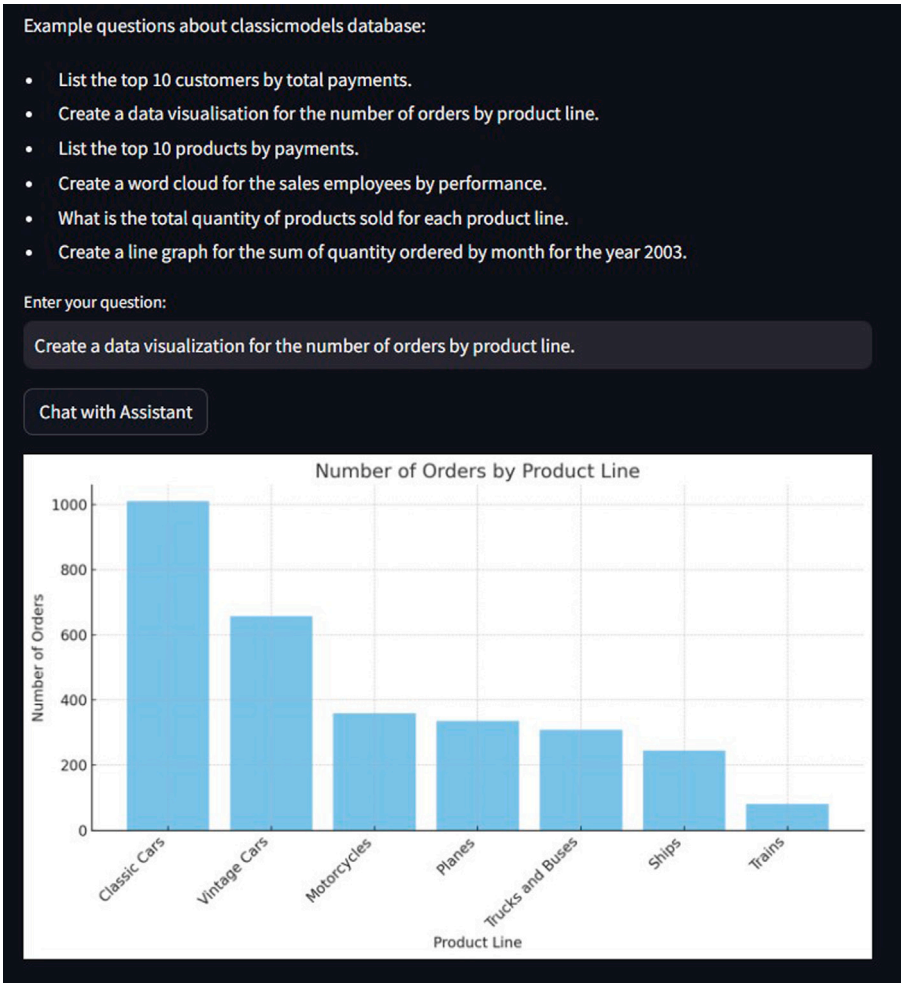

Fig. 7. Prototype's response to a question about creating a data visualization for the number of orders by product line: The figure illustrates the system's capability to convert queries into actionable insights, including visualizations, thereby enhancing decision-making.

6.2. Recommendations and future work

Further research and development in the field of artificial intelligence using LLMs and intelligent agents in business applications are essential. The integration of these technologies in various business applications will pave the way for more sophisticated, user-friendly, and efficient data retrieval systems. Organizations should consider customizing LLMs to their specific business needs as required. Also, exploring the scaling of LLM solutions to work with different sizes and types of databases is recommended to maximize their applicability across various business sectors. Future research should explore incorporating a broader range of LLMs and task-specific architectures, including open-source and parameter-efficient models, to enhance generalizability and reduce computational overhead. In-depth investigations into user experience, explainability, and trustworthiness, such as how non-technical users interpret and act upon generated SQL queries, can further refine these systems, ensuring seamless integration into existing workflows. Moreover, systematic exploration of prompt engineering methods, larger and more diverse datasets, and advanced evaluation metrics will support continual model improvement and more equitable outcomes. Ultimately, these efforts will accelerate the adoption of AI-driven decision-support tools across industries, making advanced analytics more accessible, interpretable, and beneficial for a wide range of stakeholders. Given the resource-intensive nature of LLM-driven applications, especially advanced proprietary models, organizations should balance the cost and performance benefits. Fine-tuned models might offer a more cost-effective solution for smaller enterprises or specific application use cases. Ethical and privacy concerns must be addressed efficiently

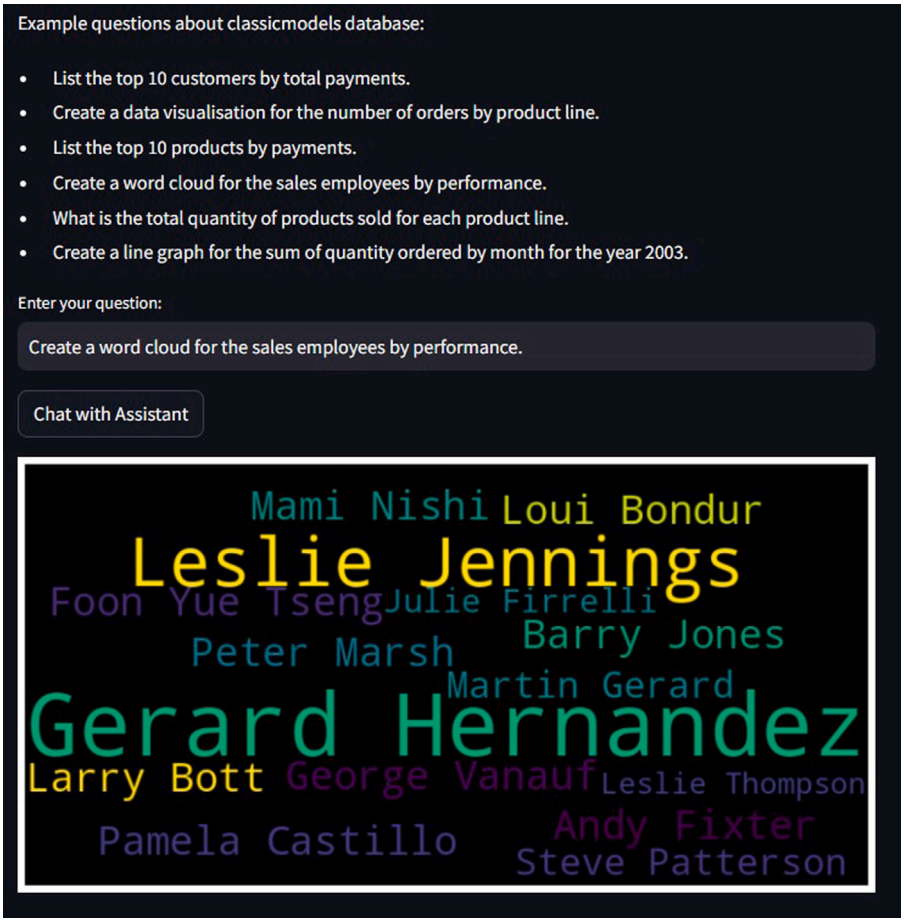

Fig. 8. Prototype's response to a question about creating a word cloud for the sales employees by performance: This example highlights the system's flexibility in generating diverse visual analytics, making complex data more accessible.

for any AI project implementation (Capraro et al., 2023). Ensuring data security, maintaining user privacy, and transparent data handling practices should be an integral part of the system design. While proprietary models currently lead in performance, the rapid advancement of open-source LLMs presents a viable alternative. Organizations should monitor developments in this space for potential future integration, particularly for cost-saving and transparency reasons.

CRediT authorship contribution statement

Samuel Ojuri: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **The Anh Han:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Methodology, Investigation, Conceptualization. **Raymond Chiong:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Alessandro Di Stefano:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

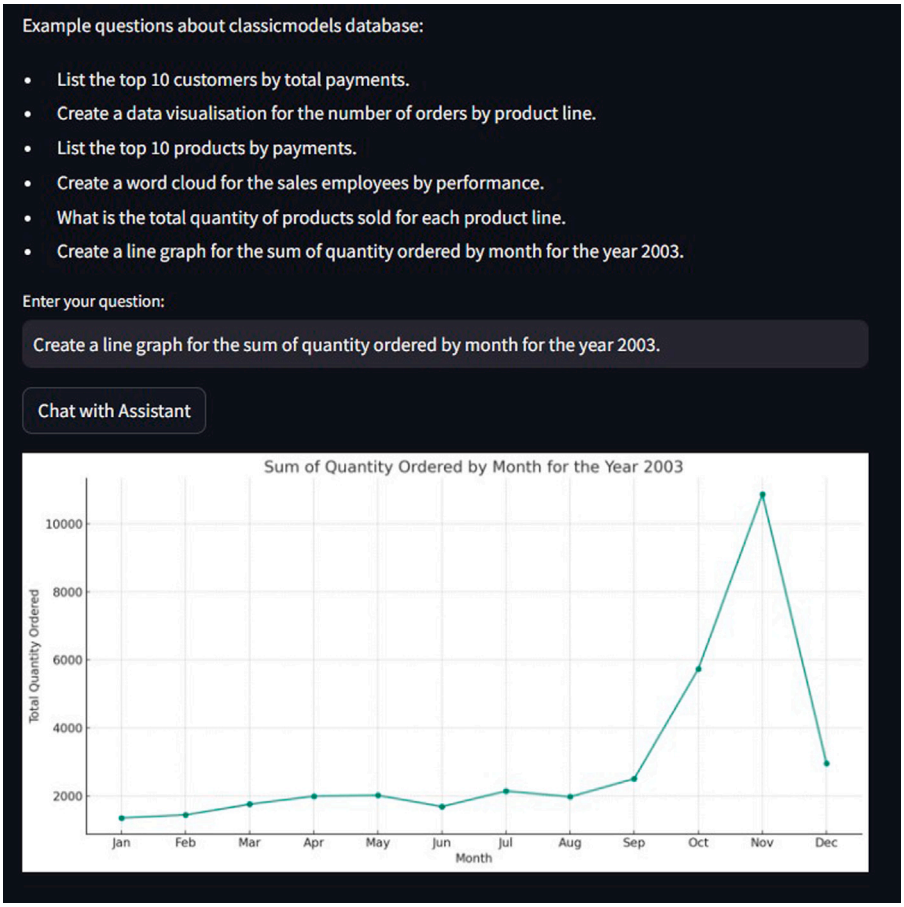

Fig. 9. Prototype's response to a question about creating a line graph for the sum of quantity ordered by month for the year 2003: This figure demonstrates the system's proficiency in handling temporal data and generating relevant visualizations, aiding in trend analysis.

Acknowledgment

The Anh Han is supported by EPSRC (grant EP/Y00857X/1).

Appendix A. Supplementary material

- A link to the code used for the system prototype built in this research is given below:
  Clickhere

Data availability

Data will be made available on request.

References

Baldazzi, T., Bellomarini, L., Ceri, S., Colombo, A., Gentili, A., & Sallinger, E. (2023). Fine-tuning large enterprise language models via ontological reasoning. In A. Fensel, A. Ozaki, D. Roman, & A. Soylu (Eds.), *Rules and reasoning* (pp. 86–94). Cham: Springer Nature Switzerland.

Boubdir, M., Kim, E., Ermiş, B. H., Hooker, S., & Fadaee, M. (2023). Elo uncovered: Robustness and best practices in language model evaluation. arXiv. arXiv:2311.17295. URL: https://api.semanticscholar.org/CorpusID:265498394.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Capraro, V., Di Paolo, R., Perc, M., & Pizziol, V. (2024). Language-based game theory in the age of artificial intelligence. *Journal of the Royal Society Interface*, *21*(212), http://dx.doi.org/10.1098/rsif.2023.0720.

Capraro, V., Lentsch, A., Acemoglu, D., Akgun, S., Akhmedova, A., Bilancini, E., Bonnefon, J.-F., Brañas-Garza, P., Butera, L., Douglas, K. M., et al. (2023). The impact of generative artificial intelligence on socioeconomic inequalities and policy making. PNAS Nexus.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), http://dx.doi.org/10.1145/3641289.

Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C.-M., Chen, W., et al. (2023). Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, *5*(3), 220–235.

Elgohary, A., Hosseini, S., & Hassan Awadallah, A. (2020). Speak to your parser: Interactive text-to-SQL with natural language feedback. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 2065–2077). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.acl-main.187, Online. URL: https://aclanthology.org/2020.acl-main.187.

Fadloun, S., Meshoul, S., Hosseini, M., & Choutri, K. (2023). Visual analytics using machine learning for transparency requirements. *Mathematics*, http://dx.doi.org/10.3390/math11143091.

Goswami, J., Prajapati, K. K., Saha, A., & Saha, A. K. (2024). Parameter-efficient fine-tuning large language model approach for hospital discharge paper summarization. *Applied Soft Computing*, *157*, Article 111531. http://dx.doi.org/10.1016/j.asoc.2024.111531, URL: https://www.sciencedirect.com/science/article/pii/S1568494624003053.

Grohs, M., Abb, L., Elsayed, N., & Rehse, J.-R. (2023). Large language models can accomplish business process management tasks. In *International conference on business process management* (pp. 453–465). Springer.

Jahan, I., Laskar, M. T. R., Peng, C., & Huang, J. X. (2024). A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. *Computers in Biology and Medicine*, *171*, 108189.

Kanbach, D. K., Heiduk, L., Blueher, G., Schreiter, M., & Lahmann, A. (2024). The GenAI is out of the bottle: generative artificial intelligence from a business model innovation perspective. *Review of Managerial Science*, *18*(4), 1189–1220. http://dx.doi.org/10.1007/s11846-023-00696-z.

Kotova, E. E., Pisarev, I. A., & Pisarev, A. S. (2023). Advantage of applying image analysis methods using intelligent agents in decision-making systems. In *2023 XXVI international conference on soft computing and measurements* (pp. 94–97). http://dx.doi.org/10.1109/SCM58628.2023.10159047.

Li, H., Zhang, J., Li, C., & Chen, H. (2023). RESDSQL: Decoupling schema linking and skeleton parsing for text-to-SQL. arXiv:2302.05965. URL: https://arxiv.org/abs/2302.05965.

Marvin, G., Hellen, N., Jjingo, D., & Nakatumba-Nabende, J. (2024). Prompt engineering in large language models. In I. J. Jacob, S. Piramuthu, & P. Falkowski-Gilski (Eds.), *Data intelligence and cognitive informatics* (pp. 387–402). Singapore: Springer Nature Singapore.

Mosbach, M., Pimentel, T., Ravfogel, S., Klakow, D., & Elazar, Y. (2023). Few-shot fine-tuning vs. In-context learning: A fair comparison and evaluation. arXiv. arXiv:2305.16938. URL: https://api.semanticscholar.org/CorpusID:258947047.

Nayem, J., Hasan, S. S., Amina, N., Das, B., Ali, M. S., Ahsan, M. M., & Raman, S. (2023). Few shot learning for medical imaging: A comparative analysis of methodologies and formal mathematical framework. arXiv. arXiv:2305.04401. URL: https://api.semanticscholar.org/CorpusID:258557608.

Powers, S. T., Linnyk, O., Guckert, M., Hannig, J., Pitt, J., Urquhart, N., Ekárt, A., Gumpfer, N., Han, T. A., Lewis, P. R., et al. (2023). The stuff we swim in: Regulation alone will not lead to justifiable trust in AI. *IEEE Technology and Society Magazine*, *42*(4), 95–106.

Puri, R., Spring, R., Patwary, M., Shoeybi, M., & Catanzaro, B. (2020). Training question answering models from synthetic data. arXiv. arXiv:2002.09599. URL: https://api.semanticscholar.org/CorpusID:211258652.

Razali, N., Mustapha, A., Aziz, A. Q. A. A., & Mostafa, S. A. (2023). Machine learning approach for Malaysia super league football match outcomes prediction based on elo rating system. In S. F. Syed Omar, M. H. A. Hassan, A. Casson, A. Godfrey, & A. P. P. Abdul Majeed (Eds.), *Innovation and technology in sports* (pp. 169–176). Singapore: Springer Nature Singapore.

Sen, J., Özcan, F., Quamar, A., Stager, G., Mittal, A. R., Jammi, M., Lei, C., Saha, D., & Sankaranarayanan, K. (2019). Natural language querying of complex business intelligence queries. In *Proceedings of the 2019 international conference on management of data*. URL: https://api.semanticscholar.org/CorpusID:195259241.

Sun, R., Arik, S. O., Muzio, A., Miculicich, L., Gundabathula, S., Yin, P., Dai, H., Nakhost, H., Sinha, R., Wang, Z., & Pfister, T. (2024). SQL-PaLM: Improved large language model adaptation for Text-to-SQL (extended). arXiv:2306.00739. URL: https://arxiv.org/abs/2306.00739.

Tan, Z., Liu, X., Shu, Q., Li, X., Wan, C., Liu, D., Wan, Q., & Liao, G. (2024). Enhancing text-to-SQL capabilities of large language models through tailored promptings. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation* (pp. 6091–6109). Torino, Italia: ELRA and ICCL, URL: https://aclanthology.org/2024.lrec-main.539.

Tan, J. C. M., & Motani, M. (2023). Large language model (LLM) as a system of multiple expert agents: An approach to solve the abstraction and reasoning corpus (ARC) challenge. arXiv. arXiv:2310.05146. URL: https://api.semanticscholar.org/CorpusID:263829220.

Tang, R., Han, X., Jiang, X., & Hu, X. (2023). Does synthetic data generation of LLMs help clinical text mining? arXiv. arXiv:2303.04360. URL: https://api.semanticscholar.org/CorpusID:257405132.

Tremblay, J., Prakash, A., Acuna, D., Brophy, M., Jampani, V., Anil, C., To, T., Cameracci, E., Boochoon, S., & Birchfield, S. (2018). Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 1082–10828). URL: https://api.semanticscholar.org/CorpusID:4929980.

Velásquez-Henao, J. D., Franco-Cardona, C. J., & Cadavid-Higuita, L. (2023). Prompt engineering: a methodology for optimizing interactions with AI-language models in the field of engineering. *DYNA*, *90*(230), 9–17.

Wang, J., Liu, K., Zhang, Y., Leng, B., & Lu, J. (2023). Recent advances of few-shot learning methods and applications. *Science China Technological Sciences*, *66*(4), 920–944.

Wu, M., & Aji, A. F. (2023). Style over substance: Evaluation biases for large language models. arXiv. arXiv:2307.03025. URL: https://api.semanticscholar.org/CorpusID:259360998.

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., ..., Yin, Z. et al. (2023). *The rise and potential of large language model based agents: A Survey*: *Technical Report*, Cornell University Library, arXiv.org.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. In *International conference on learning representations*.

Zhong, R., Yu, T., & Klein, D. (2020). Semantic evaluation for text-to-SQL with distilled test suites. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 396–411).