Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 12, 2025

ANALYSIS OF THE EFFECTIVENESS OF GENERATIVE AI MODELS FOR TEXT-TO-SQL TASKS IN BUSINESS INTELLIGENCE SYSTEMS

Mubbashir Ahmed[*1], Muhammad Zulqarnain Siddiqui[2], Muhammad Zamin Ali Khan[3]

[*1,2,3] *Department of Computer Science, Iqra University, Karachi, Pakistan*

[*1]mubbashir.21384_n@iqra.edu.pk, [2]zulqarnain@iqra.edu.pk, [3]zamin@iqra.edu.pk

DOI: https://doi.org/10.5281/zenodo.18298953

Keywords
Generative Artificial Intelligence, Large Language Models, Business Intelligence, Prompt Engineering, Text-To-Sql

Article History
Received: 03 November 2025
Accepted: 17 December 2025
Published: 31 December 2025

Corresponding Author: *
Mubbashir Ahmed

Abstract

The rising integrations of Generative Artificial Intelligence (AI) into Business Intelligence (BI) systems has transformed how organizations interact with data and enabled natural language based analytical reasoning through text-to-SQL generations. This study provides a comprehensive analysis of the effectiveness of recent Generative AI and Large Language Models (LLM) in executing text-to-SQL tasks. The study presents a comparative review of features and limitations of current and advanced text-to-SQL systems including MAC-SQL, SQL-PaLM, CHASE-SQL, and CHESS. Further, an overview of generative AI models with their architectural outline is also provided to highlight the evolution of large language model based approaches in structured data querying. Three state-of-the-art generative AI models has been selected including LLaMA 4, Qwen3 14B, Mixtral 8x7B to perform evaluations using custom evolution framework and a hybrid prompt template is also designed for text-to-SQL tasks. Evaluation metrics includes intent clarification accuracy, semantic clarification accuracy, SQL generation accuracy, execution response accuracy, and response latency. Further, the experimental results demonstrates that LLaMA 4 model acquired great overall performance across all major evaluation metrics. The findings confirms the effectiveness of Generative AI for text-to-SQL generation and highlights its potential in designing a next-generation, intelligent business intelligence systems. Future work will extend this research towards multi-agent and domain-related real-time business intelligence framework.

1.0 Introduction

The recent developments in artificial intelligence were mainly inspired by the development of large language models (LLMs) such as GPT-4, PaLM-2, LLaMA, Mixtral, and Qwen. Such advances have given rise to the fast rise of the generative artificial intelligence (GenAI), which is a paradigm that allows machines to produce natural language solutions in a manner that generates human-like responses by addressing contexts. Generative AI models feature transformer-based architectures, which enable them to compute long-range dependencies, intent, and generate structured output in real time, unlike the previous rule based or narrowly-scoped artificial intelligence systems. This change is a radical departure of deterministic and domain-specific systems in favour of flexible and general-purpose reasoning engines that can be used to support more complex analytical processes across domains.

Generative AI has proven to be effective in a broad application suited in machine translation, summarization, code generation, conversational agents, and multimodal content synthesis. In addition to a shallow language generation, contemporary LLMs are becoming more and more reasoning agents in that they can break down tasks, impose logical constraints, and refine outputs through iteration. Such features

are especially important in business intelligence (BI) systems, where the user may also need analytical information but lack the technical knowledge of querying databases or modeling information. The transformer-based models can provide contextual continuity to multi-turn interactions and, thus, are particularly useful in dialog-based analytics, query clarification, and data summarization tasks (Feuerriegel et al., 2024).

One of the main issues that has remained consistent in modern BI systems is the ability to engage with structured databases using natural language. Conventional BI applications mandate users to develop queries in Structured Query Language (SQL) which poses a major obstacle to using the systems by non technical decision makers. Text-to-SQL systems attempt to overcome this drawback by converting natural language queries into SQL query syntaxes. Nevertheless, initial systems were based on template matching, rule based parsers or supervised semantic parsers, which were difficult to extrapolate across domains, to deal with the complexity of schema, or to address the ambiguity of user intent. The introduction of generative AI into BI systems has revitalized the text-to-SQL research by highlighting models with the ability to reason semantically, link schemas, and interpret context at scale.

The increased maturity of generative AI has provided new opportunities to BI systems regarding the provision of real-time querying, automated data insight creation, and interactive exploration of data. As one example, a business user can type, "Show monthly sales of product ABC by region in the first quarter" and a generative model can guess what is being asked of it, map the needed schema elements, synthesize a valid SQL query, execute it and display the results in a summarized or detailed format. According to recent research, hybrid systems based on the integration of LLMs with schema-aware retrieval, execution-guided refinement, and rule-based validation are much more effective than previous ones (Vertsel and Rumiantsau, 2023; Busany et al., 2024).

Nevertheless, the application of generative AI in text-to-SQL tasks in practice is difficult to implement in the real world of BI activity. The degradation of model performance is usually caused by large database schema, ambiguous column names, domain-oriented names, and constraints of the execution efficiency. Besides, the multi-agent models and heavy fine-tuning pipelines have an extra computational cost, restricting access and scalability. These issues demonstrate the necessity of a systematic assessment of the contemporary generative AI models in the context of controlled but realistic BI.

This paper bridges this literature gap by providing a comparative analysis of three state-of-the-art generative AI models, including, LLaMA-4, Mixtral 8x7B and Qwen3-14B, on text-to-SQL tasks in the context of a business intelligence system. The study analyzes the model performance in several aspects based on standardized benchmarks and a hybrid framework of prompt engineering; these are intent clarification accuracy, semantic mapping accuracy, SQL generation correctness, execution accuracy, and response latency. This study provides empirical evidence by developing foundation models accessed through open access and reproducible evaluation measures by focusing on the design of the next generation intelligent BI systems which are accessible, reliable, and scalable.

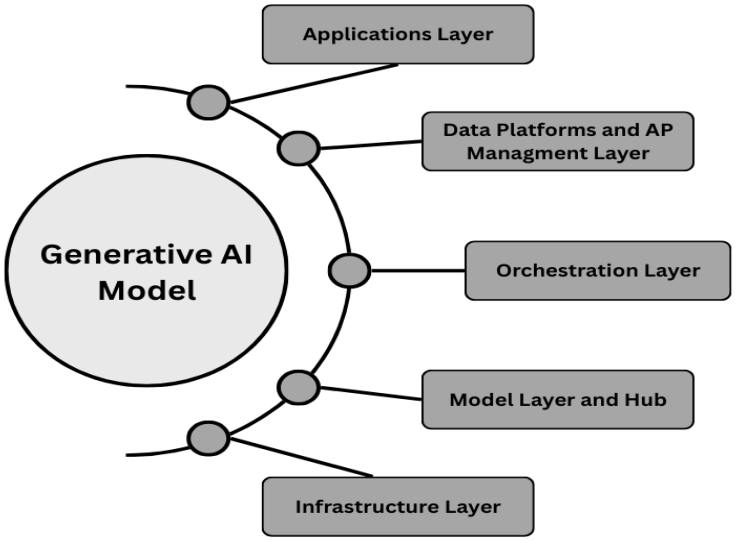


Fig. 1. GENERATIVE AI MODEL ARCHITECTURE DIAGRAM

Fig.1 demonstrates that the architecture of the generative artificial intelligence model discussed by Solulab (2024) is designed and grounded on 5 conceptual layers, and each layer is allocated to various set of responsibilities. This layered approach provides the model with modularity, scalability, and maintainability, which encourages enterprise-level organizations to design, deploy, and operate complex AI-powered solutions quite effectively. The layers are (i) Application Layer that provides interface between people and AI-driven systems like chatbots, content creation tools, virtual assistants and domain specific AI systems. In essence, the key role of this layer is to convert user inputs into structured requests which may be handled by other layers to produce outputs and the next layer is (ii) Data Platforms and AP Management Layer which is concerned with data ingestion, storage, authority and access of structured and unstructured data sources such as documents, databases, and external knowledge banks. The most important functions are preprocessing of data, metadata management, versioning, authentication, authorization and data security. (iii) Orchestration Layer manages the interrelations among data sources and data models and applications by specifying the processing logic, decision-making processes that govern the construction of the prompts, the choice of the models, and the processing and the production of the resulting output. The key functions of this layer are as follows, prompt management, agent framework, tool calling, workflow engines, and monitoring techniques with the fourth layer being (iv) Modal Layer and Hub, which contains the core generative AI models which could use hosted, open-source, fine-tuned or domain-specified models to do tasks such as multimodal generation, and reasoning. Its main functionalities are inference, multi-model architecture, model lifecycle management and performance optimization. Finally, (v) Infrastructure Layer gives the computational facilities of all the other layers mentioned above such as resources such as CPUs, GPUs, network, storage, and accessibility of the cloud or on-premises platforms. It has the mandate of making system more reliable, scalable and cost optimized.

2.0 Literature Review

2.1 Generative AI and Business Intelligence

Generative artificial intelligence has become a game changer technology in information systems and business analytics. In contrast to the previous artificial intelligence methods that focus on classification or prediction, the generative AI models are aimed at creating new content by learning probabilistic representations of language, code, and multimodal data. Storey et al. (2024) define generative AI as a sociotechnical system that is marked by the generative novelty, the high levels of apparent

intelligence, and the capacity to organize the inputs and outputs independently. Information systems Generative AI on the one hand reinvents the experience of interrelating with data, particularly by moving beyond tool-centered interface to conversational and collaborative intelligence. Feuerriegel et al. (2024) also conceptualize generative AI at model, system and application layers, highlighting its application in transforming human computer interaction in a business context. At the application level, generative AI facilitates the use of decision support systems, analytics and BI tools to aid in delegating, co-creating and hybrid intelligence. Nevertheless, the authors also point to undispersed issues such as hallucination, bias, transparency, and environmental cost, which need to be resolved to make enterprises responsible.

2.2 Text-to-SQL Systems and Multi-Agent Architectures

Recent text-to-SQL studies have changed the focus of single-pass prompting to multi-agent and multi-step reasoning systems. Wang et al. (2024) introduced MAC-SQL, a multi-agent system, which splits the text-to-SQL problem into schema selection, question decomposition, and SQL refinement. MAC-SQL is able to uses chain-of-thought reasoning and assigns specialized roles to agents, allowing it to execute both a BIRD and Spider benchmark with state of the art accuracy. Nevertheless, the use of the framework to rely on agent coordination enhances the complexity of inference and the computational expense. In the same fashion, Talaei et al. (2024) have presented CHESS, a multi-agent framework which is modular and suitable to industrial-scale databases. CHESS combines schema pruning, candidate generation and natural-language unit testing to enhance robustness in the presence of noisy and large schema. Although CHESS has good performance in limited resource environments, its multi-agent interaction and retrieval support adds to the latency and infrastructure cost. The proposed approach of the study by Pourreza et al. (2024) is called CHASE-SQL, which follows a paradigm of candidate-generation-and-selection and integrates multiple reasoning directions, execution-based ranking, and test-time refinement. This method enhances preciseness in execution at the expense of creating and assessing a massive pool of SQL applicants that can potentially induce redundancy and higher computation cost. Introduced by Sun et al. (2023), SQL-PaLM is a tuning-intensive model that integrates instruction fine-tuning, few-shot prompting, retrieval-augmented column selection, and execution-guided decoding. Although SQL-PaLM performs well with complex-schema, its pro-proprietary models and large-scale fine-tuning makes it less reproducible and accessible.

2.3 Benchmarking Datasets for Text-to-SQL Evaluation

Benchmark datasets are the main focus of the assessment of the text-to-SQL systems. The Spider dataset has been used as a classical benchmark in cross-domain semantic parsing, focusing on the generalization to unseen schemas. Nevertheless, its quite clean schemas and domain insensitivity limit its usage to real-life BI situations. To overcome this shortcoming, Li et al. (2024) proposed the BIRD dataset that comprises big, noisy, real-world databases in 37 professional fields. BIRD highlights the ambiguity of the schema, the rationality of values, and the efficiency of execution, which are the complexity of enterprise BI environments. The practical findings indicate that even developed LLMs are far behind human performance on BIRD, highlighting the challenge of real-world text-to-SQL tasks. More recently, Lei et al. (2024) introduced Spider 2.0 that targets end to end enterprise-level processes with the complex schema and with multiple SQL dialects as well as external documentation. These standards indicate an increasing disconnect between the academic assessment and real-world BI needs, which encourages additional studies of scalable and context-sensitive generative AI systems.

2.4 Prompt Engineering and Instruction Control

Prompt engineering has become a key control and optimizing tool on structured tasks of LLM. Schulhoff et al. (2024) classify the types of prompting as either instruction-based, demonstration-based, or retrieval-augmented,

and all are important to enhance the level of task alignment and output consistency. Nevertheless, Webson and Pavlick (2022) note that prompt-based advances are not always signs of in-depth semantic knowledge because model performance may be extremely sensitive to prompt surface differences. Instruction-based constraints on output and role prompting have been shown to be especially useful in domain-specific generation as in text-to-SQL generation. Wang et al. (2023) show role-guided prompting can reduce inter-domain confusion with the fine-tuning process, but does not reduce general language skills. These results indicate the application of structured, deterministic prompting strategies in evaluation-based BI applications.

3. Methodology

The proposed study uses a quantitative experimental research design, which assesses the usefulness of modern generative artificial intelligence models in text-to-SQL generation in a business intelligence (BI) system. The attempt of the methodology is such that it provides reproducibility, controlled comparison, and correspondence with real-world BI requirements. The experimental architecture applies a coordinated set of probing strategies and a set of standardized datasets to test the chosen generative AI models, which allows conducting a systematic evaluation of their performance across various performance dimensions applicable to enterprise analytics.

3.1 Research Design and Evaluation Framework

The experimental design is aimed at measuring the performance of the models to put natural language business queries into executable SQL statements. All models are evaluated with the same conditions that include homogeneous prompt templates, homogeneous decoding settings, and homogeneous evaluation measures to obtain consistency in assessments. The assessment system is intended to mirror real-life BI application requirements, whereby models need to infer the intent of the ambiguous user, match between semantic concepts and database models, generate syntactically correct SQL, and generate accurate output of the execution within reasonable latency constraints. The study does not use the subjective qualitative judgement, but uses the objective metric-based evaluation, by which the capabilities of models can be directly compared. Every generated SQL query is checked both syntactically and executing against benchmark databases to ensure that correctness is gauged by the actual query results and not apparent semblance.

3.2 Selection of Generative AI Models

Three state-of-the-art generative AI models were chosen to be evaluated: LLaMA-4, Mixtral 8x7B, and Qwen3-14B. The selection criteria depended on the architectural diversity, performance reported in the recent literature, and suitability to open-access or research deployment environments. LLaMA-4 is a high capacity model (transformer based) that is powerful in reasoning and is multilingual, so it is applicable to complex semantic interpretation tasks. The Mixtral 8x7B uses a sparse mix of experts architecture, where only a small number of parameters are activated when performing inference to trade off performance with computational efficiency. Qwen3-14B A dense, instruction-tuned model that is optimized to understand multiple languages, reason in long contexts, and produce structured output. Together, these models reflect different architectural approaches to large-scale generative AI systems that are in use today.

3.3 Datasets and Schema Configuration

In order to have academic rigor and practical relevance, the evaluation is based on two commonly used benchmark datasets, SPIDER and BIRD. Such datasets were chosen because of their complementary nature and their proven application to the research of text-to-SQL. The SPIDER data set is used to assess cross-domain generalization and schema reasoning in the zero-shot circumstances, since the training does not involve the test data bases. It consists of complicated SQL statements that can have joins, nested subqueries, aggregations, and filtering of more than one table. By contrast, the BIRD dataset is concerned with enterprise-scale, noisy and domain-specific databases, which are relevant to real-world BI environments where schema ambiguity, value reasoning and external

knowledge can frequently be needed. Through the synthesis of SPIDER and BIRD the evaluation is able to access the controlled academic contexts as well as realistic business intelligence environments allowing a holistic view of model strength and flexibility.

3.4 Hybrid Prompting Architecture

An integrated prompt engineering system was created to steer all the considered models into the creation of single-turn and executable SQL queries. The prompt architecture combines several prompting strategies such as role prompting, instruction-based constraints, schema aware context injection and deterministic decoding parameters. Such design makes responses be directly comparable across models with a minimal variation added by prompt interpretation. All of these prompts clearly identify the model as an SQL assistant and have rigid formatting of output requirements. The task of models is to produce a single SQL SELECT statement without descriptions, comments, or supporting literature. Parameters are injected into the prompt to limit column and table usage to reduce identifiers that are visualized as hallucinations and enhance schema alignment. The same parameter of deterministic decoding is used across all the models to reduce randomness and increase reproducibility. Also, failure conditions are predefined like non-SQL output, occurrence of multiple statements or unfinished queries-are factors to categorize the responses as either valid or invalid during the evaluation process.

3.5 Evaluation Metrics

Five quantitative evaluation measures are used to assess model performance, each of which indicates a significant need by BI-oriented text-to-SQL systems. Measurement Intent clarification accuracy is a measure of how the model will infer missing constraints in response to underspecified or ambiguous user queries. Semantic clarification measures the degree to which the domain terms and business concepts are successfully mapped to the appropriate schema elements. SQL generation accuracy evaluates how structural correct generated queries are compared against reference SQL statements. The accuracy of the execution response is the correctness of the results generated by the generated SQL that is executed against the database. Lastly, response latency is the cumulative time to provide a response which is practically usable in interactive BI systems. All metrics are analyzed separately and the overall model effectiveness is measured through weighted aggregation.

3.6 Experimental Procedure

The experimental process is standardized in pipeline. The hybrid prompt template is fed with the natural language query and the schema fragment that is part of the relevant dataset query. The chosen generative AI system is then used to generate a SQL query in deterministic decoding conditions. The generated queries are syntactically checked and run on the respective database. The results of the execution are contrasted with ground-truth to detect accuracy. Every experiment is carried out using the same computational and configuration parameters to be fair. Measurements of the performance are documented in all the test cases and summed to generate comparative outcomes.

I.Analysis of Experimental Results

The evaluations are conducted on three instruction-based models LLaMA 4, Mixtral 8x7B, and Qwen3 14B on end-to-end text-to-SQL generation using natural language based question prompts from SPIDER datasets and in response returns a single SQL statement. For each response generated from each model, we record for a set of evaluation metrics defined in the criteria with their defined evaluation formula, threshold and weightage including SQL Generation Accuracy (20 points), Execution Response Accuracy (30 points), Intent Clarification Accuracy (15 points), Semantic Clarification Accuracy (15 points), and Response Time or Latency (20 points).

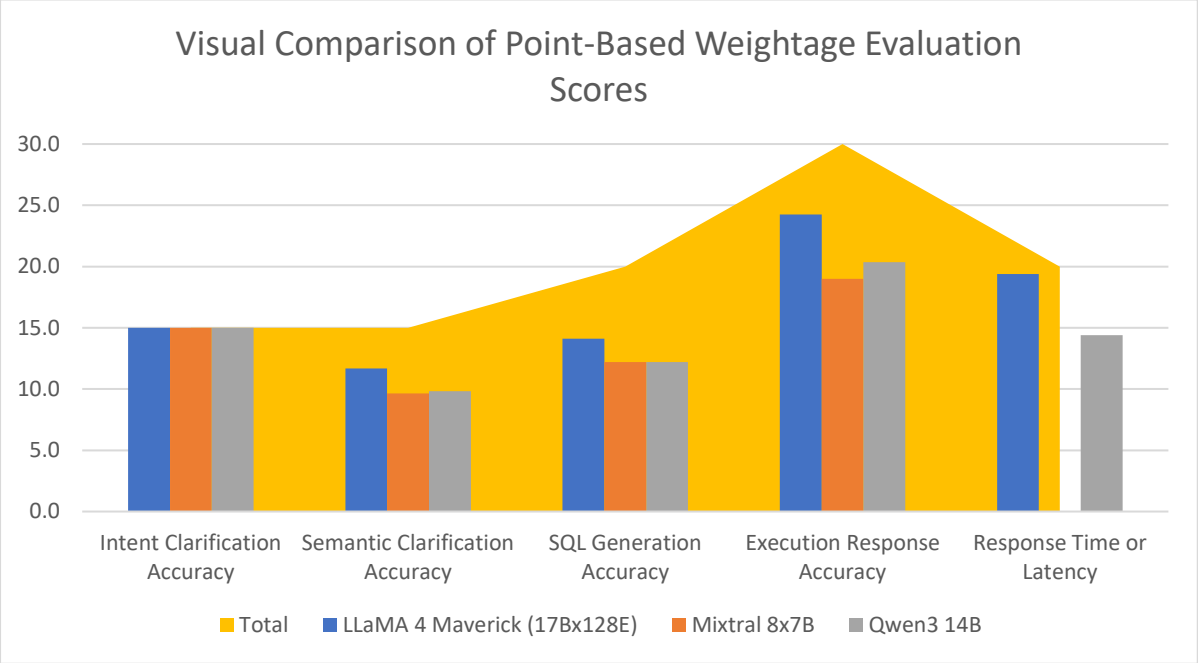


Fig. 2. VISUAL COMPARISON OF POINT-BASED WEIGHTAGE EVALUATION SCORES

The comparative analysis shows that LLaMA 4 Maverick (17Bx128E) model outperformed the other models on every defined weighted metric except Intent Clarification, where all the models have tied scoring. The high execution accuracy and low response latency balances the LLaMA 4 model and makes it the most efficient and effective model under the defined evaluation metrics and analysis. Whereas, the Mixtral 8x7B model shows deficiencies in runtime and accuracy highlights that the model is less reliable and capable for effective and efficient text-to-SQL generation in the defined environmental settings.

	Defined Weightage	*LLaMA 4 Maverick (17Bx128E)*	*Mixtral 8x7B*	*Qwen3 14B*
Intent Clarification Accuracy	15	15.0	15.0	15.0
Semantic Clarification Accuracy	15	11.7	9.6	9.8
SQL Generation Accuracy	20	14.1	12.2	12.2
Execution Response Accuracy	30	24.3	19.0	20.4
Response Time or Latency	20	19.4	0.0	14.4
Total Weightage	100	84.5	55.8	71.8

TABLE I. POINTS-BASED WEIGHTAGE EVALUATION SCORES

Further, the TABLE I. shows that by following the defined set of criteria and their weightage schemes, the collective results favour LLaMA 4 Maverick (17Bx128E) model with highest performance on our defined criteria with having score of 84.5 out of 100, where Qwen3 14B model performed with score of 71.8 out of 100. The main reason for margin gap of 12.7 points between LLaMA 4 and Qwen3 is because of the latency (19.4/20 vs. 14.4/20; +5.0) and execution response accuracy (24.3/30 vs. 20.4/30; +3.9) metrics. Further, the LLaMA 4

gets a minor gain in SQL Generation Accuracy (14.1/20 vs. 12.2/20; +1.9) and Semantic Clarification Accuracy (11.7/15 vs. 9.8/15; +1.9). Where, both models LLaMA 4 and Qwen3 accomplishes full score on Intent Clarification Accuracy (15/15) that indicates consistent production of syntactically valid SQL outputs. The LLaMA 4 model leads the competition between LLaMA 4 and Qwen3 14B models by producing correct SQL outputs in much faster response time, whereas Qwen3 14B model faces high latency and less productivity in SQL generation tasks. Furthermore, the Mixtral 8x7B model shows weaker performance across almost all defined metrics. The SQL Generation Accuracy (12.2/20) and Execution Response Accuracy (19.0/30) lagged behind the other two LLaMA 4 and Qwen3 14B models. While, the response time scored 0.0, highlighting latency issues in meeting runtime standards. These evaluations significantly dropped the overall total weightage score for Mixtral 8x7B.

1. Criteria # 1 – Intent Clarification Accuracy

The Intent Clarification Accuracy criteria analyses the ability of model to produce valid SQL queries from given prompt or to map an underspecified given prompt to an executable SQL query. In this criteria, a response from the model is marked as PASS when a non-empty SQL string is returned that satisfies the defined set of SQL syntax-based requirements. This measurement helps in capturing the ability of model to understand the intent of the question and generate an executable statement. Further, to support this evaluation, lightweight diagnostic checks are also introduced and applied including the detection of empty outputs, non-SQL outputs, and Has SELECT outputs checks are measured on the basis of percentage to provide a better in-depth additional diagnostics into the validity and reliability of the generated responses.

	LLaMA 4 Maverick (17Bx128E)	*Mixtral 8x7B*	*Qwen3 14B*
Total Count	600	600	600
Maximum Clarity (%)	100.0 %	100.0 %	100.0 %
Minimum Clarity (%)	0.0 %	0.0 %	0.0 %
Mean Clarity (%)	100.0 %	100.0 %	100.0 %
Median Clarity (%)	100.0 %	100.0 %	100.0 %
Pass Rate (%)	100.0 %	100.0 %	100.0 %

TABLE II. OVERVIEW TABLE OF INTENT CLARIFICATION ACCURACY CRITERIA

The TABLE II. provides the intent clarification results in detail for all three models across the batch of 600 evaluations for each model. Each model scored good scores across all defined metrics. The maximum, mean, and median values shows clearly the uniformity at 100.0% with a pass rate of 100.0% for each evaluated model. These findings highlight that generated SQL query statements are syntactically valid outputs that aligned with the planned execution scope of the prompt. This uniformity indicates that intent-level mapping from natural language to SQL is not a blockage factor across the evaluated large language models. However, the performance variations observed in the recent criterions are occurred due to structural accuracy or execution details rather than basic intent misalignments.

	LLaMA 4 Maverick (17Bx128E)	*Mixtral 8x7B*	*Qwen3 14B*
Count of Empty SQL Outputs	0.0 %	0.0 %	0.0 %
Count of Non-SQL Outputs	0.0 %	0.0 %	0.0 %

Has SELECT (%)	100.0 %	100.0 %	100.0 %

TABLE III. SQL OUTPUTS VALIDITY DIAGNOSTICS

The TABLE III. presents a supporting diagnostics for the validity of generated SQL responses. Where, all three models achieved great outcomes including no empty SQL outputs, no non-SQL outputs, and 100.0% of outputs consists of SELECT clause keyword. This diagnostics confirms that the prompting architecture is very efficient and effective, especially from role-based and instruction-based techniques that help the model in generating to-the-scope SQL query statements rather than adding irrelevant information into the responses or generating deformed responses.

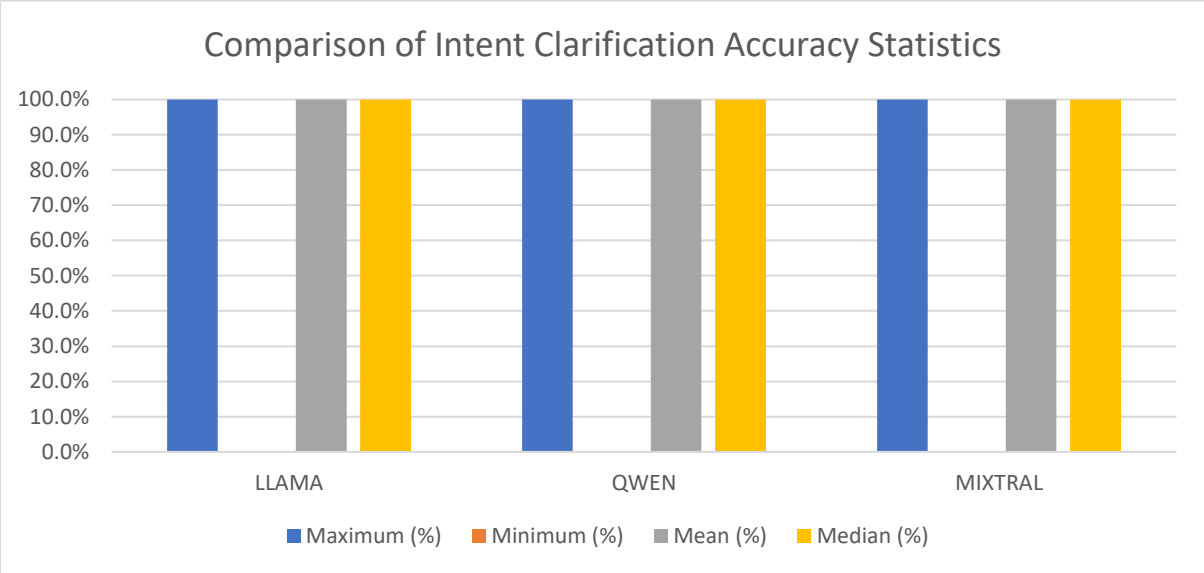


Fig. 3. COMPARISON OF INTENT CLARIFICATION ACCURACY STATISTICS

The results of this criteria indicates that all the evaluated models scored good intent clarification accuracy and validity diagnostics under the defined set of metrics. These set of findings demonstrates that the defined hybrid prompting architecture effectively marked a baseline for correctness and intent protection across models. Further, the performance differences between models does not occurred at intent level but rather in more complex features of SQL generation like structural alignment, and execution accuracy.

2. Criteria # 2 – Semantic Clarification Accuracy

By this criteria, it helps in evaluating the clause-level and operator-level alignments between base SQL queries and generated SQL queries. Regardless of the structural or execution-attentive measurements, this criteria works for models by checking generated response of model is correctly integrated and aligned with a set of defined SQL syntax-based parameters including GROUP BY, ORDER BY, PREDICATES, TABLES, and FLAGS like COUNT, SUM, AVG, DISTINCT, IN, LIKE, BETWEEN, NULL, and SUB QUERIES, where and when applicable. This correctness of semantic clarification is computed through a formula-based scoring method with values ranging from 0% to 100%. Further, a query is marked as PASS when the generated SQL query statement fulfils all the defined requirements of semantic clarification of the prompt.

	LLaMA 4 Maverick (17Bx128E)	*Mixtral 8x7B*	*Qwen3 14B*
Total Count	600	600	600

Maximum Clarity (%)	100.0 %	100.0 %	100.0 %
Minimum Clarity (%)	34.7 %	31.8 %	37.6 %
Mean Clarity (%)	86.4 %	81.2 %	83.5 %
Median Clarity (%)	90.0 %	85.0 %	87.1 %
Count of Evaluations ≥ 80%	467	385	393
Count of Evaluations < 80%	133	215	207
Pass Rate (%)	77.8 %	64.2 %	65.5%

TABLE IV. OVERVIEW TABLE OF SEMANTIC CLARIFICATION ACCURACY CRITERIA

The TABLE IV. represents the summarized results for semantic clarification across all the three evaluated models for the batch of 600 evaluations each. The LLaMA 4 model scored the maximum overall performance with mean clarity of 86.4% with 77.8% pass rate. In comparison, Qwen3 model acquired mean clarity 83.5% and a pass rate of 65.5%, while Mixtral lagged at 81.2% mean clarity and a 64.2% pass rate. The maximum clarity of 100.0% scored by every model, which confirms the ability of model to generate semantically correct SQL query statements in perfect situations. Though, disparities in minimum clarity indicates differences in robustness as LLaMA 4 performed better with score of 34.7%, while Mixtral achieved 31.8%, and Qwen3 scored the highest lowest-limit clarity with score of 37.6%, which recommends excessive constancy in avoiding strictly degraded responses. Further, LLaMA 4 also generated the most highest number of evaluations consisting of 467 out of 600 that are above the defined threshold of 80%, while Mixtral and Qwen3 only reached at 385 and 393 evaluations, respectively. These findings shows that Qwen3 model is at the median line with respect to performance, but the LLaMA 4 model gets a strong advantage in both consistency and overall performance and pass rate.

	LLaMA 4 Maverick (17Bx128E)	*Mixtral 8x7B*	*Qwen3 14B*
FLAGS Match (%)	97.7 %	95.7 %	95.7 %
GROUP BY Columns Similarity (%)	81.2 %	71.9 %	77.7 %
ORDER BY cols Similarity (%)	87.7 %	80.5 %	82.2 %
PREDICATES Similarity (%)	85.2 %	75.9 %	79.1 %
TABLES Similarity (%)	32.1 %	21.4 %	30.4 %

TABLE V. SEMANTIC ALIGNMENT DIAGNOSTICS

To diagnose the semantic alignment in-depth, the TABLE V. presents a heuristic-based similarity report across clause-level and operator-level components. Where, LLaMA 4 again gets higher scores and marks itself the strongest in the alignment across most metrics. It achieved 97.7 in FLAGS match, 81.2% in GROUP BY columns, 87.7% in ORDER BY columns, and 85.2% in PREDICATES, which shows the model outperformed both Mixtral and Qwen3 models in these dimensions. These results shows that LLaMA 4 is most effective in preserving semantic uniformity between base SQL query and the generated SQL query, especially in understanding operator-level logics. Further, Qwen3 model scored similar results in FLAGS with score of 95.7%, and GROUP BY columns with 77.7%, while Mixtral left behind in GROUP BY with 71.9% and in PREDICATES with 75.9%. With respect to TABLES similarity,

almost every model has very less performance with LLaMA 4 achieving 32.1% outperforming Qwen3 with score of 30.4% and Mixtral acquired score of 21.4%. This advises that table selection remains a mutual foundation of semantic errors, even when other clauses are aligned well with the given prompt.

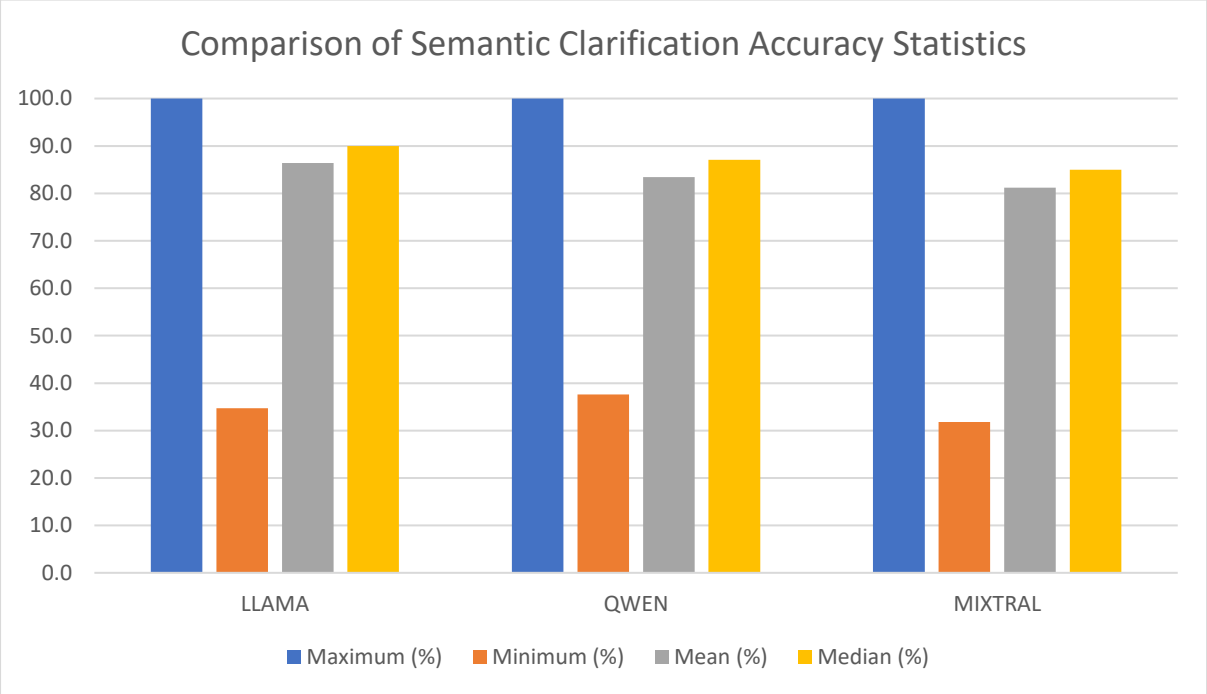


Fig. 4. COMPARISON OF SEMANTIC CLARIFICATION ACCURACY STATISTICS

The findings indicates that LLaMA 4 Maverick (17Bx128E) is the most semantically accurate model with highest mean accuracy, high pass rate, and robust alignments across main SQL components. Where, Qwen3 14B model acquired stable but a bit weaker results, whereas Mixtral 8x7B model demonstrated constant underperformance across many metrics. The TABLE similarity metric observed in all models remains at a low number for multi-entity queries, highlighting a key limitation in semantic clarification while great performance in replicating clauses and operators. These results highpoints the importance of semantic-level evaluation, as this semantic clarification exposes weaknesses in model generated outputs that are not understood by the models.

3. Criteria # 3 – SQL Generation Accuracy

By SQL Generation Accuracy criteria, it helps in measuring how closely the generated SQL query matches the corresponding base SQL query, as this metric reflects both syntactic and semantic correctness. For the evaluation of generation accuracy, a computation based on defined set of weighted parameters ranging 0% to 100% score including TABLE, JOINS, SELECTED COLUMNS, PREDICATES, and FLAGS such as COUNT, DISTINCT, GROUP BY. Further, a strict string-match rate is also defined as a secondary signal acting as a validation layer for accuracy measurement to support the first analysis of weighted parameters.

	LLaMA 4 Maverick (17Bx128E)	*Mixtral 8x7B*	*Qwen3 14B*
Total Count	600	600	600
Maximum Accuracy (%)	100.0 %	100.0 %	100.0 %
Minimum Accuracy (%)	46.7 %	40.0 %	40.0 %

Mean Accuracy (%)	86.9 %	82.8 %	82.8 %
Median Accuracy (%)	90.0 %	90.0 %	90.0 %
String Match Rate (%)	95.0 %	95 %	95.0 %
Count of Evaluations ≥ 80%	422	365	365
Count of Evaluations < 80%	178	235	235
Pass Rate (%)	70.0 %	61.0 %	61.0 %

TABLE VI. OVERVIEW TABLE OF SQL GENERATION ACCURACY CRITERIA

The TABLE VI. provides an overview of SQL generation accuracy across the three evaluated models that are LLaMA 4 Maverick (17Bx128E), Mixtral 8x7B, and Qwen3 14B based on batch of 600 query executions each. Every model achieved a maximum accuracy of 100% which indicates that under certain conditions each model is capable of generating fully correct SQL statements. Where, prominent differences shown when considering lower limit and average accuracy. LLaMA 4 acquired a minimum accuracy of 46.7% which replicates larger instability in the generation process of difficult query statements. In terms of statistical measures, the LLaMA 4 model acquired the maximum mean accuracy of 86.9%, outperforming the Mixtral and Qwen3 models having 82.8%. Where, all three models achieved the same median accuracy of 90.0% which highlights the models performance for mid-range queries but deviation in handling outliers. The string match rate are uniformly strong at 95,0% across all three models, showing the robustness of structural reproduction in most of the cases. Further, more deep breakdown shows that LLaMA 4 model successfully completed generation process correctly of 422 evaluations that scored 70.0% of pass rate that is above the declared 80% threshold, which is high as compared to other Mixtral and Qwen3 models that achieved score of 61.0% pass rate. This evaluation based on pass rate shows that all three models are capable of generating valid and correct SQL query statements but LLaMA 4 model gets superiority as per highest performance with respect to consistency and reliability across various type of prompts and queries.

	LLaMA 4 Maverick (17Bx128E) Error! Reference source not found.	*Mixtral 8x7B*	*Qwen3 14B*
TABLES Similarity (%)	32.3 %	21.5 %	30.4 %
JOINS Similarity (%)	73.0 %	46.7 %	60.7 %
SELECT Columns Similarity (%)	35.4 %	22.9 %	32.9 %
PREDICATES Similarity (%)	68.9 %	62.0 %	62.4 %
FLAGS Match (%)	96.1 %	90.6 %	92.2 %

TABLE VII. HEURISTIC BASED SQL STRUCTURE MATCHING

To show more better and in-depth view of the analysis for the generation accuracy, the TABLE VII. provides report based on heuristic-based structural similarity across different SQL components and clauses. The LLaMA 4 model outperformed in its several components and clauses like achieving similarity of 73.0% in JOIN operations, which is higher than Mixtral and Qwen3 models that acquired 46.7% and 60.7% similarity. This showcase that LLaMA 4 model is more good at handling multi-table queries that requires relational joins. Where, the LLaMA 4 model scored highest in PREDICATES and FLAGS operations of

68.9%, 96.1% as compared to Mixtral and Qwen3 models that achieved 62.0%, 90.6% and 62.4%, 92.2%. Similarly, the SELECTION operations similarity for LLaMA 4 model is 32.3%, 35.4%, showing slight better performance as compared to Mixtral model of 21.5%, 22.9% and Qwen3 model of 30.4%, 32.9%. These findings showcase that selection operations remain a challenge for all three models but LLaMA 4 shows a slight good performance and got advantage among these models.

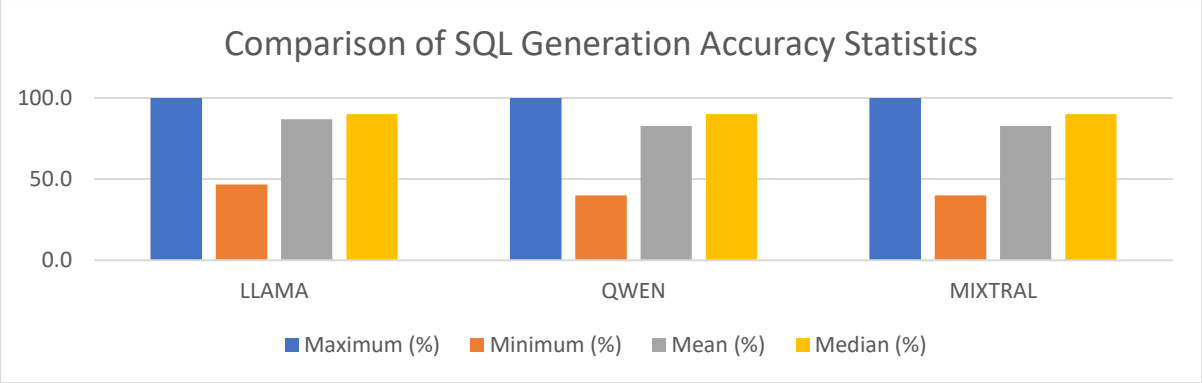


Fig. 5. COMPARISON OF SQL GENERATION ACCURACY STATISTICS

These evaluations shows that LLaMA 4 Maverick (17Bx128E) model provides the most robust performance in overall under the SQL generation accuracy metric with both aggregate scoring and structural similarity. While, Mixtral 8x7B and Qwe3 14B models competed in string match rates and median accuracies, their lower pass rate and weaker working with joins, predicates, and flags underline restrictions in constantly moving towards more complex queries. These understandings confirms that the evaluations based on structural-level with weighted scoring and string strict matching process provides a all-inclusive assessments of SQL generation accuracy.

4. Criteria # 4 – Execution Response Accuracy

The next criteria defined is of Execution Response Accuracy, measures the correctness of the SQL query statement output when executed against the target database. Unlike the SQL Generation Accuracy metric, which focused on the structural similarity, this criteria emphasises on the correctness of generated result by SQL queries are monitored through correct SQL syntaxes and execution shape by defined score range of 0% to 100% that helps in analyzing the deviation of models. The metric validates both syntactic correctness and execution-shape by analyzing the correctness of the SQL components such as aggregations, group by, order by, limit, and selected columns. By examination on these elements, the metric produces an overall insight of the ability of the models to generate queries that are not only correct syntactically but also generate the required computational outputs.

	LLaMA 4 Maverick (17Bx128E)	*Mixtral 8x7B*	*Qwen3 14B*
Total Count	600	600	600
Maximum Accuracy (%)	100.0 %	100.0 %	100.0 %
Minimum Accuracy (%)	15.0	10.0 %	25.0 %
Mean Accuracy (%)	88.8 %	83.7 %	86.2 %
Median Accuracy (%)	90.0 %	90.0 %	90.0 %
Scalar Result Match (%)	96.0 %	92.5 %	95.3 %

Count of Evaluations ≥ 90%	485	380	407
Count of Evaluations < 90%	115	220	193
Pass Rate (%)	81.0 %	63.0 %	68.0 %

TABLE VIII. OVERVIEW TABLE OF EXECUTION RESPONSE ACCURACY CRITERIA

The TABLE VIII. provides a comprehensive overview of the performance of three evaluated models across batch of 600 query executions each. All models achieved a maximum accuracy of 100% representing the ability to produce fully executable and correct queries in superlative cases. However, the variations arises in terms of minimum and mean accuracies. LLaMA 4 model attained the maximum mean accuracy at 88.8% where as the Qwen3 model at 86.2% and the Mixtral got at 83.7%. Whereas, the minimum accuracy score expose further differences where LLaMA 4 performed better with score of 15.0% than Mixtral having score of 10.0% and Qwen3 at 25.0% advised larger flexibility in avoiding extremely poor-generated responses. Regardless of same median accuracies across all the models, LLaMA 4 standout itself in terms of consistency by acquiring a scalar result match rate of 96.0% as compared to the Qwen3 model have 95.3% and Mixtral model scoring 92.5%. The evaluation threshold, LLaMA 4 model achieved 485 correct outputs with a pass rate of 81.0%, where the Qwen3 model generated 407 correct outputs with a pass rate of 68.0% and Mixtral left behind by having 380 outputs with a pass rate of 63.0%. These evaluations indicate that LLaMA 4 model generate queries that are highly accurate and precise when executed by also maintaining stability across various types of queries.

	LLaMA 4 Maverick (17Bx128E)	*Mixtral 8x7B*	*Qwen3 14B*
AGGREGATES Similarity (%)	93.1 %	88.9 %	88.9 %
DISTINCT Match (%)	94.3 %	89.0 %	89.8 %
GROUP BY Match (%)	96.5 %	90.3 %	93.8 %
ORDER BY Present Match (%)	97.0 %	95.7 %	96.3 %
ORDER BY Columns Similarity (%)	87.8 %	80.5 %	82.2 %
ORDER Direction Match (%)	97.3 %	95.8 %	96.7 %
LIMIT Match (%)	97.0 %	97.0 %	95.5 %
SELECT Columns Similarity (%)	35.4 %	22.9 %	32.9 %

TABLE IX. HEURISTIC BASED SQL BASE-AND-GENERATED QUERIES MATCHING

Further, to measure execution response accuracy, the TABLE IX. provides a heuristic-based similarity across main SQL components and clauses. LLaMA 4 constantly scored the highest performance across most metrics. Such as, in aggregations LLaMA 4 achieved 93.1%, for DISTINCT usage the score is 93.4%, ORDER BY attained at 97.0%, where direction matching acquired 97.3%, and 96.5% in GROUP BY usage, that outperformed both Mixtral and Qwen3 models. On the other hand, Mixtral and Qwen3 models showed weaker but competitive results in these metrics. Qwen3 scored 93.8% in GROUP BY, and in ORDER BY it achieved 96.3%, but in ORDER BY column selection the similarity is 82.2%, which remained under the LLaMA model's similarity 87.8%. Likewise, the Mixtral model also underperformed as

compared to other models with scoring of 80.5% in ORDER BY columns and 89.0% in DISTINCT matching, which shows the model having difficulty in handling of execution details. Furthermore, the main challenge of SELECT column similarity occurred for all models where models scores remained low, for LLaMA 4 the score was 35.4%, Qwen3 achieved 32.9%, and Mixtral got at 22.9%. This highlights that models are able to execute generated queries correctly regarding their scalar and aggregate operations but they also frequently struggle to generate the exact column selections. However, LLaMA 4 model still maintained a great advantage in the metrics among all the models.

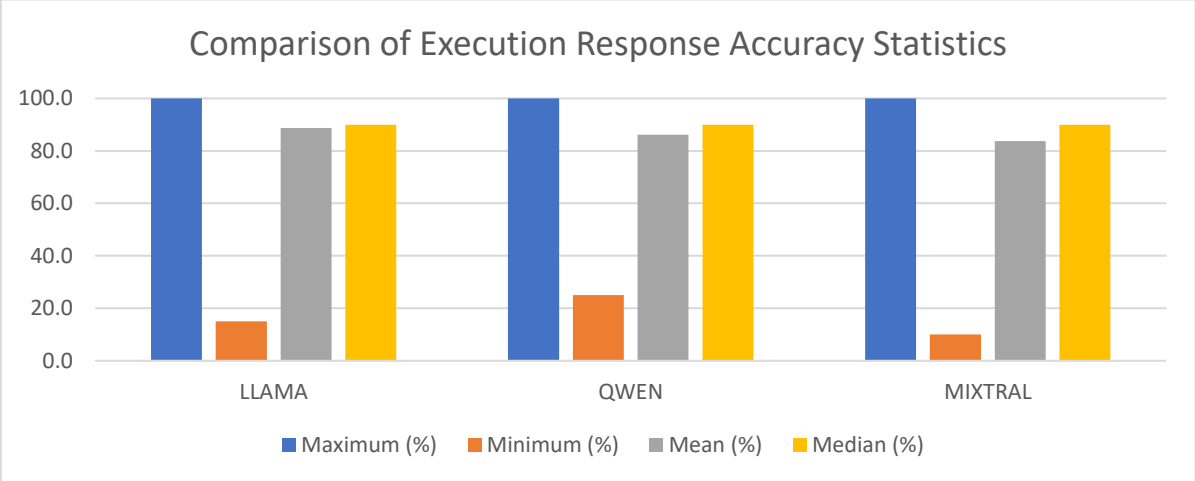


Fig. 6. COMPARISON OF EXECUTION RESPONSE ACCURACY STATISTICS

The findings of the Execution Response Accuracy confirms that LLaMA 4 Maverick (17Bx128E) constantly generate the most accurate and execution-ready SQL query statements with greater performance in both aggregate accuracy and heuristic-based structural matching accuracy measures. Qwen3 showed competitive results in maintaining higher minimum accuracy but got less score in overall pass rate and structural alignments. Where, Mixtral is capable of producing correct queries but revealed less performance across many metrics. By these understandings, it confirms the importance of execution-based evaluation, since it captures the different variations of query correctness outside the structural similarity by itself.

5. Criteria # 5 – Response Time or Latency

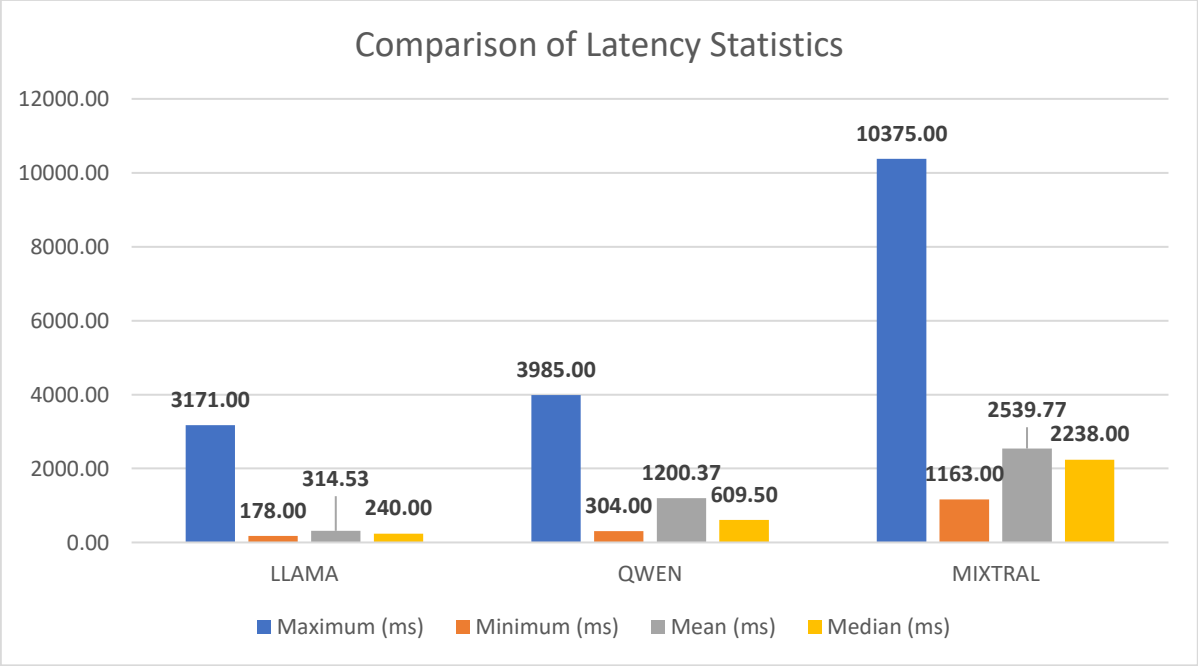
This criteria examines the end-to-end time in milliseconds (ms) taken from request submission to receiving the response from model for a single turn. This metric helps in capturing the efficiency of models under evaluation with latency defined from the moment the request is submitted to the returning of the generated output in a single turn. The performance is summarized through a set of statistical descriptors including mean, median, tail behaviours with p90 and p95 boundaries, as well as through the pass rate against a predefined threshold of ≤ 1000 ms to identify acceptable responsiveness and threshold of ≥ 3000 ms for outlier detection. By examining these values, the criteria gives a comprehensive insights on average responsiveness and stability of runtime performance across queries.

	LLaMA 4 Maverick (17Bx128E)	*Mixtral 8x7B*	*Qwen3 14B*
Total Count	600	600	600
Maximum (ms)	3171.0 ms	10375.0 ms	3985.0 ms
Minimum (ms)	178.0 ms	1163.0 ms	304.0 ms

Mean (ms)	314.5 ms	2539.7 ms	1200.3 ms
Median (ms)	240.0 ms	2238.0 ms	609.5 ms
p90 Latency (ms)	474.3 ms	3867.6 ms	3132.7 ms
p95 Latency (ms)	786.6 ms	4497.1 ms	3271.1 ms
Count ≤ 1000 ms	582	0	432
Count > 1000 ms	18	600	168
Count ≥ 3000 ms	1	151	118
Pass Rate (%)	97.0 %	0.0 %	72.0 %

TABLE X. OVERVIEW TABLE OF RESPONSE TIME OR LATENCY CRITERIA

The TABLE X. presents a comprehensive comparative statistics on response latency for all three models across a batch of 600 evaluations for each model. LLaMA 4 showed the most efficient performance with mean response time of 314.5 ms and median response time of 240.0 ms. Also its p90 is 474.3 ms and p95 is 786.6 ms that highlights more stable performance even in the tail distribution the response time rarely surpassed 1000 ms. 582 evaluations are completed within the 1000 ms threshold, acquiring a 97.0% pass rate with only one extreme outlier above the extreme limit of 3000 ms. On the other hand, Mixtral scored the slowest performance with latency of 2539.7 ms and median latency of 2238.0 ms. Its tail distribution values of p90 is at 3867.6 ms and p95 is at 4497.1 ms, that shows severe inefficiencies with all 600 evaluations surpassing 1000 ms threshold and 151 evaluations exceeding 3000 ms of extreme threshold. Mixtral acquired a pass rate of 0.0%, reflecting unsustainability for latency-sensitive applications and usages. Further, the Qwen3 model achieved a middle ground with mean latency of 1200.0 ms and median latency of 609.5 ms. The tail distribution of p90 is at 3132.7 ms and p95 is at 3271.1 ms that shows large tail delays but still lesser than Mixtral. 432 evaluations are completed within 1000 ms threshold achieving a pass rate of 72.0% while 168 evaluations surpassed the 1000 ms threshold and other 118 evaluations exceeded the extreme threshold of 3000 ms, demonstrating a lean towards unsteadiness under some query conditions.



COMPARISON OF LATENCY STATISTICS

The evaluation shows visible differences between the models. LLaMA 4 Maverick (17Bx128E) constantly acquired a great performance with minimum variance which makes it the most efficient and effective model for real-time deployment and usages. Then, Qwen3 14B model showed great tail latency but also acceptable in many cases that decreases the reliability of model for some scenarios that needs to run under strict time limits. On the other hand, the Mixtral 8x7B model scored much high latencies across all the evaluations which shows makes the model unrealistic for real-time text-to-SQL applications and response generations. These results highlights the importance of latency for response generation and its overall impact in the generation process, since efficiency is the main factor for the scalability and reliability of any generative AI system in production-ready environments.

II. Conclusion

This paper presents an overview and analysis of the recent developments on generative AIs-based text-to-SQL systems such as MAC-SQL, SQL-PaLM, CHASE-SQL, and CHESS frameworks and their characteristics and shortcomings. This discussion dwells on the increasing tendency towards multi-agent collaboration, retrieval-augmented schema processing and execution-directed refinement that characterizes the current state of the art in text-to-SQL studies. In addition, the paper provides a summary of the state-of-the-art generative AI models such as LLaMA 4, Qwen3 14B, and Mixtral 8x7B, describing their design principles in architectural design as well as outlines the effect of the large language models such as LLaMA, Qwen, and Mixtral on transformer-based methods to understand, reason, and generate context-aware query generation using SQL. More so, the experiments performed in the current paper is an analysis of LLaMA 4, Qwen3 14B, and Mixtral 8x7B text-to-SQL generation models. To ensure equal task framing and comparison of models, a hybrid prompt template (prompting methods) was developed, which integrates prompting methods (role prompting, instruction-based prompting with output formatting) together with schema prompting and few-shot prompting with decoding parameters and controls. Evaluation process is carried out with the help of the following evaluation metrics Intention Clarification Accuracy, Semantic Clarification Accuracy, SQL Generation Accuracy, Execution Response Accuracy and Response Time or Latency. The relative outcomes demonstrated that LLaMA 4 achieved excellent performance in virtually every analysis measure particularly in intent clarification, SQL generation accuracy and execution response accuracy, and response time, making it the most useful model in text-to-SQL issues within the Business Intelligence settings. Another outcome of the results also highlights that sophistication prompting and consistency of evaluation is important to enhance the capacity of model to comprehend context as well as produce correct contextual responses. Besides, the study also adds to the analytic and empirical research results of generative AI in text-to-SQL tasks. It connects academic knowledge based on existing frameworks with real-life experiments on state-of-the-art large language models. This research will be followed up by further research to extend the scope of this study to designing of multi-agent Business Intelligence framework that captures domain-related and real-time operations, context-aware text-to-SQL pipelines to enhance further the accuracy, efficiency and scale of the next generation Business Intelligence systems.

REFERENCES

Busany, N., Hadar, E., Hadad, H., Rosenblum, G., Maszlanka, Z., Akhigbe, O., et al. (2024). *Automating business intelligence requirements with generative AI and semantic search* (arXiv:2412.07668). arXiv.

Dang, H., Mecke, L., Lehmann, F., Goller, S., & Buschek, D. (2022). *How to prompt? Opportunities and challenges of zero- and few-shot learning for human-AI interaction in creative applications of generative models* (arXiv:2209.01390). arXiv.

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. *Business & Information Systems Engineering*, 66(1), 111–126.

Garg, S. (2024, May 27). Generative AI architecture: Layers and models. *Solulab Blog*.

Hugging Face. (2025, October 4). Spaces ZeroGPU: Dynamic GPU allocation for Spaces. *Hugging Face Documentation*.

Jiang, A. Q., Sablayrolles, A., Mensch, C., Bamford, C., Chaplot, D. S., de Las Casas, D., et al. (2023). *Scaling vision-language models with sparse mixture of experts* (arXiv:2310.06825). arXiv.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., et al. (2024). *Mixtral of experts* (arXiv:2401.04088). arXiv.

Lei, F., Chen, J., Ye, Y., Cao, R., Shin, D., Su, H., et al. (2025). *Spider 2.0: Evaluating language models on real-world enterprise text-to-SQL workflows* (arXiv:2411.07763). arXiv.

Li, J., Hui, B., Qu, G., Yang, J., Li, B., Li, B., et al. (2023). Can LLM already serve as a database interface? A big bench for large-scale database grounded text-to-SQLs. *Advances in Neural Information Processing Systems, 36*, 42330–42357.

Marvin, G., Nakayiza, H., 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). Springer Nature Singapore.

Meta AI. (2025, June 13). LLaMA 4: Advancing multimodal intelligence. *Meta AI Blog*.

Pourreza, M., Li, H., Sun, R., Chung, Y., Talaei, S., Kakkar, G. T., et al. (2024). *CHASE-SQL: Multi-path reasoning and preference optimized candidate selection in text-to-SQL* (arXiv:2410.01943). arXiv.

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., et al. (2024). *The Prompt Report: A systematic survey of prompt engineering techniques* (arXiv:2406.06608). arXiv.

Storey, V. C., Yue, W. T., Zhao, J. L., & Lukyanenko, R. (2025). Generative artificial intelligence: Evolving technology, growing societal impact, and opportunities for information systems research. *Information Systems Frontiers*, 1–22.

Sun, R., Arik, S. Ö., Muzio, A., Miculicich, L., Gundabathula, S., Yin, P., et al. (2023). *SQL-PaLM: Improved large language model adaptation for text-to-SQL (extended)* (arXiv:2306.00739). arXiv.

Talaei, S., Pourreza, M., Chang, Y. C., Mirhoseini, A., & Saberi, A. (2024). *CHESS: Contextual harnessing for efficient SQL synthesis* (arXiv:2405.16755). arXiv.

Vertsel, A., & Rumiantsau, M. (2024). *Hybrid LLM/rule-based approaches to business insights generation from structured data* (arXiv:2404.15604). arXiv.

Wang, B., Ren, C., Yang, J., Liang, X., Bai, J., Chai, L., et al. (2025). MAC-SQL: A multi-agent collaborative framework for text-to-SQL. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)* (pp. 540–557). Association for Computational Linguistics.

Wang, R., Mi, F., Chen, Y., Xue, B., Wang, H., Zhu, Q., et al. (2024). *Role prompting guided domain adaptation with general capability preserve for large language models* (arXiv:2403.02756). arXiv.

Webson, A., & Pavlick, E. (2022). Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2300–2344). Association for Computational Linguistics.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., et al. (2025). *Qwen3 technical report* (arXiv:2505.09388). arXiv.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., et al. (2024). *Qwen2.5 technical report* (arXiv:2412.15115). arXiv.