

Manuscript Status — Design and Architecture Paper

This manuscript presents the complete system design, architecture, and evaluation framework for the proposed bilingual NLP-based reporting system. Sections 1 through 4 are fully written and represent the current research contribution. Sections 5 through 7 (Results, Discussion, and Conclusion) are reserved for completion following system implementation and evaluation. Placeholder headings and scopes are provided for those sections.

Completed: Introduction, Related Work, System Architecture, Evaluation Framework, References.

Pending implementation: Results and Analysis, Discussion, Conclusion and Future Work.

An NLP-Based Framework for Bilingual Intelligent Report Generation

Within Oracle APEX: Architecture, Design, and Evaluation Framework

Eman Alhabashi, Merwa Idris, Alaa Humaid
Department of Artificial Intelligence
Midocean, Moroni, Comoros

Abstract

Administrative and executive reporting in enterprise and government organizations continues to depend on static, developer-built dashboards that cannot respond to ad-hoc questions and produce no narrative explanation of the data they display. This paper presents the architecture and design of a bilingual natural language processing framework embedded within Oracle Application Express (APEX) that addresses both limitations. The proposed system accepts queries in Arabic or English, translates them into validated, read-only SQL through a prompt-engineered large language model augmented with retrieval-augmented generation and a human-in-the-loop confirmation step, executes the query, and generates a concise, faithful narrative summary through a second language model stage. The framework is designed for deployment on Oracle Database 23ai using the HR and Sales History schemas as the evaluation domain. Key architectural contributions include a schema-linking and RAG pipeline using Oracle 23ai vector search, a SQL safety and validation layer, and a faithfulness-controlled narrative generation component supporting bilingual Arabic-English output. This paper presents the complete system architecture and design rationale, defines a rigorous evaluation framework covering NL2SQL accuracy, narrative faithfulness, and user task efficiency, and identifies the research gap this work addresses: no existing study integrates validated NL2SQL and controlled NLG in a bilingual administrative reporting context within a low-code enterprise platform. System implementation and experimental results will be reported in subsequent work.

Keywords: Natural Language to SQL (NL2SQL); Data-to-Text Generation; Oracle APEX; Bilingual NLP; Arabic NLP; Retrieval-Augmented Generation; Low-Code AI; Intelligent Reporting; Administrative Reporting

1. INTRODUCTION

Data-driven decision-making depends on the ability of executives and administrators to obtain accurate, timely answers from organizational data. In most enterprise and government settings, this ability remains constrained by a structural bottleneck: every new analytical question must be translated into a database query by a developer, and the result is returned as a raw table with no explanatory narrative. The global business intelligence market, valued at approximately USD 54.9 billion in 2025 and growing at 12.4% annually, reflects how heavily organizations have invested in tools to close this gap (DataStackHub, 2025). Yet only 32% of executives report confidence in making meaningful use of data (G2, 2024), suggesting that tool availability and actual data accessibility remain distinct problems.

Two capabilities from natural language processing directly address this. Natural Language to SQL (NL2SQL) systems allow a user to phrase a question in ordinary language and receive a correctly formulated database query in return. Data-to-text generation systems take structured query results and produce concise, human-readable narrative summaries. Used together, they offer a path from question to insight without developer intermediation. However, the literature shows that these two capabilities have been developed and evaluated almost entirely in isolation, predominantly in English, and rarely within a working enterprise application.

This paper presents the design and architecture of a system that combines both capabilities within Oracle Application Express (APEX), a widely deployed low-code platform that runs natively on the Oracle Database. The system supports Modern Standard Arabic and English as first-class input and output languages, directly serving the MENA region, where Oracle deployments are concentrated across government, banking, healthcare, and energy sectors but where Arabic NLP support for enterprise reporting is absent from the literature.

The contributions of this paper are fourfold. First, it proposes a complete NLP pipeline architecture integrating NL2SQL and natural language generation (NLG) inside Oracle APEX, with a schema-linking and retrieval-augmented generation (RAG) layer, SQL safety validation, and human-in-the-loop confirmation. Second, it addresses the bilingual Arabic-English dimension in enterprise reporting, which has not been studied in this context. Third, it defines a rigorous evaluation framework covering SQL accuracy, narrative faithfulness, and user task efficiency, anchored against Oracle's native Select AI as a baseline. Fourth, it introduces a methodology for constructing a bilingual administrative evaluation dataset that the research community can reuse.

The remainder of this paper is structured as follows. Section 2 reviews related work across NL2SQL, data-to-text generation, and AI integration in low-code platforms. Section 3 presents the system architecture and design. Section 4 defines the evaluation framework. Sections 5 through 7 are reserved for results, discussion, and conclusion following system implementation.

2. RELATED WORK

2.1 Natural Language to SQL

The NL2SQL problem has evolved substantially since early neural sequence-to-sequence approaches. Constrained decoding methods such as PICARD (Scholak et al., 2021) addressed the persistent problem of syntactically invalid output by filtering generation token-by-token against the target schema, achieving state-of-the-art results on Spider without retraining the underlying model. The arrival of large language models shifted the dominant approach from fine-tuning to prompt engineering. DIN-SQL (Pourreza & Rafiei, 2023) demonstrated that decomposing the task into sub-problems and adding a self-correction step significantly raised accuracy over single-shot prompting, while DAIL-SQL (Gao et al., 2024) provided a systematic analysis of which prompt design choices drive performance.

More recent benchmarks have exposed the gap between academic results and real-world enterprise conditions. BIRD (Li et al., 2023) introduced large, dirty, real-world databases requiring external knowledge, revealing a wide gap between model accuracy (around 55% for GPT-4) and human performance (around 93%). Spider 2.0 (Lei et al., 2025) pushed further into enterprise analytical workflows with production-scale schemas, where current agent frameworks solve fewer than 20% of tasks. Surveys by Hong et al. (2024) and Shi et al. (2024) organize the field along the prompting-

versus-fine-tuning axis and consistently identify robustness on messy enterprise data and multilingual coverage as the principal open challenges.

2.2 Data-to-Text Generation and Faithfulness

Data-to-text generation has been studied through a series of controlled benchmarks designed to isolate faithfulness. ToTTo (Parikh et al., 2020) introduced a highlighted-cell table-to-text task in which targets are constrained to what the selected data supports. DART (Nan et al., 2021) extended the setting to structured record inputs across domains. Both revealed that even strong models trail human writers on faithfulness, a finding consolidated in surveys by Li et al. (2022) and Ji et al. (2023), which catalogue the types of hallucination and mitigation strategies proposed to date.

Zhao et al. (2023) found that LLMs can generate more faithful table-to-text than earlier fine-tuned systems and that an LLM can serve as an evaluator of another model's faithfulness. ERATTA (Roychowdhury et al., 2024) applied retrieval-augmented generation to enterprise table question-answering with explicit hallucination checks. In the administrative domain, Yagamurthy et al. (2023) studied automated report generation from financial and business intelligence data. Oyewole et al. (2024) reviewed NLP applications to financial reporting and noted that consistency risks remain even when accuracy improves. Lin et al. (2024) survey the neural data-to-text field and identify faithfulness and bilingual coverage as the two most pressing open problems.

2.3 AI Integration in Oracle APEX and Low-Code Platforms

The literature on AI inside low-code platforms is sparse but growing. Keshireddy (2025a) proposed a modular NLP architecture for Oracle APEX covering intent recognition, entity detection, and adaptive dialogue, evaluated on task completion time and interaction error rate. Keshireddy (2025b) extended this to data transformation and validation using adaptive AI models, reporting accuracy above 92% on 980,000 enterprise records. Yeruva (2025) reviewed AI features in Oracle APEX broadly, covering predictive analytics, NLP interfaces, and REST-based external model integration. Ashraf Syed (2025) presented a proof-of-concept SQL-generating chatbot built on the APEX AI Assistant, claiming 92% NL-to-SQL accuracy and a 50% reduction in report generation time. Keshireddy (2025c) integrated LLMs into APEX for intelligent code assistance, and Keshireddy (2025d) applied RAG to APEX-based AI assistants for context-aware responses.

Outside APEX, Rao et al. (2024) compared visual and natural-language interfaces for ML pipeline construction in a low-code tool and found that natural language aided discovery but that users still

relied on the visual layer for verification, providing empirical backing for human-in-the-loop confirmation in NLP-driven applications.

Research gap. No existing study integrates validated NL2SQL and faithfulness-controlled NLG in a bilingual administrative reporting pipeline within Oracle APEX or any comparable low-code platform. The APEX-specific literature covers interaction, code generation, data processing, and single-component chatbots, but not the end-to-end path from natural-language question to faithful narrative report in Arabic and English. This paper addresses that gap through a complete system design and evaluation framework.

3. SYSTEM ARCHITECTURE AND DESIGN

3.1 Overview

The proposed system is designed as a modular NLP layer integrated into an Oracle APEX application. It follows a sequential pipeline with five stages: query intake, schema-aware NL2SQL translation, SQL validation and human confirmation, query execution, and narrative generation. All stages run within the APEX security model, ensuring each user can only access data they are already authorized to view under their APEX credentials.

3.2 Schema Linking and RAG Layer

Before a query reaches the language model, the system builds a structured prompt from two sources. First, it extracts table names, column names, data types, and primary and foreign key relationships from Oracle's data dictionary and formats them as a schema description. Then, it appends a domain glossary that maps organizational terminology to column names and values in the schema. This retrieval-augmented construction substantially reduces schema linking errors, particularly for questions that reference business concepts not directly encoded in column names.

Arabic input requires a normalization step before retrieval and prompt construction. Alef variants, ya and alef maqsura confusion, ta marbuta handling, and diacritic removal are applied to both the input question and the stored question index, ensuring orthographic variation does not fragment retrieval matches. This step applies to Arabic input and defaults to a pass-through for English.

3.3 NL2SQL Component

The NL2SQL stage uses a prompt-engineered large language model with a structured prompt containing the schema description, domain glossary, and the user's question. The prompt instructs

the model to reason step by step, identify the relevant tables and columns before writing the query, and produce only a SELECT statement. A decomposition strategy modelled on DIN-SQL (Pourreza & Rafiei, 2023) is applied for complex queries: the model identifies relevant schema elements first, drafts the query, then reviews and corrects it.

Every generated SQL statement passes through a safety validation layer before it is shown to the user. The validator parses the statement and rejects any query containing DDL or DML keywords such as CREATE, INSERT, UPDATE, DELETE, DROP, TRUNCATE, or MERGE. It verifies that all table and column references exist in the current schema. If the query passes validation, it is presented to the user alongside a plain-language explanation of what it will retrieve, generated by a separate prompt. The user must confirm before the query executes. This human-in-the-loop step is the primary mechanism for catching semantic errors that syntactic validation cannot detect.

3.4 Narrative Generation Component

Once the query executes and returns results, the NLG stage takes the original question, the query results formatted as a structured representation with column labels, values, units, and computed comparisons such as period-over-period changes, and the user's preferred output language, and produces a concise narrative summary. The prompt instructs the model to describe only what the data shows, to avoid adding interpretation or context not grounded in the results, and to reason through the key figures before writing. This chain-of-thought constraint is the primary faithfulness mechanism. Output language is selected based on the detected language of the input question.

3.5 APEX Integration

The two NLP components are exposed to the APEX application through Oracle REST Data Services (ORDS) REST endpoints backed by PL/SQL procedures that call the OCI Generative AI service. The APEX front end presents a natural-language input field, passes the input to the NL2SQL endpoint, displays the proposed query and explanation in a review panel, and on confirmation executes the query and passes the results to the NLG endpoint. The returned narrative is rendered in a report region alongside the standard tabular result. All components respect the APEX application's existing authentication and authorization scheme, including row-level security enforced through the APEX application's built-in user context.

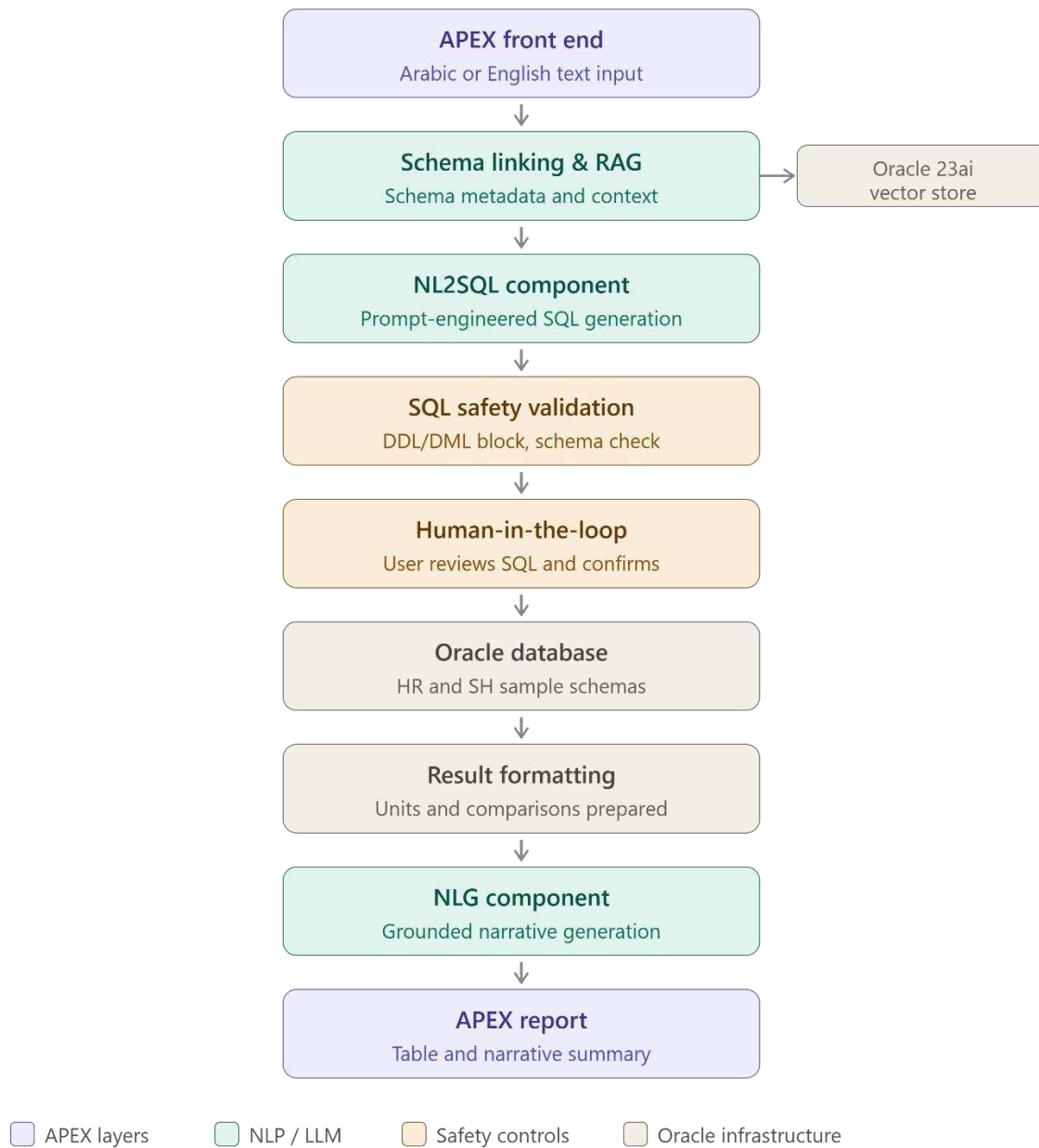


Figure 1: System architecture diagram

4. EVALUATION FRAMEWORK

4.1 Datasets

Evaluation uses three data sources. Public NL2SQL benchmarks — Spider (Yu et al., 2018), BIRD (Li et al., 2023), and Spider 2.0 (Lei et al., 2025) — provide standardized English-language baselines enabling direct comparison with published results. A custom bilingual evaluation set will be constructed in collaboration with domain administrators, comprising 150 to 200 question-SQL-narrative triples covering realistic administrative and executive questions in Arabic and English drawn from the HR and Sales History schemas. This set is the primary instrument for measuring bilingual performance. Oracle's native Select AI serves as the comparison baseline for the NL2SQL component.

4.2 NL2SQL Evaluation Metrics

NL2SQL performance will be measured on execution accuracy, which tests whether the generated query returns the correct result set; exact-set match accuracy, which checks whether the result set matches the gold standard row for row; valid SQL rate, which measures the proportion of generated statements that parse and execute without error; and query latency from question submission to result return. All metrics will be reported separately for Arabic and English inputs to quantify the bilingual performance gap.

4.3 Narrative Quality Evaluation

Narrative faithfulness will be evaluated through two complementary methods: an LLM-as-judge evaluation in which a separate model assesses whether each statement in the narrative is grounded in the query results; and human evaluation by domain administrators rating faithfulness, readability, and informativeness. Fluency will be assessed through human rating only.

5. RESULTS AND ANALYSIS PENDING IMPLEMENTATION

6. DISCUSSION PENDING IMPLEMENTATION

7. CONCLUSION AND FUTURE WORK PENDING IMPLEMENTATION

Acknowledgments

REFERENCES

- Gao, D., Wang, H., Li, Y., Sun, X., Qian, Y., Ding, B., & Zhou, J. (2024). Text-to-SQL empowered by large language models: A benchmark evaluation. *Proceedings of the VLDB Endowment*, 17(5), 1132–1145. <https://arxiv.org/abs/2308.05389>
- Hong, Z., Yuan, Z., Zhang, Q., Chen, H., Dong, J., Huang, F., & Huang, X. (2024). Next-generation database interfaces: A survey of LLM-based text-to-SQL. *arXiv:2406.08426*.
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1–38.
- Keshireddy, S. R. (2025a). Natural language processing integration in Oracle APEX for enhanced user interaction in ubiquitous systems. *International Journal of Emerging Trends in Computer Science and Informatics*, 1(3).
- Keshireddy, S. R. (2025b). Automated data transformation and validation in Oracle APEX using adaptive AI models for secure enterprise applications. *Journal of Internet Services and Information Security*, 15(2), 185–208. <https://doi.org/10.58346/JISIS.2025.I2.041>
- Keshireddy, S. R. (2025c). Low-code development enhancement integrating large language models for intelligent code assistance in Oracle APEX. [Journal pending — verify at <https://orcid.org/0009-0007-6482-4438>]
- Keshireddy, S. R. (2025d). Retrieval-augmented generation techniques in Oracle APEX: Improving contextual responses in AI assistants. [Journal pending — verify at <https://orcid.org/0009-0007-6482-4438>]
- Lei, F., Chen, J., Ye, Y., Cao, R., et al. (2025). Spider 2.0: Evaluating language models on real-world enterprise text-to-SQL workflows. *ICLR 2025*. *arXiv:2411.07763*.
- Li, J., Hui, B., Qu, G., Yang, J., Li, B., Li, B., Wang, B., Qin, B., Cao, R., Geng, R., Huo, N., Zhou, X., Ma, C., Li, G., Chang, K. C. C., Huang, F., Cheng, R., & Li, Y. (2023). Can LLM already serve as a database interface? A big bench for large-scale database grounded text-to-SQLs. *NeurIPS 2023*. *arXiv:2305.03111*.
- Li, W., Wu, W., Chen, M., Liu, J., Xiao, X., & Wu, H. (2022). Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods. *arXiv:2203.05227*.
- Lin, Y., Ruan, T., Liu, J., & Wang, H. (2024). A survey on neural data-to-text generation. *IEEE Transactions on Knowledge and Data Engineering*, 36(4), 1431–1449.
- Nan, L., Radev, D., Zhang, R., et al. (2021). DART: Open-domain structured data record to text generation. *NAACL 2021*. *arXiv:2007.02871*.
- Oyewole, A. T., Adeoye, O. B., Addy, W. A., Okoye, C. C., Ofodile, O. C., & Ugochukwu, C. E. (2024). Automating financial reporting with natural language processing: A review and case analysis.

- World Journal of Advanced Research and Reviews, 21(3), 575–589.
<https://doi.org/10.30574/wjarr.2024.21.3.0688>
- Parikh, A. P., Wang, X., Gehrmann, S., Faruqui, M., Dhingra, B., Yang, D., & Das, D. (2020). ToTTo: A controlled table-to-text generation dataset. EMNLP 2020. arXiv:2004.14373.
- Pourreza, M., & Rafiei, D. (2023). DIN-SQL: Decomposed in-context learning of text-to-SQL with self-correction. NeurIPS 2023. arXiv:2304.11015.
- Rao, N., Tsay, J., Kate, K., Hellendoorn, V. J., & Hirzel, M. (2024). AI for low-code for AI. Proceedings of the 29th International Conference on Intelligent User Interfaces (IUI 2024), 837–852.
<https://doi.org/10.1145/3640543.3645203>
- Roychowdhury, S., Krema, M., Mahammad, A., Moore, B., Mukherjee, A., & Prakashchandra, P. (2024). ERATTA: Extreme RAG for table-to-answers with large language models. arXiv:2405.03963.
- Scholak, T., Schucher, N., & Bahdanau, D. (2021). PICARD: Parsing incrementally for constrained autoregressive decoding from language models. EMNLP 2021. arXiv:2109.05093.
- Sharma, A., Giri, S., Pandey, S., Pandey, P., & Pandey, S. K. (2025). A study on natural language processing techniques for intelligent text generation. IJRASET, 13(12).
<https://doi.org/10.22214/ijraset.2025.76052>
- Shi, L., Tang, Z., Zhang, N., Zhang, X., & Yang, Z. (2024). A survey on employing large language models for text-to-SQL tasks. arXiv:2407.15186.
- Syed, A. (2025). Leveraging generative AI in Oracle APEX for intelligent applications [Preprint]. ResearchGate. <https://www.researchgate.net/publication/397555550>
- Yagamurthy, D. N., Azmeera, R., & Khanna, R. (2023). Natural language generation (NLG) for automated report generation. Journal of Technology and Systems, 5(1), 48–59.
<https://doi.org/10.47941/jts.1497>
- Yeruva, S. K. (2025). Leveraging AI features in Oracle APEX for smart application development. World Journal of Advanced Engineering Technology and Sciences, 16(1), 624–633.
<https://doi.org/10.30574/wjaets.2025.16.1.1239>
- Zhao, Y., Zhang, H., Si, S., Nan, L., Tang, X., & Cohan, A. (2023). Investigating table-to-text generation capabilities of large language models in real-world information seeking scenarios. EMNLP Findings 2023. arXiv:2305.14987.

Thesis Title: Developing an NLP-Based Model for Generating Intelligent Administrative and Executive Reports within an Oracle APEX Environment: An Applied Study

Group Members:

Eman Alhabashi

Merwa Idris

Alaa Humaid

Department: Department of Artificial Intelligence

University Name: MidOcean University

Table of Contents

Chapter 1	4
1.1 Background and context.....	4
1.2 Problem Statement	5
1.3 Research Questions.....	6
1.4 Significance of the Study	7
1.5 Scope and Delimitations	7
1.6 Approach	8
1.7 Thesis Organization	9
1.8 Expected Results (Should we remove it ?)	10
1.9 Resources Needed (Should we remove it ?)	11
Chapter 2: Literature Review	12
2.1 Introduction	12
2.2 Asking questions in natural language: NL2SQL	12
2.3 Turning data into written reports: text generation	14
2.4 Bringing AI into Oracle APEX and low-code platforms.....	16
2.5 Synthesis and research gap	18
Appendix: Synthesis Matrix.....	19
References	21

Developing an NLP-Based Model for Generating Intelligent Administrative and Executive Reports within an Oracle APEX Environment: An Applied Study

Chapter 1

1.1 Background and context

Organizations today generate and store more data than at any point in history. Relational databases sit at the center of daily operations — recording transactions, staffing movements, customer interactions, inventory levels, and financial flows — yet the ability of decision-makers to access and understand that data has not kept pace with its growth. A persistent gap has opened between data availability and data accessibility: the data exists, but turning it into an answer a manager can act on still requires specialist involvement and time that most organizations struggle to budget.

The scale of this problem is visible in how heavily organizations have invested in trying to solve it. The global business intelligence (BI) and analytics market was valued at approximately USD 54.9 billion in 2025 and is growing at a compound annual rate of 12.4% (DataStackHub, 2025). Yet investment in tools has not fully closed the access gap. Research shows that only 32% of executives feel confident making meaningful use of data (G2, 2024), and data preparation alone consumes roughly 45% of data scientists' time before any analysis can begin (SDG Group, 2024). Even among organizations that have deployed BI platforms, self-service analytics — the ability for a non-technical person to ask a new question and get an answer without developer help — has reduced report generation backlogs by 40%, suggesting that for the other 60%, developer dependency remains the norm (DataStackHub, 2025).

The language dimension compounds this problem in the MENA region. The vast majority of BI and reporting tools operate in English. Executives and administrators who work in Arabic are either excluded from self-service querying entirely or forced to operate in a second language, which slows comprehension and introduces risk of misinterpretation. This is not a niche concern: the Arabian Gulf states host large concentrations of Oracle database deployments across government, healthcare, banking, and energy, and administrative reporting in those contexts is frequently required in Arabic.

Oracle Application Express (APEX) sits at a specific intersection of this problem. It is a low-code development platform that runs directly on the Oracle Database, ships at no additional cost with Oracle licenses, and is used by hundreds of thousands of developers to build data-driven web applications for enterprise and government clients. Its tight integration with the database makes it a natural candidate for embedding intelligent querying and reporting capabilities, but as the literature reviewed in Chapter 2 shows, almost no published research has explored that combination in depth. APEX developers today build static reports and dashboards that answer yesterday's questions; this research aims to add a layer that can handle tomorrows.

1.2 Problem Statement

Decision-makers in most organizations face a consistent bottleneck when they need data-driven answers: they cannot get them without a developer. Enterprise reporting systems are built around fixed dashboards and pre-designed queries. When an executive needs a new perspective on the data — a comparison across quarters, a breakdown by department, an explanation of a sudden variance — they must request a new report, wait for a developer to design and deploy it, and then interpret a table of numbers without narrative context. This process can take days or weeks, depending on developer availability and backlog.

Three specific problems make this worse than it needs to be. First, the dependency on developers creates a structural bottleneck: executive decision-making speed is capped by developer capacity. Second, existing data is presented in raw form — rows, columns, and charts — with no system to explain what the numbers mean, highlight what is unusual, or surface the insight the reader should care about. Third, in Arabic-speaking organizations, virtually all commercial and open-source NLP tools are built for English, which means Arabic-speaking executives are excluded from any natural-language query capability, even where it theoretically exists.

This research addresses all three by proposing an NLP-based layer embedded directly inside Oracle APEX. The layer allows a decision-maker to type a question in Arabic or English,

translates that question automatically into a validated SQL query, executes it against the organization's data, and returns both the result and a concise narrative summary explaining what the data shows. The goal is to replace the developer-as-intermediary with a system that is fast, faithful to the data, and accessible to Arabic and English speakers equally.

1.3 Research Questions

This research is organized around four questions. Together they cover the accuracy of the query component, the faithfulness of the reporting component, the architecture required to make the two work together inside APEX, and how the resulting system compares to what Oracle already provides natively.

RQ1: *To what extent can a prompt-engineered large language model accurately translate Arabic and English natural-language questions into correct SQL queries against enterprise Oracle databases?*

RQ2: *How effective are LLM-based data-to-text systems at generating faithful and readable narrative summaries from SQL query results in an administrative reporting context?*

RQ3: *What architectural components are required to integrate NL2SQL and natural language generation capabilities securely and effectively within Oracle APEX?*

RQ4: *How does the proposed bilingual NLP pipeline compare to Oracle's built-in Select AI and APEX AI Assistant in terms of query accuracy and user satisfaction?*

RQ1 and RQ2 address the core technical performance of the two components. RQ3 is the integration question — it acknowledges that a working system requires more than two models and deals with schema linking, security, validation, and the user interface. RQ4 frames the evaluation against a concrete, relevant baseline rather than abstract benchmarks alone.

1.4 Significance of the Study

This research contributes at four levels.

Theoretically, the NL2SQL and data-to-text literatures are well developed but almost entirely built around English benchmarks and general-purpose datasets. Previous work has not systematically tested these two capabilities together in a bilingual administrative context, where the domain is specific, the language is mixed Arabic and English, and the constraints are real rather than idealized. This study extends both fields into that space and tests the pipeline end to end rather than each component in isolation.

Practically, the prototype gives organizations using Oracle APEX a direct path to self-service executive reporting. Decision-makers who currently wait days for a new report can instead ask a question and receive a result in seconds. Developer time is freed from repetitive report-building and redirected to work that requires genuine expertise. The bilingual capability serves organizations in the MENA region specifically, where this kind of tool does not currently exist.

Methodologically, this study creates a bilingual Arabic-English evaluation set of administrative questions paired with correct SQL and reference narrative answers, built with real domain administrators. This dataset does not exist in the published literature and is designed to be reusable for future research in Arabic NL2SQL and multilingual data-to-text evaluation.

Architecturally, the study proposes a secure, modular blueprint showing how to combine large language models, prompt engineering, SQL validation, and narrative generation inside Oracle APEX. The design decisions are documented so that other APEX developers can adapt the pattern to their own environments, and the architecture can serve as a reference model for future AI-enabled Oracle enterprise applications

1.5 Scope and Delimitations

This study is focused on a specific, well-defined slice of the broader problem it addresses. The database layer runs on Oracle Database or Autonomous Database and uses the standard HR and Sales History sample schemas, which provide a clean relational foundation with well-defined tables, columns, and key relationships.

All queries are strictly read-only — data integrity is maintained throughout and no write operations are possible. The system supports Modern Standard Arabic and English, both for user input and generated output.

The NLP capabilities cover two tasks: translating natural-language questions into SQL, and generating short, accurate narrative summaries of the results. Both are handled through prompt engineering and in-context learning with large language models, without fine-tuning any model.

A human-in-the-loop step sits between generation and execution — every SQL query is shown to the user with a plain-language explanation before it runs, and access is limited to data the user is already authorized to see under their APEX credentials.

Several boundaries are worth stating explicitly:

- The system is not designed to work with other database platforms such as PostgreSQL or SQL Server, and it does not handle analytical tasks outside the HR and Sales History schemas.
- The system does not support real-time, streaming, or unstructured data, and write operations are explicitly blocked.
- From a language standpoint, dialectical Arabic is out of scope, as are all other languages.
- The interface is text-based — there is no voice input.
- The prototype processes single-turn queries and does not maintain conversational context across turns, nor does it learn from user interactions over time.
- On the security and evaluation side, the system relies on APEX's native security framework. Additional measures such as end-to-end encryption or GDPR compliance are not implemented, since the system runs on synthetic sample schemas rather than live organizational data.
- The user evaluation draws on a small, targeted group of participants to gather qualitative insight rather than broad statistical results.

These are deliberate choices that keep the study feasible and its findings honest about what the prototype can and cannot do.

1.6 Approach

This study follows four stages: preparing the data, getting it into a form the model can use, building the two core NLP components, and evaluating the result. Each stage feeds the next, and together they produce both a working prototype and the evidence needed to answer the four research questions.

The data stage assembles two kinds of input — the organizational database the system reports on, built from Oracle's HR and Sales History schemas, and the public benchmarks and custom bilingual evaluation set used for testing.

The preparation stage pulls schema metadata out of Oracle's data dictionary, normalizes Arabic text, and sets up a vector-search retrieval store so that relevant history and terminology can be injected into each prompt at query time.

The build stage produces the NL2SQL component, which generates and validates SQL with human confirmation before execution, and the narrative generation component, which turns clean query results into a short faithful summary, both integrated into an APEX application through PL/SQL and REST.

The evaluation stage measures SQL accuracy on benchmarks and the bilingual set, assesses narrative faithfulness through automatic metrics and human review, compares the system against Oracle's native Select AI, and captures user feedback through a task-based study. Full technical detail for each stage is provided in Chapter 3.

1.7 Thesis Organization

This thesis is organized into six chapters that move the reader from the initial identification of the problem through to the final conclusions.

Chapter 1 — **Introduction** — establishes the research context. It presents the challenge of inaccessible ad-hoc data querying in Oracle APEX environments and introduces the proposed solution: an NLP layer that translates natural-language questions into SQL and generates narrative summaries. The chapter states the research questions, describes the significance of the work, and sets out the scope and boundaries of the study.

Chapter 2 — **Literature Review** — examines prior research across the three areas this thesis draws on. It reviews the development of NL2SQL methods, key benchmarks and their limitations, and the challenges of faithful data-to-text generation. It also surveys the emerging literature on AI integration in low-code platforms and Oracle APEX specifically, and closes by identifying the research gap this thesis occupies: no existing study combines validated NL2SQL, faithfulness-controlled NLG, and bilingual capability inside an APEX reporting application.

Chapter 3 — **Methodology** — describes the technical blueprint for building the system. It covers the Oracle database and dataset setup, the preparation of data for model readability, the selection and design of the NL2SQL and NLG modules, and the evaluation

framework including the construction of the bilingual test set and the metrics used to assess SQL accuracy and narrative faithfulness.

Chapter 4 — **System Implementation** — documents the actual construction of the prototype, from setting up the Oracle APEX application to building the NL2SQL and NLG modules with prompt engineering and safety controls. It walks through the complete user flow from question to narrative summary, and describes the configuration of Oracle's native AI features — Select AI and the APEX AI Assistant — as baseline comparators.

Chapter 5 — **Evaluation and Results** — presents the empirical findings. It reports SQL accuracy on public benchmarks and the custom bilingual test set, assesses narrative quality using both automatic metrics and human evaluation, and compares the proposed system against Oracle's built-in AI features. Qualitative feedback from end-users is included, and results are discussed in relation to each research question.

Chapter 6 — **Conclusion and Future Work** — draws the thesis together. It summarizes the contributions, reflects on what was learned during development, acknowledges the limitations of the prototype, and outlines directions for future work — including extending schema support, enabling multi-turn conversations, and developing a production-grade version of the system.

1.8 Expected Results (Should we remove it ?)

By the end of the project, the main expectation is to have a working tool inside Oracle APEX where the manager should be able to type a question in Arabic or English, and the system will generate the correct query, run it, and display both the numbers and a short written summary explaining their meaning.

We also expect to have a real measuring on the efficiency of query generation on both the public test sets and my own Arabic/English set, and how accurate and trustworthy the written summaries are.

The project should also produce some useful comparisons. We would like to see the difference between Oracle AI and our approach, especially in Arabic, since Arabic has far less support.

Finally, the project is meant to leave something behind that others can build on: a clear blueprint for adding this kind of natural-language reporting to other APEX applications, a bilingual Arabic–English test set for administrative questions that did not exist before, and a set of practical lessons on how to keep the system accurate and secure.

1.9 Resources Needed (Should we remove it ?)

1. Oracle platform: Oracle Database 23ai or Autonomous Database (OCI Always Free tier is sufficient for development), Oracle APEX (apex.oracle.com or a local instance).
2. LLM access: an API to a commercial model (e.g., OpenAI, Cohere, or OCI Generative AI) and, optionally, open models (Llama 3, Qwen).
3. Development tools: VS Code, SQL Developer / SQLcl, Python (transformers, datasets, evaluation libraries), Git.
4. Datasets: Spider, BIRD, Spider 2.0; ToTTo, DART, WebNLG; Oracle sample schemas; plus the custom bilingual set.

Chapter 2: Literature Review

2.1 Introduction

Most organizations already hold the data they need to run themselves. It lives in relational databases behind their day-to-day systems: finance, human resources, operations, case records. The difficulty is rarely a shortage of data. It is getting an answer out of that data and into a shape a manager can use without waiting for a developer to build yet another report. Two capabilities from natural language processing speak to this. One lets a person ask a question in plain language and have the system write the database query for them. The other takes the rows that come back and turns them into a short written account instead of leaving the reader to interpret a table. This thesis works at the join between those two, and adds a condition most studies leave out: keeping the whole thing inside a low-code platform, Oracle APEX, where the data and the application already sit together.

This chapter reviews the work that bears on that goal across three areas. The first is converting natural language into SQL, which has the most active research record. The second is generating text from data, where the recurring worry is whether the words stay true to the numbers. The third, and youngest, is the effort to put AI to work inside low-code tools and APEX in particular. I have kept the focus on work from 2020 onward, since the field moved sharply once large language models arrived and older results no longer describe what these systems can do.

2.2 Asking questions in natural language: NL2SQL

The dominant way to frame this problem is as translation: take a question and the structure of the database it refers to, and produce a SQL statement that runs and returns the right answer. A fluent model will gladly produce SQL that does not parse, so some of the early progress came from controlling the output. Scholak, Schucher, and Bahdanau (2021), in PICARD, constrained the model's decoding so that at each step it could only add tokens that keep the query valid, which cut down on broken SQL without retraining the model underneath.

Large language models then shifted the methods. Instead of training a dedicated parser, researchers found they could hand a general model the schema and a few examples and get usable SQL back, provided they structured the prompt well. Pourreza and Rafiei (2023), in DIN-SQL, showed that splitting the task into smaller steps and letting the model review and repair its own draft raised accuracy well above plain prompting. Gao et al. (2024), behind DAIL-SQL, stepped back and asked which prompt choices actually matter, comparing how the question and schema are represented and how in-context examples are picked and ordered. Their contribution reads less like a new model and more like a careful account of what works, which is what a builder needs.

The benchmarks, meanwhile, became more honest about real conditions. Li et al. (2023), in BIRD, assembled large databases with messy values and a need for outside knowledge, and they measured efficiency rather than only whether the answer was right. The gap they reported between the best models and human experts was wide. Lei et al. (2025), in Spider 2.0, pushed further toward the sprawling schemas and multi-step analysis an analyst meets at work, and current models clear only a modest fraction of those tasks. Two surveys, Hong et al. (2024) and Shi et al. (2024), organize this fast-moving area, usually along one line: approaches that steer a general model through prompting, and approaches that specialize one through fine-tuning, each trading off cost, control, and data.

Read together, a few things stay unresolved for anyone building an administrative reporting tool. The benchmarks are almost all English and drawn from generic domains, so a high score says little about a specific organization's tables and vocabulary. Performance on messy, enterprise-scale data is still weak, which is the BIRD and Spider 2.0 result. The safety of a generated query, whether it only reads, whether it respects a user's permissions, and how a person checks it before it runs, is treated as an implementation detail in most academic work even though it decides whether the thing can be deployed at all. And the bilingual case, Arabic next to English, barely appears.

2.3 Turning data into written reports: text generation

The second capability is the reverse direction: start from structured data and produce readable text. It helps to set the scene with the broader shift in text generation that Sharma et al. (2025) describe. Their overview traces the move from handcrafted rules to statistical methods to neural models and then transformers, and it illustrates the modern end of that line with a small experiment in which Google's Gemini turns a user's background into career insights and draft documents. The paper is introductory and runs no real evaluation, but it captures the point that matters here: contextual, transformer-based generation has largely replaced the older template and feature-engineering approaches.

Within data-to-text specifically, the work has circled one problem: keeping the text faithful to the source. Parikh et al. (2020), in ToTTo, built a controlled task in which particular table cells are highlighted and the target sentence is produced by trimming a real Wikipedia sentence down to only what the data supports, which gives a cleaner test of whether a model stays grounded. Nan et al. (2021), in DART, framed generation from sets of structured records into sentences. Two surveys then consolidate the concern itself. Li et al. (2022) review how faithfulness is analyzed, measured, and improved, and Ji et al. (2023) give a broad map of hallucination across generation tasks, data-to-text among them, with the mitigation strategies tried so far.

Large language models reopened the question of how best to chase faithfulness. Zhao et al. (2023) studied table-to-text in realistic information-seeking settings and found that LLMs can write statements more faithful than earlier fine-tuned systems, and further that an LLM can act as a reasonable judge of faithfulness and give step-by-step feedback on another model's draft. Lin et al. (2024) survey the neural data-to-text field and give it a clear shape, separating the planning of what to say from the wording of how to say it. On the applied side, Roychowdhury et al. (2024), in ERATTA, paired retrieval over enterprise tables with explicit checks for hallucinated answers, which is close to what a trustworthy reporting tool needs.

One paper here sits squarely in the domain this thesis cares about. Oyewole et al. (2024) review the use of NLP to automate financial reporting, covering sentiment and thematic

analysis, summarization, and the regulatory side, and they report gains in efficiency and accuracy alongside real worries about consistency. They cite, for instance, a query-based summarization system reaching roughly a ROUGE-1 score of 46%, and tools built specifically to test whether text-based financial forecasts stay logically consistent. It is a secondary review rather than a built system, and its prose leans heavily on the promotional register, but it earns its place for one reason: in finance and administration the cost of an unfaithful sentence is high, which is exactly why faithfulness cannot be an afterthought.

Yagamurthy, Azmeera, and Khanna (2023) take this a step closer to the administrative domain. Their study targets the automated transformation of structured organizational data — financial records, business-intelligence figures, healthcare KPIs, and customer feedback — into human-readable reports and narratives. The pipeline they describe runs through data cleaning, standardization, and data-to-text mapping before generation, and the challenges they name — domain-specific language, contextual understanding, data quality, and ethical handling — are exactly the obstacles an executive-reporting tool faces. It is not an APEX study, and it does not address bilingual output, but it is the clearest existing precedent for automated narrative report generation from the kind of data this thesis works with.

One paper here sits squarely in the domain this thesis cares about. Oyewole et al. (2024) review the use of NLP to automate financial reporting, covering sentiment and thematic analysis, summarization, and the regulatory side, and they report gains in efficiency and accuracy alongside real worries about consistency. They cite, for instance, a query-based summarization system reaching roughly a ROUGE-1 score of 46%, and tools built specifically to test whether text-based financial forecasts stay logically consistent. It is a secondary review rather than a built system, and its prose leans heavily on the promotional register, but it earns its place for one reason: in finance and administration the cost of an unfaithful sentence is high, which is exactly why faithfulness cannot be an afterthought.

The gaps mirror the query side. Faithfulness is still the open risk, and it matters most in executive reporting, where one wrong figure in a summary can sink trust in the whole tool. Most of the work is English and built on Wikipedia tables or generic sources, with little on

administrative or financial narratives, less on bilingual generation, and little that ties the written summary back to a live query result inside a working application rather than a fixed benchmark table.

2.4 Bringing AI into Oracle APEX and low-code platforms

The third area is the youngest and the thinnest. Low-code platforms let people with limited programming background build working applications through visual, declarative tools, and Oracle APEX is a prominent case, set apart by running directly on the Oracle Database and shipping with it. Oracle has recently added generative AI features to the platform, and a handful of papers have begun to study what that enables.

Yeruva (2025) gives a wide review of AI inside APEX, covering predictive analytics, sentiment analysis, recommendation engines, NLP-driven interfaces, in-database machine learning, and the pattern of calling external models through REST APIs. It is broad rather than deep, with no single system built and tested end to end, and it closes on familiar challenges around integration, explainability, and scalability. Its value is as a map of the territory.

Keshireddy (2025c) go a step further by embedding a large language model directly into the APEX development environment to provide intelligent code assistance. The system helps developers generate, understand, and optimize code, and the reported outcomes include faster development cycles, fewer coding errors, and better support for the low-code model overall. Like Yeruva's review, the focus is on building applications rather than querying or reporting on their data, but it contributes something useful: it shows that LLM assistance inside APEX is not just possible at a proof-of-concept level but measurably accelerates real development work.

Keshireddy(2025d) present an AI assistant built inside Oracle APEX using Retrieval-Augmented Generation (RAG). Rather than relying on a model's fixed training knowledge, the system retrieves relevant content from organizational documents and records at query time and uses it to produce context-aware, accurate responses. The reported gains are higher response accuracy, better context understanding, and improved user satisfaction. RAG is a method this thesis relies on directly — using Oracle 23ai's vector search to pull relevant

history and schema documentation into each prompt — so this paper is direct evidence that the integration is viable inside APEX, not just in stand-alone LLM settings.

Keshireddy (2025a) is the most thorough of the APEX papers. It proposes a modular NLP architecture for APEX built around intent recognition and context-aware responses, with lightweight pipelines for intent detection, entity recognition, and adaptive dialogue. It implements four use cases: triggering predefined workflows from natural language, retrieving data on request, filling forms intelligently, and answering user queries through a conversational agent. And it evaluates with metrics a committee will recognize as real: task completion time, interaction error rate, intent-recognition precision, latency, and throughput. The focus is on interaction and usability rather than report generation, but as a template for how to structure and measure an APEX-plus-NLP system, it is the strongest of the set.

Keshireddy (2025b), in a separate study published in the Journal of Internet Services and Information Security (JISIS), shifts from interaction to data integrity. The paper proposes an adaptive AI framework inside Oracle APEX that automates data transformation and validation using semantic, statistical, and constraint-based techniques rather than the rigid PL/SQL rules most APEX systems rely on. It was tested against 980,000 records across HRMS, CRM, and Accounting modules, reporting transformation accuracy above 92% in HRMS and CRM and 94.6% in Accounting. The framework also incorporates encryption, access control, and audit logging, and it adapts automatically when schemas change. This connects to two objectives in this thesis: the need to handle real enterprise data reliably, and the need to ensure that the system's operations are secure and auditable. It is the most evidence-grounded of the APEX papers and the most explicitly enterprise-scale.

Ashraf Syed (2025) is the closest in subject to this thesis. It centers on the APEX AI Assistant introduced in APEX 24.1 and presents a proof-of-concept chatbot that turns natural-language prompts into executable SQL, reporting around 92% accuracy and a 50% cut in the time to produce a report, with an enterprise example to back the claim. It names data privacy, model accuracy, and skill gaps as the obstacles, and points to prompt engineering

and Oracle's cloud infrastructure as the answers. The subject is right, but the work is a single-author proof of concept, and the headline figures come without a transparent evaluation protocol, so the numbers are better read as encouraging than as established.

A broader lesson comes from outside APEX. Rao et al. (2024) built LowCoder, which pairs a visual interface with a natural-language interface for assembling machine-learning pipelines, and ran a user study comparing the two. Their finding carries well beyond their setting: the natural-language interface helped people discover operators and options they would not have found in the menus, but they still leaned on the visual layer to read and correct what the system produced. The implication for a reporting tool is direct. A natural-language front end works best when the person can still see and fix what the machine generated, which is the argument for letting a user inspect the SQL before it runs and weigh the summary against the numbers.

Taken together, the AI-in-APEX literature is small, recent, aimed more at building applications or smoothing interaction than at generating reports from data, and mostly at the proof-of-concept stage. Like the other two areas, it is almost entirely English.

2.5 Synthesis and research gap

Each area is in decent health on its own. NL2SQL has capable methods and increasingly realistic benchmarks, but it is tested mostly on clean, English, generic data, and the practical questions of validating queries on messy enterprise schemas, and of handling Arabic, are barely addressed. Data-to-text generation has reached the point where large models can write grounded summaries, yet faithfulness remains the unsolved risk, and administrative or bilingual reporting has had little attention. AI inside low-code platforms is just starting, and the work closest to this thesis, a SQL-generating chatbot in APEX, is a proof of concept without a rigorous evaluation, and it does not address narrative reporting or two languages.

The space between these is where this thesis sits. As far as this review found, no existing study brings together a validated natural-language-to-SQL front end and a faithfulness-controlled data-to-text back end, inside a low-code Oracle APEX application, serving

administrative and executive reporting in both Arabic and English. The pieces exist separately, usually tested far from a working enterprise system. Building and evaluating the whole pipeline in its real setting is the contribution this thesis aims for, and it is the kind of contribution an applied study, backed by direct APEX experience, is well placed to make.

Appendix: Synthesis Matrix

Work	Area	What it does	Language	Limitation for this thesis
Scholak et al. (2021), PICARD	NL2SQL	Constrained decoding for valid SQL	English	Validity only; not permissions or safety
Pourreza & Rafiei (2023), DIN-SQL	NL2SQL	Decomposition + self-correction (LLM)	English	No bilingual or deployment study
Gao et al. (2024), DAIL-SQL	NL2SQL	Systematic prompt-strategy evaluation	English	Benchmark focus; not an applied setting
Li et al. (2023), BIRD	NL2SQL	Realistic, dirty-database benchmark	English	Wide human-model gap remains
Lei et al. (2025), Spider 2.0	NL2SQL	Enterprise analytical workflows	English	Low solve rate; no Arabic
Hong et al. (2024)	NL2SQL	Survey	English	Confirms gaps; no system
Shi et al. (2024)	NL2SQL	Survey	English	Confirms gaps; no system
Sharma et al. (2025)	Text generation	Overview + Gemini career-text demo	English	Introductory; no evaluation
Parikh et al. (2020), ToTTo	Data-to-text	Controlled, highlighted-cell generation	English	Open-domain Wikipedia tables
Nan et al. (2021), DART	Data-to-text	Record / triple-to-text generation	English	Not administrative reporting
Li et al. (2022)	Data-to-text	Faithfulness survey	English	Frames the risk; no system

Work	Area	What it does	Language	Limitation for this thesis
Ji et al. (2023)	Data-to-text	Hallucination survey	English	Frames the risk; no system
Zhao et al. (2023)	Data-to-text	LLM table-to-text + LLM-as-evaluator	English	No bilingual or live-query link
Lin et al. (2024)	Data-to-text	Survey	English	Confirms gaps; no system
Roychowdhury et al. (2024), ERATTA	Data-to-text	RAG over enterprise tables + hallucination checks	English	Q&A focus, not narrative reports
Yagamurthy et al. (2023)	Data-to-text	NLG for automated report generation from structured data	English	No APEX integration; no bilingual coverage
Oyewole et al. (2024)	Data-to-text (finance)	Review of NLP for financial reporting	English	Secondary review; no built system
Yeruva (2025)	AI + APEX	Review of AI features in APEX	English	Broad; no evaluated system
Keshireddy (2025c)	AI + APEX	LLM code assistance for APEX development	English	Developer productivity focus, not reporting
Keshireddy (2025d)	AI + APEX	RAG-based AI assistant inside APEX	English	Q&A / conversational focus, not reporting
Keshireddy (2025a)	AI + APEX	Modular NLP architecture; four use cases; evaluated	English	Interaction focus, not reporting
Keshireddy (2025b, JISIS)	AI + APEX	Adaptive AI for data transformation & validation in APEX	English	Data processing focus; not NL querying or reporting
Ashraf Syed (2025)	AI + APEX	SQL-generating chatbot proof of concept (APEX 24.1)	English	Not rigorously evaluated; not narrative or bilingual
Rao et al. (2024), LowCoder	AI + low-code	Visual + NL interface for ML pipelines; user study	English	ML pipelines, not database reporting

References

- Yu, T., Zhang, R., Yang, K., et al. (2018). Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task. EMNLP 2018.
- Li, J., Hui, B., Qu, G., et al. (2023). Can LLM Already Serve as a Database Interface? A Big Bench for Large-Scale Database Grounded Text-to-SQLs (BIRD). NeurIPS 2023.
- Lei, F., Chen, J., Ye, Y., et al. (2025). Spider 2.0: Evaluating Language Models on Real-World Enterprise Text-to-SQL Workflows. ICLR 2025.
- Hong, Z., Yuan, Z., Zhang, Q., et al. (2024). Next-Generation Database Interfaces: A Survey of LLM-Based Text-to-SQL. arXiv:2406.08426.
- Shi, L., Tang, Z., Zhang, N., et al. (2024). A Survey on Employing Large Language Models for Text-to-SQL Tasks. arXiv:2407.15186.
- Gao, D., Wang, H., Li, Y., et al. (2024). Text-to-SQL Empowered by Large Language Models: A Benchmark Evaluation (DAIL-SQL). Proceedings of the VLDB Endowment.
- Lin, Y., Ruan, T., Liu, J., & Wang, H. (2024). A Survey on Neural Data-to-Text Generation. IEEE Transactions on Knowledge and Data Engineering, 36(4), 1431–1449.
- Li, W., Wu, W., Chen, M., et al. (2022). Faithfulness in Natural Language Generation: A Systematic Survey of Analysis, Evaluation and Optimization Methods. arXiv:2203.05227.
- Zhao, Y., et al. (2023). Investigating Table-to-Text Generation Capabilities of LLMs in Real-World Information Seeking Scenarios. arXiv:2305.14987.
- Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS 2020.
- Oracle. Configuring Generative AI in Oracle APEX using Select AI and DBMS_CLOUD_AI documentation. Oracle APEX Blog / Oracle Documentation.
- Oracle. Announcing Oracle APEX 24.2 General Availability — Generative Development (GenDev), AI Configurations, and RAG Sources. Oracle APEX Blog, 2024–2025.
- Oracle. Oracle APEX AI Platform. <https://apex.oracle.com/en/platform/ai/>
- Gao, D., Wang, H., Li, Y., Sun, X., Qian, Y., Ding, B., & Zhou, J. (2024). Text-to-SQL empowered by large language models: A benchmark evaluation. *Proceedings of the VLDB Endowment*, 17(5), 1132–1145. <https://arxiv.org/abs/2308.05389>
- Hong, Z., Yuan, Z., Zhang, Q., Chen, H., Dong, J., Huang, F., & Huang, X. (2024). *Next-generation database interfaces: A survey of LLM-based text-to-SQL*. arXiv. <https://arxiv.org/abs/2406.08426>

- Lei, F., Chen, J., Ye, Y., Cao, R., Shin, D., Su, H., Suo, Z., Gao, H., Hu, W., Yin, P., Zhong, V., Xiong, C., Sun, R., Liu, Q., Wang, S., & Yu, T. (2025). Spider 2.0: Evaluating language models on real-world enterprise text-to-SQL workflows. *International Conference on Learning Representations (ICLR)*. <https://arxiv.org/abs/2411.07763>
- Li, J., Hui, B., Qu, G., Yang, J., Li, B., Li, B., Wang, B., Qin, B., Cao, R., Geng, R., Huo, N., Zhou, X., Ma, C., Li, G., Chang, K. C. C., Huang, F., Cheng, R., & Li, Y. (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 (NeurIPS)*, 36. <https://arxiv.org/abs/2305.03111>
- Pourreza, M., & Rafiei, D. (2023). DIN-SQL: Decomposed in-context learning of text-to-SQL with self-correction. *Advances in Neural Information Processing Systems (NeurIPS)*, 36. <https://arxiv.org/abs/2304.11015>
- Scholak, T., Schucher, N., & Bahdanau, D. (2021). PICARD: Parsing incrementally for constrained auto-regressive decoding from language models. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9895–9901. <https://arxiv.org/abs/2109.05093>
- Shi, L., Tang, Z., Zhang, N., Zhang, X., & Yang, Z. (2024). A survey on employing large language models for text-to-SQL tasks. *arXiv*. <https://arxiv.org/abs/2407.15186>
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1–38. <https://arxiv.org/abs/2202.03629>
- Li, W., Wu, W., Chen, M., Liu, J., Xiao, X., & Wu, H. (2022). *Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods*. *arXiv*. <https://arxiv.org/abs/2203.05227>
- Lin, Y., Ruan, T., Liu, J., & Wang, H. (2024). A survey on neural data-to-text generation. *IEEE Transactions on Knowledge and Data Engineering*, 36(4), 1431–1449. <https://doi.org/10.1109/TKDE.2023.3304385>
- Nan, L., Radev, D., Zhang, R., Rau, A., Sivaprasad, A., Hsieh, C., Tang, X., Vyas, A., Verma, N., Krishna, P., Liu, Y., Irwanto, N., Pan, J., Rahman, F., Zaidi, A., Mutuma, M., Tarabar, Y., Gupta, A., Yu, T., ... Radev, D. (2021). DART: Open-domain structured data record to text generation. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 432–447. <https://arxiv.org/abs/2007.02871>
- Oyewole, A. T., Adeoye, O. B., Addy, W. A., Okoye, C. C., Ofodile, O. C., & Ugochukwu, C. E. (2024). Automating financial reporting with natural language processing: A review and case analysis. *World Journal of Advanced Research and Reviews*, 21(3), 575–589. <https://doi.org/10.30574/wjarr.2024.21.3.0688>

- Parikh, A. P., Wang, X., Gehrmann, S., Faruqui, M., Dhingra, B., Yang, D., & Das, D. (2020). ToTTo: A controlled table-to-text generation dataset. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1173–1186. <https://arxiv.org/abs/2004.14373>
- Roychowdhury, S., Krema, M., Mahammad, A., Moore, B., Mukherjee, A., & Prakashchandra, P. (2024). ERATTA: Extreme RAG for table-to-answers with large language models. *IEEE International Conference on Big Data*. <https://arxiv.org/abs/2405.03963>
- Sharma, A., Giri, S., Pandey, S., Pandey, P., & Pandey, S. K. (2025). A study on natural language processing techniques for intelligent text generation. *International Journal for Research in Applied Science & Engineering Technology*, 13(12). <https://doi.org/10.22214/ijraset.2025.76052>
- Zhao, Y., Zhang, H., Si, S., Nan, L., Tang, X., & Cohan, A. (2023). Investigating table-to-text generation capabilities of large language models in real-world information seeking scenarios. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 160–175. <https://arxiv.org/abs/2305.14987>
- Keshireddy, S. R. (2025). Natural language processing integration in Oracle APEX for enhanced user interaction in ubiquitous systems. *International Journal of Emerging Trends in Computer Science and Informatics*, 1(3).
- Rao, N., Tsay, J., Kate, K., Hellendoorn, V. J., & Hirzel, M. (2024). AI for low-code for AI. *Proceedings of the 29th International Conference on Intelligent User Interfaces (IUI '24)*, 837–852. <https://doi.org/10.1145/3640543.3645203>
- Syed, A. (2025). *Leveraging generative AI in Oracle APEX for intelligent applications* [Preprint]. ResearchGate. <https://www.researchgate.net/publication/397555550>
- Yeruva, S. K. (2025). Leveraging AI features in Oracle APEX for smart application development. *World Journal of Advanced Engineering Technology and Sciences*, 16(1), 624–633. <https://doi.org/10.30574/wjaets.2025.16.1.1239>
- Yagamurthy, D. N., Azmeera, R., & Khanna, R. (2023). Natural language generation (NLG) for automated report generation. *Journal of Technology and Systems*, 5(1), 48–59. <https://doi.org/10.47941/jts.1497>
- Keshireddy, S. R. (2025). Automated data transformation and validation in Oracle APEX using adaptive AI models for secure enterprise applications. *Journal of Internet Services and Information Security (JISIS)*, 15(2), 185–208. <https://doi.org/10.58346/JISIS.2025.12.041>
- Keshireddy, S.R. (2025) Retrieval-Augmented Generation Techniques in Oracle APEX: Improving Contextual Responses in AI Assistants. Available at: <https://orcid.org/0009-0007-6482-4438> (Accessed: 27 June 2026).

