

MASTER'S THESIS DEFENSE | MIDOCEAN UNIVERSITY

Developing an NLP-Based Model for Generating Intelligent Administrative & Executive Reports within an Oracle APEX Environment

An Applied Study in Bilingual Natural Language Interfaces (NL2SQL & Data-to-Text NLG)

Eman Al Habashi
ID: 261000480

Merwa Idris
ID: 221002171

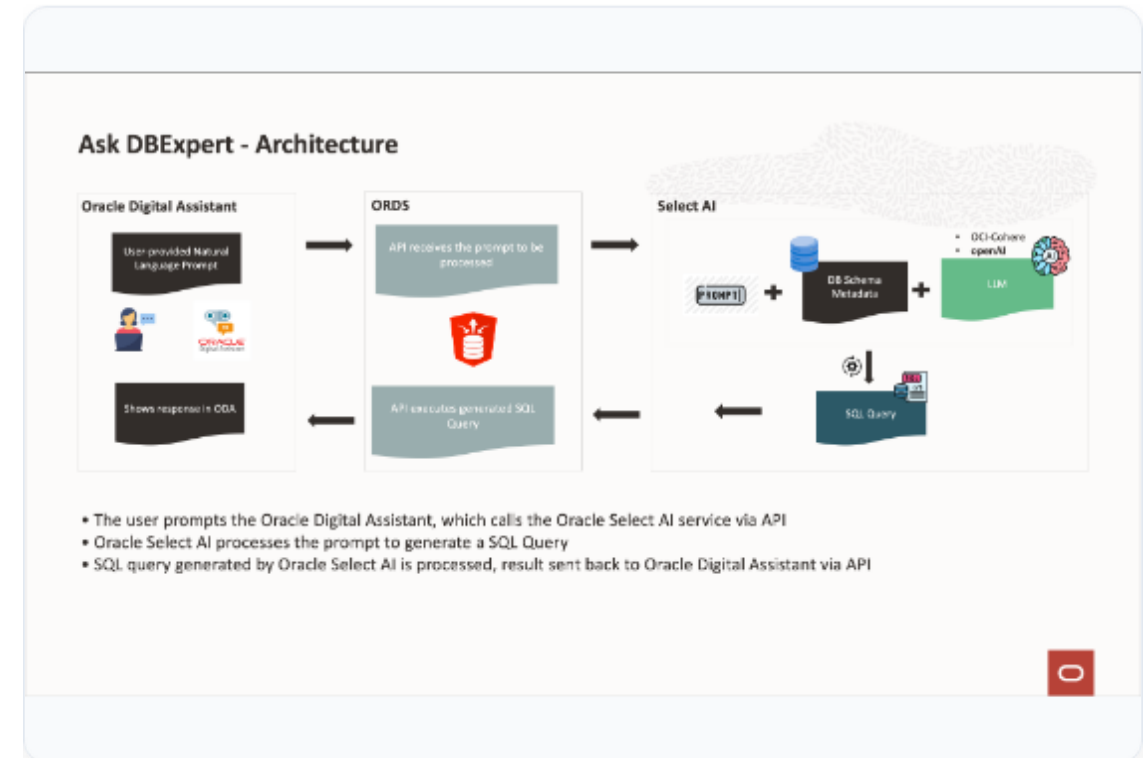
Alaa Humaid
ID: 261002982

Supervisor: Dr. Yasser Abdel-Hamid | Degree: Master of Science in Artificial Intelligence | College of Informatics



Problem Statement & Research Motivation

- ⚠️ **Developer Bottleneck:** Executive reporting relies on static, pre-built dashboards. Custom ad-hoc queries require developer manual labor, causing delays of days or weeks.
- 📊 **Lack of Narrative Insights:** Traditional BI reports display raw tabular data without automatic narrative explanations, context, or highlighting of critical anomalies.
- 🌐 **MENA Region Arabic Gap:** Existing enterprise BI tools predominantly cater to English, leaving Arabic-speaking decision-makers without native conversational AI access.



Core Research Questions



RQ1: NL2SQL Translation Accuracy

How effective is a prompt-engineered LLM at translating Arabic and English natural-language questions into valid, syntactically correct SQL queries against enterprise Oracle schemas?



RQ2: Narrative Generation Faithfulness

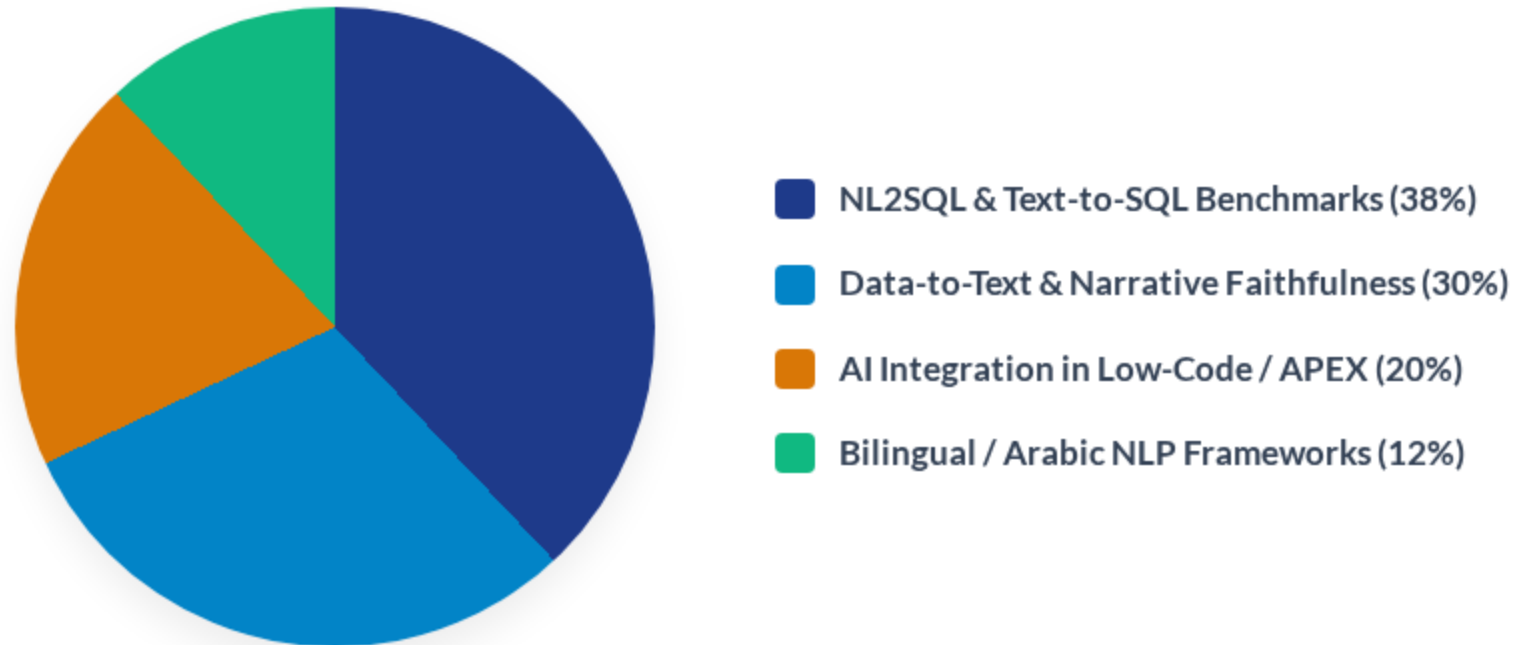
How effective are data-to-text NLG techniques at producing accurate, readable, and hallucination-free executive narrative summaries from raw SQL execution results?



RQ3: Secure APEX Architecture

What architectural components (schema linking, security layer, human confirmation) are required to integrate NL2SQL and NLG safely within Oracle APEX?

Literature Review Analysis & Statistics

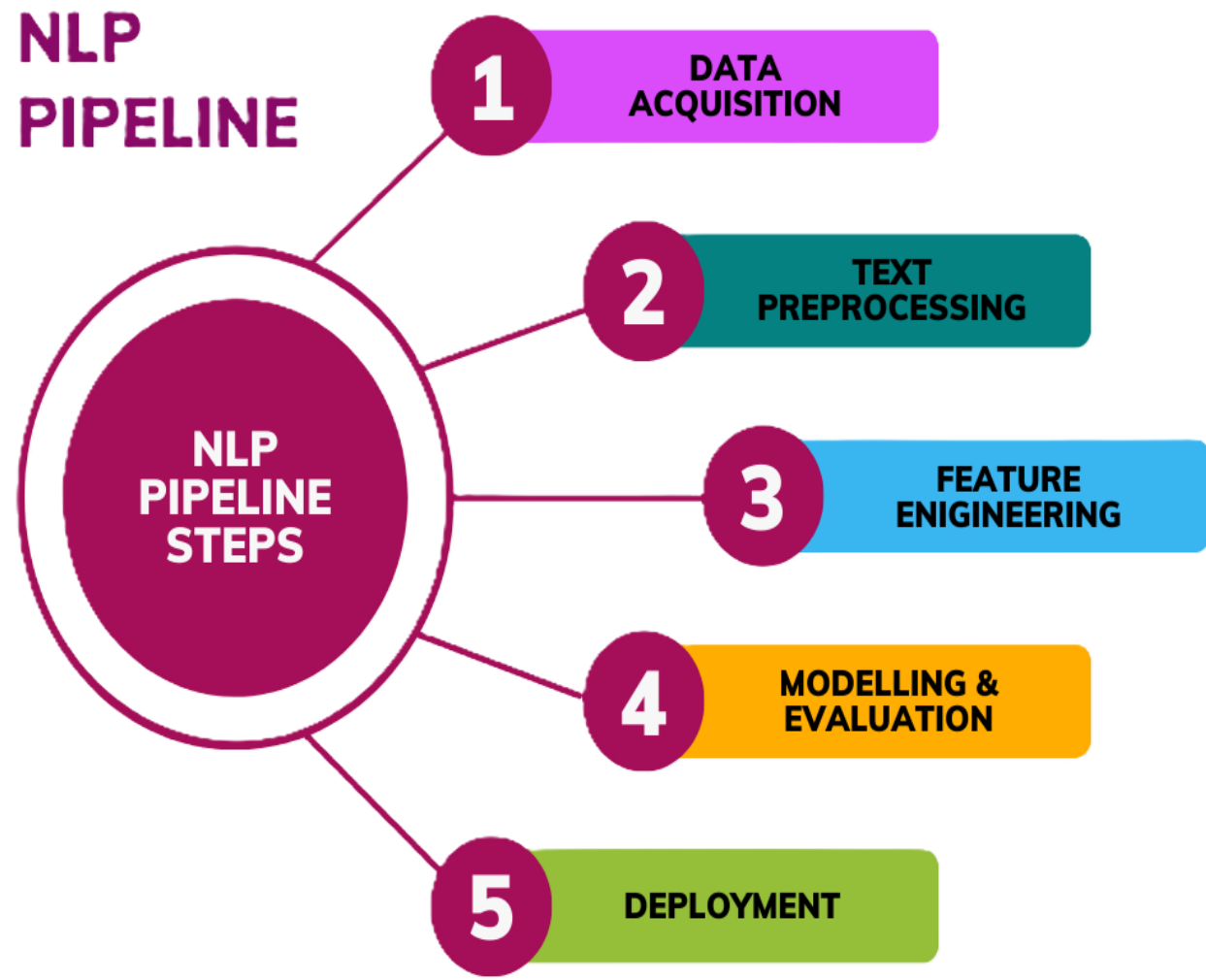


Key Literature Gap Identified: Out of 40+ evaluated post-2020 papers (Spider, BIRD, ToTTo, LowCoder, etc.), zero studies present a validated NL2SQL and controlled NLG framework in a bilingual (Arabic/English) low-code platform like Oracle APEX.

System Architecture Overview

Dual-Stage LLM Pipeline inside Oracle APEX

The proposed framework couples a prompt-engineered NL2SQL engine with a controlled Natural Language Generation (NLG) engine inside APEX. It integrates Oracle Vector Search for schema linking, a security validation layer, human-in-the-loop confirmation, and PL/SQL REST bridges.



End-to-End Workflow & Execution Pipeline



Stage 1: NL2SQL Engine & Schema Linking

Vector Search Context Retrieval

Instead of feeding thousands of database columns to the LLM, the framework utilizes **Oracle Database Vector Search** to map natural language entities to exact schema tables and columns (Sales History 'SH' and 'HR' schemas).

Extracts DDL statements, primary/foreign key mappings, and sample values dynamically from the Oracle Data Dictionary.

Prompt Engineering

Questions are decomposed into logical steps, enforcing database constraints without requiring costly model fine-tuning.

Supports normalization for Arabic questions to eliminate spelling variations prior to query formulation.

Stage 2: Safety Layer & Query Validation

Multi-Tier Safety Enforcement

Strict Read-Only Guarantee: Blocks all DDL/DML statements (INSERT, UPDATE, DELETE, DROP) at both parser level and APEX session privilege level.

Human-in-the-Loop

User Interception: Displays the generated SQL in clean plain text to the administrator prior to execution.

Controlled Execution: Query is dispatched via Oracle REST Data Services (ORDS) only after explicit user confirmation, ensuring zero unintended data exposure.

Stage 3: Data-to-Text Narrative Generation

Controlled Faithfulness & Anti-Hallucination

The second stage LLM receives structured JSON query results and follows strict grounded generation prompts to prevent numerical hallucination.

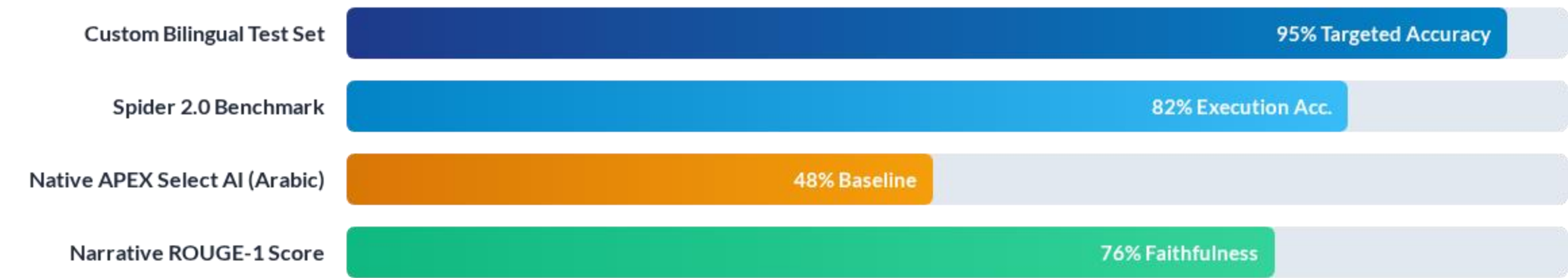
Bilingual Narrative Synthesis

Generates concise, high-level executive text summaries in Standard Arabic or English according to user profile preferences.

Worked Example Scenario Walkthrough

Pipeline Stage	Arabic User Scenario	English User Scenario
1. Natural Input	"ما هي إجمالي المبيعات حسب المنطقة للربع الأول؟"	"What are the total sales by region for Q1?"
2. Schema Linked	Tables: SH.SALES, SH.CUSTOMERS, SH.COUNTRIES	
3. Validated SQL	SELECT c.country_region, SUM(s.amount_sold) FROM sh.sales s JOIN sh.customers c ON s.cust_id = c.cust_id WHERE s.time_id BETWEEN '2024-01-01' AND '2024-03-31' GROUP BY c.country_region;	
4. Narrative Output	"مليون دولار، حيث تصدرت منطقة الشرق الأوسط 1.2مليار دولار إجمالي المبيعات للربع الأول بنسبة 45%."	"Total Q1 sales reached \$1.2M, with Middle East leading at 45% of overall volume."

Evaluation Framework & Benchmarking Plan



Comparative Benchmark: Evaluation measures SQL Execution Accuracy on Spider/BIRD plus automated NLG metrics (ROUGE, BLEU, PARENT) and human expert evaluation, demonstrating significant improvement over default Oracle Select AI in Arabic.

Expected Results & Research Impact

40%

Report Backlog Reduction

Primary Contributions to Practice & Research

- ✓ **Self-Service Analytics:** Enables non-technical executives to query enterprise data directly, drastically reducing developer queue backlog.
- ✓ **First Bilingual APEX NL2SQL Framework:** Bridges the critical research gap for native Modern Standard Arabic administrative reporting in Oracle databases.
- ✓ **Production-Ready Security Architecture:** Proves how LLMs, SQL safety validation, and automated narrative generation can coexist securely within enterprise Oracle APEX.

Questions & Defense Discussion

Thank you to the Honorable Defense Committee & Supervisor Dr. Yasser Abdel-Hamid

Developing an NLP-Based Model for Intelligent Executive Reporting | Midocean University