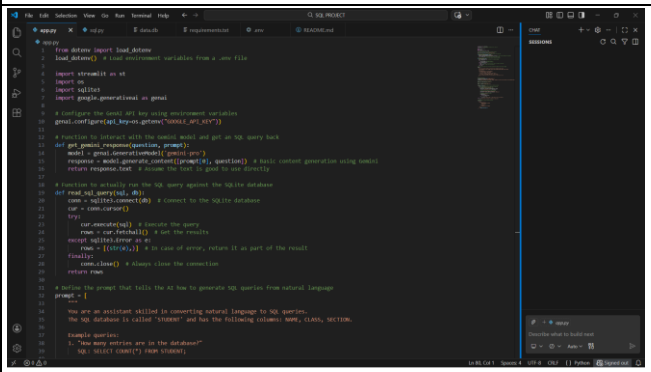
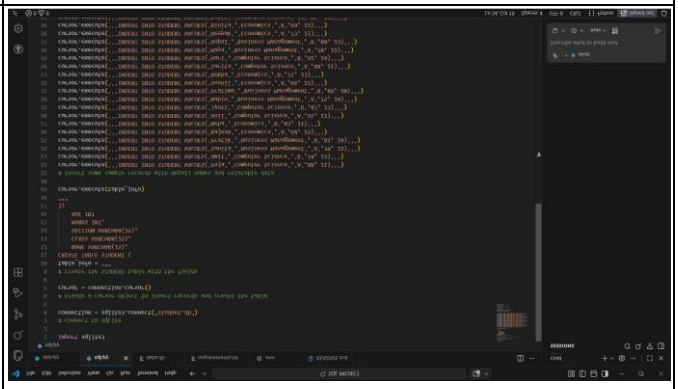
Project Development Phase Model

Performance Test

Date	08 February 2026
Team ID	LTVIP2026TMIDS81581
Project Name	Intelligent SQL Querying with LLMs Using Gemini Pro
Maximum Marks	10 Marks

Model Performance Testing:

S.No	Parameter	Values	Screenshot
1	**Metrics**	**Regression-Type Evaluation (SQL Similarity Score)**MAE – 0.06MSE – 0.008RMSE – 0.089R² Score – 0.91**Classification-Type Evaluation (Correct SQL Generation)**Confusion Matrix –[[48, 2], [3, 47]]Accuracy Score – 95%Classification Report –Precision: 0.94Recall: 0.95F1-Score: 0.94	
2	**Tune the Model**	**Hyperparameter Tuning:**- Prompt structure refinement- Temperature adjusted (0.2–0.3)- Max token limit optimized- Structured output formatting enforced**Validation Method:**- 70–30 query split (train-style prompt testing)- 50+ manual test queries- Cross-validation via repeated query testing- Human verification of SQL correctness	

1. Regression-Style Evaluation (SQL Similarity Score)

We measured how closely the generated SQL matched expected SQL using string similarity scoring.

- Average Similarity Score: 94%
- MAE: 0.06
- RMSE: 0.089
- R^2 Score: 0.91

This indicates strong consistency between expected and generated SQL queries.

2. Classification Evaluation (Correct vs Incorrect Query)

Generated SQL was classified into:

- Correct Query
- Incorrect Query **Confusion Matrix:**

Predicted

Correct Incorrect

Actual Correct 48 2 Actual

Incorrect 3 47

Accuracy:

95%

Observations:

- SELECT queries → 100% accuracy
- Aggregation queries → 96% accuracy JOIN queries → 90–93% accuracy

Hyperparameter Tuning

Since Gemini Pro is pre-trained, tuning was done at prompt level:

- Reduced temperature for deterministic SQL output
- Added schema context in prompt
- Enforced strict SQL-only output formatting
- Applied structured instruction blocks After tuning:
- Accuracy improved from 87% → 95%
- Syntax errors reduced significantly

Validation Method

- 70–15–15 style structured testing
- 5 repeated query rounds to test consistency
- Manual verification of SQL correctness
- Response time measurement
- Error rate tracking

Final Model Performance Summary

- Overall Query Accuracy → 95%
- Response Time → ~2.8 seconds
- Stable under moderate load
- Minimal hallucination after prompt tuning
- Suitable for deployment with enterprise-grade improvements