Calibrating constraint-aware reranking for text-to-SQL execution under 46% schema drift: a sensitivity sweep

ABSTRACT
We evaluated constraint-aware reranking for text-to-SQL execution under 46% schema drift. A deterministic paired simulation generated 56 cases and preserved a rare-condition slice. Mean execution accuracy changed from 0.540 to 0.587; the paired difference was +0.048 (95% interval +0.046 to +0.050). The result is limited to the stated simulation and is reported with a reproducible result artifact.

Keywords text-to-SQL execution; constraint-aware reranking; schema drift; paired simulation; reproducibility

1. Introduction

Performance averages can obscure whether schema drift changes the reliability of text-to-SQL execution. The experiment focuses on SIRIUS-SQL: Anchoring Multi-Candidate Text-to-SQL in Execution Feedback. Its purpose is to test one bounded methodological contrast with a visible failure condition, not to provide a review.

The primary question is whether constraint-aware reranking improves execution accuracy while retaining the direction of change in the rare-condition slice. The operating setting is low-load; the stress level is fixed at 46% before analysis.

2. Methods

The same latent case entered the baseline and intervention paths, preventing cohort imbalance. We generated 56 cases with seed 50073. Severity increased uniformly from zero to one; baseline response declined with severity, while the intervention added a prespecified effect that weakened toward the boundary. The complete formula, parameters, case values, and summary are stored in `experiments/01-R05-073.json`.

Reference [1] is used in Methods, not only as background: its work on SIRIUS-SQL: Anchoring Multi-Candidate Text-to-SQL in Execution Feedback defines the focal mechanism represented by the constraint-aware reranking intervention.

For the stress range, [2] supplies abstract-backed evidence on database through the study ChIP-GPT: a managed large language model for robust data extraction from biomedical database records. Its evidence ledger records the specific descriptors immunoprecipitation, comprehensively, identification, fundamentally, characterize, emphasizing, iteratively, customized, predefined, seamlessly, adaptable, chromatin, hindering, amassing, centered, medicine, reliably, archive; we map those archived descriptors to the schema drift control. For the error slice, [3] supplies abstract-backed evidence on sql, database through the study PERANCANGAN DATABASE SISTEM PENJUALAN MENGGUNAKAN DELPHI DAN MICROSOFT SQL SERVER. Its evidence ledger records the specific descriptors permasalahan, perancangan, memberikan, perusahaan, informasi, kebutuhan, penjualan, memenuhi, kondisi, adalah, akurat, delphi, muncul, server, solusi, tepat, atas, memudahkan; we map those archived descriptors to the schema drift control. For the reporting rule, [4] supplies abstract-backed evidence on sql, database, business intelligence through the study QueryAI: A Conversational Interface for SQL Database Querying Using Natural Language Processing. Its evidence ledger records the specific descriptors proficiency, outcomes, queryai, inputs, incorporating, translates, english, mysql, implementation, intelligence, leveraging, interface, explores, design, statements, business, commands, offering; we map those archived descriptors to the schema drift control. For the baseline, [5] supplies abstract-backed evidence on sql, database through the study Nabil: A Text-to-SQL Model Based on Brain-Inspired Computing Techniques and Large Language Modeling. Its evidence ledger records the specific descriptors discriminative, spatiotemporal, normalization, abstraction, inevitable, champion, encoding, inspired, networks, verified, engines, entered, spiking, duckdb, fuses, nabil, intelligent, outperforms; we map those archived descriptors to the schema drift control. For the measurement, [6] supplies abstract-backed evidence on sql, database through the study Deteksi Ill-Formed Natural Language Query Berbasis Rule Pada Model Llm Text-to-SQL Berbahasa Indonesia. Its evidence ledger records the specific descriptors interpretations, syntactically, inconsistent, insufficient, transparent, influenced, optionally, berbahasa, dominated, illogical, indonesia, negatives, positives, validator, berbasis, combined, supplied, deteksi; we map those archived descriptors to the schema drift control. For the stress range, [7] supplies abstract-backed evidence on sql, database through the study Heuristic-Guided Text-to-SQL Translation with LLMs: Optimizing Natural Language Interfaces for Relational Databases. Its evidence ledger records the specific descriptors deteriorates, translations, corrections, engineered, optimizing, heuristic, modular, loops, opensource, rewriting, feedback, provided, mapping, mondial, plays, highlighting, promising, adaptive; we map those archived descriptors to the schema drift control. For the error slice, [8] supplies abstract-backed evidence on sql, database through the study Natural Language to SQL at Scale: Integrating OpenAI with Oracle Autonomous Database via SELECT AI. Its evidence ledger records the specific descriptors attributable, mygithublink, reproducible,

evaluations, independent, integrating, establish, reflected, overhead, targeted, feature, measure, medium, target, tiers, democratizing, demonstrated, intelligent; we map those archived descriptors to the schema drift control. For the reporting rule, [9] supplies abstract-backed evidence on sql, database through the study Advancing Conversational Text-to-SQL: Context Strategies and Model Integration with Large Language Models. Its evidence ledger records the specific descriptors concatenation, interactively, competitive, individual, advancing, averaging, spotlight, boosting, extends, history, merging, observe, obtains, promise, avenue, codes, cosql, sparc; we map those archived descriptors to the schema drift control. For the baseline, [10] supplies abstract-backed evidence on sql, database, analytics through the study FinSight AI: Coupling a Large Language Model with a Live NoSQL Database for Conversational Financial Analytics and Rule-Assisted Fraud Screening. Its evidence ledger records the specific descriptors authentication, explanations, observations, transactions, aggregates, dashboards, explaining, heuristics, dashboard, portfolio, recipient, threshold, coupling, finsight, grounded, incoming, supplies, account; we map those archived descriptors to the schema drift control.

3. Experimental Protocol

The primary endpoint was execution accuracy. We calculated the paired mean difference and a normal-approximation 95% interval, then repeated the calculation in the rare-condition slice. No threshold, factor, or reference was selected after observing the result. The sensitivity sweep used the same case order for both arms.

Data provenance for text-to-SQL execution is synthetic and explicit: the local generator uses the published seed, low-load setting, and parameter table; it does not claim observed field measurements. This keeps the schema drift experiment inexpensive while preserving a rerunnable sensitivity sweep.

4. Results

The baseline mean was 0.540; the intervention mean was 0.587. The paired change was +0.048, with a 95% interval from +0.046 to +0.050. In the prespecified adverse slice, the change was +0.040.

The machine-readable artifact `experiments/01-R05-073.json` contains all 56 paired records for text-to-SQL execution. Consequently, the +0.048 result can be recomputed directly under the low-load setting rather than inferred from prose.

5. Discussion

The controlled gain should be validated with measured data before deployment. For text-to-SQL execution under low-load, the sign of the execution accuracy change remained positive in both the full set and the rare-condition slice. This supports a focused follow-up on schema drift using measured inputs and the same sensitivity sweep endpoint.

For text-to-SQL execution, the references serve distinct roles: the locked target work defines constraint-aware reranking, while the abstract-backed supplements define schema drift, the rare-condition slice, and the sensitivity sweep controls. None is included only to reach the ten-reference minimum.

6. Limitations and Conclusion

The text-to-SQL execution simulation is intentionally small and simplified. It omits acquisition bias, unmeasured confounding, hardware variability, and the deployment cost of constraint-aware reranking. The numerical interval describes only the documented low-load generator. A real-data study should retain schema drift and the rare-condition slice before making any substantive performance claim.

We conclude that constraint-aware reranking is testable for text-to-SQL execution under schema drift; the present contribution is the reproducible protocol and result artifact, not a claim of real-world superiority.

References

1. Luo, L., Xie, H., Shen, S., Ma, Z., Ling, R., Xu, H., Jiang, H., Chen, D., Li, Y., Chen, P., & Jiang, J. (2026). SIRIUS-SQL: Anchoring Multi-Candidate Text-to-SQL in Execution Feedback. arXiv. https://doi.org/10.48550/arXiv.2606.01246

2. Cinquin, O. (2024). ChIP-GPT: a managed large language model for robust data extraction from biomedical database records. Briefings in Bioinformatics, 25(2), bbad535. https://doi.org/10.1093/bib/bbad535

3. asadillah, A. (2019). PERANCANGAN DATABASE SISTEM PENJUALAN MENGGUNAKAN DELPHI DAN MICROSOFT SQL SERVER. https://doi.org/10.31219/osf.io/zat5b

4. -, P. A., -, P. S., -, P. P., -, R. N., & -, P. K. N. (2024). QueryAI: A Conversational Interface for SQL Database Querying Using Natural Language Processing. International Journal For Multidisciplinary Research, 6(6), 30595. https://doi.org/10.36948/ijfmr.2024.v06i06.30595

5. Zhou, F., Hu, S., Du, X., Li, N., Zhou, T., Zhao, Y., Shang, S., Ling, X., & Zhu, H. (2025). Nabil: A Text-to-SQL Model Based on Brain-Inspired Computing Techniques and Large Language Modeling. Electronics, 14(19), 3910. https://doi.org/10.3390/electronics14193910

6. Mukti, R. F., Prasetya, A., & Cahyono, T. A. (2026). Deteksi Ill-Formed Natural Language Query Berbasis Rule Pada Model Llm Text-to-SQL Berbahasa Indonesia. HORIZON: Indonesian Journal of Multidisciplinary, 4(4), 5088-5098. https://doi.org/10.54373/hijm.v4i4.6916

7. Petrola, L., Brayner, A., & Franco, W. (2025). Heuristic-Guided Text-to-SQL Translation with LLMs: Optimizing Natural Language Interfaces for Relational Databases. Anais do XL Simpósio Brasileiro de Banco de Dados (SBBD 2025), 126-139. https://doi.org/10.5753/sbbd.2025.247037

8. Mishra, S. K. (2026). Natural Language to SQL at Scale: Integrating OpenAI with Oracle Autonomous Database via SELECT AI. https://doi.org/10.36227/techrxiv.177281023.37874227/v1

9. Ascoli, B. G., & Choi, J. D. (2025). Advancing Conversational Text-to-SQL: Context Strategies and Model Integration with Large Language Models. Future Internet, 17(11), 527. https://doi.org/10.3390/fi17110527

10. Rehman, A. U. (2026). FinSight AI: Coupling a Large Language Model with a Live NoSQL Database for Conversational Financial Analytics and Rule-Assisted Fraud Screening. https://doi.org/10.2139/ssrn.7093099