Measuring constraint-aware reranking for text-to-SQL execution under 13% schema drift: a threshold audit

ABSTRACT
We evaluated constraint-aware reranking for text-to-SQL execution under 13% schema drift. A deterministic paired simulation generated 64 cases and preserved a low-signal stratum. Mean execution accuracy changed from 0.481 to 0.526; the paired difference was +0.045 (95% interval +0.042 to +0.047). 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

Method comparisons in text-to-SQL execution need an adverse slice as well as an overall average. The experiment focuses on SiriusDeliver: Automating Data Warehouse Delivery at Tencent. 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 low-signal stratum. The operating setting is resource-limited; the stress level is fixed at 13% before analysis.

2. Methods

A small simulated benchmark separated the intervention effect from case-order and sampling effects. We generated 64 cases with seed 50142. 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-R05-142.json`.

Reference [1] is used in Methods, not only as background: its work on SiriusDeliver: Automating Data Warehouse Delivery at Tencent defines the focal mechanism represented by the constraint-aware reranking intervention.

For the stress range, [2] supplies abstract-backed evidence on sql, database through the study Efficient schema-less text-to-SQL conversion using large language models. Its evidence ledger records the specific descriptors infrastructure, lessening, corpora, revolve, smaller, around, notion, paired, defog, doing, turbo, flan, less, much, size, outperforming, considerable, increasingly; 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 Large Language Model Method as a Translator Indonesian Into SQL Language. Its evidence ledger records the specific descriptors padangsidimpuan, administrative, automatically, intuitively, background, encouraged, supporting, translated, translator, lecturers, flexibly, intended, obstacle, becomes, certain, command, massive, quickly; we map those archived descriptors to the schema drift control. For the reporting rule, [4] supplies abstract-backed evidence on database through the study Large language model-based information extraction from free-text radiology reports: a scoping review protocol. Its evidence ledger records the specific descriptors dissemination, presentation, publications, radiological, standardized, informatics, publication, collection, conduction, diagnostic, guidelines, identified, predictive, according, appraised, contained, continues, dedicated; we map those archived descriptors to the schema drift control. For the baseline, [5] supplies abstract-backed evidence on sql through the study Confidence Estimation for Text-to-SQL in Large Language Models. Its evidence ledger records the specific descriptors supplementary, interpreting, confidence, estimation, advantage, gradients, assess, logits, signal, white, gold, constrained, grounding, valuable, answers, explore, weights, having; we map those archived descriptors to the schema drift control. For the measurement, [6] supplies abstract-backed evidence on sql, database through the study Pengembangan Model Migrasi Database Relational ke NoSQL Memanfaatkan Metadata SQL. Its evidence ledger records the specific descriptors dikembangkan, memanfaatkan, transformasi, berdasarkan, penyimpanan, terstruktur, diperlukan, diterapkan, eksperimen, mengajukan, pendekatan, perbendaan, dilakukan, kecepatan, kelemahan, melakuakn, menyimpan, merupakan; we map those archived descriptors to the schema drift control. For the stress range, [7] supplies abstract-backed evidence on sql through the study Pengembangan Aplikasi Chatbot dengan Large Language Model untuk Text-to-SQL Generation. Its evidence ledger records the specific descriptors pengeksekusian, menerjemahkan, mengembangkan, menunjukkan, menyediakan, pemeriksaan, pemrograman, acceptable, pengecekan, pertanyaan, antarmuka, pemilihan, penulisan, pertanian, usability, aplikasi, kegunaan, kriteria; we map those archived descriptors to the schema drift control. For the error slice, [8] supplies abstract-backed evidence on sql through the study SQL/PGQ data model and graph schema. Its evidence ledger records the specific descriptors describing, documents, property, neo4j, mapping, graph, three, schema; 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 Analyzing the Impact of Database Architecture on Performance: SQL vs NoSQL. Its evidence ledger records the specific descriptors representative, characterized, transactional, availability, exponential, intensified, persistence, universally, conversely, emphasizes, horizontal, analyzing, cassandra, integrity, analyzes, flexible, polyglot, depend; we map those archived descriptors to the schema drift control. For the baseline, [10] 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.

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 low-signal stratum. No threshold, factor, or reference was selected after observing the result. The threshold audit 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, resource-limited setting, and parameter table; it does not claim observed field measurements. This keeps the schema drift experiment inexpensive while preserving a rerunnable threshold audit.

4. Results

The baseline mean was 0.481; the intervention mean was 0.526. The paired change was +0.045, with a 95% interval from +0.042 to +0.047. In the prespecified adverse slice, the change was +0.038.

The machine-readable artifact `experiments/01-R05-142.json` contains all 64 paired records for text-to-SQL execution. Consequently, the +0.045 result can be recomputed directly under the resource-limited setting rather than inferred from prose.

5. Discussion

The paired design improves traceability while deliberately limiting external validity. For text-to-SQL execution under resource-limited, the sign of the execution accuracy change remained positive in both the full set and the low-signal stratum. This supports a focused follow-up on schema drift using measured inputs and the same threshold audit 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 low-signal stratum, and the threshold audit 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 resource-limited generator. A real-data study should retain schema drift and the low-signal stratum 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. Xie, H., Zhou, X., Yang, J., Shen, S., Wang, Z., Zheng, Y., Xu, T., Shi, Y., Zong, Z., Li, Y., Chen, P., Jiang, J., He, D., Yan, X., & Jiang, J. (2026). SiriusDeliver: Automating Data Warehouse Delivery at Tencent. arXiv. https://doi.org/10.48550/arXiv.2608.09185

2. Mellah, Y., Kocaman, V., Ul Haq, H., & Talby, D. (2024). Efficient schema-less text-to-SQL conversion using large language models. Artificial Intelligence in Health, 1(2), 96. https://doi.org/10.36922/aih.2661

3. Putra, C., Arlis, S., & Nurcahyo, G. W. (2025). Large Language Model Method as a Translator Indonesian Into SQL Language. Jurnal KomtekInfo, 124-130. https://doi.org/10.35134/komtekinfo.v12i3.658

4. Reichenpfader, D., Müller, H., & Denecke, K. (2023). Large language model-based information extraction from free-text radiology reports: a scoping review protocol. https://doi.org/10.1101/2023.07.28.23292031

5. Maleki, S. E., Pourreza, M., & Rafiei, D. (2026). Confidence Estimation for Text-to-SQL in Large Language Models. Proceedings of the AAAI Conference on Artificial Intelligence, 40(38), 32474-32482. https://doi.org/10.1609/aaai.v40i38.40523

6. Utomo, M. N. Y. (2020). Pengembangan Model Migrasi Database Relational ke NoSQL Memanfaatkan Metadata SQL. Jurnal Teknologi Elekterika, 4(2), 1. https://doi.org/10.31963/elekterika.v4i2.2212

7. Nugraha, G. P., Suadaa, L. H., Wilantika, N., & Maghfiroh, L. R. (2024). Pengembangan Aplikasi Chatbot dengan Large Language Model untuk Text-to-SQL Generation. Seminar Nasional Official Statistics, 2024(1), 831-840. https://doi.org/10.34123/semnasoffstat.v2024i1.2252

8. Furniss, P., & Green, A. (2023). SQL/PGQ data model and graph schema. https://doi.org/10.54285/ldbc.qzsk3559

9. Jawale, D., Shelke, S., & Yadav, S. (2026). Analyzing the Impact of Database Architecture on Performance: SQL vs NoSQL. International Journal of Mathematics And Computer Research, 14(03). https://doi.org/10.47191/ijmcr/v14ispc3.13

10. -, 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