Measuring constraint-aware reranking for text-to-SQL execution under 20% schema drift: a sensitivity sweep

ABSTRACT
We evaluated constraint-aware reranking for text-to-SQL execution under 20% schema drift. A deterministic paired simulation generated 64 cases and preserved a rare-condition slice. Mean execution accuracy changed from 0.554 to 0.608; the paired difference was +0.054 (95% interval +0.052 to +0.056). 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 SQLGovernor: An LLM-powered SQL Toolkit for Real World Application. 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 shifted-distribution; the stress level is fixed at 20% before analysis.

2. Methods

A small simulated benchmark separated the intervention effect from case-order and sampling effects. We generated 64 cases with seed 50014. 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-R05-014.json`.

Reference [1] is used in Methods, not only as background: its work on SQLGovernor: An LLM-powered SQL Toolkit for Real World Application 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 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 error slice, [3] 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 reporting rule, [4] supplies abstract-backed evidence on sql, database through the study Comparison between Database Search Algorithms (SQL and no SQL). Its evidence ledger records the specific descriptors differences, timization, underlying, emergence, document, includes, overview, thorough, weakness, careful, demands, stores, tabase, tional, flexi, newer, rdbms, value; we map those archived descriptors to the schema drift control. For the baseline, [5] supplies abstract-backed evidence on sql, database through the study Experimental Evaluation: Is NoSQL better than SQL Database?. Its evidence ledger records the specific descriptors experimentation, instantiating, proliferation, partitioning, behavioural, graphically, represented, apparently, functional, operation, tolerance, indexing, sharding, picture, ranking, stacked, tabular, giving; we map those archived descriptors to the schema drift control. For the measurement, [6] supplies abstract-backed evidence on sql, database through the study BENCHMARK DE MODELOS DE LINGUAGEM DE CÓDIGO ABERTO PARA TEXT-TO-SQL EM DADOS ONCOLÓGICOS BRASILEIROS. Its evidence ledger records the specific descriptors complemented, brasileiros, formulation, influential, superficial, linguistic, portuguese, suggesting, brazilian, latencies, linguagem, executed, oncology, presence, strongly, adopted, degrade, locally; 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 CRED-SQL: Enhancing Real-World Large Scale Database Text-to-SQL Parsing Through Cluster Retrieval and Execution Description. Its evidence ledger records the specific descriptors reformulation, alleviating, description, exacerbated, spiderunion, decomposes, ultimately, birdunion, deviation, mismatch, performs, pinpoint, cluster, smduan, stages, drift, cred, sota; 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 Large Language Model Based Chatbot for Database Interaction through Natural Language. Its evidence ledger records the specific descriptors completeness, complexities, naturalness, empowering, interprets, usefulness, interpret, barriers, cohesion, fluency,

number, today, back, democratizing, translates, prototype, relevance, informed; we map those archived descriptors to the schema drift control. For the reporting rule, [9] supplies abstract-backed evidence on sql, database, business intelligence through the study Natural Language to SQL (NL2SQL): A Comprehensive Study of Text-to-SQL Systems, Conversational AI for Databases, Enterprise Architectures, Challenges, and Future Directions. Its evidence ledger records the specific descriptors differentiator, pharmaceutical, orchestration, organizations, hallucinated, exemplified, illustrates, interrogate, responsibly, accumulate, components, embeddings, executives, glossaries, multimodal, prescriber, reflection, validators; we map those archived descriptors to the schema drift control. For the baseline, [10] supplies abstract-backed evidence on sql, database through the study Enhanced Financial Text-to-SQL Generation via Fine-Grained SQL Refinement. Its evidence ledger records the specific descriptors substantially, degradation, incorporate, refinement, alignment, diversity, necessity, agnostic, captures, dialects, implicit, reflects, dialect, grained, jointly, largely, adapt, logic; 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, shifted-distribution 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 0.554; the intervention mean was 0.608. The paired change was +0.054, with a 95% interval from +0.052 to +0.056. In the prespecified adverse slice, the change was +0.049.

The machine-readable artifact `experiments/01-R05-014.json` contains all 64 paired records for text-to-SQL execution. Consequently, the +0.054 result can be recomputed directly under the shifted-distribution setting rather than inferred from prose.

5. Discussion

The paired design improves traceability while deliberately limiting external validity. For text-to-SQL execution under shifted-distribution, 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 shifted-distribution 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. Jiang, J., Shen, S., Xie, H., Li, Y., Shen, Y., Huang, D., Qian, B., Wu, Y., Zhang, W., Cui, B., & Chen, P. (2025). SQLGovernor: An LLM-powered SQL Toolkit for Real World Application. arXiv. https://doi.org/10.48550/arXiv.2509.08575

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

3. 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

4. Amel Abdyssalam A Alhaag (2025). Comparison between Database Search Algorithms (SQL and no SQL). □□□□ □□□□□□ □□□□□□□, 9(□□□□ 36), 1810-1830. https://doi.org/10.65405/bc8atc11

5. Jain, K. (2024). Experimental Evaluation: Is NoSQL better than SQL Database?. https://doi.org/10.22541/au.172498858.84181818/v1

6. Mota, F. D. C., Silva, W. D. V. R. D., & Soares, J. D. N. (2026). BENCHMARK DE MODELOS DE LINGUAGEM DE CÓDIGO ABERTO PARA TEXT-TO-SQL EM DADOS ONCOLÓGICOS BRASILEIROS. REMUNOM, 13(14), 1-67. https://doi.org/10.66104/zvzdrv85

7. Duan, S., Wang, Z., Liu, C., Zhu, Z., Zhang, Y., Han, P., Yan, L., & Peng, Z. (2025). CRED-SQL: Enhancing Real-World Large Scale Database Text-to-SQL Parsing Through Cluster Retrieval and Execution Description. Frontiers in Artificial Intelligence and Applications. https://doi.org/10.3233/faia251337

8. Souto Prego, B., Vilares, D., & Cabado Lousa, B. (2024). Large Language Model Based Chatbot for Database Interaction through Natural Language. VII Congreso XoveTIC: impulsando el talento científico, 335-342. https://doi.org/10.17979/spudc.9788497498913.47

9. Shamal Chavan and Prof. Sandeep Vishwakarma (2026). Natural Language to SQL (NL2SQL): A Comprehensive Study of Text-to-SQL Systems, Conversational AI for Databases, Enterprise Architectures, Challenges, and Future Directions. International Journal of Advanced Research in Science Communication and Technology, 380. https://doi.org/10.48175/ijarsct-37344

10. Wang, Y., Chen, Y., Chen, R., Shu, H., Liu, P., & Xu, W. (2026). Enhanced Financial Text-to-SQL Generation via Fine-Grained SQL Refinement. https://doi.org/10.2139/ssrn.6502095