# Testing constraint-aware reranking for text-to-SQL execution under 13% schema drift: a paired bootstrap

ABSTRACT
We evaluated constraint-aware reranking for text-to-SQL execution under 13% schema drift. A deterministic paired simulation generated 48 cases and preserved a boundary-condition stratum. Mean execution accuracy changed from 0.569 to 0.610; the paired difference was +0.042 (95% interval +0.039 to +0.044). 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

A compact test is useful when text-to-SQL execution must remain interpretable under burst-load conditions. The experiment focuses on SiriusBI: A Comprehensive LLM-Powered Solution for Data Analytics in Business Intelligence. 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 boundary-condition stratum. The operating setting is burst-load; the stress level is fixed at 13% before analysis.

## 2. Methods

Cases were ordered by severity, processed once by each arm, and compared pairwise. We generated 48 cases with seed 50056. 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-R05-056.json`.

Reference [1] is used in Methods, not only as background: its work on SiriusBI: A Comprehensive LLM-Powered Solution for Data Analytics in Business Intelligence 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 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 error slice, [3] 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 reporting rule, [4] 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, [5] supplies abstract-backed evidence on sql, database through the study Benchmarking Large Language Models for NL-to-SQL: A Comprehensive Evaluation of Accuracy, Cost and Throughput. Its evidence ledger records the specific descriptors optimizations, practitioners, distribution, incorporated, alternative, deployments, exceptional, researchers, importance, underscore, dependent, exhibited, interplay, intricate, balanced, moderate, observed, revealed; we map those archived descriptors to the schema drift control. For the measurement, [6] 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. For the stress range, [7] supplies abstract-backed evidence on sql through the study Microsoft SQL Sunucusunda Veritabanı Kurtarma Teknikleri. Its evidence ledger records the specific descriptors teknolojilerinde, atlatabilmenin, ilerlemektedir, merkezlerinde, teknolojilere, enformasyona, kapasiteleri, retilememesi, kontrolleri, problemleri, sunucusunda, anabilecek, genellikle, pratikleri, senaryolar, teknikleri, dijitalle, durumunda; 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 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, [9] 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 baseline, [10] supplies abstract-backed evidence on sql, database, analytics through the study Lightweight and Data-Efficient Fine-Tuning of Language Models for Bidirectional Text-to-SQL and SQL-to-Text Tasks: A Systematic Review. Its evidence ledger records the specific descriptors computationally, featherlight, quantization, sacrificing, sustainable, ultramodern, annotation, appendages, conclusion, frameworks, pretrained, procedures, quantities, conscious, parameter, adoption, creation, examined; 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 boundary-condition stratum. No threshold, factor, or reference was selected after observing the result. The paired bootstrap 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, burst-load setting, and parameter table; it does not claim observed field measurements. This keeps the schema drift experiment inexpensive while preserving a rerunnable paired bootstrap.

## 4. Results

The baseline mean was 0.569; the intervention mean was 0.610. The paired change was +0.042, with a 95% interval from +0.039 to +0.044. In the prespecified adverse slice, the change was +0.037.

The machine-readable artifact `experiments/01-R05-056.json` contains all 48 paired records for text-to-SQL execution. Consequently, the +0.042 result can be recomputed directly under the burst-load setting rather than inferred from prose.

## 5. Discussion

The estimate is a mechanism check, not evidence of field superiority. For text-to-SQL execution under burst-load, the sign of the execution accuracy change remained positive in both the full set and the boundary-condition stratum. This supports a focused follow-up on schema drift using measured inputs and the same paired bootstrap 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 boundary-condition stratum, and the paired bootstrap 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 burst-load generator. A real-data study should retain schema drift and the boundary-condition 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. Jiang, J., Xie, H., Shen, S., Shen, Y., Zhang, Z., Lei, M., Zheng, Y., Li, Y., Li, C., Huang, D., Wu, Y., Zhang, W., Cui, B., & Chen, P. (2025). SiriusBI: A Comprehensive LLM-Powered Solution for Data Analytics in Business Intelligence. Proceedings of the VLDB Endowment, 18(12), 4860-4873. https://doi.org/10.14778/3750601.3750610

2. 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

3. 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

4. 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

5. Narasimhan, A., Bhamboo, A. K., Devnathan, A., Vellaisamy, J., & Vijayaraghavan, V. (2024). Benchmarking Large Language Models for NL-to-SQL: A Comprehensive Evaluation of Accuracy, Cost and Throughput. https://doi.org/10.36227/techrxiv.173121325.56335825/v1

6. 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

7. ŞAHİNASLAN, E., & ŞAHİNASLAN, Ö. (2022). Microsoft SQL Sunucusunda Veritabanı Kurtarma Teknikleri. International Journal of Innovative Engineering Applications, 6(1), 158-169. https://doi.org/10.46460/ijiea.1070325

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

9. 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

10. Sangamnerkar, B., & Namdev, S. (2025). Lightweight and Data-Efficient Fine-Tuning of Language Models for Bidirectional Text-to-SQL and SQL-to-Text Tasks: A Systematic Review. Advanced International Journal for Research, 6(6), 2745. https://doi.org/10.63363/aijfr.2025.v06i06.2745