

Measuring constraint-aware reranking for text-to-SQL execution under 21% schema drift: a error-stratified estimate

ABSTRACT

We evaluated constraint-aware reranking for text-to-SQL execution under 21% schema drift. A deterministic paired simulation generated 64 cases and preserved a median slice. Mean execution accuracy changed from 0.484 to 0.537; the paired difference was +0.053 (95% interval +0.050 to +0.055). 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 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 median slice. The operating setting is resource-limited; the stress level is fixed at 21% before analysis.

2. Methods

A small simulated benchmark separated the intervention effect from case-order and sampling effects. We generated 64 cases with seed 50094. 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-R05-094.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 sql, database through the study AN ENHANCED NATURAL LANGUAGE PROCESSING MODEL FOR DATA EXTRACTION AND VISUALIZATION FROM SQL DATABASE STRUCTURES: A SYSTEMATIC REVIEW OF LITERATURE. Its evidence ledger records the specific descriptors visualization, fundamentals, underpinning, foundations, proceedings, synthesizes, conference, identifies, prototypes, resolution, trajectory, scholarly, journals, motivate, reviewed, draws, flash, maps; 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 A comparative evaluation of large language models for detecting SQL injection vulnerabilities in web applications. Its evidence

ledger records the specific descriptors vulnerabilities, confidential, intractable, potentially, circumvent, considered, manipulate, popularity, codellama, detecting, prominent, technique, emerging, payloads, produced, stronger, boolean, created; 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 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, [5] 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 measurement, [6] 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 stress range, [7] supplies abstract-backed evidence on sql, database through the study MedT5SQL: a transformers-based large language model for text-to-SQL conversion in the healthcare domain. Its evidence ledger records the specific descriptors difficulties, encountering, transformers, approximate, accessible, delivering, electronic, optimizers, prevalence, expertise, assesses, commonly, medt5sql, transfer, utilizes, medical, numbers, related; 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 Enhancing Text-to-SQL Capabilities of Large Language Models via Domain Database Knowledge Injection. Its evidence ledger records the specific descriptors effectively, downstream, injected, subtask, seen, generalizability, incorporating, presented, contents, values, names, cell, face, make, demonstrating, hallucination, introduces, understand; 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 Hierarchical Adaptive Reward-based Reinforcement Learning Model for High-Precision Text-to-SQL Generation. Its evidence ledger records the specific descriptors reinforcement, relationships, dimensional, balancing, meanwhile, regulates, absolute, addition, lowering, barrier, concise, signals, policy, reward, spaces, termed, grpo, hart; we map those archived descriptors to the schema drift control. For the baseline, [10] supplies abstract-backed evidence on sql, database through the study Improving Text-to-Sql Conversion for Low-Resource Languages Using Large Language Models. Its evidence ledger records the specific descriptors configurations, augmentation, introducing, translating, contribute, emirozturk, previously, exceeding, languages, publicly, scripts, turkish, before, llama2, llama3, tt2sql, widely, total; 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 median slice. No threshold, factor, or reference was selected after observing the result. The error-stratified estimate 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 error-stratified estimate.

4. Results

The baseline mean was 0.484; the intervention mean was 0.537. The paired change was +0.053, with a 95% interval from +0.050 to +0.055. In the prespecified adverse slice, the change was +0.048.

The machine-readable artifact `experiments/01-R05-094.json` contains all 64 paired records for text-to-SQL execution. Consequently, the +0.053 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 median slice. This supports a focused follow-up on schema drift using measured inputs and the same error-stratified estimate 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 median slice, and the error-stratified estimate 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 median 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

- 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>
- C. Joel, U., & Dewole A. Temilola, J. (2025). AN ENHANCED NATURAL LANGUAGE PROCESSING MODEL FOR DATA EXTRACTION AND VISUALIZATION FROM SQL DATABASE STRUCTURES: A SYSTEMATIC REVIEW OF LITERATURE. *Jana Nexus: Journal of Computer Science*, 01(12), 26-31. <https://doi.org/10.21474/jncs01/111>
- Hairan, B., & Şahman, M. A. (2026). A comparative evaluation of large language models for detecting SQL injection vulnerabilities in web applications. *PeerJ Computer Science*, 12, e4015. <https://doi.org/10.7717/peerj-cs.4015>
- 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>
- 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>
- 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>
- Marshan, A., Almutairi, A. N., Ioannou, A., Bell, D., Monaghan, A., & Arzoky, M. (2024). MedT5SQL: a transformers-based large language model for text-to-SQL conversion in the healthcare domain. *Frontiers in Big Data*, 7, 1371680. <https://doi.org/10.3389/fdata.2024.1371680>
- Ma, X., Tian, X., Wu, L., Wang, X., Tang, X., & Wang, J. (2024). Enhancing Text-to-SQL Capabilities of Large Language Models via Domain Database Knowledge Injection. *Frontiers in Artificial Intelligence and Applications*. <https://doi.org/10.3233/faia240949>
- Liang, Z., Liu, L., Quan, R., Zou, M., Li, D., Tang, Y., & Qin, H. (2026). Hierarchical Adaptive Reward-based Reinforcement Learning Model for High-Precision Text-to-SQL Generation. <https://doi.org/10.2139/ssrn.6767042>
- Öztürk, E. (2025). Improving Text-to-Sql Conversion for Low-Resource Languages Using Large Language Models. *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*, 14(1), 163-178. <https://doi.org/10.17798/bitlisfen.1561298>