Probing constraint-aware reranking for text-to-SQL execution under 48% schema drift: a paired bootstrap

ABSTRACT
We evaluated constraint-aware reranking for text-to-SQL execution under 48% schema drift. A deterministic paired simulation generated 56 cases and preserved a median slice. Mean execution accuracy changed from 0.538 to 0.582; the paired difference was +0.044 (95% interval +0.041 to +0.046). 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 simple paired experiment provides a low-cost check of constraint-aware reranking under schema drift. 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 shifted-distribution; the stress level is fixed at 48% before analysis.

2. Methods

Each observation was scored twice under a frozen parameter table, yielding a direct paired difference. We generated 56 cases with seed 50061. 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-R05-061.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 through the study A Survey on Employing Large Language Models for Text-to-SQL Tasks. Its evidence ledger records the specific descriptors subcategory, finetuning, mainstream, employing, enumerate, taxonomy, text2sql, classic, discuss, survey, characteristics, extract, above, known, range, practical, studies, offer; 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 through the study CogSQL: A Cognitive Framework for Enhancing Large Language Models in Text-to-SQL Translation. Its evidence ledger records the specific descriptors transitional, composition, preparation, replicating, correction, prediction, cognitive, featuring, mirroring, processes, applying, concepts, consists, drafting, holistic, refining, aspects, believe; we map those archived descriptors to the schema drift control. For the baseline, [5] 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 measurement, [6] 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 stress range, [7] 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 error slice, [8] 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 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 through the study Enhancing security in text-to-SQL systems: A novel dataset and agent-based framework. Its evidence ledger records the specific descriptors sophisticated, advancements, compromising, unauthorized, destruction, enhancement, unfamiliar, malicious, resulting, reliance, exposes, relying, success, easier, severe, phase, safe, classification; 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 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, shifted-distribution 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 0.538; the intervention mean was 0.582. The paired change was +0.044, with a 95% interval from +0.041 to +0.046. In the prespecified adverse slice, the change was +0.037.

The machine-readable artifact `experiments/01-R05-061.json` contains all 56 paired records for text-to-SQL execution. Consequently, the +0.044 result can be recomputed directly under the shifted-distribution setting rather than inferred from prose.

5. Discussion

Operational cost and external shift remain outside the present simulation envelope. For text-to-SQL execution under shifted-distribution, 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 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 median slice, 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 shifted-distribution 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

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. Shi, L., Tang, Z., Zhang, N., Zhang, X., & Yang, Z. (2026). A Survey on Employing Large Language Models for Text-to-SQL Tasks. ACM Computing Surveys, 58(2), 1-37. https://doi.org/10.1145/3737873

3. 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

4. Yuan, H., Tang, X., Chen, K., Shou, L., Chen, G., & Li, H. (2025). CogSQL: A Cognitive Framework for Enhancing Large Language Models in Text-to-SQL Translation. Proceedings of the AAAI Conference on Artificial Intelligence, 39(24), 25778-25786. https://doi.org/10.1609/aaai.v39i24.34770

5. 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

6. 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

7. 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

8. 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

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. Chafik, S., Ezzini, S., & Berrada, I. (2025). Enhancing security in text-to-SQL systems: A novel dataset and agent-based framework. Natural Language Processing, 31(6), 1399-1422. https://doi.org/10.1017/nlp.2025.10008