Measuring constraint-aware reranking for text-to-SQL execution under 16% schema drift: a error-stratified estimate

ABSTRACT
We evaluated constraint-aware reranking for text-to-SQL execution under 16% schema drift. A deterministic paired simulation generated 64 cases and preserved a upper-severity quartile. Mean execution accuracy changed from 0.557 to 0.608; the paired difference was +0.051 (95% interval +0.049 to +0.054). 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 ZhuJiu: A Multi-dimensional, Multi-faceted Chinese Benchmark for Large Language Models. 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 upper-severity quartile. The operating setting is shifted-distribution; the stress level is fixed at 16% before analysis.

2. Methods

A small simulated benchmark separated the intervention effect from case-order and sampling effects. We generated 64 cases with seed 50038. 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-R05-038.json`.

Reference [1] is used in Methods, not only as background: its work on ZhuJiu: A Multi-dimensional, Multi-faceted Chinese Benchmark for Large Language Models 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, 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. For the reporting rule, [4] 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 baseline, [5] supplies abstract-backed evidence on sql, database through the study A COMPARATIVE STUDY OF LLM - POWERED DATABASE INTERFACES VERSUS TRADITIONAL SQL SYSTEMS FOR INVENTORY MANAGEMENT. Its evidence ledger records the specific descriptors deterministic, unprecedented, prioritizing, emonstrate, guaranteed, llmpowered, sqlalchemy, identical, inventory, penalties, averaged, backends, degraded, parallel, mission, slower, versus, incur; we map those archived descriptors to the schema drift control. For the measurement, [6] 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. For the stress range, [7] supplies abstract-backed evidence on sql, database, analytics through the study FinSight AI: Coupling a Large Language Model with a Live NoSQL Database for Conversational Financial Analytics and Rule-Assisted Fraud Screening. Its evidence ledger records the specific descriptors authentication, explanations, observations, transactions, aggregates, dashboards, explaining, heuristics, dashboard, portfolio, recipient, threshold, coupling, finsight, grounded, incoming, supplies, account; we map those archived descriptors to the schema drift control. For the error slice, [8] 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. For the reporting rule, [9] 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 baseline, [10] 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.

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 upper-severity quartile. 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, shifted-distribution 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 0.557; the intervention mean was 0.608. The paired change was +0.051, with a 95% interval from +0.049 to +0.054. In the prespecified adverse slice, the change was +0.044.

The machine-readable artifact `experiments/01-R05-038.json` contains all 64 paired records for text-to-SQL execution. Consequently, the +0.051 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 upper-severity quartile. 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 upper-severity quartile, 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 shifted-distribution generator. A real-data study should retain schema drift and the upper-severity quartile 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. Zhang, B., Xie, H., Du, P., Chen, J., Cao, P., Chen, Y., Liu, S., Liu, K., & Zhao, J. (2023). ZhuJiu: A Multi-dimensional, Multi-faceted Chinese Benchmark for Large Language Models. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 479-494. https://doi.org/10.18653/v1/2023.emnlp-demo.44

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. -, 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

4. 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

5. Guo, M., & Sun, Y. (2026). A COMPARATIVE STUDY OF LLM - POWERED DATABASE INTERFACES VERSUS TRADITIONAL SQL SYSTEMS FOR INVENTORY MANAGEMENT. NLP & Text Mining, 11-21. https://doi.org/10.5121/csit.2026.160602

6. 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

7. Rehman, A. U. (2026). FinSight AI: Coupling a Large Language Model with a Live NoSQL Database for Conversational Financial Analytics and Rule-Assisted Fraud Screening. https://doi.org/10.2139/ssrn.7093099

8. 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

9. 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

10. 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