Probing constraint-aware reranking for text-to-SQL execution under 43% schema drift: a paired bootstrap

ABSTRACT
We evaluated constraint-aware reranking for text-to-SQL execution under 43% schema drift. A deterministic paired simulation generated 56 cases and preserved a upper-severity quartile. Mean execution accuracy changed from 0.497 to 0.535; the paired difference was +0.038 (95% interval +0.036 to +0.040). 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 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 mixed-load; the stress level is fixed at 43% 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 50005. 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-R05-005.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 sql, database through the study ETL pipelines and SQL database management. Its evidence ledger records the specific descriptors indispensable, authenticate, consolidated, contemporary, continuously, centralized, elucidating, responsible, configured, immediate, processed, frontend, visitors, display, visitor, manner, serves, page; we map those archived descriptors to the schema drift control. For the error slice, [3] supplies abstract-backed evidence on sql through the study Confidence Estimation for Text-to-SQL in Large Language Models. Its evidence ledger records the specific descriptors supplementary, interpreting, confidence, estimation, advantage, gradients, assess, logits, signal, white, gold, constrained, grounding, valuable, answers, explore, weights, having; 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 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 baseline, [5] 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 measurement, [6] supplies abstract-backed evidence on sql, database through the study Deteksi Ill-Formed Natural Language Query Berbasis Rule Pada Model Llm Text-to-SQL Berbahasa Indonesia. Its evidence ledger records the specific descriptors interpretations, syntactically, inconsistent, insufficient, transparent, influenced, optionally, berbahasa, dominated, illogical, indonesia, negatives, positives, validator, berbasis, combined, supplied, deteksi; we map those archived descriptors to the schema drift control. For the stress range, [7] 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, [8] supplies abstract-backed evidence on sql through the study Text-to-SQL Empowered by Large Language Models: A Benchmark Evaluation. Its evidence ledger records the specific descriptors investigations, disadvantages, additionally, explorations, organization, systematical, leaderboard, supervised, designing, elaborate, empowered, refreshes, economic, inhibits, inspires, compare, example, towards; 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 GRASP-SQL: Graph Retrieval and Agentic Schema Pruning for Recall-First Text-to-SQL. Its evidence ledger records the specific descriptors discriminability, discriminatively, representational, preprocessing, statistically, concatenates, connectivity, conditioned, unnecessary, adaptively, ensembling, orthogonal, recruiting, ablations, embedding, expansion, generator, necessary; we map those archived descriptors to the schema drift control. For the baseline, [10] 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.

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 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, mixed-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 0.497; the intervention mean was 0.535. The paired change was +0.038, with a 95% interval from +0.036 to +0.040. In the prespecified adverse slice, the change was +0.031.

The machine-readable artifact `experiments/01-R05-005.json` contains all 56 paired records for text-to-SQL execution. Consequently, the +0.038 result can be recomputed directly under the mixed-load 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 mixed-load, 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 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 upper-severity quartile, 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 mixed-load 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. Muppala, M. (2025). ETL pipelines and SQL database management. SQL Database Mastery: Relational Architectures, Optimization Techniques,and Cloud-Based Applications, 84-101. https://doi.org/10.70593/978-93-7185-191-6_5

3. Maleki, S. E., Pourreza, M., & Rafiei, D. (2026). Confidence Estimation for Text-to-SQL in Large Language Models. Proceedings of the AAAI Conference on Artificial Intelligence, 40(38), 32474-32482. https://doi.org/10.1609/aaai.v40i38.40523

4. 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

5. 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

6. Mukti, R. F., Prasetya, A., & Cahyono, T. A. (2026). Deteksi Ill-Formed Natural Language Query Berbasis Rule Pada Model Llm Text-to-SQL Berbahasa Indonesia. HORIZON: Indonesian Journal of Multidisciplinary, 4(4), 5088-5098. https://doi.org/10.54373/hijm.v4i4.6916

7. 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

8. Gao, D., Wang, H., Li, Y., Sun, X., Qian, Y., Ding, B., & Zhou, J. (2024). Text-to-SQL Empowered by Large Language Models: A Benchmark Evaluation. Proceedings of the VLDB Endowment, 17(5), 1132-1145. https://doi.org/10.14778/3641204.3641221

9. Kim, H., Kim, W., & Kim, W. (2026). GRASP-SQL: Graph Retrieval and Agentic Schema Pruning for Recall-First Text-to-SQL. https://doi.org/10.2139/ssrn.7194451

10. 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