Probing constraint-aware reranking for text-to-SQL execution under 35% schema drift: a robust mean contrast

ABSTRACT
We evaluated constraint-aware reranking for text-to-SQL execution under 35% schema drift. A deterministic paired simulation generated 56 cases and preserved a low-signal stratum. Mean execution accuracy changed from 0.498 to 0.547; the paired difference was +0.049 (95% interval +0.047 to +0.051). 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 low-signal stratum. The operating setting is mixed-load; the stress level is fixed at 35% 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 50053. 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-R05-053.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 Text-to-SQL Conversion by using DeepLearning/Machine Learning: IntegratingNatural Language with Database Queries.. Its evidence ledger records the specific descriptors integratingnatural, databaseresponses, languagestructure, revolutionizedata, usersunfamiliar, involvingjoins, andinnovation, plainlanguage, professionals, deeplearning, democratizes, productivity, significance, userfriendly, underscores, benefiting, datadriven, industries; 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 FGCSQL: A Three-Stage Pipeline for Large Language Model-driven Chinese Text-to-SQL. Its evidence ledger records the specific descriptors breakthroughs, culminating, propagation, confronted, harnessing, indicating, irrelevant, mitigating, eliminate, outstrips, showcases, typically, uniquely, validity, cspider, decoder, lingual, fgcsql; 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 Nabil: A Text-to-SQL Model Based on Brain-Inspired Computing Techniques and Large Language Modeling. Its evidence ledger records the specific descriptors discriminative, spatiotemporal, normalization, abstraction, inevitable, champion, encoding, inspired, networks, verified, engines, entered, spiking, duckdb, fuses, nabil, intelligent, outperforms; 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 Lightweight and Explainable Conversational AI Framework for Natural Language SQL Learning without Large Language Models. Its evidence ledger records the specific descriptors computational, conversations, interpretable, progressively, explainable, facilitates, interactive, multilayer, beginners, developed, similarly, consumes, empowers, execute, labeled, operate, pattern, merges; we map those archived descriptors to the schema drift control. For the measurement, [6] supplies abstract-backed evidence on sql, database through the study Combining Text-to-SQL and Large Language Models for Maintenance Decision Support. Its evidence ledger records the specific descriptors deterministically, communication, inappropriate, establishing, transparency, contributes, suggestions, artificial, documented, historical, verifiable, vocabulary, personnel, replicate, traceable, ablation, bridging, combines; 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 CIR-SQL: A Dual-Model Intent Recognition Framework for Chinese Text-to-SQL. Its evidence ledger records the specific descriptors backtracking, interference, containing, decoupling, monitoring, operators, frequent, speaking, control, status, leads, seven, sort, representations, classification, hierarchical, intermediate, maintenance; 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 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 baseline, [10] 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.

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 low-signal stratum. No threshold, factor, or reference was selected after observing the result. The robust mean contrast 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 robust mean contrast.

4. Results

The baseline mean was 0.498; the intervention mean was 0.547. The paired change was +0.049, with a 95% interval from +0.047 to +0.051. In the prespecified adverse slice, the change was +0.043.

The machine-readable artifact `experiments/01-R05-053.json` contains all 56 paired records for text-to-SQL execution. Consequently, the +0.049 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 low-signal stratum. This supports a focused follow-up on schema drift using measured inputs and the same robust mean contrast 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 low-signal stratum, and the robust mean contrast 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 low-signal 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. 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. Amar Kaygude, Onkar Rajguru, Sandesh Karad, & G.T.Avhad (2025). Text-to-SQL Conversion by using DeepLearning/Machine Learning: IntegratingNatural Language with Database Queries. international journal of engineering technology and management sciences, 9(3), 48-52. https://doi.org/10.46647/ijetms.2025.v09i03.009

3. Jiang, G., Li, W., Yu, C., Zhu, Z., & LI, W. (2025). FGCSQL: A Three-Stage Pipeline for Large Language Model-driven Chinese Text-to-SQL. https://doi.org/10.20944/preprints202502.1059.v1

4. Zhou, F., Hu, S., Du, X., Li, N., Zhou, T., Zhao, Y., Shang, S., Ling, X., & Zhu, H. (2025). Nabil: A Text-to-SQL Model Based on Brain-Inspired Computing Techniques and Large Language Modeling. Electronics, 14(19), 3910. https://doi.org/10.3390/electronics14193910

5. Nayanakantha, B., Vidanage, K., Nirkhi, S., & Bhattacharyya, S. (2026). A Lightweight and Explainable Conversational AI Framework for Natural Language SQL Learning without Large Language Models. https://doi.org/10.21203/rs.3.rs-9823032/v1

6. Beckmann, S., Wiesner, K., Tebruegge, C., & Grum, M. (2026). Combining Text-to-SQL and Large Language Models for Maintenance Decision Support. https://doi.org/10.2139/ssrn.7103027

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. Wang, Y., Lv, H., & Qian, Y. (2026). CIR-SQL: A Dual-Model Intent Recognition Framework for Chinese Text-to-SQL. AI, 7(3), 91. https://doi.org/10.3390/ai7030091

9. 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

10. 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