Measuring constraint-aware reranking for text-to-SQL execution under 8% schema drift: a threshold audit

ABSTRACT
We evaluated constraint-aware reranking for text-to-SQL execution under 8% schema drift. A deterministic paired simulation generated 64 cases and preserved a boundary-condition stratum. Mean execution accuracy changed from 0.555 to 0.596; the paired difference was +0.041 (95% interval +0.039 to +0.044). 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 boundary-condition stratum. The operating setting is shifted-distribution; the stress level is fixed at 8% before analysis.

2. Methods

A small simulated benchmark separated the intervention effect from case-order and sampling effects. We generated 64 cases with seed 50086. 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-R05-086.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 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 error slice, [3] 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 reporting rule, [4] supplies abstract-backed evidence on sql, database through the study Pengembangan Model Migrasi Database Relational ke NoSQL Memanfaatkan Metadata SQL. Its evidence ledger records the specific descriptors dikembangkan, memanfaatkan, transformasi, berdasarkan, penyimpanan, terstruktur, diperlukan, diterapkan, eksperimen, mengajukan, pendekatan, perbendaan, dilakukan, kecepatan, kelemahan, melakuakn, menyimpan, merupakan; we map those archived descriptors to the schema drift control. For the baseline, [5] supplies abstract-backed evidence on sql, database through the study Comparison between Database Search Algorithms (SQL and no SQL). Its evidence ledger records the specific descriptors differences, timization, underlying, emergence, document, includes, overview, thorough, weakness, careful, demands, stores, tabase, tional, flexi, newer, rdbms, value; we map those archived descriptors to the schema drift control. For the measurement, [6] supplies abstract-backed evidence on sql through the study SQL/PGQ data model and graph schema. Its evidence ledger records the specific descriptors describing, documents, property, neo4j, mapping, graph, three, schema; 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 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 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 ETM: Modern Insights into Perspective on Text-to-SQL Evaluation in the Age of Large Language Models. Its evidence ledger records the specific descriptors stylistically, counterparts, misrepresent, overestimate, shortcomings, perspective, problematic, community, overlooks, elements, ignoring, inherent, negative, release, anyone, script, rates, rigid; we map those archived descriptors to the schema drift control. For the baseline, [10] 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.

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 boundary-condition stratum. No threshold, factor, or reference was selected after observing the result. The threshold audit 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 threshold audit.

4. Results

The baseline mean was 0.555; the intervention mean was 0.596. The paired change was +0.041, with a 95% interval from +0.039 to +0.044. In the prespecified adverse slice, the change was +0.035.

The machine-readable artifact `experiments/01-R05-086.json` contains all 64 paired records for text-to-SQL execution. Consequently, the +0.041 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 boundary-condition stratum. This supports a focused follow-up on schema drift using measured inputs and the same threshold audit 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 boundary-condition stratum, and the threshold audit 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 boundary-condition 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. 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

3. 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

4. Utomo, M. N. Y. (2020). Pengembangan Model Migrasi Database Relational ke NoSQL Memanfaatkan Metadata SQL. Jurnal Teknologi Elekterika, 4(2), 1. https://doi.org/10.31963/elekterika.v4i2.2212

5. Amel Abdyssalam A Alhaag (2025). Comparison between Database Search Algorithms (SQL and no SQL). □□□□ □□□□□□ □□□□□□□, 9(□□□□ 36), 1810-1830. https://doi.org/10.65405/bc8atc11

6. Furniss, P., & Green, A. (2023). SQL/PGQ data model and graph schema. https://doi.org/10.54285/ldbc.qzsk3559

7. 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

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. Ascoli, B. G., Kandikonda, Y. S. R., & Choi, J. D. (2025). ETM: Modern Insights into Perspective on Text-to-SQL Evaluation in the Age of Large Language Models. Future Internet, 17(8), 325. https://doi.org/10.3390/fi17080325

10. 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