Probing constraint-aware reranking for text-to-SQL execution under 39% schema drift: a threshold audit

ABSTRACT
We evaluated constraint-aware reranking for text-to-SQL execution under 39% schema drift. A deterministic paired simulation generated 56 cases and preserved a boundary-condition stratum. Mean execution accuracy changed from 0.502 to 0.556; the paired difference was +0.054 (95% interval +0.051 to +0.057). 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 boundary-condition stratum. The operating setting is mixed-load; the stress level is fixed at 39% 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 50029. 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-R05-029.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 BENCHMARK DE MODELOS DE LINGUAGEM DE CÓDIGO ABERTO PARA TEXT-TO-SQL EM DADOS ONCOLÓGICOS BRASILEIROS. Its evidence ledger records the specific descriptors complemented, brasileiros, formulation, influential, superficial, linguistic, portuguese, suggesting, brazilian, latencies, linguagem, executed, oncology, presence, strongly, adopted, degrade, locally; 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 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 reporting rule, [4] 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 baseline, [5] 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 measurement, [6] 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 stress range, [7] 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, [8] 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 reporting rule, [9] supplies abstract-backed evidence on sql, database through the study Natural Language to SQL at Scale: Integrating OpenAI with Oracle Autonomous Database via SELECT AI. Its evidence ledger records the specific descriptors attributable, mygithublink, reproducible, evaluations, independent, integrating, establish, reflected, overhead, targeted, feature, measure, medium, target, tiers, democratizing, demonstrated, intelligent; we map those archived descriptors to the schema drift control. For the baseline, [10] 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.

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, mixed-load 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 0.502; the intervention mean was 0.556. The paired change was +0.054, with a 95% interval from +0.051 to +0.057. In the prespecified adverse slice, the change was +0.045.

The machine-readable artifact `experiments/01-R05-029.json` contains all 56 paired records for text-to-SQL execution. Consequently, the +0.054 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 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 mixed-load 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. Mota, F. D. C., Silva, W. D. V. R. D., & Soares, J. D. N. (2026). BENCHMARK DE MODELOS DE LINGUAGEM DE CÓDIGO ABERTO PARA TEXT-TO-SQL EM DADOS ONCOLÓGICOS BRASILEIROS. REMUNOM, 13(14), 1-67. https://doi.org/10.66104/zvzdrv85

3. 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

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

5. 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

6. 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

7. 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

8. 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

9. Mishra, S. K. (2026). Natural Language to SQL at Scale: Integrating OpenAI with Oracle Autonomous Database via SELECT AI. https://doi.org/10.36227/techrxiv.177281023.37874227/v1

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