Stabilizing constraint-aware reranking for text-to-SQL execution under 24% schema drift: a blocked comparison

ABSTRACT
We evaluated constraint-aware reranking for text-to-SQL execution under 24% schema drift. A deterministic paired simulation generated 72 cases and preserved a boundary-condition stratum. Mean execution accuracy changed from 0.497 to 0.551; the paired difference was +0.054 (95% interval +0.052 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

The central concern is whether constraint-aware reranking remains measurable when schema drift increases. 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 resource-limited; the stress level is fixed at 24% before analysis.

2. Methods

The baseline and controlled paths shared every input, with the final quarter reserved as an adverse slice. We generated 72 cases with seed 50119. 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-R05-119.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 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 error slice, [3] supplies abstract-backed evidence on sql, database through the study Enhancing security in text-to-SQL systems: A novel dataset and agent-based framework. Its evidence ledger records the specific descriptors sophisticated, advancements, compromising, unauthorized, destruction, enhancement, unfamiliar, malicious, resulting, reliance, exposes, relying, success, easier, severe, phase, safe, classification; 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 Large Language Model Method as a Translator Indonesian Into SQL Language. Its evidence ledger records the specific descriptors padangsidimpuan, administrative, automatically, intuitively, background, encouraged, supporting, translated, translator, lecturers, flexibly, intended, obstacle, becomes, certain, command, massive, quickly; we map those archived descriptors to the schema drift control. For the baseline, [5] 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 measurement, [6] supplies abstract-backed evidence on sql through the study Microsoft SQL Sunucusunda Veritabanı Kurtarma Teknikleri. Its evidence ledger records the specific descriptors teknolojilerinde, atlatabilmenin, ilerlemektedir, merkezlerinde, teknolojilere, enformasyona, kapasiteleri, retilememesi, kontrolleri, problemleri, sunucusunda, anabilecek, genellikle, pratikleri, senaryolar, teknikleri, dijitalle, durumunda; 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 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 error slice, [8] supplies abstract-backed evidence on sql, database through the study Experimental Evaluation: Is NoSQL better than SQL Database?. Its evidence

ledger records the specific descriptors experimentation, instantiating, proliferation, partitioning, behavioural, graphically, represented, apparently, functional, operation, tolerance, indexing, sharding, picture, ranking, stacked, tabular, giving; 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 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, [10] 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.

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 blocked comparison 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, resource-limited setting, and parameter table; it does not claim observed field measurements. This keeps the schema drift experiment inexpensive while preserving a rerunnable blocked comparison.

4. Results

The baseline mean was 0.497; the intervention mean was 0.551. The paired change was +0.054, with a 95% interval from +0.052 to +0.057. In the prespecified adverse slice, the change was +0.046.

The machine-readable artifact `experiments/01-R05-119.json` contains all 72 paired records for text-to-SQL execution. Consequently, the +0.054 result can be recomputed directly under the resource-limited setting rather than inferred from prose.

5. Discussion

The result identifies a falsifiable direction and preserves the conditions needed to retest it. For text-to-SQL execution under resource-limited, 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 blocked comparison 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 blocked comparison 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 resource-limited 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. 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

3. Chafik, S., Ezzini, S., & Berrada, I. (2025). Enhancing security in text-to-SQL systems: A novel dataset and agent-based framework. Natural Language Processing, 31(6), 1399-1422. https://doi.org/10.1017/nlp.2025.10008

4. Putra, C., Arlis, S., & Nurcahyo, G. W. (2025). Large Language Model Method as a Translator Indonesian Into SQL Language. Jurnal KomtekInfo, 124-130. https://doi.org/10.35134/komtekinfo.v12i3.658

5. 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

6. ŞAHİNASLAN, E., & ŞAHİNASLAN, Ö. (2022). Microsoft SQL Sunucusunda Veritabanı Kurtarma Teknikleri. International Journal of Innovative Engineering Applications, 6(1), 158-169. https://doi.org/10.46460/ijiea.1070325

7. 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

8. Jain, K. (2024). Experimental Evaluation: Is NoSQL better than SQL Database?. https://doi.org/10.22541/au.172498858.84181818/v1

9. 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

10. 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