Probing constraint-aware reranking for text-to-SQL execution under 36% schema drift: a factor ablation

ABSTRACT
We evaluated constraint-aware reranking for text-to-SQL execution under 36% schema drift. A deterministic paired simulation generated 56 cases and preserved a rare-condition slice. Mean execution accuracy changed from 0.541 to 0.591; the paired difference was +0.050 (95% interval +0.047 to +0.052). 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 SQLGovernor: An LLM-powered SQL Toolkit for Real World Application. 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 rare-condition slice. The operating setting is shifted-distribution; the stress level is fixed at 36% 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 50133. 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-R05-133.json`.

Reference [1] is used in Methods, not only as background: its work on SQLGovernor: An LLM-powered SQL Toolkit for Real World Application 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 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 error slice, [3] 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 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 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 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 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 error slice, [8] 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 reporting rule, [9] supplies abstract-backed evidence on sql, database through the study Enhanced Financial Text-to-SQL Generation via Fine-Grained SQL Refinement. Its evidence ledger records the specific descriptors substantially, degradation, incorporate, refinement, alignment, diversity, necessity, agnostic, captures, dialects, implicit, reflects, dialect, grained, jointly, largely, adapt, logic; we map those archived descriptors to the schema drift control. For the baseline, [10] 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.

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 rare-condition slice. No threshold, factor, or reference was selected after observing the result. The factor ablation 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 factor ablation.

4. Results

The baseline mean was 0.541; the intervention mean was 0.591. The paired change was +0.050, with a 95% interval from +0.047 to +0.052. In the prespecified adverse slice, the change was +0.041.

The machine-readable artifact `experiments/01-R05-133.json` contains all 56 paired records for text-to-SQL execution. Consequently, the +0.050 result can be recomputed directly under the shifted-distribution 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 shifted-distribution, the sign of the execution accuracy change remained positive in both the full set and the rare-condition slice. This supports a focused follow-up on schema drift using measured inputs and the same factor ablation 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 rare-condition slice, and the factor ablation 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 rare-condition slice 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. Jiang, J., Shen, S., Xie, H., Li, Y., Shen, Y., Huang, D., Qian, B., Wu, Y., Zhang, W., Cui, B., & Chen, P. (2025). SQLGovernor: An LLM-powered SQL Toolkit for Real World Application. arXiv. https://doi.org/10.48550/arXiv.2509.08575

2. 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

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

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. 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

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. 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

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

9. Wang, Y., Chen, Y., Chen, R., Shu, H., Liu, P., & Xu, W. (2026). Enhanced Financial Text-to-SQL Generation via Fine-Grained SQL Refinement. https://doi.org/10.2139/ssrn.6502095

10. 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