Calibrating constraint-aware reranking for text-to-SQL execution under 41% schema drift: a sensitivity sweep

ABSTRACT
We evaluated constraint-aware reranking for text-to-SQL execution under 41% schema drift. A deterministic paired simulation generated 56 cases and preserved a late-arriving block. Mean execution accuracy changed from 0.499 to 0.544; the paired difference was +0.046 (95% interval +0.044 to +0.048). 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

Performance averages can obscure whether schema drift changes the reliability of text-to-SQL execution. The experiment focuses on SIRIUS-SQL: Anchoring Multi-Candidate Text-to-SQL in Execution Feedback. 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 late-arriving block. The operating setting is sparse-observation; the stress level is fixed at 41% before analysis.

2. Methods

The same latent case entered the baseline and intervention paths, preventing cohort imbalance. We generated 56 cases with seed 50017. 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-R05-017.json`.

Reference [1] is used in Methods, not only as background: its work on SIRIUS-SQL: Anchoring Multi-Candidate Text-to-SQL in Execution Feedback 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 MIGRATION OF LEGACY DATABASE ORACLE/MS SQL TO HANA DATABASE TO IMPROVE REAL-TIME ONLINE ANALYTICAL PROCESSING AND ONLINE TRANSACTION PROCESSING FROM ONE DATA MODEL. Its evidence ledger records the specific descriptors possibilities, transitioning, streamlining, assessment, migrating, migration, proposing, explored, keywords, vitality, migrate, legacy, online, paved, hana, olap, oltp, path; 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 Heuristic-Guided Text-to-SQL Translation with LLMs: Optimizing Natural Language Interfaces for Relational Databases. Its evidence ledger records the specific descriptors deteriorates, translations, corrections, engineered, optimizing, heuristic, modular, loops, opensource, rewriting, feedback, provided, mapping, mondial, plays, highlighting, promising, adaptive; 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 Improving Text-to-Sql Conversion for Low-Resource Languages Using Large Language Models. Its evidence ledger records the specific descriptors configurations, augmentation, introducing, translating, contribute, emirozturk, previously, exceeding, languages, publicly, scripts, turkish, before, llama2, llama3, tt2sql, widely, total; we map those archived descriptors to the schema drift control. For the baseline, [5] 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 measurement, [6] 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 stress range, [7] 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 error slice, [8] 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. For the reporting rule, [9] supplies abstract-backed evidence on database through the study Steering veridical large language model analyses by correcting and enriching generated database queries: first steps toward ChatGPT bioinformatics. Its evidence ledger records the specific descriptors bioinformatics, authoritative, manipulations, facilitating, detrimental, encountered, extensively, functioning, identifiers, mitigations, pretraining, redirecting, synthesized, gatekeeper, middleware, performing, scientific, unmodified; we map those archived descriptors to the schema drift control. For the baseline, [10] 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.

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 late-arriving block. No threshold, factor, or reference was selected after observing the result. The sensitivity sweep 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, sparse-observation setting, and parameter table; it does not claim observed field measurements. This keeps the schema drift experiment inexpensive while preserving a rerunnable sensitivity sweep.

4. Results

The baseline mean was 0.499; the intervention mean was 0.544. The paired change was +0.046, with a 95% interval from +0.044 to +0.048. In the prespecified adverse slice, the change was +0.039.

The machine-readable artifact `experiments/01-R05-017.json` contains all 56 paired records for text-to-SQL execution. Consequently, the +0.046 result can be recomputed directly under the sparse-observation setting rather than inferred from prose.

5. Discussion

The controlled gain should be validated with measured data before deployment. For text-to-SQL execution under sparse-observation, the sign of the execution accuracy change remained positive in both the full set and the late-arriving block. This supports a focused follow-up on schema drift using measured inputs and the same sensitivity sweep 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 late-arriving block, and the sensitivity sweep 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 sparse-observation generator. A real-data study should retain schema drift and the late-arriving block 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. Luo, L., Xie, H., Shen, S., Ma, Z., Ling, R., Xu, H., Jiang, H., Chen, D., Li, Y., Chen, P., & Jiang, J. (2026). SIRIUS-SQL: Anchoring Multi-Candidate Text-to-SQL in Execution Feedback. arXiv. https://doi.org/10.48550/arXiv.2606.01246

2. Rapolu, N. K. (2023). MIGRATION OF LEGACY DATABASE ORACLE/MS SQL TO HANA DATABASE TO IMPROVE REAL-TIME ONLINE ANALYTICAL PROCESSING AND ONLINE TRANSACTION PROCESSING FROM ONE DATA MODEL. International Scientific Journal of Engineering and Management, 02(03), 1-7. https://doi.org/10.55041/isjem00215

3. Petrola, L., Brayner, A., & Franco, W. (2025). Heuristic-Guided Text-to-SQL Translation with LLMs: Optimizing Natural Language Interfaces for Relational Databases. Anais do XL Simpósio Brasileiro de Banco de Dados (SBBD 2025), 126-139. https://doi.org/10.5753/sbbd.2025.247037

4. Öztürk, E. (2025). Improving Text-to-Sql Conversion for Low-Resource Languages Using Large Language Models. Bitlis Eren Üniversitesi Fen Bilimleri Dergisi, 14(1), 163-178. https://doi.org/10.17798/bitlisfen.1561298

5. 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

6. 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

7. 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

8. 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

9. Cinquin, O. (2024). Steering veridical large language model analyses by correcting and enriching generated database queries: first steps toward ChatGPT bioinformatics. Briefings in Bioinformatics, 26(1), bbaf045. https://doi.org/10.1093/bib/bbaf045

10. 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