Stress-testing constraint-aware reranking for text-to-SQL execution under 19% schema drift: a robust mean contrast

ABSTRACT
We evaluated constraint-aware reranking for text-to-SQL execution under 19% schema drift. A deterministic paired simulation generated 64 cases and preserved a late-arriving block. Mean execution accuracy changed from 0.479 to 0.516; the paired difference was +0.037 (95% interval +0.035 to +0.040). 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

This study isolates one operational question in text-to-SQL execution: whether a transparent control survives long-tail inputs. 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 long-tail; the stress level is fixed at 19% before analysis.

2. Methods

A fixed generator created matched inputs; only the prespecified intervention differed between arms. We generated 64 cases with seed 50106. 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-R05-106.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 PERANCANGAN DATABASE SISTEM PENJUALAN MENGGUNAKAN DELPHI DAN MICROSOFT SQL SERVER. Its evidence ledger records the specific descriptors permasalahan, perancangan, memberikan, perusahaan, informasi, kebutuhan, penjualan, memenuhi, kondisi, adalah, akurat, delphi, muncul, server, solusi, tepat, atas, memudahkan; 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 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 reporting rule, [4] supplies abstract-backed evidence on sql, database through the study Querying Databases with Natural Language: The use of Large Language Models for Text-to-SQL tasks. Its evidence ledger records the specific descriptors descriptions, dissertation, openly, poorly, views, evaluates, although, confirms, enhances, involves, mondial, simpler, joins, known, experiment, friendly, struggle, given; we map those archived descriptors to the schema drift control. For the baseline, [5] 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. 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, business intelligence through the study Natural Language to SQL (NL2SQL): A Comprehensive Study of Text-to-SQL Systems, Conversational AI for Databases, Enterprise Architectures, Challenges, and Future Directions. Its evidence ledger records the specific descriptors differentiator, pharmaceutical, orchestration, organizations, hallucinated, exemplified, illustrates, interrogate, responsibly, accumulate, components, embeddings, executives, glossaries, multimodal, prescriber, reflection, validators; 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 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 baseline, [10] supplies abstract-backed evidence on sql, database through the study CRED-SQL: Enhancing Real-World Large Scale Database Text-to-SQL Parsing Through Cluster Retrieval and Execution Description. Its evidence ledger records the specific descriptors reformulation, alleviating, description, exacerbated, spiderunion, decomposes, ultimately, birdunion, deviation, mismatch, performs, pinpoint, cluster, smduan, stages, drift, cred, sota; 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 robust mean contrast 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, long-tail setting, and parameter table; it does not claim observed field measurements. This keeps the schema drift experiment inexpensive while preserving a rerunnable robust mean contrast.

4. Results

The baseline mean was 0.479; the intervention mean was 0.516. The paired change was +0.037, with a 95% interval from +0.035 to +0.040. In the prespecified adverse slice, the change was +0.032.

The machine-readable artifact `experiments/01-R05-106.json` contains all 64 paired records for text-to-SQL execution. Consequently, the +0.037 result can be recomputed directly under the long-tail setting rather than inferred from prose.

5. Discussion

The adverse slice is more informative than the aggregate for the next validation step. For text-to-SQL execution under long-tail, 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 robust mean contrast 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 robust mean contrast 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 long-tail 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. asadillah, A. (2019). PERANCANGAN DATABASE SISTEM PENJUALAN MENGGUNAKAN DELPHI DAN MICROSOFT SQL SERVER. https://doi.org/10.31219/osf.io/zat5b

3. 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

4. Nascimento, E. R. S., & Casanova, M. A. (2024). Querying Databases with Natural Language: The use of Large Language Models for Text-to-SQL tasks. Anais Estendidos do XXXIX Simpósio Brasileiro de Banco de Dados (SBBD Estendido 2024), 196-201. https://doi.org/10.5753/sbbd_estendido.2024.240552

5. 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

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

7. Shamal Chavan and Prof. Sandeep Vishwakarma (2026). Natural Language to SQL (NL2SQL): A Comprehensive Study of Text-to-SQL Systems, Conversational AI for Databases, Enterprise Architectures, Challenges, and Future Directions. International Journal of Advanced Research in Science Communication and Technology, 380. https://doi.org/10.48175/ijarsct-37344

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. 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

10. Duan, S., Wang, Z., Liu, C., Zhu, Z., Zhang, Y., Han, P., Yan, L., & Peng, Z. (2025). CRED-SQL: Enhancing Real-World Large Scale Database Text-to-SQL Parsing Through Cluster Retrieval and Execution Description. Frontiers in Artificial Intelligence and Applications. https://doi.org/10.3233/faia251337