Stabilizing constraint-aware reranking for text-to-SQL execution under 36% schema drift: a hold-out check

ABSTRACT
We evaluated constraint-aware reranking for text-to-SQL execution under 36% schema drift. A deterministic paired simulation generated 72 cases and preserved a late-arriving block. Mean execution accuracy changed from 0.506 to 0.551; the paired difference was +0.046 (95% interval +0.043 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

The central concern is whether constraint-aware reranking remains measurable when schema drift increases. The experiment focuses on SiriusBI: A Comprehensive LLM-Powered Solution for Data Analytics in Business Intelligence. 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 resource-limited; the stress level is fixed at 36% 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 50047. 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-R05-047.json`.

Reference [1] is used in Methods, not only as background: its work on SiriusBI: A Comprehensive LLM-Powered Solution for Data Analytics in Business Intelligence 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 MedT5SQL: a transformers-based large language model for text-to-SQL conversion in the healthcare domain. Its evidence ledger records the specific descriptors difficulties, encountering, transformers, approximate, accessible, delivering, electronic, optimizers, prevalence, expertise, assesses, commonly, medt5sql, transfer, utilizes, medical, numbers, related; 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 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. 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, 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 measurement, [6] supplies abstract-backed evidence on sql, database through the study Large Language Model Based Chatbot for Database Interaction through Natural Language. Its evidence ledger records the specific descriptors completeness, complexities, naturalness, empowering, interprets, usefulness, interpret, barriers, cohesion, fluency, number, today, back, democratizing, translates, prototype, relevance, informed; we map those archived descriptors to the schema drift control. For the stress range, [7] 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 error slice, [8] 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, [9] supplies abstract-backed evidence on sql through the study CogSQL: A Cognitive Framework for Enhancing Large Language Models in Text-to-SQL Translation. Its evidence ledger records the specific descriptors transitional, composition, preparation, replicating, correction, prediction, cognitive, featuring, mirroring, processes, applying, concepts, consists, drafting, holistic, refining, aspects, believe; we map those archived descriptors to the schema drift control. For the baseline, [10] 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.

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 hold-out check 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 hold-out check.

4. Results

The baseline mean was 0.506; the intervention mean was 0.551. The paired change was +0.046, with a 95% interval from +0.043 to +0.048. In the prespecified adverse slice, the change was +0.038.

The machine-readable artifact `experiments/01-R05-047.json` contains all 72 paired records for text-to-SQL execution. Consequently, the +0.046 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 late-arriving block. This supports a focused follow-up on schema drift using measured inputs and the same hold-out check 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 hold-out check 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 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. Jiang, J., Xie, H., Shen, S., Shen, Y., Zhang, Z., Lei, M., Zheng, Y., Li, Y., Li, C., Huang, D., Wu, Y., Zhang, W., Cui, B., & Chen, P. (2025). SiriusBI: A Comprehensive LLM-Powered Solution for Data Analytics in Business Intelligence. Proceedings of the VLDB Endowment, 18(12), 4860-4873. https://doi.org/10.14778/3750601.3750610

2. Marshan, A., Almutairi, A. N., Ioannou, A., Bell, D., Monaghan, A., & Arzoky, M. (2024). MedT5SQL: a transformers-based large language model for text-to-SQL conversion in the healthcare domain. Frontiers in Big Data, 7, 1371680. https://doi.org/10.3389/fdata.2024.1371680

3. 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

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. 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

6. Souto Prego, B., Vilares, D., & Cabado Lousa, B. (2024). Large Language Model Based Chatbot for Database Interaction through Natural Language. VII Congreso XoveTIC: impulsando el talento científico, 335-342. https://doi.org/10.17979/spudc.9788497498913.47

7. 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

8. 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

9. Yuan, H., Tang, X., Chen, K., Shou, L., Chen, G., & Li, H. (2025). CogSQL: A Cognitive Framework for Enhancing Large Language Models in Text-to-SQL Translation. Proceedings of the AAAI Conference on Artificial Intelligence, 39(24), 25778-25786. https://doi.org/10.1609/aaai.v39i24.34770

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