

Calibrating constraint-aware reranking for text-to-SQL execution under 15% schema drift: a paired bootstrap

ABSTRACT

We evaluated constraint-aware reranking for text-to-SQL execution under 15% schema drift. A deterministic paired simulation generated 56 cases and preserved a median slice. Mean execution accuracy changed from 0.534 to 0.574; the paired difference was +0.040 (95% interval +0.038 to +0.042). 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 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 median slice. The operating setting is low-load; the stress level is fixed at 15% before analysis.

2. Methods

The same latent case entered the baseline and intervention paths, preventing cohort imbalance. We generated 56 cases with seed 50001. 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-R05-001.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 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 error slice, [3] supplies abstract-backed evidence on sql, database through the study AN ENHANCED NATURAL LANGUAGE PROCESSING MODEL FOR DATA EXTRACTION AND VISUALIZATION FROM SQL DATABASE STRUCTURES: A SYSTEMATIC REVIEW OF LITERATURE. Its evidence ledger records the specific descriptors visualization, fundamentals, underpinning, foundations, proceedings, synthesizes, conference, identifies, prototypes,

resolution, trajectory, scholarly, journals, motivate, reviewed, draws, flash, maps; 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 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 baseline, [5] 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, perbandingan, dilakukan, kecepatan, kelemahan, melakuakn, menyimpan, merupakan; we map those archived descriptors to the schema drift control. For the measurement, [6] 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 stress range, [7] supplies abstract-backed evidence on sql, database, analytics through the study FinSight AI: Coupling a Large Language Model with a Live NoSQL Database for Conversational Financial Analytics and Rule-Assisted Fraud Screening. Its evidence ledger records the specific descriptors authentication, explanations, observations, transactions, aggregates, dashboards, explaining, heuristics, dashboard, portfolio, recipient, threshold, coupling, finsight, grounded, incoming, supplies, account; 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 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 reporting rule, [9] 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, [10] 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 padangsimpulan, 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.

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 median slice. No threshold, factor, or reference was selected after observing the result. The paired bootstrap 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, low-load setting, and parameter table; it does not claim observed field measurements. This keeps the schema drift experiment inexpensive while preserving a rerunnable paired bootstrap.

4. Results

The baseline mean was 0.534; the intervention mean was 0.574. The paired change was +0.040, with a 95% interval from +0.038 to +0.042. In the prespecified adverse slice, the change was +0.034.

The machine-readable artifact `experiments/01-R05-001.json` contains all 56 paired records for text-to-SQL execution. Consequently, the +0.040 result can be recomputed directly under the low-load 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 low-load, the sign of the execution accuracy change remained positive in both the full set and the median slice. This supports a focused follow-up on schema drift using measured inputs and the same paired bootstrap 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 median slice, and the paired bootstrap 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 low-load generator. A real-data study should retain schema drift and the median 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

- 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>
- Furniss, P., & Green, A. (2023). SQL/PGQ data model and graph schema. <https://doi.org/10.54285/ldbc.qzsk3559>
- C. Joel, U., & Dewole A. Temilola, J. (2025). AN ENHANCED NATURAL LANGUAGE PROCESSING MODEL FOR DATA EXTRACTION AND VISUALIZATION FROM SQL DATABASE STRUCTURES: A SYSTEMATIC REVIEW OF LITERATURE. *Jana Nexus: Journal of Computer Science*, 01(12), 26-31. <https://doi.org/10.21474/jncs01/111>
- asadillah, A. (2019). PERANCANGAN DATABASE SISTEM PENJUALAN MENGGUNAKAN DELPHI DAN MICROSOFT SQL SERVER. <https://doi.org/10.31219/osf.io/zat5b>
- Utomo, M. N. Y. (2020). Pengembangan Model Migrasi Database Relational ke NoSQL Memanfaatkan Metadata SQL. *Jurnal Teknologi Elektroika*, 4(2), 1. <https://doi.org/10.31963/elektroika.v4i2.2212>
- 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>
- Rehman, A. U. (2026). FinSight AI: Coupling a Large Language Model with a Live NoSQL Database for Conversational Financial Analytics and Rule-Assisted Fraud Screening. <https://doi.org/10.2139/ssrn.7093099>
- 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>
- 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
- 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>