

# Measuring constraint-aware reranking for text-to-SQL execution under 33% schema drift: a factor ablation

## ABSTRACT

We evaluated constraint-aware reranking for text-to-SQL execution under 33% schema drift. A deterministic paired simulation generated 64 cases and preserved a low-signal stratum. Mean execution accuracy changed from 0.483 to 0.528; 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

Method comparisons in text-to-SQL execution need an adverse slice as well as an overall average. 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 low-signal stratum. The operating setting is resource-limited; the stress level is fixed at 33% before analysis.

## 2. Methods

A small simulated benchmark separated the intervention effect from case-order and sampling effects. We generated 64 cases with seed 50022. 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-R05-022.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 CIR-SQL: A Dual-Model Intent Recognition Framework for Chinese Text-to-SQL. Its evidence ledger records the specific descriptors backtracking, interference, containing, decoupling, monitoring, operators, frequent, speaking, control, status, leads, seven, sort, representations, classification, hierarchical, intermediate, maintenance; we map those archived descriptors to the schema drift control. For the error slice, [3] supplies abstract-backed evidence on sql through the study Microsoft SQL Sunucusunda Veritabanı Kurtarma Teknikleri. Its evidence ledger records the specific descriptors teknolojilerinde, atlatılabilmenin, 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 reporting rule, [4] 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, [5] supplies abstract-backed evidence on sql, database through the study Pengembangan Model Migrasi Database Relasional 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 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 stress range, [7] supplies abstract-backed evidence on sql, database through the study DB-GPT: Large Language Model Meets Database. Its evidence ledger records the specific descriptors tsiinghuadatabasegroup, demonstration, distributions, revolutionize, instructions, equivalence, preliminary, characters, relatively, automatic, ensuring, optimize, overcome, physical, privacy, rewrite, utilize, vision; 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 Enhancing Text-to-SQL Capabilities of Large Language Models via Domain Database Knowledge Injection. Its evidence ledger

records the specific descriptors effectively, downstream, injected, subtask, seen, generalizability, incorporating, presented, contents, values, names, cell, face, make, demonstrating, hallucination, introduces, understand; 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 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 baseline, [10] supplies abstract-backed evidence on sql, database through the study Efficient schema-less text-to-SQL conversion using large language models. Its evidence ledger records the specific descriptors infrastructure, lessening, corpora, revolve, smaller, around, notion, paired, defog, doing, turbo, flan, less, much, size, outperforming, considerable, increasingly; 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 low-signal stratum. 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, resource-limited 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 0.483; the intervention mean was 0.528. 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-R05-022.json` contains all 64 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 paired design improves traceability while deliberately limiting external validity. For text-to-SQL execution under resource-limited, the sign of the execution accuracy change remained positive in both the full set and the low-signal stratum. 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 low-signal stratum, 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 resource-limited generator. A real-data study should retain schema drift and the low-signal stratum 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>
- Wang, Y., Lv, H., & Qian, Y. (2026). CIR-SQL: A Dual-Model Intent Recognition Framework for Chinese Text-to-SQL. *AI*, 7(3), 91. <https://doi.org/10.3390/ai7030091>
- 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>
- 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), bbafo45. <https://doi.org/10.1093/bib/bbafo45>
- 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>
- 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>
- Zhou, X., Sun, Z., & Li, G. (2024). DB-GPT: Large Language Model Meets Database. *Data Science and Engineering*, 9(1), 102-111. <https://doi.org/10.1007/s41019-023-00235-6>
- Ma, X., Tian, X., Wu, L., Wang, X., Tang, X., & Wang, J. (2024). Enhancing Text-to-SQL Capabilities of Large Language Models via Domain Database Knowledge Injection. *Frontiers in Artificial Intelligence and Applications*. <https://doi.org/10.3233/faia240949>
- 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>
- Mellah, Y., Kocaman, V., Ul Haq, H., & Talby, D. (2024). Efficient schema-less text-to-SQL conversion using large language models. *Artificial Intelligence in Health*, 1(2), 96. <https://doi.org/10.36922/aih.2661>