Stress-testing constraint-aware reranking for text-to-SQL execution under 23% schema drift: a factor ablation

ABSTRACT
We evaluated constraint-aware reranking for text-to-SQL execution under 23% schema drift. A deterministic paired simulation generated 64 cases and preserved a low-signal stratum. Mean execution accuracy changed from 0.479 to 0.525; 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

This study isolates one operational question in text-to-SQL execution: whether a transparent control survives long-tail inputs. 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 long-tail; the stress level is fixed at 23% before analysis.

2. Methods

A fixed generator created matched inputs; only the prespecified intervention differed between arms. We generated 64 cases with seed 50082. 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-R05-082.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, business intelligence through the study QueryAI: A Conversational Interface for SQL Database Querying Using Natural Language Processing. Its evidence ledger records the specific descriptors proficiency, outcomes, queryai, inputs, incorporating, translates, english, mysql, implementation, intelligence, leveraging, interface, explores, design, statements, business, commands, offering; 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 DB-GPT: Large Language Model Meets Database. Its evidence ledger records the specific descriptors tsinghuadatabasegroup, 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 reporting rule, [4] supplies abstract-backed evidence on sql, database through the study ETM: Modern Insights into Perspective on Text-to-SQL Evaluation in the Age of Large Language Models. Its evidence ledger records the specific descriptors stylistically, counterparts, misrepresent, overestimate, shortcomings, perspective, problematic, community, overlooks, elements, ignoring, inherent, negative, release, anyone, script, rates, rigid; we map those archived descriptors to the schema drift control. For the baseline, [5] 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 measurement, [6] 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 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,

analytics through the study Lightweight and Data-Efficient Fine-Tuning of Language Models for Bidirectional Text-to-SQL and SQL-to-Text Tasks: A Systematic Review. Its evidence ledger records the specific descriptors computationally, featherlight, quantization, sacrificing, sustainable, ultramodern, annotation, appendages, conclusion, frameworks, pretrained, procedures, quantities, conscious, parameter, adoption, creation, examined; 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 Analyzing the Impact of Database Architecture on Performance: SQL vs NoSQL. Its evidence ledger records the specific descriptors representative, characterized, transactional, availability, exponential, intensified, persistence, universally, conversely, emphasizes, horizontal, analyzing, cassandra, integrity, analyzes, flexible, polyglot, depend; 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 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.

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, long-tail 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 0.479; the intervention mean was 0.525. The paired change was +0.046, with a 95% interval from +0.043 to +0.048. In the prespecified adverse slice, the change was +0.041.

The machine-readable artifact `experiments/01-R05-082.json` contains all 64 paired records for text-to-SQL execution. Consequently, the +0.046 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 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 long-tail 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

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. -, P. A., -, P. S., -, P. P., -, R. N., & -, P. K. N. (2024). QueryAI: A Conversational Interface for SQL Database Querying Using Natural Language Processing. International Journal For Multidisciplinary Research, 6(6), 30595. https://doi.org/10.36948/ijfmr.2024.v06i06.30595

3. 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

4. Ascoli, B. G., Kandikonda, Y. S. R., & Choi, J. D. (2025). ETM: Modern Insights into Perspective on Text-to-SQL Evaluation in the Age of Large Language Models. Future Internet, 17(8), 325. https://doi.org/10.3390/fi17080325

5. 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

6. 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

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. Sangamnerkar, B., & Namdev, S. (2025). Lightweight and Data-Efficient Fine-Tuning of Language Models for Bidirectional Text-to-SQL and SQL-to-Text Tasks: A Systematic Review. Advanced International Journal for Research, 6(6), 2745. https://doi.org/10.63363/aijfr.2025.v06i06.2745

9. Jawale, D., Shelke, S., & Yadav, S. (2026). Analyzing the Impact of Database Architecture on Performance: SQL vs NoSQL. International Journal of Mathematics And Computer Research, 14(03). https://doi.org/10.47191/ijmcr/v14ispc3.13

10. 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