# When Text-to-SQL Evaluation Misleads: Rethinking Benchmarking Practices

Oktie Hassanzadeh, Yotam Perlitz, Nhan Pham, Timothy Dinger, Tanvi Kaple, Long Vu, Michael Glass, and Dharmashankar Subramanian

## I. Background / Question / Methods

Text-to-SQL (Text2SQL) systems have made significant progress in recent years, and some results on public benchmarks have even led to the perception that the problem is close to being solved. However, observations from practical deployments and from new, more realistic benchmarks indicate that evaluation remains a major open challenge [1]. Existing metrics often fail to compare systems fairly, penalizing harmless stylistic differences such as generating more concise queries or selecting additional non-essential columns, and are sensitive to dataset artifacts and benchmark design limitations [2]. These issues introduce inconsistencies that obscure meaningful differences between methods, as also discussed in prior surveys [3], [4].

We present a modular open-source evaluation toolkit designed to provide more robust and reproducible assessment of Text2SQL pipelines [5]. The toolkit includes: (1) unified execution adapters across multiple SQL engines, (2) lightweight semantic-aware execution metrics to reduce false penalties arising from structural variations, (3) an LLM-as-judge component for intent-based correctness assessment, and (4) automated error-analysis utilities for identifying common sources of metric distortion. Using a number of public benchmarks and commercial datasets, we examine the following questions: To what extent do current metrics distort comparisons between methods? How sensitive are rankings to harmless variations? And how can we achieve more stable and interpretable evaluation?

## II. Results / Conclusions

Our analysis shows that evaluation artifacts frequently dominate the apparent performance differences between Text2SQL methods. In several datasets, systems with similar execution accuracy diverge by up to 40 percentage points under semantic or LLM-based evaluation, indicating that many execution "errors" arise from stylistic variation or equivalence-related phenomena rather than substantive failures. Conversely, methods that appear close under execution accuracy sometimes separate under semantic evaluation, revealing reasoning differences that strict execution metrics fail to capture.

We also document concrete instances where existing metrics penalize shorter SQL, optional column selection, or engine-dependent behaviors, leading to unstable or misleading system rankings. These findings underscore the need for evaluation approaches that explicitly account for semantic equivalence and structural variability.

**Take-home message:** Current Text2SQL evaluation practices often obscure, rather than reveal, the true differences between systems. More semantic, reproducible, and fine-grained evaluation provides significantly more reliable comparisons. Our open-source toolkit aims to facilitate such evaluation for both research and applied settings.

## References

[1] C. Renggli, I. F. Ilyas, and T. Rekatsinas, "Fundamental Challenges in Evaluating Text2SQL Solutions and Detecting Their Limitations," arXiv:2501.18197, 2025. https://arxiv.org/abs/2501.18197

[2] A. Mitsopoulou and G. Koutrika, "Analysis of Text-to-SQL Benchmarks: Limitations, Challenges and Opportunities," in *Proc. EDBT*, 2025. https://openproceedings.org/2025/conf/edbt/paper-41.pdf

[3] H. Kim, B.-H. So, W.-S. Han, and H. Lee, "Natural Language to SQL: Where Are We Today?" *Proc. VLDB Endowment*, vol. 13, no. 10, pp. 1737–1750, 2020. https://doi.org/10.14778/3401960.3401970

[4] N. Deng, Y. Chen, and Y. Zhang, "Recent Advances in Text-to-SQL: A Survey of What We Have and What We Expect," in *Proc. COLING*, 2022, pp. 2166–2187. https://aclanthology.org/2022.coling-1.190

[5] Text2SQL Evaluation Toolkit: https://github.com/IBM/text2sql-eval-toolkit

All authors are with IBM Research. Emails: hassanzadeh@us.ibm.com. Y.Perlitz@ibm.com, nhp@ibm.com, dinger@us.ibm.com, Tanvi.Kaple@ibm.com,

2375-026X/26/$31.00 ©2026 IEEE
DOI 10.1109/ICDE65706.2026.00364