BENCHPRESS: An Annotation System for Rapid Text-to-SQL Benchmark Curation

Fabian Wenz (MIT), Peter Baile Chen (MIT), Moe Kayali (UW), Nesime Tatbul (Intel Labs & MIT), Çağatay Demiralp (AWS AI Labs & MIT), Michael Stonebraker (MIT)

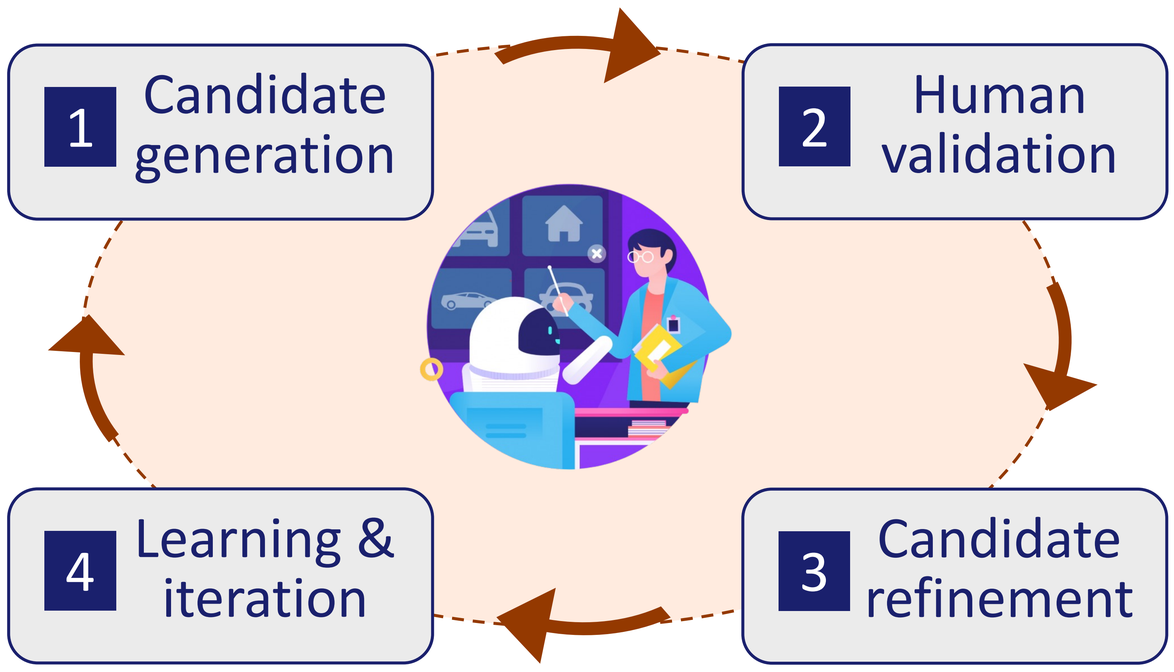
Do you think **PUBLIC BENCHMARKS** are **NOT REALISTIC**? Need a **QUICK & EASY WAY TO CREATE BENCHMARKS** based on **YOUR OWN DATA**?

The Need for New Benchmarks & Annotation at Scale

- Increasing demand for LLM-based Text-to-SQL capabilities (and others) across enterprises
- Benchmarks based on public data (e.g., Spider [2]) are insufficient to capture the needs of enterprise data workloads
- Creating new benchmarks calls for scalable data curation tooling support

BENCHPRESS: An Active Learning Approach via Bi-Directional Query Translation

- Combine input from LLMs and Human Experts to iteratively generate Text-to-SQL and SQL-to-Text translations

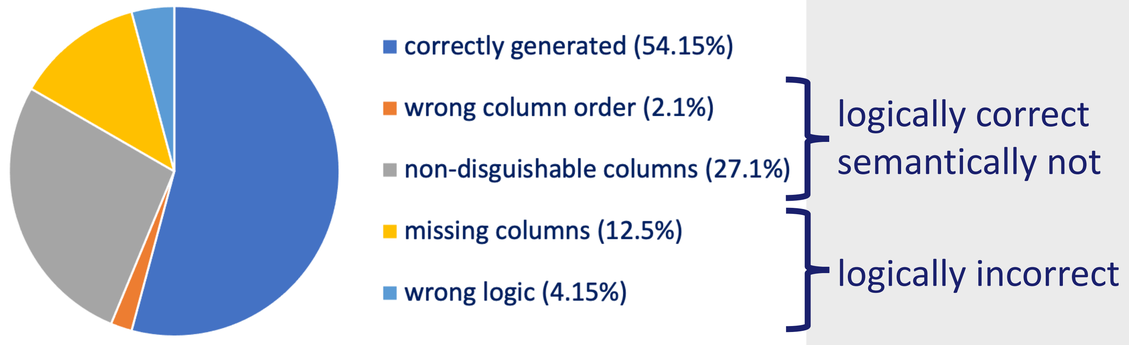


Semantic Context Generation

Leveraging contextual metadata enhances the accuracy of LLM-generated natural language (NL) descriptions

- *Annotation integration:* Annotators can refine auto-generated metadata for domain-specific precision
- *Knowledge base linking: S*emantic context can be enriched via knowledge base integration

SQL-to-Text Generation Accuracy for BEAVER Queries [1]



Optimization via Back-Translation

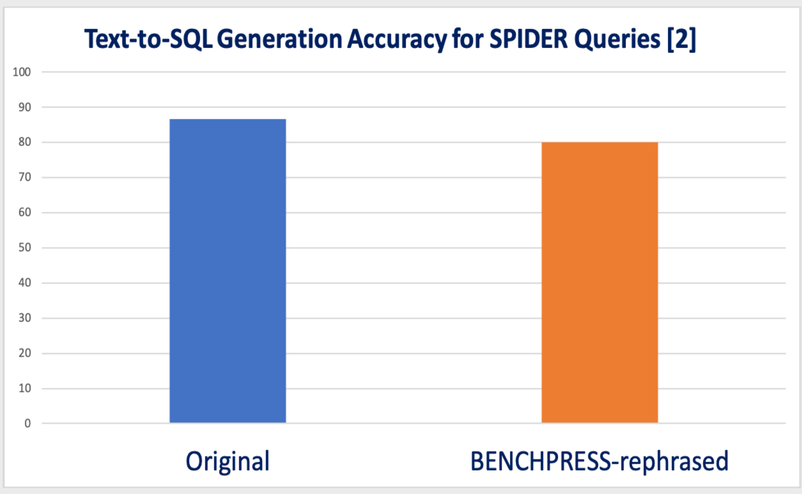
- *Quality assurance:* Comparing regenerated SQL to original SQL for assuring translation quality
- *Iterative improvement:* A back-translation process as feedback loop to optimize LLM performance by prioritizing candidates that maintain high fidelity

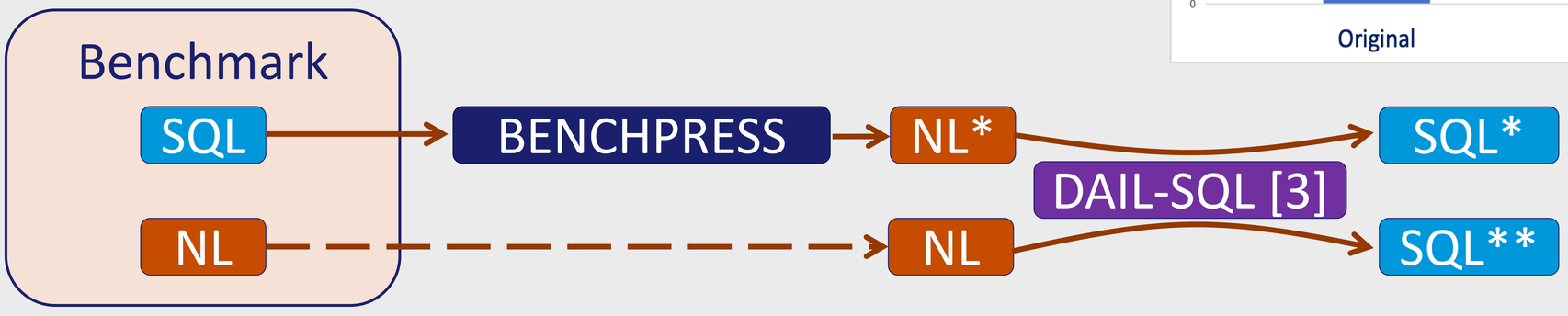
Robustness Evaluation

Robustness: the ability to handle variations in input (e.g., rephrased questions, typos)

- A single NL question can have dozens of valid, semantically equivalent rephrasings
- Noise can increase this variation further to hundreds of possibilities
- Enterprise solutions must handle these variations without performance drops
- Current benchmarks do not systematically test for robustness

Text-to-SQL Generation Accuracy for SPIDER Queries [2]





References

[1] Chen et al, “BEAVER: An Enterprise Benchmark for Text-to-SQL”, arXiv:2409.02038, 2024.
[2] Yu et al, “Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task”, EMNLP 2018.
[3] Gao et al, “Text-to-SQL Empowered by Large Language Models: A Benchmark Evaluation”, PVLDB 17(5), 2024.

http://dsg-burton.csail.mit.edu:5000/