FALCON: A Model and Dataset Agnostic Framework for Synthetic Data Generation for NL2SQL Pairs

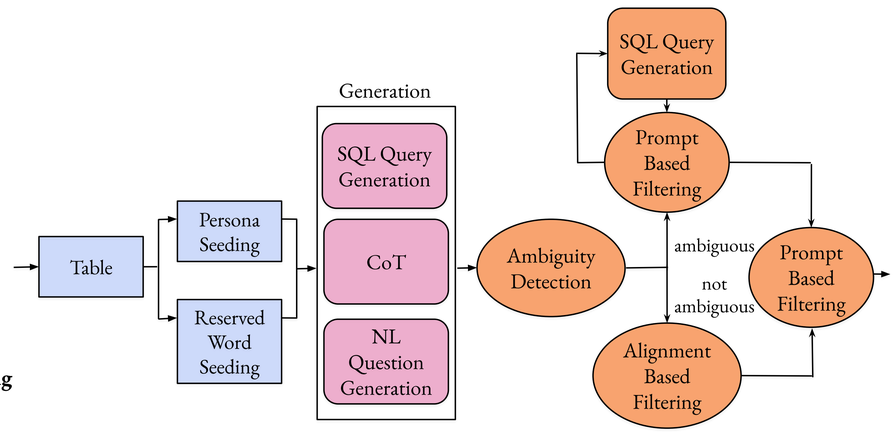
Darian Lee, Shannon Rumsey, Jack St. Clair, Xinyi Tang, Aditya Bansal, Yuanming Shi

Motivation

- Previous synthetic NL-to-SQL datasets fail to capture the diversity and complexity common to human-generated datasets
  - They are **not 100% synthetic**
  - Don't take into account **ambiguous queries**
  - **Not difficult** enough
- We introduce a **fully synthetic** NL-to-SQL generation framework that:
  - Is **model- and dataset-agnostic**
  - Generates **diverse** SQL queries and NL questions
  - Incorporates **ambiguity detection**
  - Employs **two-stage filtering** to ensure data quality

Framework

- **SQL and Natural Language Generation**
  - Enforces diversity through **reserved word** and **persona seeding**
  - Uses the SQL query to generate four NL questions with CoT
  - **LLM-as-a-judge** chooses the best question
- **Ambiguity Detection**
  - Routes SQL query-NL pairs to appropriate filtering based on presence of ambiguity
- **Prompt Based Filtering for Ambiguous Questions**
  - An LLM identifies the ambiguity type and **generates a second SQL query corresponding to the additional interpretation**

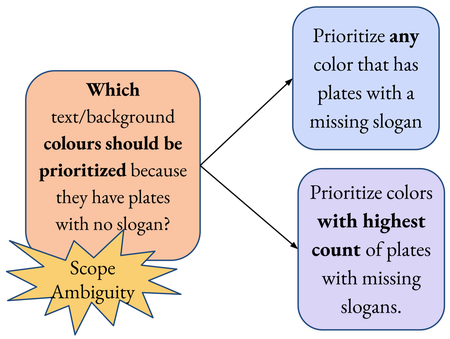


- **Alignment Based Filtering**
  - **Filters** (SQL query, NL question) **pairs** that are semantically plausible but logically misaligned
- **Final Prompt Based Filtering**
  - Checks that:
    - The SQL validly interprets the NL
    - The SQL is sensible and non-redundant
    - The NL resembles something a human would ask

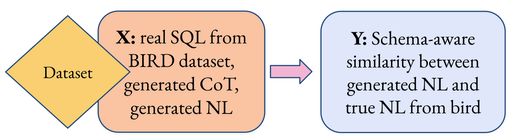
Methodology

Handling Real-World Noise

- Ambiguous natural language questions have two possible SQL query interpretations generated
- Filters using an LLM to capture additional nuances



Enforcing NL-SQL Alignment



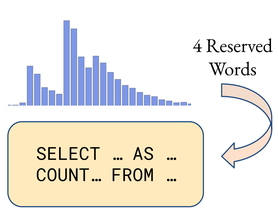
- A model that scores alignment between NL and SQL
- Reward Model: Fine-tuned ModernBERT on 9k examples
- **MSE on test: 0.02**

Emulating Real-World Diversity

Persona

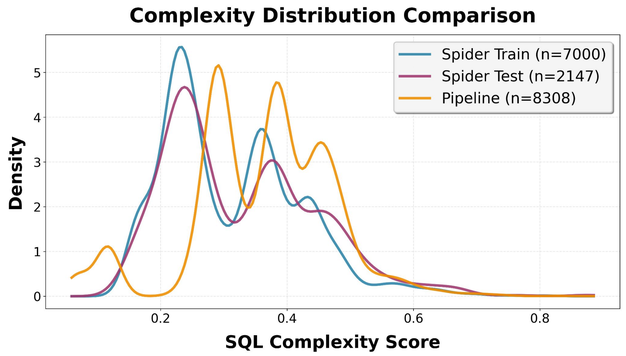


Number of Reserved Words



Results

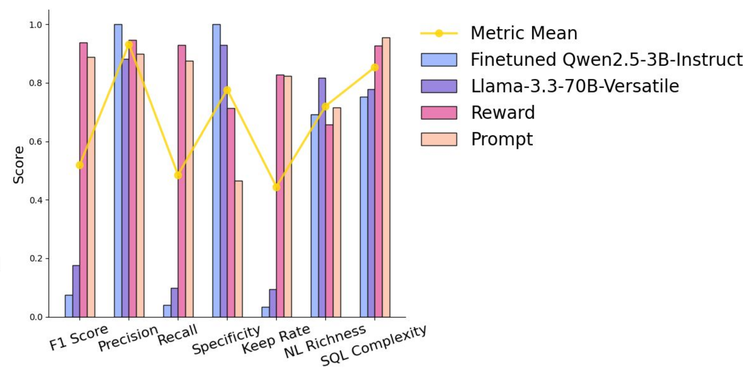
Matching Human Complexity



- FALCON data (Pipeline) results in a higher density of complex queries when compared to the Spider data

Outperforming Baselines

- The Reward filter **outperforms baselines** on F1
- The Reward filter yields **8-9x higher Recall** with **~90% Precision**, successfully capturing more valid data without introducing noise



- **Baselines**:
  - **Qwen2.5-3B:** Fine-tuned on unfiltered data; filters examples not reproduced in 3 attempts
  - **Llama-3-70B:** Teacher-based SQL regeneration (3 attempts)
  - **Prompt-based LLM-as-Judge:** Llama-3-70B scores NL–SQL alignment via chain-of-thought

Key Findings

- **High-Quality Synthetic Data:** Human evaluation across four models yielded scores ≥ 0.92 for SQL quality, NL quality, and SQL–NL alignment
- **High Complexity:** The generated data effectively mirrors the complexity of human generated counterparts

Future Work

- Improve efficiency
  - Main bottlenecks
    - Numerous LLM calls
    - Latency of each LLM-API call
- Investigate pipeline performance on larger tables
  - Use only table schemas or a sample of rows vs. an entire table
- Extend the pipeline beyond SQLite to improve compatibility with other database systems
- Evaluate and refine the pipeline using proprietary databases to address complexities not captured by publicly available benchmarks

Selective References

Alisia Maria Lupidi, Carlos Gemmell, Nicola Cancedda,Jane Dwivedi-Yu, Jason E. Weston, Jakob Nicolaus Foerster, Roberta Raileanu, and Maria Lomeli. 2025. **Source2synth: Synthetic data generation and curation grounded in real data sources**. In Workshop on Synthetic Data × Data Access Problem at the International Conference on Learning Representations (ICLR).

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga,Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2018. **Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task**. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3911–3921, Brussels, Belgium. Association for Computational Linguistics.

Saparina, I., & Lapata, M. (2024). **AMBROSIA: A Benchmark for Parsing Ambiguous Questions into Database Queries**. arXiv. https://doi.org/10.48550/arXiv.2406.19073

Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Geng, Nan Huo, Xuanhe Zhou, Chenhao Ma, Guoliang Li, Kevin C.C. Chang, Fei Huang, Reynold Cheng, and Yongbin Li. 2023. **Can LLM already serve as a database interface? a big bench for large-scale database grounded text-to-SQLs**. In Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23). Curran Associates Inc., Red Hook, NY, USA, Article 1835, 42330–42357.