 **Evaluation Metrics Limitations:**

* Discusses two key metrics:
  + **Exact set match accuracy:** Checks for lexical match between generated and reference SQL queries.
  + **Execution accuracy:** Verifies if the outputs of the generated and reference queries match.
* Highlights unreliability of these metrics as model performance approaches human levels, due to:
  + Ambiguities in natural language questions.
  + Multiple valid SQL interpretations.

 **Benchmark Limitations:**

* Identifies issues in popular benchmarks like Spider and BIRD:
  + **Ties in output:** Arise from clauses (LIMIT, GROUP BY, ORDER BY) leading to multiple valid results.
  + **Ambiguity in schema matching:** Different database columns can represent the same semantic meaning, resulting in valid but distinct SQL queries.
  + **Incorrect assumptions about database content:** Some queries may not hold true across all database instances.

 **Query Rewriting and Human Evaluation:**

* Researchers rewrote 16-20% of reference queries to address tie issues.
* Conducted a manual evaluation of 102 question-query pairs, finding many automated "incorrect" labels were accepted by human annotators.

 **Error Analysis:**

* Categorized errors into:
  + Schema errors (most common).
  + Condition errors.
  + Nested query issues.
  + GROUP BY and LIMIT problems.

 **Standard SQL Validation:**

* Migration from SQLite to PostgreSQL exposed syntax and compatibility issues, highlighting the need for standardized SQL practices in benchmarks.

 **Model Performance:**

* Notably, a GPT-4-based model (DIN-SQL) outperformed gold standard reference queries in human evaluations, suggesting existing benchmarks may underestimate advanced model performance.

 **Implications and Recommendations:**

* Achieving perfect accuracy on benchmarks is deemed impossible due to inherent ambiguities.
* Authors propose using multiple SQL queries as ground truth to represent various valid interpretations of natural language questions.
* Emphasizes the necessity for more comprehensive assessment methods, including improved human evaluations and benchmarks that account for multiple valid interpretations.

Link to the paper[: https://ppl-ai-file-upload.s3.amazonaws.com/web/direct-files/34786601/6e47b1c1-da58-40e0-8bed-5620c8c0f33c/4819\_Evaluating\_Cross\_Domain\_T.pdf](:%20https:/ppl-ai-file-upload.s3.amazonaws.com/web/direct-files/34786601/6e47b1c1-da58-40e0-8bed-5620c8c0f33c/4819_Evaluating_Cross_Domain_T.pdf)