Audit a Text-to-SQL harness before you retrain

A short audit of your evaluation harness, to run before you spend anything on retraining.

- [ ] 01 150 to 300 real questions, asked by analysts in the course of work
- [ ] 02 Two annotators and one written policy
- [ ] 03 Treat the questions where the annotators disagree as the specification your harness is missing
- [ ] 04 Build test pairs from those disagreements

WHY IT MATTERS

End-to-end accuracy conflates two specification failures with different fixes. Most Text to SQL evaluation harnesses score correct queries as wrong because nobody wrote down what correct means.

From: *Your Text to SQL Model Didn't Fail. Your Evaluation Metric Did.*
jordanbrowne-moore.com/articles/text2sql-evaluation-harness

NOTES

Jordan Browne-Moore
ML consultant for fintechs
jordan@jordanbrowne-moore.com
jordanbrowne-moore.com
Book a call
cal.com/jordan-browne-moore-juluar