Problems of Manual Evaluation

- Expensive in terms of time/money.
- Subjective (some judges are more careful/better at guessing).
- Not quite consistent judgments from different people.
- Not quite consistent judgments from a single person!
- Not reproducible (too easy to solve a task for the second time).
- Experiment design is critical!

- Black-box evaluation important for users/sponsors.
- Gray/Glass-box evaluation important for the developers.
- SRC-based allows to compare with humans.
- Sentence-level no longer relevant for large language pairs.