Translation Quality in Multilingual LLM Evaluation – Revisions from ARR May 2025

July 29, 2025

1 Changes

In the last ARR cycle, the Meta Reviewer suggested the following revisions:

1. Improve coherence by using the same datasets to compare several models
2. Using llm-as-a-judge with a single model introduces a strong dependency and bias toward this model which is not considered nor evaluated
3. Reorganise the paper to make it easier to follow
4. Improve research limitations

Regarding point 1, we perform two additional analyses relying only on EU20 datasets, in addition to refining the reference-based comparison with Global-MMLU. We apply χ^2 tests and logistic regression to assess the association between detected translation errors and LLM accuracy errors on EU20 ARC, MMLU and GSM8K which offer dichotomous accuracy metrics on the sample level.

We adress point 2 by including three annotator models in addition to GPT-4o-mini, conducting meta-evaluation with three and statistical testing and logistic regression with two, due to time and budget constraints.

Regarding 3 and 4, we have fully restructured the paper for improved clarity – removing the cumbersome split between methodology and results subsections, and instead presenting each analysis in its own (sub-)section.

We have reorganized the discussion of limitations and moved future work to a separate section. In addition, we extend the translation quality meta-evaluation by leveraging the high-quality, multi-parallel FLORES-200 dataset and using error detection rates (false-positive rate per sentence and

error rate per 1,000 words) as a proxies to compare the error reporting tendency of different annotator LLMs, partially closing an evaluation gap left by Span-ACES.