Author Response of Resubmission:
Is Human-Like Text Liked by Humans?
Multilingual Human Detection and Preference Against AI

Dear Reviewers, Area Chair and Program Chair,

We sincerely thank the reviewers for their insightful feedback, which has helped us improve our work.
The meta review highlights two major weaknesses of our paper (Q1-Q2), and we further summarize another two comments of the reviewers (Q3-Q4) and respond all below.

Black text: reviewer comment and our responses;
Blue: the excerpts that we have added to the amended manuscript.

Response to the Reviewers Comments and Suggestions
—----------

> Q1. Clarification of Annotation Scope and Purpose

We clarified in the introduction that **the study targets the upper bound of human detection capability by recruiting expert annotators**. The human annotation sample size (~11.5K) is now explicitly stated early in the paper.
We have acknowledged annotator bias and the need for broader demographic coverage in future work, including laypersons with diverse demographics to complement expert-based findings.

> Q2: Lack of linguistic analysis using automatic tools.

Appendix D provides detailed linguistic analysis summarized by annotators for each language and dataset. We did not perform tool-based analysis because previous studies have conducted lots of linguistic analysis based on automatic tools [1]. See more in Appendix A.2. **This work aims to analyze features from human intuitions, which is also how humans perceive the quality of AI generations in real-life human-AI interactions.** Generally, we leverage our intuition rather than calling tools to judge an output.

> Q3: Human Preference Judgments Clarification

We clarified that **the preference study aimed to investigate the impact of human detection accuracy on their preferences (line 479-481),** rather than general human preferences for AI content.
The study scope is explained more clearly: 10 expert annotators, 3 high-resource languages, and 6 datasets were selected to span a range of detection accuracy (50% to 100%)**, rather than covering all datasets **(line 482-492)**. The rationale behind using the same annotators for both detection and preference tasks is now stated explicitly to ensure consistency in evaluating the relationship.

> Q4: Paper Organization Improvements

Following reviewer suggestions, we:

- Moved details of the annotation tool** to the Appendix B.
- Brought prompt examples and human intuition-based findings into the main content.
- Clarified average calculations (e.g., in Table 3) and labeled them as simple averages to avoid misinterpretation.
- Adjusted formatting in Tables 1 and 2 for improved readability.

We acknowledge the value of deeper demographic analysis (e.g., by age, education, or technical background). However, this is beyond the primary scope of our current work and will be considered in future research.
Regarding the suggestion to expand the language coverage, we believe the nine languages included — Arabic, Chinese, English, Hindi, Italian, Japanese, Kazakh, Russian, and Vietnamese—already offer broad linguistic diversity.
As for the advice to conduct linguistic analysis using automatic tools, we note that many prior studies have already taken this approach, making it less novel. Moreover, our study is human-centered, focusing on both detection and preference. As discussed, the goal is to assess how humans perceive the quality of AI-generated outputs directly, rather than relying on automated tools to determine user preference.

Overall, these changes strengthen the **paper's contributions across its three pillars**:

1. This is the first study investigating the upper bound of human detection capability of machine-generated text in the multilingual setting, covering nine languages, nine domains by 16 datasets. The manually-identified distinguishable features for each language and each dataset between human-written and machine-generated text (MGT) can benefit LLM generation improvement in multilingual settings.

2. We performed a preliminary attempt to bridge the gap by improving prompting. Experiments show that the improved prompts can either fully or partially address the gap over more than 50% cases. Cultural nuances and diversity in structure, sentiment and style are more challenging than providing concrete details, formatting, and mixture of other languages in a response. In the evaluation, we conducted the second round of MGT detection over the new generations from the objective perspective, and also perform a survey to intuitively assess whether the generations by new prompts mitigate the gap by asking annotators case by case. Evaluations from two perspectives make the results more solid and convincing.

3. We further explored the impact of human detection capability of MGTs on their preference judgements. Results show that people are more likely to choose machine text as their preference when they cannot clearly identify which text is human-written. Otherwise, they tend to favor peer-written text.

Reference

[1] How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection: https://arxiv.org/pdf/2301.07597