Explanation of Revisions

1. We included in our evaluation the ALDi metric (Keleg et al., 2023) to evaluate dialect authenticity.
2. We completed the missing human evaluations in Table 3 (and Table 10 in the appendix) to cover **all the model-dialect pairs.**
3. We added 3-way overlap for all the dialects except Emirati (2-way overlap instead) due to the lack of annotators.
4. We included an Error Analysis carried out by paper co-authors that are native speakers, covering 600 predictions that examines:
    a. **Cross-model** error patterns: highlighting the types and frequency of errors made by each model across dialects.
    b. **Cross-dialect** error patterns: identifying which dialects or cultural contexts are more challenging across models.
5. We expanded the "Dialectal Authenticity" abbreviation into *DAuth* to avoid confusion with "Dialectal Arabic" (DA).
6. We revised section 4 for clarity.

References

Amr Keleg, Sharon Goldwater, and Walid Magdy. 2023. ALDi: Quantifying the Arabic level of dialectness of text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10597–10611, Singapore. Association for Computational Linguistics.