# Human Evaluation Guideline

## Goal

Human evaluations were conducted alongside automatic metrics such as BLEU and ChrF to better capture the linguistic nuances that normalization might obscure. Native speakers of each language assessed translations produced using different normalization settings: **No Normalization, H-Only Normalization, and HSL Normalization**.

## Evaluation Setup

- **Annotators:** Native speakers of the target language.
- **Samples:** 50 randomly selected sentences.
- **Inputs shown to annotators:**
    1. Source sentence (English)
    2. Reference sentence ( Amharc/Geez/Tigrinya)
    3. Model outputs from different normalization settings

## Evaluation Measurements

### 1. Overall Translation Quality

- **Task:** Judge whether the translations are acceptable in meaning and fluency.
- **Options:**
    - *All translations are good*
    - *Two translations are good* (specify which)
    - *Only one translation is good* (specify which)

    - *All translations are bad*

### 2. Comparative Preference

- **Task:** If more than one translation is acceptable, choose the one that is **clearly better** overall.

- **Instruction:** Preference should be based on meaning preservation, fluency, and appropriate character usage.

### 3. Error Identification

- **Task:** Record specific issues observed in translations.

- **Examples:**

    - Mistranslations (wrong meaning, cross-language contamination, such as Amharic words in Tigrinya outputs)

    - Morphological errors (gender, plural, agreement mistakes)

    - Repetition or unnatural phrasing

    - Homophone handling (collapsing distinct Ge’ez characters into a normalized form)

## Annotation Format

For each sentence, annotators record:

1. **Rate translation** (all good / two good / one good / all bad)

2. **Select if two are good** (list which outputs)

3. **Select if one is better** (list which output)

4. **Other** (text notes on mistranslations, homophone issues, repetitions, etc.)