

Human Evaluation Template: Machine Translation Comparison

Project 7: Analysis and comparison of translation errors and biases in LLMs

Metadata

- Language Pairs: ENG → GER, ENG → SLO, GER → SLO
- Sentence ID
- Evaluator name / ID
- 3 evaluation criteria
- Include “comment” section
- 2/3 sentences:
 - Original source sentence
 - Machine translation output
 - (Gold standard / reference (human) translation)

Evaluation Criteria

1. **Formality** – How formal is the machine translation output?
 - Scale:
 - -3 = very informal
 - -2 = informal
 - -1 = somewhat informal
 - 0 = neutral
 - +1 = somewhat formal
 - +2 = formal
 - +3 = very formal
2. **Fluency** – How fluent and natural is the machine translation output in the target language?
 - Scale:
 - 5 = Perfect (completely natural)
 - 4 = comprehensible (minor, errors, still easy to understand)
 - 3 = somewhat comprehensible (noticeable errors, but understandable)
 - 2 = Incomprehensible (hard or impossible to understand)
 - 1 = Other (leave a comment)
3. **Meaning Preservation** – How well does the machine translation preserve the meaning of the source sentence?
 - Scale:
 - 6 = Completely equivalent
 - 5 = Mostly equivalent
 - 4 = Roughly equivalent
 - 3 = Not equivalent, but share some details
 - 2 = Not equivalent, but on the same topic
 - 1 = Completely dissimilar

