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:
  + 0 = very informal
  + 0.17 = informal
  + 0.33 = somewhat informal
  + 0.5= neutral
  + 0.67 = somewhat formal
  + 0.83 = formal
  + 1 = very formal

1. **Fluency** – How fluent and natural is the machine translation output in the target language?

* Scale:
  + 1 = Perfect (completely natural)
  + 0.67 = comprehensible (minor, errors, still easy to understand)
  + 0.33 = somewhat comprehensible (noticeable errors, but understandable)
  + 0 = Incomprehensible (hard or impossible to understand)

1. **Meaning Preservation** – How well does the machine translation preserve the meaning of the source sentence?

* Scale:
  + 1 = Completely equivalent
  + 0.8 = Mostly equivalent
  + 0.6 = Roughly equivalent
  + 0.4 = Not equivalent, but share some details
  + 0.2 = Not equivalent, but on the same topic
  + 0 = Completely dissimilar