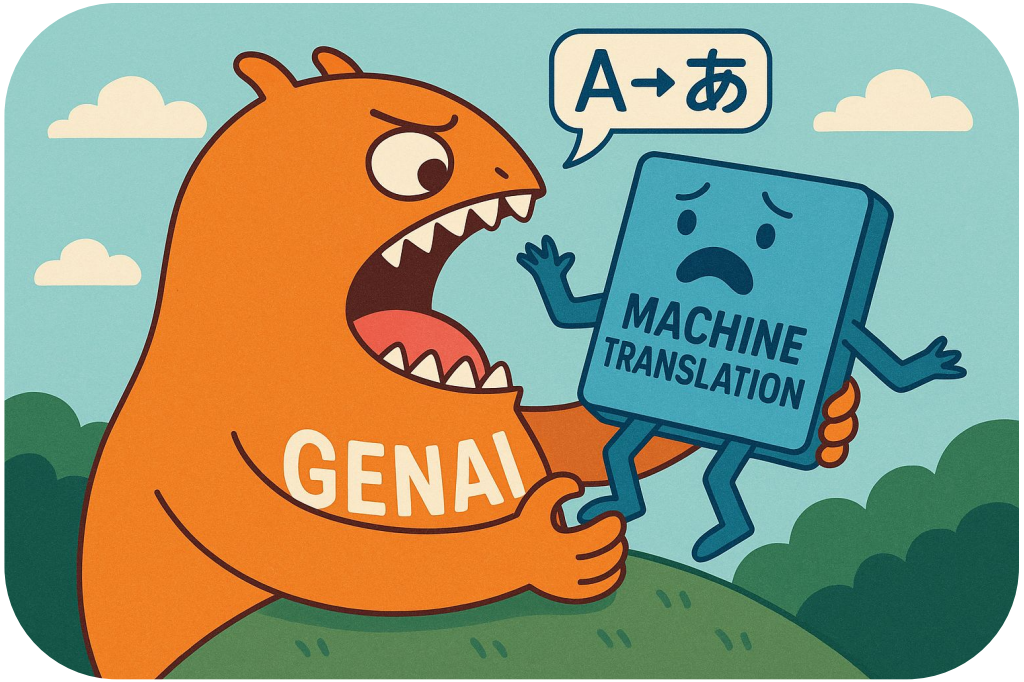
From MT to Multilingual Gen AI: What to Expect and Evaluate

Mariya Shmatova
Toloka AI, WMT

What’s up?



GenAI absorbs MT as a function / service

From Isolated Tool to Integrated Capability

- translation is one of the instructs
- translation can be a part of a broader conversational and reasoning system:
  - translate AND reason
  - translate AND explain
  - translate AND reformulate
  - …
- translation becomes more manageable
- quality expectations grow

What's then?

To provide expected quality for expected capabilities AI models should be really multilingual (not English-centric) that's why we need:

- more multilingual training data of high quality (especially for low resourced languages)
- more multilingual benchmarks and evaluation criteria adjusted for language and cultural specificity

To highlight these needs and to focus on multilinguality in LLMs we launch **Multilingual Instruction Shared Task** at WMT25.

What do we Test at MIST?

- machine translation by LLMs
- linguistic reasoning
- open-ended generation
- LLM-as-a-Judge for multilinguality

28 languages (both high and low resourced)

https://www2.statmt.org/wmt25/multilingual-instruction.html

New Challenges for Evaluation

- output variability
  - evaluate more than one run
- broader context sensitivity
  - evaluate long contexts
- hallucinations
  - detect hallucination patterns
- task specificity
  - every subtask needs its evaluation approach / benchmarks / criteria

Methods to Use for Evaluation

- Custom human evaluation
  - Pros: expert quality, direct insights
  - Cons: expensive, time consuming
  - Examples: MQM-based approach, ESA-based approach, side by side comparison
- LLM-as-a-judge evaluation
  - Pros: when set up, could be run fast with good quality
  - Cons: needs predefined benchmark & trusted LLM-as-a-judge
  - Examples: reference-free judgements, reference-based judgements, rubric-based judgements

Thank you!