intento

Comparative Evaluation of Large Language Models for Linguistic Quality Assessment in Machine Translation

Daria Sinitsyna
Intento

Konstantin Savenkov
Intento

Summary

1. Research goals
2. Background
3. Methodology
4. Data
5. Choosing the approach
6. Model Analysis
7. Cost Analysis
8. Conclusion

Goals

Research goals

1. How well can LLMs perform Linguistic Quality Assessment (LQA)?
2. What is the **best LLM solution design** for automatic LQA (zero-shot, CoT, multi-agent)?
3. **Which LLM is the best today** for identifying and classifying translation errors?
4. How do they compare cost-wise?

Background

Complexity of MT post-editing largely depends on quality requirements.

Reaching perfect automatic translation (no edits) requires a way to automatically assess and improve translation quality.

This research is dedicated to finding the best design and building blocks for such solution.

For that, we evaluate design choices and LLMs on a simple and well-defined task - automatic LQA based on the MQM error typology.

Last year we have assessed GPT-4 capabilities for Linguistic Quality Assessment using DQF-MQM error typology*

German

78% accuracy**

76% agreement with linguists

Spanish

80% accuracy**

82% agreement with linguists

* review was done on samples of 50 segments from Annual State of MT Report 2023 and presented at TAUS 2023

** precision

Methodology

The Approach

1. Choose a quality estimation agent design for the multi-agent LQA (with GPT-4o as the baseline & Reviewer Agent) with the MQM error typology.
    a. Zero-shot LLM agent for all MQM dimensions
    b. A system of one agent per MQM dimension
    c. Zero-shot Chain of Thought agent
2. Compare 6 LLMs as a model for the QE agent of quality estimation agent combined with a reviewer agent (based on Claude 3.5 Sonnet)
3. Check correlation of different multi-agent systems with each other to understand whether they have similar biases and limitations
4. Analyze all issues found by all multi-agent LQA systems with human linguists
5. Compare LLMs in terms of false alarms and found issues

In a nutshell

Models:

- OpenAI GPT-4o
- OpenAI GPT-4o mini
- Anthropic Claude 3.5 Sonnet
- Gemini 1.5 Flash
- Gemini 1.5 Pro
- Llama 3.1 405B

Prompting techniques:

- Zero-shot LLM agent for all MQM dimensions
- A system of one agent per MQM dimension
- Zero-shot Chain of Thought agent

Language pairs:

- English-Spanish
- English-German
- English-Chinese

Data

Notes on Data Preparation

- 500 longest, or contextually rich, segments taken from our Annual State of MT Report with no reference translations
- Chosen using stratified sampling based on COMET score
- Larger proportion of segments with lower score where issues are more prominent
    - 250 low-scoring segments
    - 150 segments with middle scores
    - 100 high-scoring segments

Choosing the approach

We use MQM dimensions that LLMs can work around without additional contextual knowledge

- Terminology
- Mistranslation
- Over-translation
- Under-translation
- Addition
- Omission
- Untranslated text
- Culture-specific reference
- Markup

- Awkward style
- Unidiomatic style
- Inconsistent style
- Formatting
- Grammar
- Spelling
- Punctuation
- Character encoding

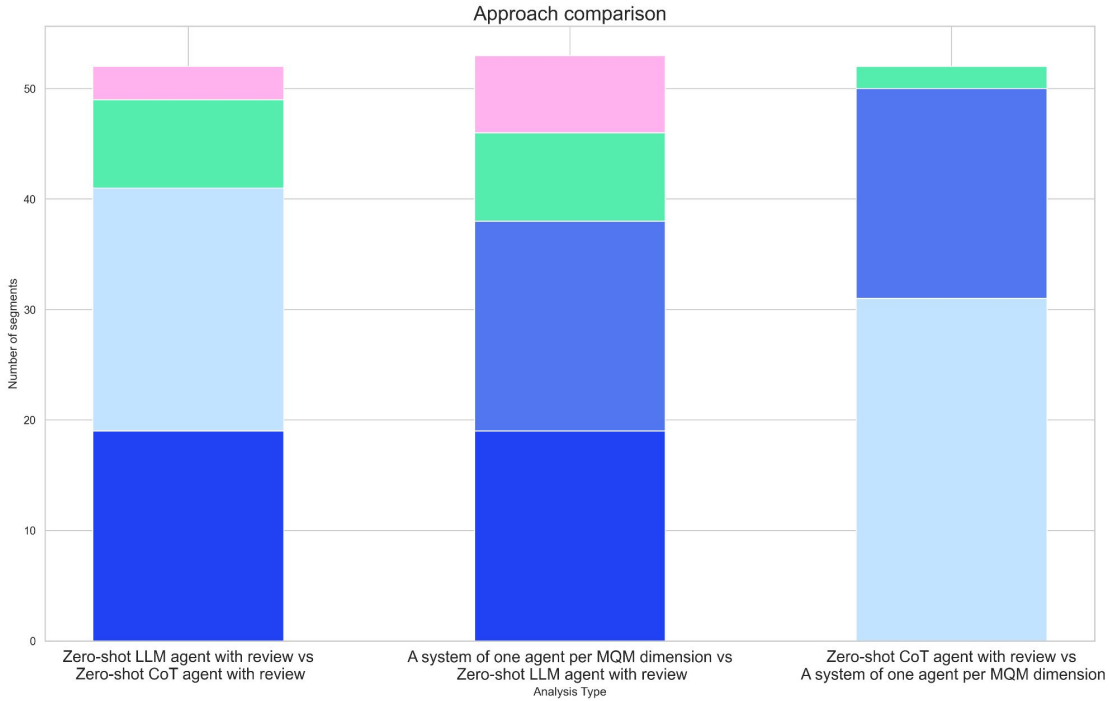
Choosing the approach

When choosing the multi-agent solution architecture, we assess the differences in judgment, not the absolute LQA accuracy

- We focus on segments where LLM solutions with different architecture disagree about the translation quality
- For each solution, we select segments where there's a disagreement (one setup finds much less issues/less critical issues than another)
- A human assessor determines:
  - Which approach led to a more accurate analysis
  - Whether both, neither, or only one of the analyses is correct

Reviewer has assessed CoT as the more correct when it comes to approach disagreement but states CoT and zero-shot LLM agent are of nearly the same quality

- Zero-shot LLM agent
- A system of one agent per MQM dimension
- Zero-shot CoT agent
- Both approaches are correct
- Neither approach is correct



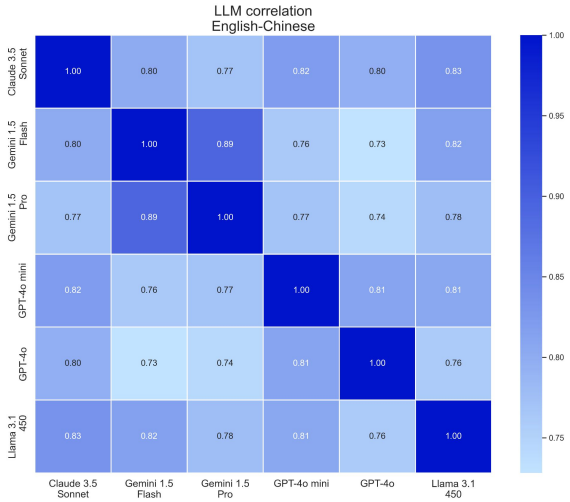
We proceed with one prompt MQM with a reviewer model

- Between the compared models, zero-shot LLM agent is the fastest:
  - Zero-shot LLM agent takes ~20 minutes per 500 segments
  - Zero-shot CoT takes ~30 minutes per 500 segments
  - A system of one agent per MQM dimension takes ~150 minutes per 500 segments
- It is also the least expensive due to having the least tokens in the system message
- Quality-wise, zero-shot LLM agent and CoT have nearly identical results however, since one-prompt LLM agent is first by other parameters, we use it in the final setup
- We proceed with zero-shot LLM agent with the same reviewer model for all models - Claude 3.5 Sonnet

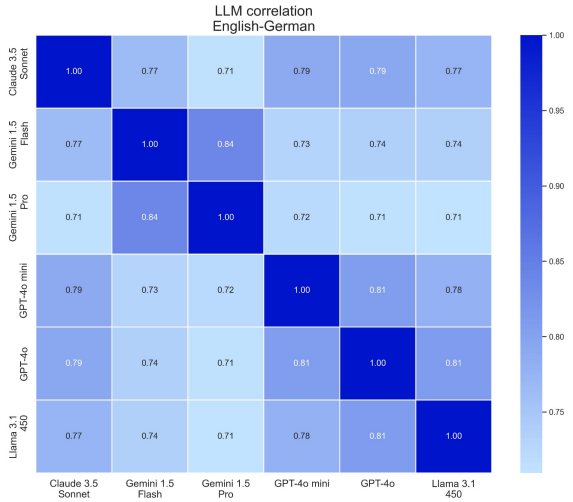
Model Analysis

There is high degree of agreement and consistency between different LLMs

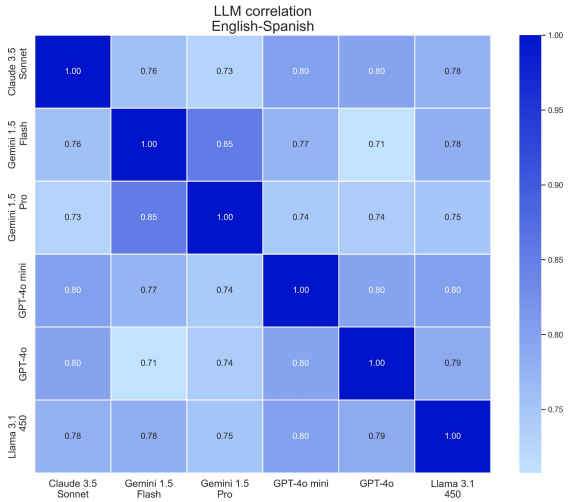
Chinese



German



Spanish



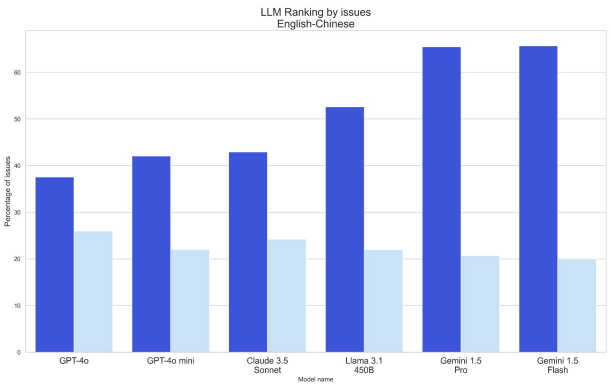
The correlation is shown between multi-agents: systems where the LQA agent is one of the 6 LLMs and reviewer agent is Claude 3.5 Sonnet

While Gemini models produce the least false alarms, GPT-4o tends to find the most major and critical relevant issues

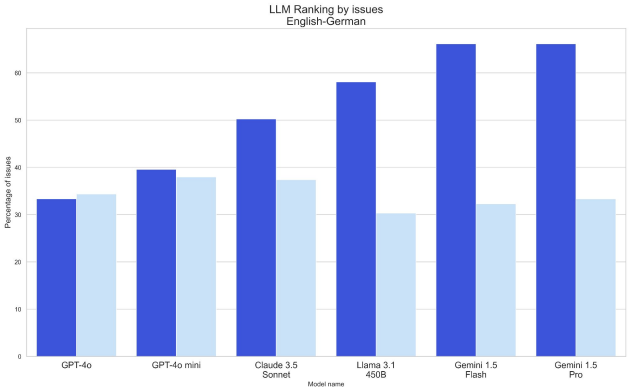
False alarms

Missed issues

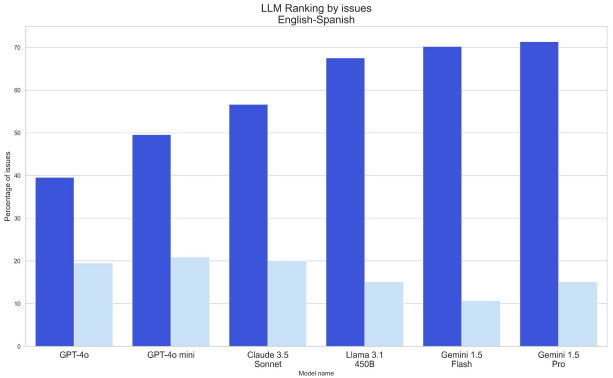
Chinese



German



Spanish

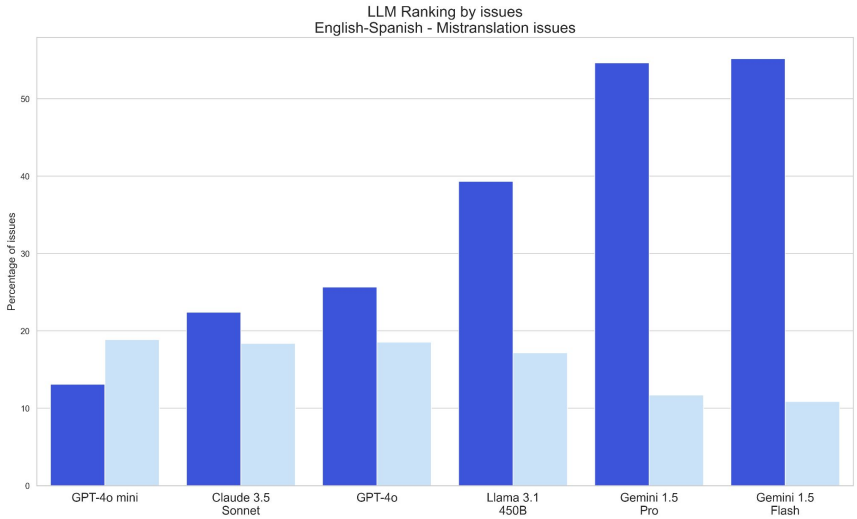


Different LLMs excel in detection of different issues, and higher results could be achieved with model ensembling

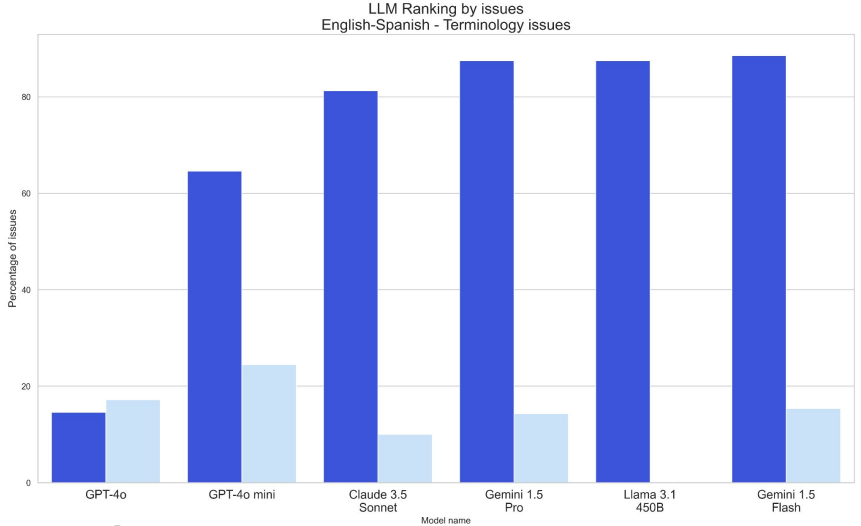
False alarms

Missed issues

Mistranslation detection



Terminology detection



GPT-4o mini has the lowest number of missed terminology errors while GPT-4o nearly excels in domain terminology issue detection in Spanish

GPT-4o shows the best results in terms of identified major and critical issues

- GPT-4o perform the best in all language pairs, showing the best harmony between the number of false alarms and identifying correct issues
- Gemini models produce the least false alarms in all languages, as on average, only 18% of all issues Gemini models identified were false alarms
- GPT4o and GPT-4o mini models find the most issues compared to the rest of LLMs, as between all languages pairs, they identify nearly major and critical 70% issues
- The latest Llama 3.1 with 450B parameters shows comparable results to commercially available models
- The highest rate of identified issues can be achieved with model ensembling due to different LLMs excelling in different issues' detection

Cost analysis

Cost analysis

- Since we used Llama-3.1 model through OctaAI, costs for this model were calculated using pricing on the official website:
- https://octo.ai/docs/getting-started/pricing-and-billing.

LQA pricing

Prices for each model and language pair per 1 million characters
with Claude 3.5 Sonnet as the reviewer model

Language pair	GPT-4o	GPT-4o mini	Gemini 1.5 Pro	Gemini 1.5 Flash	Claude 3.5 Sonnet	Llama 3.1 450B
English-German	19.00	9.38	16.08	9.13	18.00	15.00
English-Spanish	13.42	6.62	11.36	6.45	6.36	10.06
English-Chinese	25.33	12.50	21.43	12.18	24.00	20.01

Conclusion

Conclusion

1. Between several approaches, zero-shot LLM agent is the fastest, least expensive prompting method, comparable in quality with Chain of Thought only, with Reviewer agent being the key to achieving higher quality results.
2. GPT-4o showcases the best harmony between the comparatively low number of false alarms and identifying correct issues, proving to be the best at Linguistic Quality Assessment among all language pairs.
3. Cost-wise, GPT-4o mini and Gemini 1.5 Flash are the cheapest, although all models are comparable in price.
4. We generally see even better results when it comes to client data LQA due to the possibility of adding more information and shots to the LQA and reviewer agents.

Thank you!

ks@inten.to

daria.sinitsyna@inten.to

An independent multi-domain evaluation of MT engines

Commercially available pre-trained MT models

2261 Market St, #4273
San Francisco, CA 94114

inten.to