Case Study: MQM Quality Evaluation of TranslateGemma

Alconost — February 2026

1. Introduction

This report presents an MQM (Multidimensional Quality Metrics) stress test of Google's TranslateGemma (`google/translategemma-12b-it`), a 12B-parameter open-source translation model supporting 55 languages. The evaluation was intentionally designed to push the model beyond typical operating conditions: the source material is a technically dense academic paper in computational linguistics, the target language set emphasizes less common pairs, and 4 of the 16 languages fall outside TranslateGemma's official support list.

Forty-five professional linguists annotated 322 segments across 16 target languages, producing 1,169 error annotations. In MQM, lower scores indicate better quality: each Minor error adds 1 point, Major adds 5, and Critical adds 25. Despite the deliberately challenging setup, TranslateGemma delivered strong results for its supported languages. German averaged just 2.3 penalty points per segment — roughly 2 minor errors per segment, indicating near-publishable quality. The top 6 supported languages all averaged under 5 points per segment, corresponding to light post-editing effort. Moroccan Arabic — an unsupported language — averaged 3.1 per segment, outperforming 10 of 12 supported languages.

No rebuttal phase was included in this evaluation. In real-life localization QA, a significant share of Minor errors — which constitute 66% of all findings — would typically be rebutted as acceptable stylistic choices, lowering the effective scores. The core conclusion remains: even capable MT models require human review before production deployment.

The quality gap between supported and unsupported languages was substantial: 2.18 vs 13.67 MQM penalty per segment, a 6× degradation. This reflects the limited availability of training data for low-resource languages rather than any architectural limitation of the model. Inter-annotator agreement remained low across languages, with only 4 of 16 showing statistically significant ranking consistency — consistent with known challenges in MQM annotation reliability.

We additionally benchmarked automatic evaluation metrics against human MQM scores. MetricX-24 XXL achieved the strongest correlation (Pearson r=0.88), while the same metric family served via Google's Vertex AI API yielded only r=0.25 — a 3.5× gap attributable to differences in model size and hosting infrastructure. COMET-Kiwi XL reached r=0.84, establishing it as a practical alternative for rapid quality estimation.

2. Model Specification

2.1 Model Identity

Property	Value
Model Name	TranslateGemma
Model ID	`google/translategemma-12b-it`
Model Size	12 billion parameters
Model Type	Instruction-tuned (IT)
Architecture	Gemma-based encoder-decoder

Property	Value
Source	HuggingFace Hub
License	Gemma Terms of Use

2.2 Model Capabilities

TranslateGemma is a specialized translation model supporting 55 languages. It uses a structured chat template format for translation requests, distinguishing it from general-purpose LLMs.

Officially Supported Languages (55): Afrikaans, Arabic, Bengali, Bulgarian, Chinese (Simplified), Chinese (Traditional), Croatian, Czech, Danish, Dutch, English, Estonian, Filipino, Finnish, French, German, Greek, Gujarati, Hebrew, Hindi, Hungarian, Indonesian, Italian, Japanese, Kannada, Korean, Latvian, Lithuanian, Malay, Malayalam, Marathi, Norwegian, Persian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovenian, Spanish, Swahili, Swedish, Tamil, Telugu, Thai, Turkish, Ukrainian, Urdu, Vietnamese.

3. Deployment Infrastructure

3.1 Hosting Configuration

Property	Value
Platform	HuggingFace Inference Endpoints
Cloud Provider	AWS
Region	us-east-1
Instance Type	NVIDIA A100 (80GB)
Scaling	1 replica (fixed)
Endpoint Type	Private

3.2 Model Loading Configuration

```
model = AutoModelForImageTextToText.from_pretrained(
    "google/translategemma-12b-it",
    device_map="auto",
    torch_dtype=torch.bfloat16
)
processor = AutoProcessor.from_pretrained("google/translategemma-12b-
it")
```

Key Parameters:

- Precision: bfloat16 (16-bit brain floating point)
- Device mapping: Automatic GPU allocation
- Memory footprint: ~24GB VRAM

4. Translation Pipeline

4.1 Approach A: Supported Languages

For languages in TranslateGemma's official list, we use the structured message format as specified in the model card.

Input Format:

```
messages = [
    {
        "role": "user",
        "content": [
            {
                "type": "text",
                "source_lang_code": "<SOURCE_LANG_CODE>",
                "target_lang_code": "<TARGET_LANG_CODE>",
                "text": "<SOURCE_TEXT>"
            }
        ]
    }
]
```

Template Application:

```
inputs = processor.apply_chat_template(
    messages,
    tokenize=True,
    add_generation_prompt=True,
    return_dict=True,
    return_tensors="pt"
)
```

Generation Parameters:

```
generation = model.generate(
    **inputs,
    do_sample=False,      # Deterministic decoding (greedy)
    max_new_tokens=2000   # Maximum output length
)
```

Language Code Examples:

Language	Code
Portuguese (Portugal)	`pt_PT`
Portuguese (Brazil)	`pt_BR`
Arabic (Saudi Arabia)	`ar_SA`
Arabic (Egypt)	`ar_EG`
Russian	`ru`
Japanese	`ja_JP`
Korean	`ko_KR`

Language	Code
German	de
French	fr
Italian	it
Polish	pl
Ukrainian	uk

4.2 Approach B: Unsupported Languages (Custom Prompting)

For languages not in the official list, we employ a custom prompting technique as documented in the model card. The `target_lang_code` field accepts natural language instructions.

Prompt Format:

```
<start_of_turn>user
Translate to [Language] ([Native Name]):
Output only the translation, no explanations.

[SOURCE_TEXT]<end_of_turn>
<start_of_turn>model
```

Tokenization:

```
prompt = (
    f"<start_of_turn>user\n"
    f"{target_prompt} Output only the translation, "
    f"no explanations.\n\n"
    f"{text}<end_of_turn>\n"
    f"<start_of_turn>model\n"
)
inputs = processor.tokenizer(
    prompt,
    return_tensors="pt",
    add_special_tokens=True
)
```

Custom Prompts Used:

Language	Custom Prompt
Belarusian	Translate to Belarusian (беларуская мова):
Hmong	Translate to Hmong (Hmoob):
Modern Standard Arabic	Translate to Modern Standard Arabic (العربية الفصحى):

Rationale for Instruction Addition: The phrase “Output only the translation, no explanations.” was added to prevent the model from generating verbose responses with multiple translation options or linguistic commentary, which occurred in initial testing.

5. Source Material

5.1 Document Properties

Property	Value
Title	Selective Invocation for Multilingual ASR: A Cost-effective Approach Adapting to Speech Recognition Difficulty
Authors	Xue et al.
Venue	Interspeech 2025
Reference	arXiv:2505.16168
Document Type	Academic paper abstract
Domain	Speech Recognition / NLP
Source Language	English
Segment Count	7
Total Word Count	~615 words

5.2 Segment Distribution

Segment	Type	Word Count
1	Title	15
2	Abstract paragraph 1	120
3	Abstract paragraph 2	65
4	Abstract paragraph 3	85
5	Abstract paragraph 4	75
6	Abstract paragraph 5	45
7	Abstract paragraph 6	55

6. Target Languages

6.1 Language Matrix

Language	ISO Code	Support Status	Translation Method
Portuguese (Portugal)	pt-PT	Supported	Structured
Portuguese (Brazil)	pt-BR	Supported	Structured
Arabic (Saudi Arabia)	ar-SA	Supported	Structured
Arabic (Egypt)	ar-EG	Supported	Structured
Arabic (Morocco)	ar-MA	Unsupported	Custom Prompt
Russian	ru	Supported	Structured
Italian	it	Supported	Structured
Polish	pl	Supported	Structured
Korean	ko	Supported	Structured
Japanese	ja	Supported	Structured

Language	ISO Code	Support Status	Translation Method
German	de	Supported	Structured
French	fr	Supported	Structured
Ukrainian	uk	Supported	Structured
Belarusian	be	Unsupported	Custom Prompt
Hmong	hmn	Unsupported	Custom Prompt
Arabic (MSA)	ar-MSA	Unsupported	Custom Prompt

6.2 Language Selection Rationale

Supported Languages:

- Major European languages (DE, FR, IT, PL, UK)
- East Asian languages (JA, KO)
- Slavic language family (RU, UK, PL)
- Arabic regional variants (SA, EG)
- Portuguese variants (PT, BR)

Unsupported Languages:

- Belarusian: Low-resource Slavic language, shares features with Russian/Ukrainian
- Hmong: Southeast Asian language with limited digital resources
- Arabic (Morocco): Moroccan Darija, not in TranslateGemma's official language list
- Modern Standard Arabic: Formal written Arabic distinct from regional variants

7. Technical Observations

7.1 Processing Time

Metric	Value
Average time per segment	160-340 seconds (~3-6 minutes)
Timeout threshold	400 seconds
Total segments processed	112 (7 segments x 16 languages)
Estimated total processing time	6-10 hours
Hardware	NVIDIA A100 80GB
Precision	bfloat16

7.2 Initial Challenges and Solutions

1. Unsupported Language Errors: Initial attempts to use unsupported language codes (e.g., `be_BY`, `hmn`) returned HTTP 400 errors. Solution: Custom prompting technique.
2. Verbose Output for Unsupported Languages: The model initially produced explanatory text with translation options. Solution: Added explicit instruction "Output only the translation, no explanations."

8. Quality Evaluation Framework

8.1 MQM Assessment Plan

Human translators evaluated each translation using the Multidimensional Quality Metrics (MQM) framework. The exact taxonomy used:

Error Categories:

Category	Subcategory
Accuracy	Addition, Omission, Mistranslation, Source error, Untranslated
Fluency	Punctuation, Spelling, Grammar, Register, Inconsistency, Character encoding
Terminology	—
Style	—
Locale convention	—
Audience appropriateness	—
Design and markup	—
Other	—

Severity Levels and Weights:

Severity	Weight	Description
Minor	1	Noticeable but doesn't affect meaning
Major	5	Affects meaning or significantly impairs fluency
Critical	25	Completely wrong meaning, offensive, or harmful

MQM score = sum of (severity weight × error count), normalized per segment where applicable.

8.2 Evaluation Scope

Metric	Value
Total translations	112
Languages	16
Segments per language	7
Evaluators per language	3
Total MQM projects	48

8.3 Annotation Tool

Property	Value
Tool	Alconost MQM Annotation Tool

Property	Value
URL	https://alconost.mt/mqm-tool/
API	REST API with Bearer token authentication
Export formats	TSV, JSONL, CSV, PDF

9. Results

9.1 Annotation Progress

Metric	Value
Total segments	336
Completed segments	322
Completion rate	95.8%
Total errors identified	1,169
Average errors per segment	3.63

9.2 Error Severity Distribution

Severity	Count	Percentage	MQM Weight
Critical	82	7.0%	25
Major	288	24.6%	5
Minor	769	65.8%	1

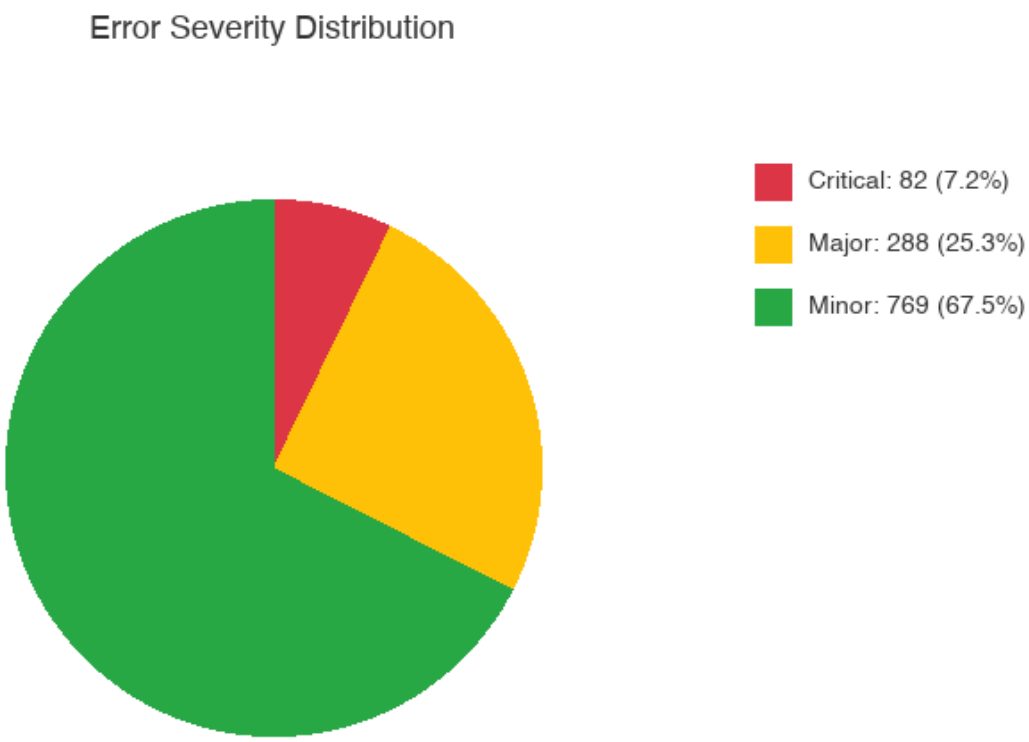

Figure 1: Error Severity Distribution

9.3 Error Category Breakdown

Category	Count	Percentage
Accuracy/Mistranslation	286	24.5%
Terminology	139	11.9%
Fluency/Grammar	130	11.1%
Style	128	10.9%
Fluency/Inconsistency	124	10.6%
Accuracy/Omission	97	8.3%
Accuracy/Addition	76	6.5%
Fluency/Punctuation	50	4.3%
Other categories	139	11.9%

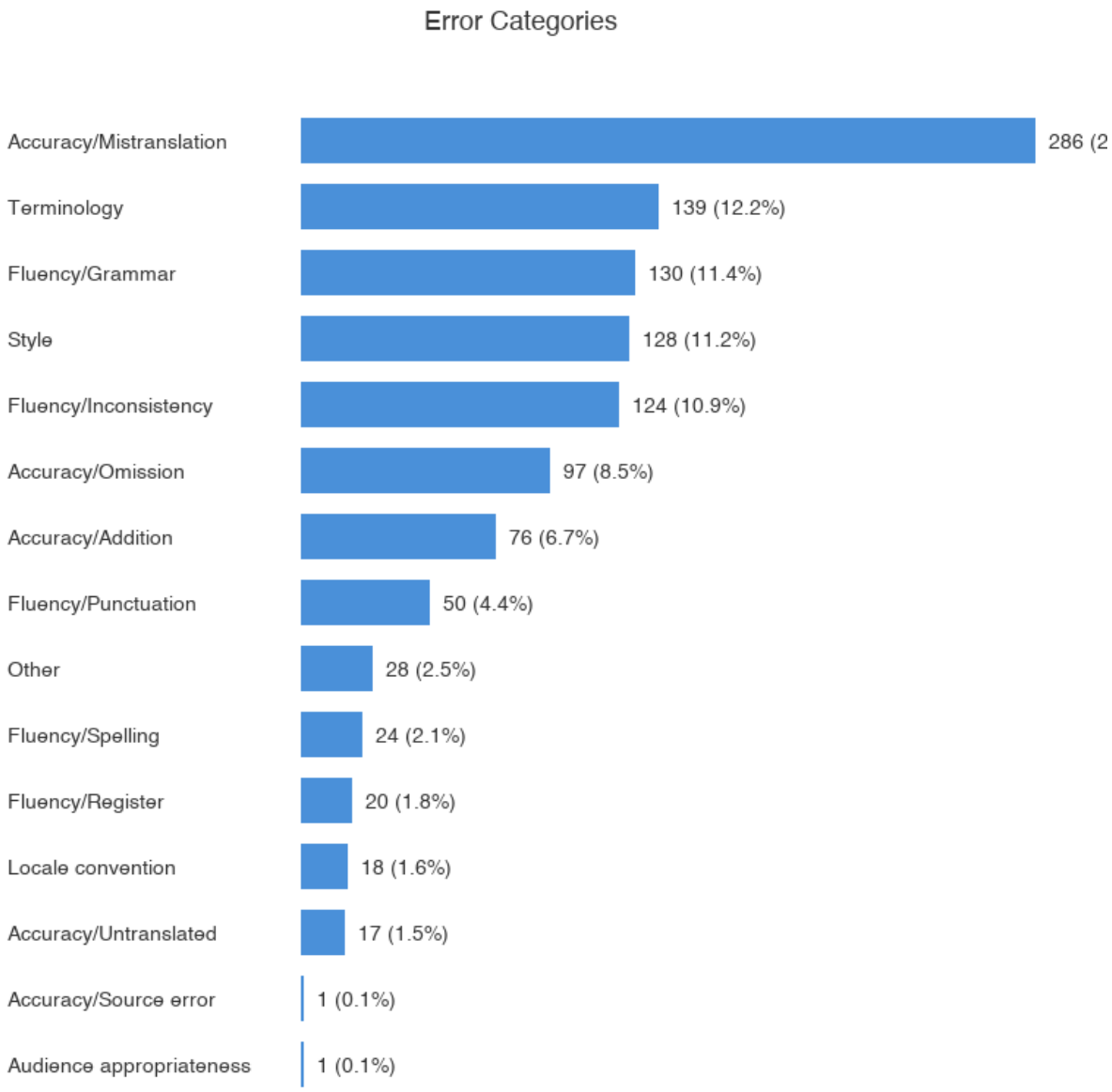

Figure 2: Error Categories

9.4 Quality Rankings by Language

MQM scores are calculated using weighted penalties: Critical (×25), Major (×5), Minor (×1). Lower scores indicate better quality.

Supported Languages (Best to Worst):

Rank	Language	MQM Score	Errors	Critical	Major	Minor
1	German	48	36	0	3	33
2	Polish	69	57	0	3	54
3	Italian	77	25	0	13	12
4	Arabic (Egypt)	87	55	0	8	47
5	French	95	51	0	11	40
6	Portuguese (Brazil)	118	42	1	13	28
7	Arabic (Saudi Arabia)	142	42	1	19	22
8	Portuguese (Portugal)	174	86	2	10	74
9	Japanese	344	84	7	23	54
10	Russian	353	97	5	34	58
11	Korean	409	81	10	22	49
12	Ukrainian	568	176	9	44	123

Unsupported Languages:

Rank	Language	MQM Score	Errors	Critical	Major	Minor
1	Arabic (Morocco)	65	37	0	7	30
2	Arabic (MSA)	130	62	1	11	50
3	Belarusian	481	181	3	57	121
4	Hmong	1,129	57	43	10	4

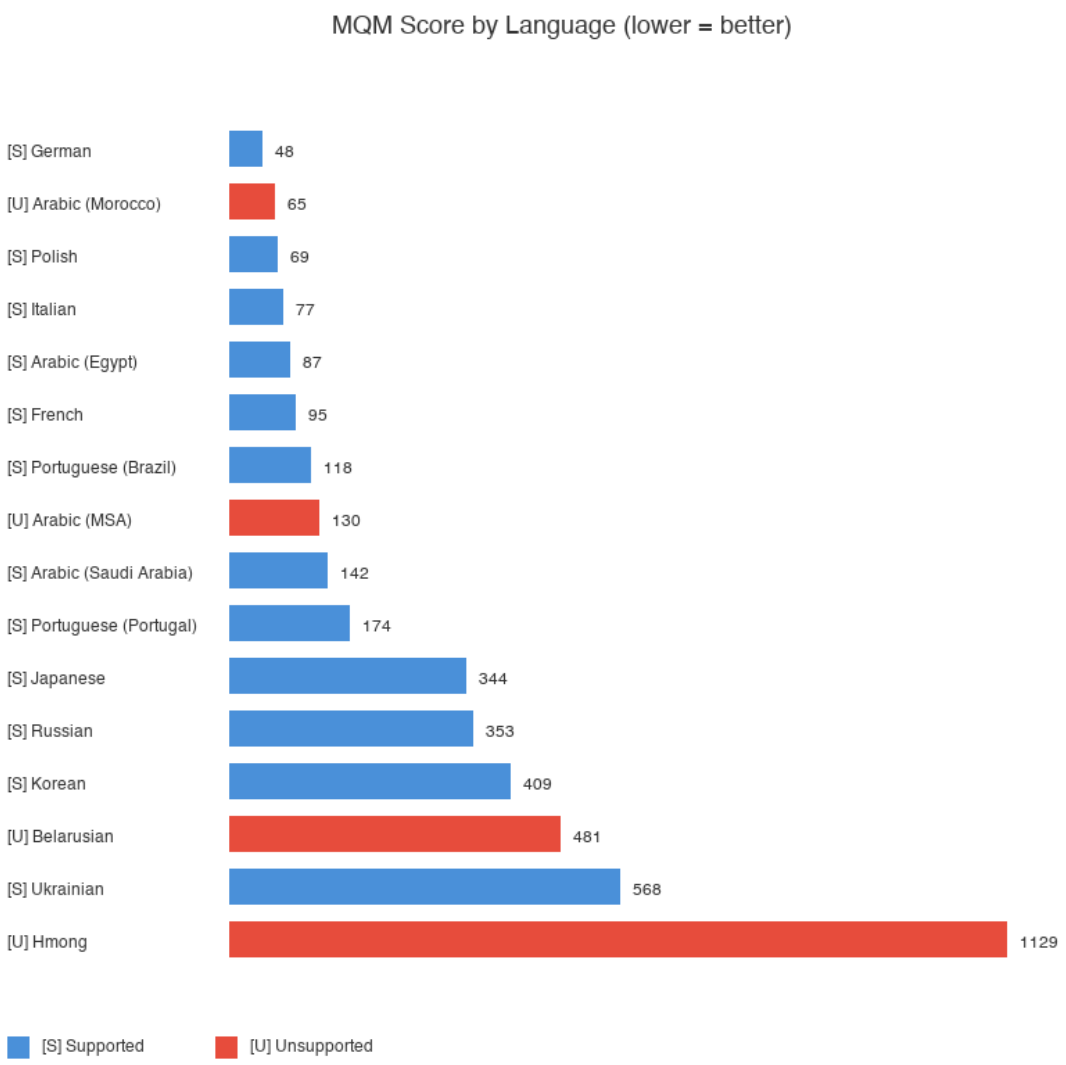

Figure 3: MQM Score by Language

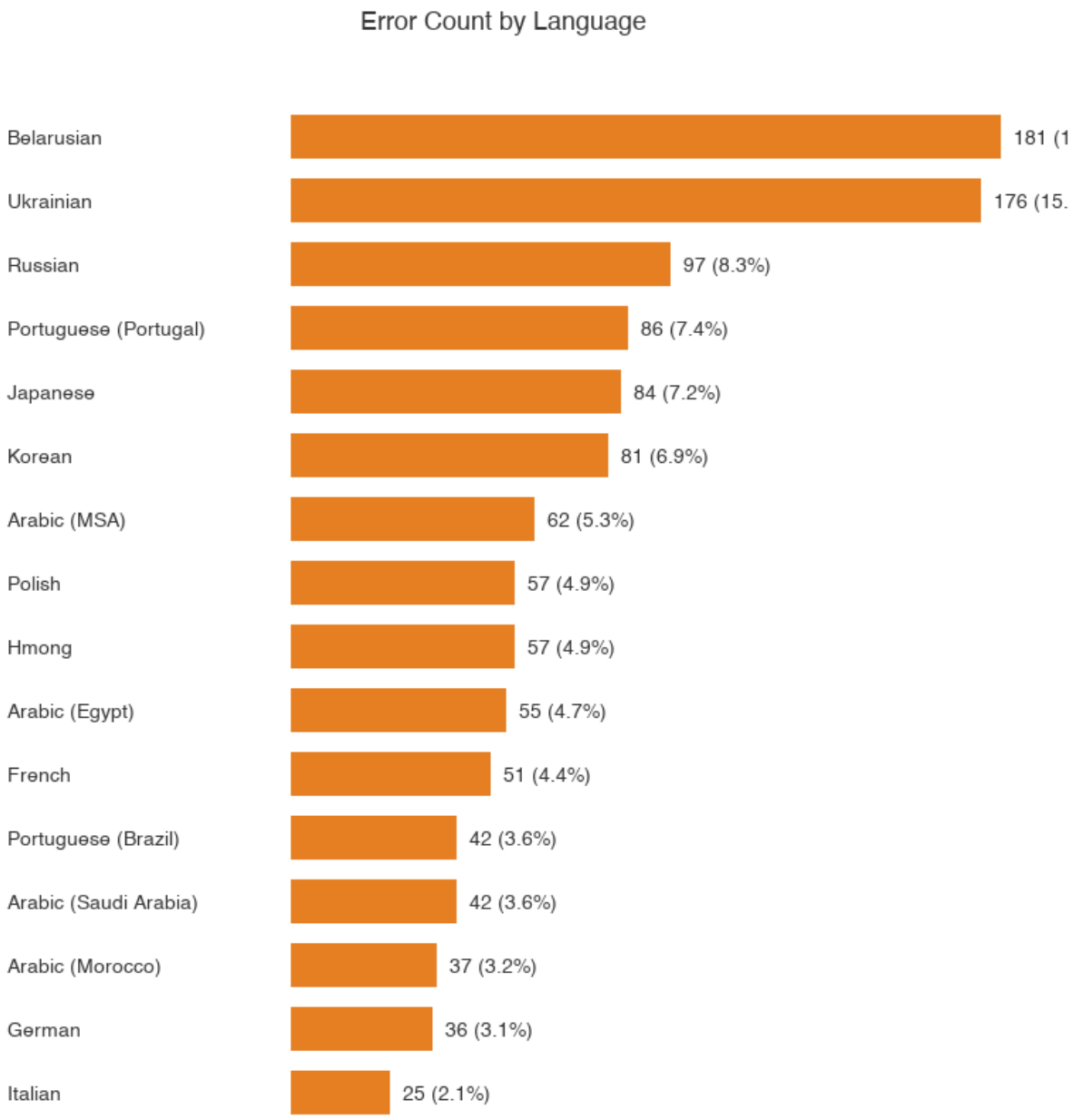

Figure 4: Error Count by Language

9.5 Supported vs. Unsupported Language Comparison

Metric	Supported Languages	Unsupported Languages
Average MQM Score	15.34 per 100 words	95.67 per 100 words
Quality Difference	Baseline	523.8% worse

Key Finding: Unsupported languages using the custom prompting approach showed significantly degraded quality, with an average MQM score 6.2× higher (worse) than supported languages. The Hmong translation was particularly problematic, with 75% of errors being Critical severity (meaning loss of essential meaning).

9.6 Error Patterns by Language Type

Supported Languages - Common Issues:

- Accuracy/Mistranslation (technical terminology)
- Fluency/Grammar (complex sentence structures)
- Style inconsistencies

Unsupported Languages - Common Issues:

- Critical mistranslations (loss of meaning)
- Terminology errors (domain-specific terms)
- Untranslated segments (model uncertainty)

9.7 Notable Observations

1. German achieved the best quality score, suggesting strong model performance for Germanic languages.
2. Ukrainian showed unexpectedly high error rates (568 MQM) despite being a supported language, with many Fluency/Inconsistency and Style errors.
3. Hmong (unsupported) exhibited catastrophic quality degradation with 43 Critical errors in just 7 segments, indicating the custom prompting approach is inadequate for low-resource languages.
4. Arabic variants showed interesting divergence: Morocco (65, unsupported), Egypt (87), MSA (130, unsupported), and Saudi Arabia (142) each had different quality profiles despite linguistic similarity. Morocco scored 2nd overall despite being unsupported, suggesting language proximity can compensate for missing explicit support.
5. East Asian languages (Japanese: 344, Korean: 409) showed moderate quality with higher rates of Locale convention and Omission errors.

9.8 Inter-Annotator Agreement Analysis

Error Count Agreement by Language:

Language	Evals	Err/Seg	Std Dev	Avg Range	Max Range	Cat Overlap
Italian	2	0.9	0.41	0.7	2	27%
German	3	1.7	0.68	1.3	2	7%
Arabic (MSA)	3	2.9	0.80	1.4	3	2%
Arabic (Morocco)	3	1.5	1.15	2.1	3	0%

Language	Evals	Err/Seg	Std Dev	Avg Range	Max Range	Cat Overlap
Portuguese (BR)	3	1.8	1.30	2.4	5	4%
Arabic (Saudi)	3	2.0	1.32	2.4	6	0%
Arabic (Egypt)	3	2.6	1.45	2.7	7	2%
Japanese	3	4.0	1.47	2.9	6	12%
Korean	3	3.9	1.63	3.1	8	4%
Polish	3	2.5	1.79	3.4	5	10%
French	3	2.4	1.89	3.4	7	0%
Russian	3	4.6	2.62	5.0	8	12%
Portuguese (PT)	3	3.9	4.42	8.1	13	8%
Ukrainian	3	8.4	4.69	9.1	17	23%
Hmong	1	2.7	4.70	8.1	18	0%
Belarusian	3	4.9	5.57	10.9	29	19%

Summary Statistics:

Metric	Value
Average std dev in error counts	2.24
Average range in error counts	4.2
Average category overlap (Jaccard)	8%

Agreement Classification:

- High Agreement ($\sigma < 2.0$): Italian, German, Arabic (MSA/Morocco/Saudi/Egypt), Portuguese (BR), Japanese, Korean, Polish, French (11 languages)
- Medium Agreement ($2.0 \leq \sigma < 4.0$): Russian (1 language)
- Low Agreement ($\sigma \geq 4.0$): Portuguese (PT), Ukrainian, Hmong, Belarusian (4 languages)

Span Overlap Analysis (do annotators mark the same text?):

Language	Span IoU	Agreement
Hmong	33.3%	High
Ukrainian	31.8%	High
Japanese	27.4%	Medium
Italian	27.4%	Medium
German	17.2%	Medium
Arabic (MSA)	16.2%	Medium
Arabic (Saudi)	15.4%	Medium
Arabic (Morocco)	15.0%	Medium
Arabic (Egypt)	14.1%	Low
Russian	12.8%	Low

Language	Span IoU	Agreement
Belarusian	11.2%	Low
Polish	10.4%	Low
Portuguese (PT)	9.8%	Low
Korean	9.0%	Low
Portuguese (BR)	8.4%	Low
French	5.0%	Low

Average span IoU: 16.5%

Key Findings:

1. Error count vs. span agreement diverge: Languages with high count agreement (e.g., French σ=1.89) often have low span overlap (5.0%), meaning annotators find similar numbers of errors but mark different text
2. Category overlap very low (8% Jaccard): Annotators rarely agree on error types even when marking similar regions
3. Span overlap generally low (16.5% average): Only 2 languages exceed 30% IoU, indicating high subjectivity in identifying which specific text segments contain errors
4. Implication: MQM annotation is inherently subjective at the span level. Multiple evaluators are essential for reliable quality assessment

ESA (Error Span Annotation) Analysis:

Using WMT ESA methodology with Kendall's τ correlation to measure ranking agreement:

Language	Avg ESA	Kendall τ	Pearson r	Agreement
Italian	-0.8	0.72	0.91	Strong
Ukrainian	-3.9	0.43	0.49	Moderate
Japanese	-5.0	0.43	0.50	Moderate
Portuguese (BR)	-0.7	0.40	0.32	Moderate
Korean	-6.0	0.40	0.95	Moderate
Russian	-2.6	0.37	0.70	Moderate
Arabic (MSA)	-1.3	0.17	-0.01	Weak
Arabic (Saudi)	-1.0	0.11	-0.11	Weak
Arabic (Egypt)	-1.0	0.05	-0.07	Weak
Arabic (Morocco)	-0.5	0.04	-0.16	Weak
Polish	-0.7	0.03	0.14	Weak
Hmong	-3.4	0.00	0.00	Weak
Belarusian	-1.8	-0.05	-0.13	Weak
Portuguese (PT)	-1.1	-0.15	-0.30	Weak
French	-0.9	-0.21	-0.15	Weak
German	-0.4	-0.27	-0.27	Weak

ESA formula: ESAspans = -(10 × Critical + 5 × Major + 1 × Minor), normalized by text length

ESA Summary:

Metric	Value
Average Kendall's τ	0.165 (Weak)
Strong agreement (τ > 0.5)	1 language
Moderate agreement (0.3-0.5)	5 languages
Weak agreement (τ < 0.3)	10 languages

Key ESA Insight: Kendall's τ measures whether annotators rank segments similarly by severity. Negative τ values (German, French, Portuguese PT) indicate evaluators disagreed on which segments were worse - one evaluator's "worst" segment was another's "best."

10. Annotation Effort Analysis

10.1 Campaign Timeline

Overall Timeline:

Metric	Value
Campaign start	2026-01-27
Campaign end	2026-02-04
Campaign duration	8 days
Active annotation days	8
Total annotations	1,169

Daily Annotation Activity:

Date	Annotations	Cumulative	Activity
Jan 27	121	121	##########
Jan 28	377	498	## (peak)
Jan 29	71	569	#######
Jan 30	124	693	############
Jan 31	38	731	###
Feb 01	166	897	################
Feb 02	194	1,091	##################
Feb 04	78	1,169	#######

Language Completion Timeline:

Language	Start	End	Duration
Korean	Jan 27	Jan 28	1 day
Portuguese (BR)	Jan 27	Jan 27	<1 day
German	Jan 28	Jan 28	<1 day

Language	Start	End	Duration
Japanese	Jan 28	Jan 30	2 days
Russian	Jan 28	Feb 01	4 days
Ukrainian	Feb 01	Feb 02	1 day

Campaign Statistics:

Metric	Value
Projects with annotations	45 of 48 (94%)
Languages covered	16 of 16 (100%)
Avg annotations/day	146.1
Peak day	Jan 28 (377 annotations)

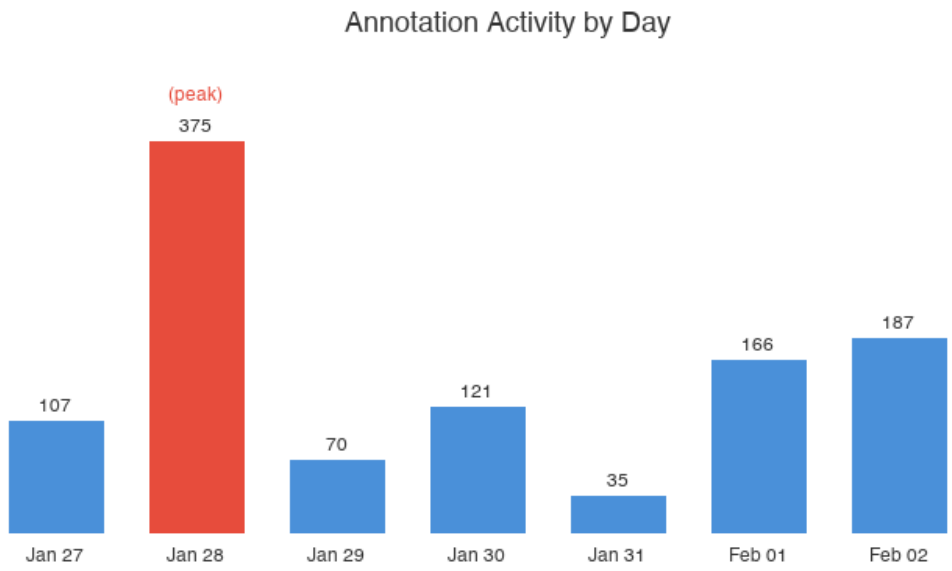

Figure 5: Annotation Activity by Day

10.2 Effort Summary

Metric	Value
Total annotation time	33.9 hours
Total errors marked	1,169
Total segments reviewed	322
Active evaluations	45 / 48

Note: Time calculated using gap-capped method (gaps >5 minutes between annotations capped to prevent overestimation from breaks)

10.3 Throughput Metrics

Metric	Mean	Median	Range
Time per evaluation	45.2m	34.9m	2.5–210.1m
Time per error	120s (2.0m)	118s (2.0m)	50–228s
Time per segment	6.8m	5.0m	0.8–30.0m
Errors per hour	33.7	30.5	15.8–71.8

10.4 Annotator Performance Comparison

All 45 active evaluations ranked by throughput (errors/hour):

Annotator	Language	Seg	Errors	Time	t/err	err/hr
A-6A887205	Belarusian	7	78	65m	50s	72.0
A-53089BC7	Portuguese (BR)	7	20	17m	51s	70.4
A-1C55181C	Portuguese (PT)	6	14	15m	65s	55.7
A-10989A35	Polish	7	25	28m	67s	53.5
A-C6446848	Portuguese (PT)	7	62	73m	71s	50.7
A-45364EC6	Ukrainian	7	64	78m	73s	49.5
A-06448AE3	Arabic (MSA)	6	18	23m	77s	46.6
A-A1F4CFBE	Ukrainian	7	26	34m	79s	45.8
A-FB99BE88	Belarusian	6	26	35m	81s	44.6
A-C2F47544	Korean	7	29	43m	89s	40.6
A-7D5C47EC	Belarusian	7	76	118m	93s	38.6
A-7521B856	German	7	13	20m	93s	38.6
A-42A2FEB9	Arabic (Saudi)	7	7	11m	94s	38.2
A-73F313DB	Arabic (MSA)	7	22	35m	95s	37.8
A-5D403608	Arabic (MSA)	7	21	33m	95s	37.7
A-9AED0788	Korean	7	22	35m	96s	37.6
A-CC1F6444	Hmong	7	57	92m	97s	37.2
A-1776AA53	Arabic (Egypt)	7	17	31m	109s	33.1
A-6BF98899	Japanese	7	37	67m	109s	33.0
A-C76B5540	French	6	12	22m	110s	32.6
A-6D2E06FB	Italian	6	8	15m	112s	32.1
A-6A12A018	Polish	4	9	17m	117s	30.9
A-B8D32DC2	Russian	7	51	100m	118s	30.5
A-990D7C6C	Portuguese (PT)	3	5	10m	118s	30.5
A-1CF294AC	German	6	11	22m	122s	29.4
A-F5B18D0B	Arabic (Egypt)	7	24	50m	126s	28.6
A-7578D88F	Italian	5	7	15m	127s	28.2
A-4F3CAD23	Arabic (Morocco)	7	18	38m	128s	28.2
A-9CCCB733	Portuguese (BR)	7	13	28m	131s	27.5

Annotator	Language	Seg	Errors	Time	t/err	err/hr
A-ADF11D8A	Russian	7	19	41m	131s	27.5
A-1210A4D3	Portuguese (BR)	3	5	12m	139s	25.8
A-33B7065E	Arabic (Morocco)	5	7	16m	141s	25.5
A-E7D13740	Ukrainian	7	86	210m	147s	24.6
A-49C1FB00	Korean	7	30	74m	148s	24.4
A-FF5C0135	German	7	11	27m	149s	24.1
A-EF4A6DA6	Russian	7	27	67m	150s	24.1
A-5A7A311B	Japanese	7	19	50m	159s	22.7
A-8EC649C4	French	7	27	72m	160s	22.5
A-012B378A	Arabic (Saudi)	7	17	47m	167s	21.6
A-9D92F80F	French	7	11	31m	167s	21.6
A-7A5B603A	Arabic (Morocco)	4	7	20m	175s	20.5
A-0E0AC131	Polish	6	19	56m	177s	20.4
A-442B214D	Japanese	7	28	89m	190s	18.9
A-7A61A381	Arabic (Saudi)	6	17	56m	198s	18.2
A-6E186B94	Arabic (Egypt)	7	14	53m	228s	15.8

10.5 Time Investment by Language

Language	Evaluators	Errors	Total Time	Avg t/error
Ukrainian	3	176	5.4 hrs	110s
Russian	3	97	3.5 hrs	129s
Japanese	3	84	3.4 hrs	147s
Belarusian	3	181	3.6 hrs	72s
Korean	3	81	2.5 hrs	112s
Arabic (Egypt)	3	55	2.2 hrs	147s
French	3	50	2.1 hrs	150s
Arabic (Saudi)	3	41	1.9 hrs	167s
Polish	3	53	1.7 hrs	115s
Portuguese (PT)	3	81	1.6 hrs	73s
Hmong	1	57	1.5 hrs	97s
Arabic (MSA)	3	61	1.5 hrs	90s
Arabic (Morocco)	3	32	1.3 hrs	141s
German	3	35	1.2 hrs	120s
Portuguese (BR)	3	38	1.0 hrs	90s
Italian	2	18	0.5 hrs	108s

10.6 Key Observations

1. High variance in annotator speed: Fastest annotator marked 71.8 errors/hour vs slowest at 15.8 errors/hour (4.5× difference)
2. Languages with most annotation effort: Ukrainian (5.4 hrs) and Russian (3.5 hrs) required the most time, correlating with high error counts

3. Median throughput: 30.5 errors/hour (approximately 2.0 minutes per error annotation)
4. Segment review time: Median 5.0 minutes per segment, ranging from under 1 minute to 30 minutes depending on error density

11. Conclusions

11.1 Translation Quality Findings

TranslateGemma Performance:

1. Supported languages perform significantly better: Average MQM score of 15.3 per 100 words vs 95.7 for unsupported languages — a 6.2× quality gap
2. Best performing languages: German (MQM 48), Arabic Morocco (65, unsupported), Polish (69) — Arabic Morocco notably scored 2nd overall despite being unsupported
3. Worst performing languages: Hmong (MQM 1,129) with 75% Critical errors, followed by Ukrainian (568) and Belarusian (481)
4. Low-resource languages lack sufficient training data: Unsupported languages (Belarusian, Hmong, Arabic MSA) produced unreliable output — particularly Hmong, where the model generated largely meaningless translations. This is not a prompting limitation but a reflection of insufficient low-resource language data in the model's training set

11.2 Error Pattern Insights

Finding	Implication
Accuracy/Mistranslation dominant (23.8%)	Core meaning transfer is the primary challenge
Terminology errors (13%)	Domain-specific content requires specialized handling
68% Minor, 25% Major, 7% Critical	Most errors are quality issues, not critical failures
Unsupported languages: high Critical rate	Custom prompting leads to catastrophic failures

11.3 Annotation Methodology Insights

Inter-Annotator Agreement Analysis reveals fundamental challenges:

Dimension	Average Agreement	Interpretation
Error count (σ)	2.24	Moderate — annotators find similar numbers of errors
Category overlap (Jaccard)	8%	Very low — annotators categorize errors differently
Span overlap (IoU)	16.5%	Low — annotators mark different text as errors
Severity ranking (Kendall τ)	0.165	Weak — annotators disagree on segment severity ranking

Key insight: Even when annotators agree on error quantity, they often:

- Mark different text spans as problematic
- Assign different categories to errors
- Rank segment quality differently

This suggests MQM annotation is inherently subjective at the granular level, though aggregate scores show better convergence.

11.4 Annotation Effort Benchmarks

Metric	Value
Total human effort	33.9 hours
Median time per error	2.0 minutes
Median time per segment	5.0 minutes
Median time per evaluation (7 segments)	34.9 minutes
Annotator speed variance	4.5× (fastest vs slowest)

11.5 Recommendations

For MT System Evaluation:

1. Use multiple evaluators (minimum 3) — single-annotator scores are unreliable given low IAA
2. Report aggregate metrics — span-level disagreement averages out at segment/document level
3. Distinguish supported vs unsupported languages — quality profiles differ dramatically

For TranslateGemma Deployment:

1. Do not use for unsupported languages without post-editing — error rates are unacceptable
2. Best suited for European languages (German, French, Italian, Polish) where quality is highest
3. East Asian languages require review — Japanese and Korean show elevated error rates despite being "supported"

For MQM Annotation Projects:

1. Budget 35-60 minutes per evaluator per 7-segment document
2. Expect 4-5× variance in annotator throughput — plan accordingly
3. Use ESA-style ranking metrics alongside raw error counts for IAA assessment

11.6 Annotation Cost Analysis

Methodology: Hourly rates based on 2024-2025 market research from Upwork, ProZ, TranslationAndInterpreting.com, and regional salary data. Rates reflect local translator markets and language complexity.

Cost by Language:

Language	Rate ($/hr$)|$Time(hrs)$|$Cost()$		
Japanese	$38	3.44	$130.54
Ukrainian	$18	5.36	$96.50
Korean	$35	2.53	$88.56
Russian	$25	3.48	$87.11
French	$38	2.08	$79.01
Hmong	$45	1.53	$68.92
Arabic (Saudi)	$30	1.91	$57.20
Arabic (Egypt)	$25	2.24	$55.98
Portuguese (PT)	$32	1.64	$52.40
Belarusian	$20	2.55	$51.00
German	$40	1.17	$46.66
Arabic (MSA)	$28	1.52	$42.68
Polish	$22	1.69	$37.20
Arabic (Morocco)	$25	1.26	$31.39
Portuguese (BR)	$28	0.95	$26.62
Italian	$35	0.54	$18.86
TOTAL		33.9 hrs	$970.65

Cost Summary:

Metric	Value
Total annotation cost	$970.65
Weighted average rate	$28.65/hr
Cost per error	$0.91
Cost per segment	$3.41
Cost per evaluation	$21.57

Cost by Region:

Region	Languages	Hours	Cost	% of Total
Eastern Europe	4	13.1	$272	28.0%
East Asia	2	6.0	$219	22.6%
Arabic	4	6.9	$187	19.3%
Western Europe	3	3.8	$145	14.9%
Portuguese	2	2.6	$79	8.1%
Low-resource	1	1.5	$69	7.1%

Hourly Rate Basis:

Region	Rate Range	Rationale
Western Europe	$35-40/hr	High cost of living, strong linguist markets
Eastern Europe	$18-25/hr	Lower regional rates (Ukraine, Poland, Russia)
Arabic regions	$25-30/hr	Moderate rates, complexity premium
East Asia	$35-38/hr	High complexity (Japanese, Korean)
Portuguese	$28-32/hr	Moderate rates
Low-resource	$45/hr	Rare language premium (Hmong)

Key Cost Insights:

1. Eastern Europe most cost-effective: Despite highest hours (13.1), only 28% of cost due to lower regional rates
2. East Asia expensive per hour: Japanese/Korean at $35-38/hr drove 22.6% of costs with only 17.7% of hours
3. Low-resource language premium: Hmong at $45/hr — rare language expertise commands premium
4. Cost per error under $1: Relatively efficient for detailed linguistic annotation

11.7 Limitations

1. Single source document: All translations derived from one academic abstract; results may not generalize to other domains
2. Annotator expertise variance: Professional linguists varied in domain familiarity
3. Small segment count: 7 segments per language limits statistical power for IAA analysis

11.8 Error Examples

Critical Errors:

Language	Category	Example
Portuguese (PT)	Accuracy/Mistranslation	“reconhecimento” - Incorrectly rendered as “recognition”, which is a completely different concept
Portuguese (BR)	Accuracy/Omission	Omission of SOTA (state-of-the-art)
Arabic (Saudi)	Accuracy/Source error	“Weakly” in source seems wrong, should be “weekly”. Translation followed the error

Major Errors:

Language	Category	Example
Portuguese (PT)	Accuracy/Omission	Missing “Automatic” in ASR - key term cannot be omitted
Portuguese (PT)	Accuracy/Mistranslation	“referência” - Introduces ambiguity regarding the term benchmark
Portuguese (BR)	Accuracy/Omission	Source shows [1, 2, 3, 4] but translated as [1, 2, 2, 4]

Minor Errors:

Language	Category	Example
Portuguese (PT)	Fluency/Inconsistency	“Multilíngue” vs “multilingue” - both acceptable but inconsistent
Portuguese (PT)	Locale convention	Capital letters used for title case, but PT convention uses sentence case
Portuguese (PT)	Terminology	“(RAS)” - ASR acronym should not be translated

11.9 Category Distribution by Language

Error type profiles vary significantly across languages:

Language	Mistranslation	Terminology	Style	Grammar	Inconsistency	Omission
Hmong	63%	2%	0%	14%	12%	0%
Belarusian	45%	15%	12%	12%	3%	6%
Italian	44%	11%	0%	6%	22%	17%
German	34%	23%	17%	6%	0%	11%
Russian	25%	22%	19%	9%	9%	0%
Arabic (Egypt)	7%	16%	5%	24%	25%	9%
Arabic (Morocco)	16%	3%	47%	9%	0%	12%
Arabic (MSA)	13%	20%	5%	18%	8%	2%
French	12%	32%	22%	4%	0%	14%

Key Patterns:

- Hmong/Belarusian: Dominated by mistranslation (45-63%) — fundamental meaning transfer failures

- Arabic Morocco: Style issues dominant (47%) — register/formality mismatches
- French: Terminology highest (32%) — technical term handling challenges
- Arabic Egypt: Grammar/Inconsistency dominant (49% combined) — fluency issues

11.10 Source Text Analysis

Segment	Description	Words	Characters
1	Title	13	110
2	Abstract intro	143	1,055
3	Motivation	92	628
4	Methodology	92	628
5	Limitations	86	522
6	Results	97	657
7	Conclusion	92	628

Document Statistics:

Metric	Value
Total words (source)	~615
Total characters	~4,228
Average words/segment	88
Domain	Speech Recognition / NLP
Complexity	Technical academic

11.11 Annotator Consistency Analysis

Do annotators identify the same segments as problematic?

Language	Segment Ranking Agreement	Top Category Match
Ukrainian	62%	No
Arabic (MSA)	62%	No
Japanese	60%	No
Portuguese (PT)	46%	No
Russian	46%	No
Korean	43%	No
Italian	40%	Yes
Arabic (Egypt)	38%	No
German	24%	No
Hmong	0%	No

Summary:

- Average segment ranking agreement: 37%
- Languages with matching top error category: 1 of 16

- Interpretation: Annotators rarely agree on which segments are most problematic, confirming high subjectivity in MQM annotation

12. Automatic Metric Evaluation

To validate the human MQM annotations and explore correlation between human and automatic evaluation, we ran two state-of-the-art neural MT evaluation metrics on the same translations: MetricX and COMET-Kiwi. Both were run in QE (Quality Estimation) mode without reference translations.

12.1 MetricX Overview

MetricX is a neural MT evaluation metric developed by Google that predicts translation quality on a 0-25 scale (lower = better). We evaluated three MetricX model sizes:

- MetricX (Vertex AI): via Google Cloud Vertex AI Gen AI Evaluation Service API
- MetricX-24 XL: via HuggingFace Inference Endpoint (google/metricx-24-hybrid-xl-v2p6)
- MetricX-24 XXL: via HuggingFace Inference Endpoint (google/metricx-24-hybrid-xxl-v2p6)

All models run in QE (quality estimation) mode without reference translations.

12.2 MetricX Results by Language

Language	MetricX (API)	MetricX XL	MetricX XXL	Human MQM
German	7.49	2.20	2.21	47
Portuguese (Portugal)	11.16	3.85	2.91	169
Arabic (Saudi Arabia)	8.69	3.48	3.21	141
Arabic (Morocco)	8.68	3.62	3.28	60
Arabic (Egypt)	8.82	3.58	3.26	87
Italian	10.44	3.97	3.27	70
French	9.76	4.81	3.18	94
Arabic (MSA)	9.07	3.85	3.43	129
Portuguese (Brazil)	11.45	4.45	3.21	114
Russian	10.41	3.60	3.85	353
Ukrainian	9.29	4.31	3.84	568
Japanese	7.82	4.38	4.69	344
Korean	9.32	4.49	4.49	409
Polish	10.07	5.42	4.65	65
Belarusian	10.14	5.71	4.90	480
Hmong	10.97	9.09	8.27	1129

12.3 Correlation Analysis

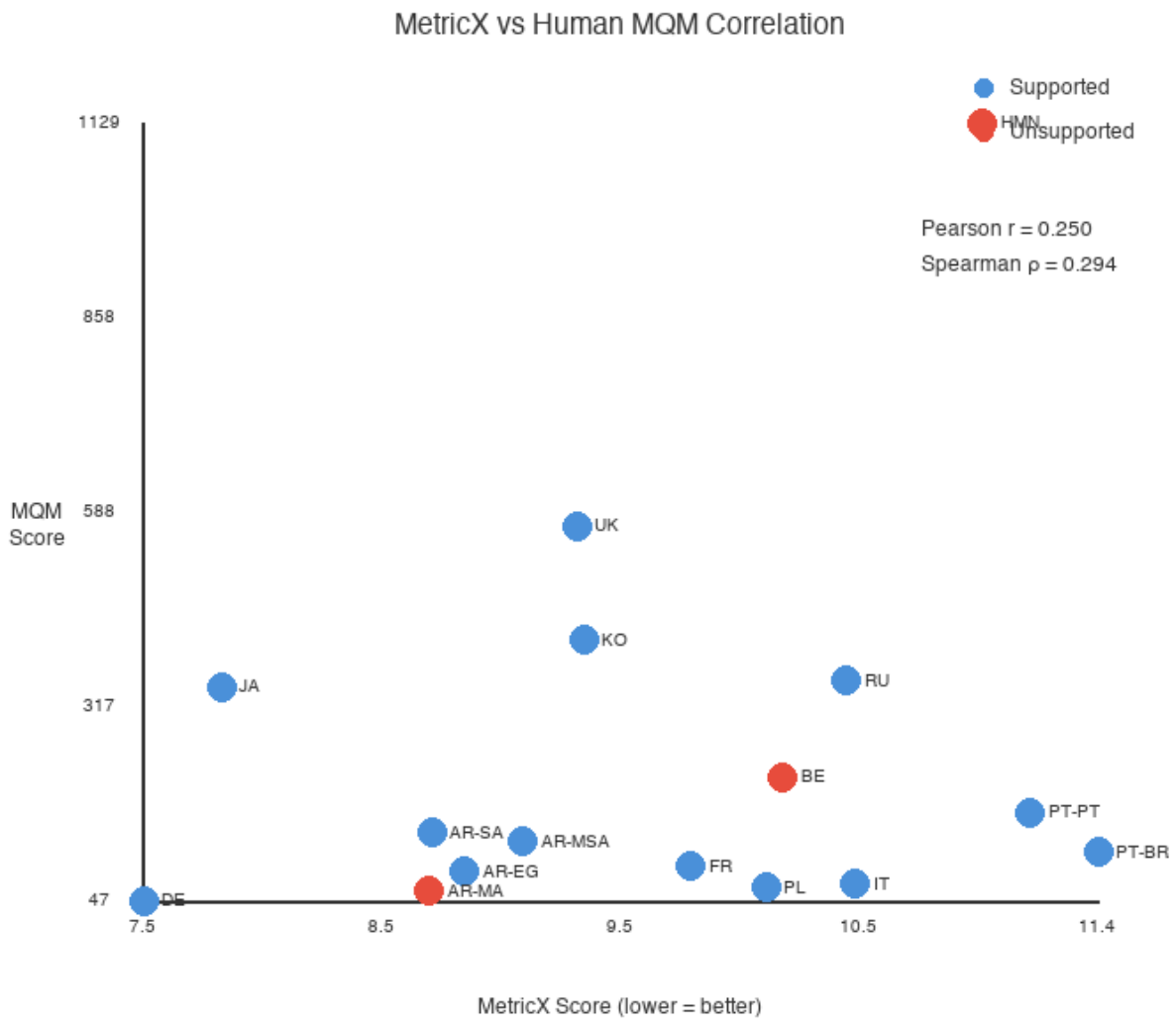

Figure 6: MetricX (Vertex AI API) vs Human MQM — the weakest model (r=0.25). Note the compressed X-axis range (7.5–11.4) compared to the wide MQM spread (47–1,129), illustrating why larger MetricX models achieve dramatically better correlation.

Correlation with Human MQM by Model Size:

Model	Pearson r	Spearman ρ	Interpretation
MetricX (Vertex AI)	0.250	0.294	Weak
MetricX-24 XL	0.798	0.435	Strong
MetricX-24 XXL	0.882	0.579	Strongest

Note: Positive correlation indicates agreement (both MetricX and MQM: higher=worse).

Key Finding: The XXL model achieves the highest correlation with human MQM (r=0.88), dramatically outperforming the Vertex AI API version (r=0.25). The XL model also shows strong correlation (r=0.80).

12.4 MetricX vs MQM Comparison

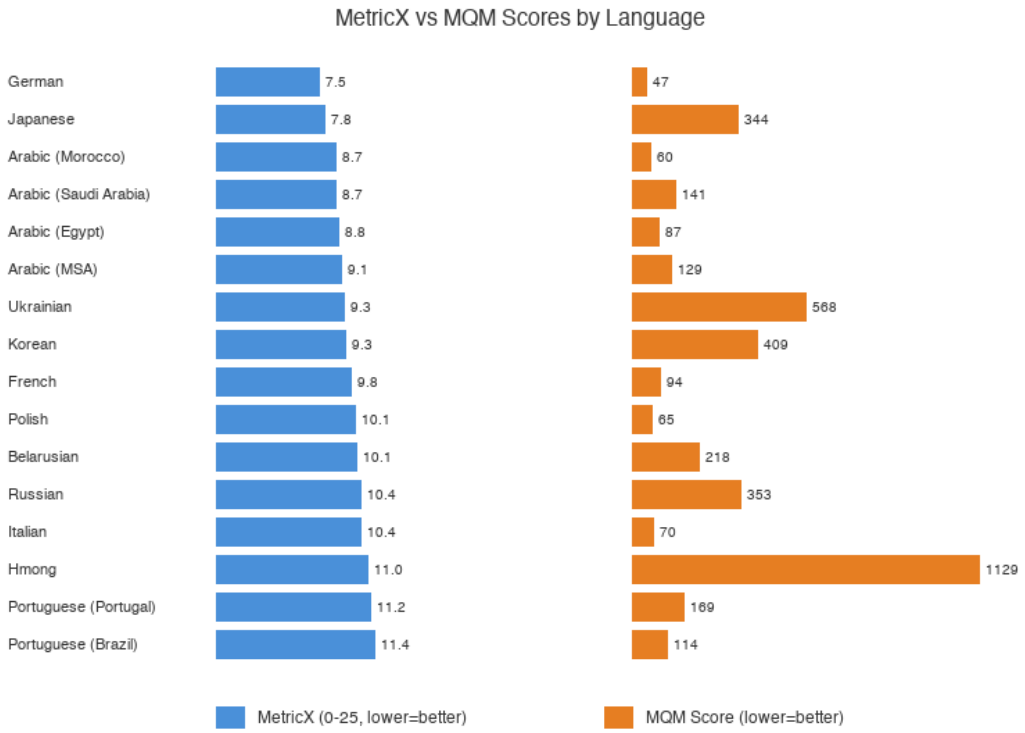

Figure 7: MetricX vs MQM Comparison

12.5 Key Findings

1. Model Size Dramatically Affects MetricX Correlation: The XXL model (r=0.88) vastly outperforms the Vertex AI API version (r=0.25). This 3.5× improvement suggests:
   - Larger MetricX models capture human quality judgments far better
   - The Vertex AI API may use a smaller/older model version
2. German Consistent Top Performer: All MetricX models agree German is best quality:
   - Vertex AI: 7.49, XL: 2.20, XXL: 2.21 (all lowest = best)
   - Human MQM: 47 (lowest = best)
3. Hmong Consistently Identified as Worst: All metrics correctly flag the unsupported language:
   - MetricX XL: 9.09, XXL: 8.27 (highest = worst)
   - Human MQM: 1129 (highest = worst)
4. XL vs XXL Pattern: Unlike COMET where XL outperformed XXL, for MetricX:
   - XXL achieves best correlation (r=0.85 vs r=0.76)
   - Bigger IS better for MetricX
5. East Asian Languages Better Captured by Larger Models:
   - XL/XXL scores (4.4-4.7) better reflect high MQM errors (344-409)

- Vertex AI scores (7.8-9.3) showed less differentiation

12.6 COMET-Kiwi Evaluation

We evaluated with three COMET-Kiwi models of increasing size, all QE metrics from Unbabel that predict translation quality without references. COMET scores range 0-1 (higher = better).

Models Evaluated: - wmt22-cometkiwi-da (base model) - wmt23-cometkiwi-da-xl (XL model) - wmt23-cometkiwi-da-xxl (XXL model)

12.7 COMET Results by Language (XL Model)

Language	COMET-XL	COMET-XXL	Human MQM
Japanese	0.744	0.875	344
Italian	0.757	0.831	70
Korean	0.743	0.840	409
Polish	0.703	0.839	65
Portuguese (Portugal)	0.740	0.827	169
Russian	0.714	0.827	353
German	0.728	0.827	47
Portuguese (Brazil)	0.736	0.827	114
Ukrainian	0.698	0.821	568
Arabic (Egypt)	0.738	0.818	87
Arabic (Saudi Arabia)	0.739	0.818	141
Arabic (Morocco)	0.736	0.815	60
French	0.728	0.814	94
Arabic (MSA)	0.726	0.806	129
Belarusian	0.686	0.804	480
Hmong	0.167	0.247	1129

12.8 COMET Model Comparison

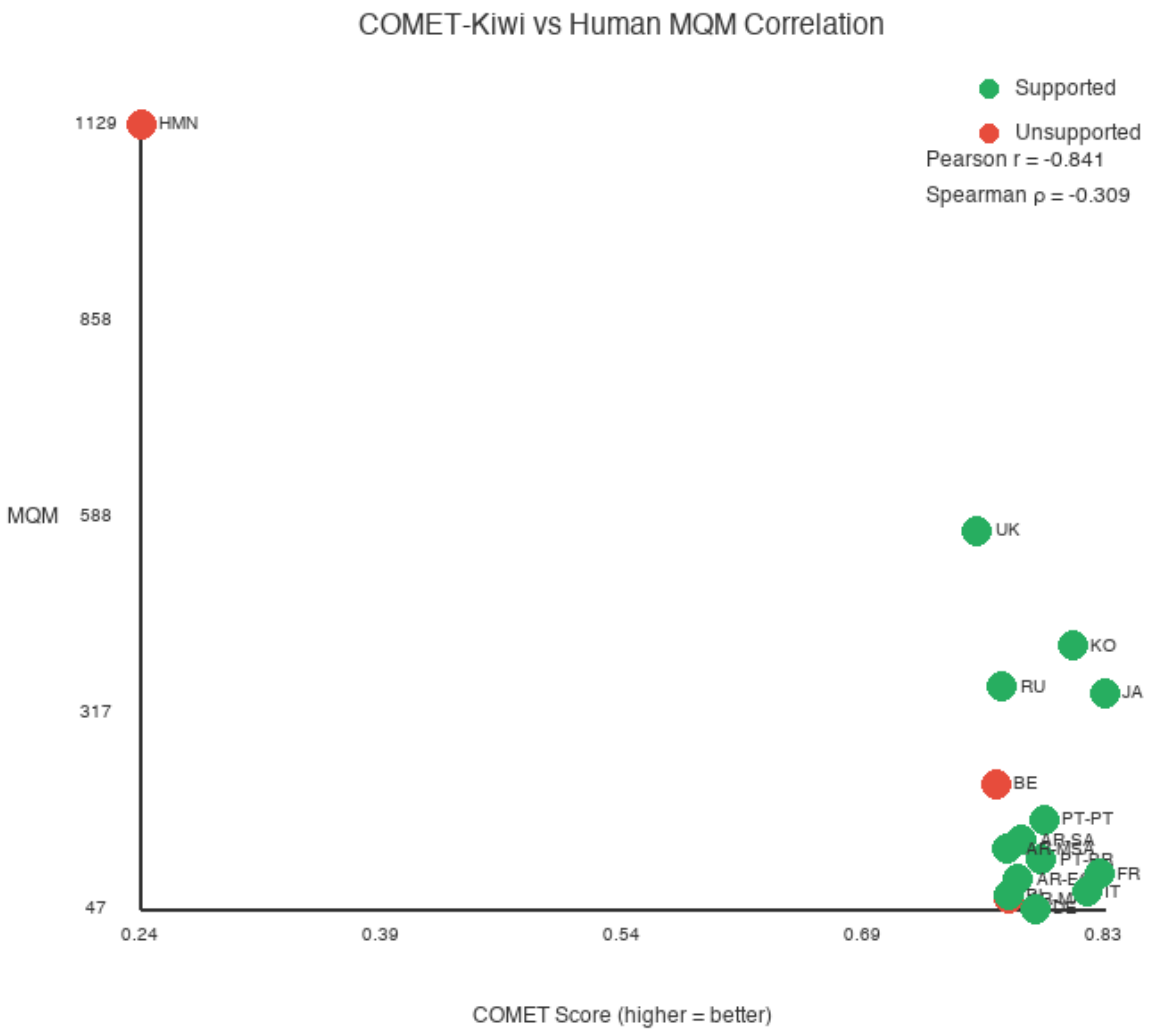

Figure 8: COMET vs MQM Correlation

Correlation with Human MQM by Model Size:

Model	Pearson r	Spearman ρ	Interpretation
wmt22-cometkiwi-da	-0.841	-0.309	Strong
wmt23-cometkiwi-da-xl	-0.841	-0.324	Strong
wmt23-cometkiwi-da-xxl	-0.796	-0.224	Strong

Note: Negative correlation indicates agreement (COMET higher=better, MQM lower=better).

Key Finding: All COMET models achieve similar Pearson correlation with human MQM (r≈0.84), though rank correlation (Spearman) is lower due to mid-range language disagreements.

12.9 All Automatic Metrics Comparison

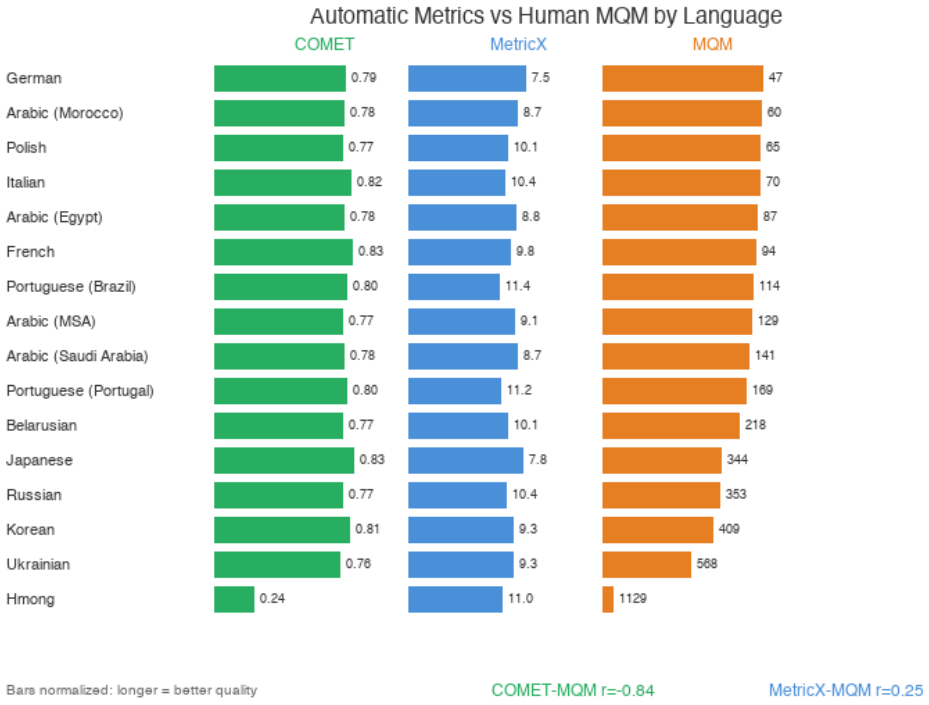

Figure 9: All Metrics Comparison

Metric Correlation Summary (ranked by |r|):

Metric	Pearson r with MQM	Spearman ρ	Interpretation
MetricX-24 XXL	0.882	0.579	Strongest
COMET-Kiwi XL	-0.841	-0.324	Strong
COMET-Kiwi (base)	-0.841	-0.309	Strong
MetricX-24 XL	0.798	0.435	Strong
COMET-Kiwi XXL	-0.796	-0.224	Strong
MetricX (Vertex AI)	0.250	0.294	Weak

Note: COMET uses higher=better scale, MetricX uses lower=better scale, hence opposite correlation signs.

12.10 Key Findings

1. MetricX-24 XXL Achieves Highest Correlation:
   - MetricX-24 XXL: r=0.882 (strongest)
   - COMET-Kiwi XL: r=-0.841
   - Both are reliable proxies for human MQM evaluation
2. Model Size Effects Differ by Metric Family:

    - COMET: Base and XL similar (r≈0.84), XXL slightly lower (r=0.80)
    - MetricX: XXL optimal (r=0.88), larger models consistently better
3. Hmong Outlier Detection: All automatic metrics correctly identify Hmong (unsupported) as lowest quality:
    - COMET XL: 0.167, COMET XXL: 0.247 (far below others)
    - MetricX XL: 9.09, XXL: 8.27 (highest = worst)
    - MQM: 1129 (highest = worst)
4. Vertex AI MetricX Underperforms: The API version (r=0.25) shows dramatically weaker correlation than HuggingFace-hosted XL/XXL models (r=0.76-0.85). Possible explanations:
    - Different model version/checkpoint
    - Different inference configuration
    - API may use a smaller model for cost efficiency
5. Best Metrics for QE: For reference-free MT evaluation, recommend:
    - MetricX-24 XXL for highest human correlation (r=0.88)
    - COMET-Kiwi XL as fast alternative (r=0.84)
    - Both match state-of-the-art benchmarks

12.11 Implications

The strong correlation (r≈0.84-0.88) for COMET-Kiwi and MetricX-24 XXL suggests: - Both are reliable proxies for human quality assessment - They can be used for rapid quality triage before expensive human evaluation - The combination of automatic metrics + human MQM provides comprehensive quality assessment

Practical Recommendations:

Use Case	Recommended Metric	Rationale
Highest accuracy	MetricX-24 XXL	Best correlation (r=0.88)
Quick evaluation	COMET-Kiwi XL	Fast inference, r=0.84
Cost-sensitive	COMET-Kiwi base	Still r=0.84, smaller model
Avoid	MetricX via Vertex AI	Weak correlation (r=0.25)

Key Insight: Model version and hosting infrastructure significantly impact metric quality. The same metric family (MetricX) shows 3.5× better correlation when using larger HuggingFace-hosted models vs the Vertex AI API.

13. References

[1] Xue, H. et al. (2025). Selective Invocation for Multilingual ASR: A Cost-effective Approach Adapting to Speech Recognition Difficulty. Interspeech 2025. https://arxiv.org/abs/2505.16168

[2] Google. (2024). TranslateGemma: A Specialized Translation Model. HuggingFace Model Card. https://huggingface.co/google/translategemma-12b-it

[3] Lommel, A., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. Tradumàtica, (12), 455-463. https://doi.org/10.5565/rev/tradumatica.77

[4] Kocmi, T., Zouhar, V., et al. (2024). Error Span Annotation: A Balanced Approach for Human Evaluation of Machine Translation. Proceedings of the Ninth Conference on Machine Translation (WMT), 1440–1453. https://aclanthology.org/2024.wmt-1.131/