Translation Quality in Multilingual LLM Evaluation

Anonymous ACL submission

Abstract

Machine-translated evaluation benchmarks are widely used to assess the multilingual capabilities of large language models (LLMs). However, translation errors in such benchmarks remain underexplored, raising concerns about the reliability and comparability of multilingual evaluation. This study examines the types of translation errors that occur in benchmark translations and how they affect LLM performance. We analyze five widely used English benchmarks translated into 20 European languages, using a validated LLM-based method to identify span-level translation errors at scale. To assess the impact of these errors, we apply three complementary analyses: comparing model accuracy on corrected vs. erroneous translations, testing statistical associations between error types and model performance, and estimating how strongly they affect model outcomes. Across all methods, meaning-related errors (mistranslations) lead to lower model performance, while other accuracy errors and fluency issues show weaker and more variable effects. Our results motivate translation-aware evaluation practices and enable scalable detection and analysis of translation artifacts.

1 Introduction

In multilingual evaluation, machine-translated datasets are commonly relied on as reference data, yet their translation quality is often overlooked, undermining the reliability and comparability of results (Choenni et al., 2024; Artetxe et al., 2020; Plaza et al., 2024).

Translation quality can be assessed through multiple paradigms (Zhao et al., 2024). Human protocols, such as Direct Assessment (DA), Multidimensional Quality Metrics (MQM; Lommel et al., 2013, 2024), and Error Span Annotation (ESA; Kocmi et al., 2024), set the standard for translation evaluation, offering increasing levels of diagnostic granularity (Freitag et al., 2021).

More recently, researchers have treated LLMs themselves as translation judges ("LLM-as-a-judge"; Kocmi and Federmann, 2023b), using zero- or few-shot prompting to tag MQM-style error spans. This trend is exemplified by GEMBA and GEMBA-ESA (Kocmi and Federmann, 2023a; Kocmi et al., 2024), and by GPT-based evaluators such as AutoMQM (Huang et al., 2024) for inline span detection, or MQM-APE (Lu et al., 2025), which uses automatic post-editing to refine translations.

Prior work investigating the effects of translation artifacts on model performance relies either on manual inspection of small samples (Artetxe et al., 2020; Plaza et al., 2024), which provides qualitative insights but does not scale, or on heuristics (Park et al., 2024; Choenni et al., 2024) such as sentence length ratios or learned quality estimation scores (e.g., COMET-QE (Rei et al., 2020)), both of which lack precision in identifying the type and location of translation errors. In addition, most of these studies are limited to single benchmarks or languages (e.g. Spanish MMLU (Plaza et al., 2024) or XNLI (Artetxe et al., 2020)).

In this work, we bridge these two strands of research by combining automated span-level MQM annotation of machine-translated benchmarks with large-scale analyses that examine both the types of translation errors that occur and how these errors affect the performance of multilingual LLMs. Our study analyzes the EU20 benchmark suite (Thellmann et al., 2024), which comprises five widely used evaluation datasets translated into 20 European languages and covers diverse task types such as logical reasoning, factual knowledge, and truthfulness.

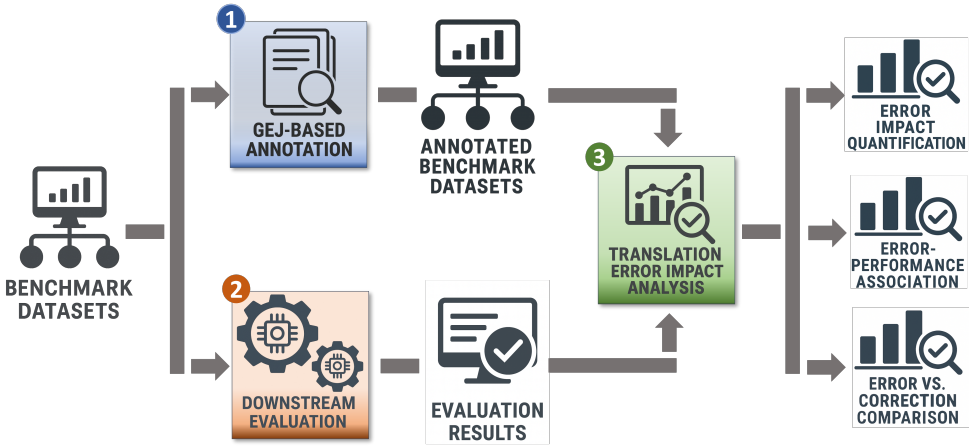

Figure 1: LLM-based TQE pipeline combining span-level MQM annotation and model performance analysis. The workflow involves (1) automatic error annotation (GEJ), (2) downstream evaluation of model accuracy, and (3) targeted analysis of how translation errors affect performance.

Our main contributions are as follows:

1. **Validated LLM-based MQM annotation for multilingual benchmarks:** We introduce GEJ (GEMBA-ESA-JSON), an LLM-as-a-judge method based on Kocmi et al. (2024) with improved prompting and structured span-level MQM outputs. We validate GEJ on Span-ACES (Moghe et al., 2025), a benchmark for span-level error detection, using annotations from four different LLMs acting as GEJ-annotators and assess false-positive plausibility using high-quality FLORES-200 reference translations (Team et al., 2022).

2. **Impact analysis of translation errors on LLM predictions:** Following a three-step pipeline (see Figure 1), we first annotate all 20 language versions of the EU20 suite using GEJ with three LLMs as annotators. We then evaluate eight multilingual LLMs on the EU20 benchmarks to obtain model predictions. Finally, we combine the predictions with the GEJ annotations to analyze how translation errors affect model performance. We focus on key MQM categories such as *Accuracy* and *Fluency*, with particular attention to *mistranslations*. Our analysis includes: (i) comparing model performance on EU20 vs. alternative human-corrected translations, (ii) testing for statistical associations between detected errors and model accuracy, and (iii) estimating how strongly detected errors of different types affect LLM accuracy.

2 Related Work

Translation artifacts and their effects. Several studies have shown that translation artifacts can undermine the reliability of model evaluation: Choenni et al. (2024) found that MT-generated test sets may overestimate model capabilities, especially in low-resource languages; Artetxe et al. (2020) demonstrated that subtle "translationese" can bias cross-lingual benchmarks like XNLI; Plaza et al. (2024) reported that mistranslations in Spanish MMLU data cause 6–13% accuracy loss for GPT-4, with up to 60% of failures directly linked to translation errors; and Park et al. (2024) observed similar effects for VQA models. While these findings underscore the need for rigorous quality control, prior work remains limited in scale and granularity.

Multilingual benchmarks. Recent multilingual benchmarks range from carefully curated, manually translated datasets (e.g., SuperGLEBer (Pfister and Hotho, 2024), ScandEval (Nielsen, 2023), IberoBench (Baucells et al., 2025), FrenchBench (Faysse et al., 2025), BenCzechMark (Fajcik et al., 2025)) to large-scale resources generated via machine translation. While manually constructed benchmarks offer high quality, they are costly and difficult to scale, prompting the use of machine translation for broader coverage (e.g., Global MMLU (Singh et al., 2025), XNLI (Conneau

et al., 2018), OKAPI (Lai et al., 2023), and LAMBADA (Paperno et al., 2016)). However, many such resources lack transparent quality control. Our work advances the field by combining automated span-level error annotation with statistical analysis to assess translation error impact on model performance.

3 GEJ and Meta-Evaluation

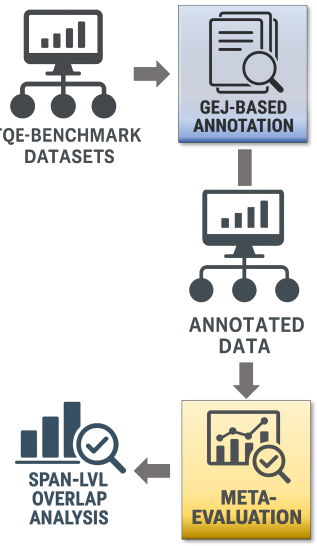

Figure 2: Meta-evaluation: TQE-benchmarks (Span-ACES and FLORES) are annotated by LLMs using GEJ. Span overlap analysis yields recall and F1 scores and false positive rates.

This section introduces two key components of our approach: (1) the GEJ annotator, (2) the meta-evaluation of GEJ, as illustrated in Figure 2.

GEJ is a GEMBA-ESA[1] adaptation, extended with curated multilingual few-shot examples, covering a broad range of error types and both structured content (e.g., multiple-choice questions) and general-purpose text (see Table 5, Appendix A.2). We include GEMBA-ESA as baseline for comparison with GEJ.

We employ four LLMs as GEJ annotators, all prompted identically – GPT-4o-mini[2], DeepSeek R1[3], Llama-4 Scout[4], and Mistral-Large-Instruct-2411[5] – to identify error spans in a few-shot setting, without assigning quality scores. The LLM-annotators are prompted as shown in Figure 4 to produce JSON-structured output containing an array of identified error spans and error types from the MQM categories detailed in Table 7 of Appendix A.1.

L	N	GPT	DeepSeek	LLaMA	Mistral
		B/GEJ	B/GEJ	B/GEJ	B/GEJ
Mistrans.					
DA	24	.12/.58	.00/.71	.25/.42	.71/.33
DE	1044	.15/.20	.00/.41	.14/.09	.29/.21
ES	29	.24/.69	.00/.55	.17/.34	.38/.38
ET	15	.27/.47	.00/.67	.27/.47	.60/.40
FR	225	.12/.16	.00/.21	.12/.05	.16/.07
HU	34	.09/.23	.00/.43	.20/.20	.56/.54
LT	2	1.0/.50	.00/1.0	1.0/.00	.50/.50
NL	23	.09/.35	.00/.39	.22/.13	.35/.48
PL	8	.38/.75	.00/.62	.12/.12	.12/.12
PT	21	.09/.59	.00/.59	.36/.36	.52/.50
RO	33	.06/.30	.00/.82	.09/.12	.64/.42
SK	9	.00/.44	.00/.44	.11/.22	.56/.33
SL	19	.00/.32	.00/.21	.11/.11	.53/.37
SV	21	.00/.38	.00/.67	.10/.14	.48/.33

Table 1: Span-Recall for MQM mistranslation., by model and language, for GEMBA-ESA (B) and GEJ.

For meta-evaluation, we focus on two TQE-benchmark datasets: Span-ACES and FLORES. We employ a subset of Span-ACES [6], a contrastive dataset of 36,476 samples over 146 language pairs, each annotated for 68 error categories. Table 8 in Appendix A.1 summarizes the subset used in our meta-evaluation, which covers MQM *Accuracy* subtypes (e.g., *mistranslation*, *addition*, *over/under-translation*) and *Fluency* subtypes (e.g., *grammar*).

In contrast, WMT introduced error-span subtasks for quality estimation in 2023 (Blain et al., 2023; Zerva et al., 2024), but so far only English–German (2023) provides relevant span-based gold labels, while for English–Spanish (2024) only input data are available. In Span-ACES, only one error type is annotated per sample; any additional errors are not labeled. As a consequence, extra correctly identified errors are counted as false positives, leading to an underestimation of precision and Span-F1. We therefore report Span-Recall as our primary metric, as it measures the recovery of annotated gold spans without being affected by unannotated errors. Formally, Span-Recall is defined as the proportion of reference error spans (including error type labels) that are

[1] github.com/MicrosoftTranslator/GEMBA
[2] platform.openai.com/docs/models/gpt-4o-mini
[3] api-docs.deepseek.com/guides/reasoning_model
[4] HF: meta-llama/Llama-4-Scout-17B-16E-Instruct
[5] HF: mistralai/Mistral-Large-Instruct-2411
[6] github.com/EdinburghNLP/ACES

L	N	GPT	DeepSeek	LLaMA	Mistral
		B/GEJ	B/GEJ	B/GEJ	B/GEJ
Accuracy					
BG	10	.10/.10	.00/.60	.40/.30	.50/.40
CS	356	.00/.01	.00/.61	.01/.01	.01/.31
DA	11	.00/.17	.00/.58	.17/.08	.18/.42
DE	311	.00/.01	.00/.16	.01/.00	.02/.07
EL	6	.00/.14	.00/.43	.14/.14	.50/.43
ES	328	.00/.00	.00/.79	.01/.00	.01/.09
ET	5	.00/.00	.00/.60	.00/.00	.00/.40
FI	6	.00/.17	.00/.83	.00/.00	.67/.50
FR	11	.00/.09	.00/.55	.18/.09	.55/.36
HU	4	.00/.00	.00/.75	.25/.25	.25/.75
IT	5	.00/.40	.20/.80	.40/.20	.20/.40
LT	6	.00/.00	.00/.43	.00/.00	.33/.29
LV	8	.00/.11	.00/.67	.11/.11	.38/.56
NL	15	.07/.00	.00/.47	.20/.00	.33/.20
PL	324	.00/.01	.00/.75	.00/.01	.01/.29
PT	11	.00/.09	.00/.82	.18/.36	.82/.45
RO	10	.00/.10	.00/.80	.20/.10	.50/.60
SK	7	.00/.00	.00/.38	.00/.00	.29/.12
SL	13	.00/.00	.00/.69	.08/.08	.23/.54
SV	13	.08/.08	.00/.77	.08/.15	.38/.46
Fluency					
DE	700	.00/.00	.00/.42	.03/.04	.02/.51

Table 2: Span-Recall by model and language for GEMBA-ESA (B) and GEJ, for MQM: (a) *Accuracy*, (b) *Fluency*.

correctly predicted: Span-Recall $= \frac{|S_{\text{pred}} \cap S_{\text{ref}}|}{|S_{\text{ref}}|}$ where S_{pred} and S_{ref} denote the sets of predicted and reference error spans, respectively, each represented as tuples of (span, error type label).

Since GEJ produces MQM-style tags but Span-ACES uses a different annotation scheme, we mapped Span-ACES labels to MQM categories, considering only error types that unambiguously map to our MQM subset (Appendix A.1, Table 6). To complement Span-ACES, we use FLORES-200[7], a high-quality parallel dataset with about 7,000 samples per language (145,252 total). On FLORES, we estimate false positive rates using the mean sentence-level false positive rate, $FPR_s = S_{\text{err}}/N$, and the mean number of error spans per 1,000 words, $ER_w = (n_{\text{err}}/W) \cdot 1000$, where S_{err} is the number of sentences with at least one predicted error, N is the total number of sentences, n_{err} is the total predicted error spans, and W is the total word count (based on the English reference).

Table 1 and Figure 2 compare Span-Recall for GEMBA-ESA baseline (B) and GEJ across all LLM-annotators and languages. We achieve about 90% extraction rate of about 93% for Span-ACES and 91% for FLORES with GPT. Across nearly all languages and models, GEJ achieves higher Span-Recall than the baseline, with the largest improvements for DeepSeek and Mistral – for example, DeepSeek's recall on German for *mistranslation* rises from .00 to .41, on Polish for *Accuracy* from .00 to .75, and for Mistral on German for *Fluency* from .02 to .51. On average across all languages and models, GEJ achieves a Span-Recall improvement of 0.23 for *mistranslation*, 0.29 for *Accuracy*, and 0.23 for *Fluency* compared to the baseline.

	FPR_s	ER_w
	B/GEJ	B/GEJ
Mistrans.		
GPT	.29/.38	13.82/20.52
DeepSeek	.00/.42	.21/23.88
LLaMA	.42/.14	23.66/6.79
Mistral	.49/.56	34.28/31.53
Accuracy		
GPT	.00/.07	.04/3.20
DeepSeek	.00/.50	.04/28.35
LLaMA	.01/.18	.52/9.37
Mistral	.06/.21	3.29/10.48
Fluency		
GPT	.28/.02	13.16/1.04
DeepSeek	.00/.09	.03/4.34
LLaMA	.06/.00	2.98/.23
Mistral	.17/.12	8.24/6.46

Table 3: Mean FPR_s and ER_w for GEMBA-ESA (B) and GEJ on FLORES, for MQM categories *Accuracy*, *Fluency*, and *mistranslation*.

Table 3 reports mean FPR_s and ER_w for GEMBA-ESA (B) and GEJ on FLORES. Across most categories and models, GEJ exhibits higher FPR_s and ER_w values than the baseline, indicating a greater tendency to label errors in reference translations. On average across all categories and models, GEJ marks reference translations as erroneous about 0.07 more often per sentence (FPR_s) and produces 3.4 more error spans per 1,000 words (ER_w) than the baseline (a 39% relative rise). The significance of this increase should be taken into account when interpreting downstream performance analyses based on error counts.

[7] huggingface.co/datasets/facebook/flores

4 Performance Impact Analysis

To assess the impact of translation errors on LLM performance, we follow the three-step TQE pipeline illustrated in Figure 1.

All EU20 benchmark datasets are automatically annotated for translation errors using GEJ with three state-of-the-art LLMs (GPT-4o-mini, Llama-4 Scout, and Mistral-Large-Instruct-2411; see Sections 3). The benchmark datasets annotated in this process are detailed in Section 4.1.

We evaluated eight recent instruction-tuned multilingual LLMs (see Appendix B.2, Table 12) on the EU20 benchmark datasets using a variant of EleutherAI's LM Evaluation Harness[8] recording binary prediction outcomes (correct/incorrect) for each sample.

Finally, we combine model predictions with GEJ span-level MQM error annotations to analyze the effect of translation errors on model accuracy, as detailed in Sections 4.2, 4.3 and 4.4.

4.1 Benchmark Datasets

To quantify the impact of automatically detected translation errors, we applied GEJ to detect translation errors on translated versions of five LLM benchmark datasets: MMLU (Hendrycks et al., 2021), ARC (Clark et al., 2018), HellaSwag (Zellers et al., 2019), GSM8K (Cobbe et al., 2021), and TruthfulQA (Lin et al., 2022). Translations were sourced from EU20 (Thellmann et al., 2024), translated via DeepL[9] into the 20 official EU languages supported by DeepL, and Global-MMLU (Singh et al., 2025), combining Google Translate[10] with professional and community post-editing across 42 languages. Global-MMLU features high-quality translations for Spanish, French, German, and Italian, and community translations for Czech and Romanian. We excluded Portuguese as it is of the Brazilian variety and our focus is on European languages.

Before applying GEJ, we verified the structural integrity of all translated datasets by aligning each sample with its English original and checking for completeness, split and subset assignment[11], and consistency of correct answers across languages. Samples with missing or incomplete translations in one or more of the languages under investigation were excluded from our analysis. The results of these checks are summarized in Appendix B.1 in Table 10, while the final cleaned dataset statistics are shown in Table 11.

4.2 Error vs. Correction Comparison

In our first analysis, we leverage of the fact that Global-MMLU and EU20-MMLU are alternative translations of the same dataset. Specifically, we match EU20-MMLU samples where GEJ, with GPT-4o-mini as annotator LLM, detected errors of type *mistranslation* but not other types of errors, with the corresponding Global-MMLU samples in languages with human-reviewed translations (see Dataset Section 4.1), if GEJ detected no errors in the latter. We denote the total number of such samples N_C (for "corrected", since the EU20 version is considered incorrect and the Global-MMLU version is correct according to GEJ).

We construct cross-tabulations of the binary Global-MMLU and EU20-MMLU downstream evaluation outcomes (correct/incorrect) per model, focusing on differences in prediction correctness. For this purpose, we consider the the number N_G (for *Gain*) of samples where the model is incorrect on EU20-MMLU but correct on Global-MMLU and the converse number N_L (for *Loss*) of samples where the model is incorrect on Global-MMLU but correct on EU20-MMLU. Thus, $N_A = N_L + N_G$ (for *Affected*) is the number of samples for which the prediction correctness differs between the two translations. To measure the effect of translation corrections on model accuracy, we define the *Net Accuracy Impact* (NAI) as the difference between the proportions of losses and gains among the N_C "corrected" samples: $\text{NAI} = \frac{N_L - N_G}{N_C}$.

A negative NAI potentially indicates that model performance suffers under incorrect translations, in line with the expectation that *mistranslations* reduce model performance.

The results of this analysis are presented succinctly in Table 4 and more elaborately in Table 13 in Appendix B.3. The comparison across six languages and eight models reveals

[8] github.com/EleutherAI/lm-evaluation-harness
[9] developers.deepl.com/docs
[10] cloud.google.com/translate/docs/reference/api-overview
[11] Automated when unique IDs are available.

#Affected (NAI)												
Lang.	CS		DE		ES		FR		IT		RO	
N_C	132		104		106		101		146		97	
Gemma	22	(−.045)	9	(−.029)	12	(−.057)	9	(−.030)	18	(+.055)	14	(−.021)
Mistral	18	(+.000)	11	(−.048)	15	(−.028)	14	(−.059)	14	(+.041)	11	(−.010)
Pharia	25	(+.038)	17	(+.010)	16	(+.019)	16	(+.000)	24	(+.000)	18	(−.021)
Phi	28	(+.030)	16	(−.058)	14	(−.019)	16	(−.040)	40	(−.027)	15	(+.010)
Qwen	23	(−.038)	11	(−.048)	13	(−.047)	13	(−.050)	15	(−.048)	12	(−.021)
Salamandra	13	(−.038)	17	(−.010)	11	(−.028)	12	(−.079)	10	(+.000)	9	(−.010)
Aya	18	(−.061)	9	(−.010)	13	(−.028)	9	(−.010)	16	(+.027)	9	(−.010)
Command-A	20	(−.030)	12	(−.077)	15	(−.009)	10	(−.079)	12	(+.014)	10	(−.041)

Table 4: EU20-MMLU "only mistranslations" vs. Global-MMLU "no errors"

a consistent pattern: models are more likely to predict incorrectly on EU20-MMLU samples with detected mistranslation errors than on their error-free Global-MMLU counterparts, so the gain at least slightly exceeds the loss across nearly all models and languages, with Italian and Pharia being the exception. For instance, in Romanian, Command-A shows a 7.2% gain from corrected translations vs. 3.1% losses, and in French a 8.9% gain against a 1% loss, resulting in an NAI of -4.1% and -7.9%, respectively. The positive or near-zero NAI values which Pharia exhibits in most languages might be explained by the low overall model performance observed in Section 4.4.

Although the number of affected samples per model is relatively low (ranging from 9 to 40 out of 97 to 146 samples), the commonly observed negative Net Accuracy Impact (NAI) across most models and all six languages indicates that further investigation may have merit.

4.3 Error/Performance Association

To extend our analysis beyond MMLU and human-verified translations, we examined whether the detection of translation errors of specific MQM error categories (*Accuracy*, *Fluency*) and error types (*mistranslation*) by GEJ is associated with a difference in LLM prediction correctness in the machine-translated EU20 benchmarks.

Specifically, we consider the accuracy-based multiple-choice benchmarks ARC, Hellaswag, and MMLU, which have dichotomous accuracy metrics on the sample level. For each combination of model, language, task, and error type, we perform a χ^2 independence test to assess whether LLM prediction correctness systematically differs depending on the detection of specific error types. The χ^2 test aims to refute the null hypothesis "H_0: The prediction correctness of the LLM considered is independent of the detection of translation errors in the test samples evaluated" at a significance level of 0.05. This method is well suited to our use case, as both the detection of errors and LLM prediction correctness for multiple-choice tasks can be represented as binary variables at the sample level.

	mistrans.	**acc.**	**fluency**
ARC			
GPT	31	22	8
Llama	60	43	54
Mistral	83	36	18
Hellaswag			
GPT	7	3	1
Llama	21	6	1
Mistral	19	4	5
MMLU			
GPT	78	25	32
Llama	46	28	11
Mistral	91	17	47

Table 5: Percentage of LLMs with significant ($p < 0.05$) χ^2 results for accuracy independence between samples without errors and those with only the specified error category, across languages, models, annotators, and benchmarks.

We summarize the independence results of the χ^2 analysis in Table 5: for each annotator model, task, and error type, we computed the proportion of models with a significant independence result between translation error detection and model prediction correctness. The purpose of this analysis is to identify annotators and categories of detected translation errors that are likely to impact model accuracy. The results indicate that mistranslation errors are signifi-

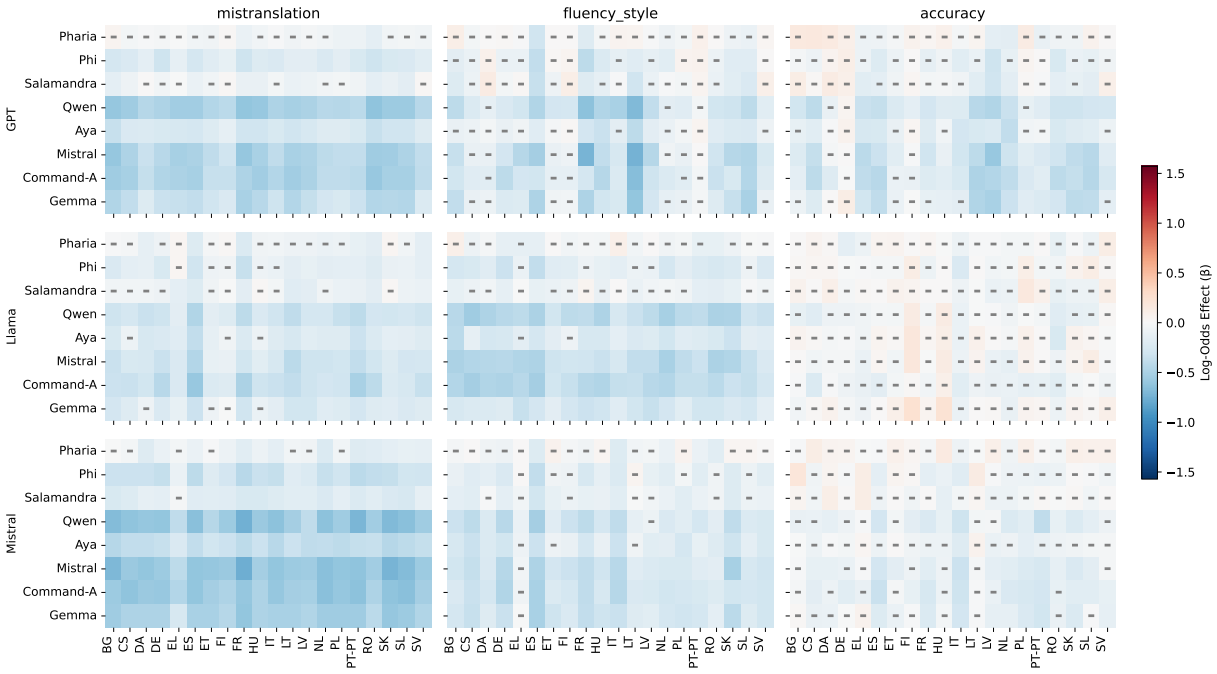

Figure 3: Effect of translation error types on model accuracy (Log-Odds β_e). Each heatmap cell shows the estimated effect of a given error type on the probability of a correct model response, for each model (row) and language (column). Cell color intensity reflects the effect size; cells marked with "–" indicate non-significant effects based on the 95% bootstrap confidence interval.

cantly associated with model accuracy differences in 78-91% of model-language combinations for MMLU, but 21% or less for Hellaswag. For ARC and MMLU, non-mistranslation Accuracy ("acc.") errors show moderate effects in 17-43% of model-language combinations. The impact of translation errors varies across tasks in our evaluation: mistranslation errors have a strong negative effect on MMLU, while in Hellaswag, such phenomena are less pronounced. It is important to note that this χ^2 analysis can capture only associations, not causal effects, and does not indicate effect strength. Nevertheless, the method provides a useful initial quantitative indicator of the relevance of individual error types.

4.4 Error Impact Quantification

To complement the initial Chi-Square analysis, we conducted a logistic regression analysis to quantify the impact of different error types on model accuracy. For each model–language pair, we fitted a logistic regression model predicting binary sample correctness based on the presence of error types:

$$\text{logit}(P(\text{correct})) = \beta_0 + \sum_{e \in \text{Errors}} \beta_e \cdot \text{I}_e$$

where *Errors* = {accuracy, fluency, mistranslation}, and I_e is an indicator variable detected errors of type e. The reference category no_errors is captured in the intercept β_0. Each coefficient β_e estimates the change in log-odds of a correct response relative to this baseline. Negative coefficients indicate a reduction in accuracy, positive coefficients an increase.

We computed nonparametric bootstrap confidence intervals: for each model–language pair, we performed 2500 bootstrap resamples, refitted the model, and computed percentile-based 95% confidence intervals (CIs) for all coefficients. A coefficient with a CI entirely below zero indicates a robust negative effect; a CI overlapping zero indicates an uncertain effect. This regression analysis complements the Chi-Square tests by explicitly estimating, for each error type, both the direction (positive or negative effect) and the magnitude (size) of the effect on the log-odds of a correct model response, with bootstrap CIs providing a more robust assessment of these effects (coefficient stability).

Figure 3 provides a visual summary of the estimated impact of translation errors on model performance across tasks. These heatmaps il-

lustrate that *mistranslation* errors consistently lead to strong and significant accuracy drops for high-performing models such as Qwen, Command-A, and Mistral across nearly all languages (e.g., Qwen: 20/20 languages significant, strongest effect –0.63 for Romanian; Mistral: up to –0.77 for French). In contrast, weaker models like Pharia and Salamandra show predominantly non-significant results for these errors.

To quantitatively compare the impact of different translation error types on model performance, we compute two key statistics for each error type f across all models, languages, and annotators: (i) the proportion of significant effects and (ii) the average absolute effect size.

The **proportion of significant effects** is defined as

$$P_{\text{sig}}(f) = \frac{1}{N_f} \sum_{i=1}^{N_f} \mathbb{I}\left[\text{CI}_i^{\text{upper}} < 0\right], \qquad (1)$$

where N_f is the total number of evaluated combinations (models, languages, annotators) for error type f, $\text{CI}_i^{\text{upper}}$ denotes the upper bound of the 95% bootstrap confidence interval for the i-th log-odds coefficient, and $\mathbb{I}[\cdot]$ is the indicator function.

The **average absolute effect size** (restricted to significant cases) is computed as

$$\bar{\beta}(f) = \frac{1}{N_{f,\text{sig}}} \sum_{i=1}^{N_f} |\beta_i| \cdot \mathbb{I}\left[\text{CI}_i^{\text{upper}} < 0\right], \qquad (2)$$

where β_i is the estimated log-odds effect in the i-th cell, and $N_{f,\text{sig}} = \sum_{i=1}^{N_f} \mathbb{I}\left[\text{CI}_i^{\text{upper}} < 0\right]$ is the number of significant cases. Errors of type *mistranslation* have the largest and most frequent impact on model performance, with $P_{\text{sig}} = 0.86$ and $\bar{\beta} = 0.35$. *Fluency* errors have an intermediate effect ($P_{\text{sig}} = 0.68$, $\bar{\beta} = 0.30$), while *Accuracy* errors show only a weak and infrequent influence ($P_{\text{sig}} = 0.30$, $\bar{\beta} = 0.26$).

5 Conclusion

This work investigates the effects of translation errors on model performance in multilingual benchmarks, with a focus on the EU20 benchmark suite. To enable detailed analysis, we applied an LLM-as-a-judge method (GEJ) for span-level MQM annotation. Compared to the GEMBA-ESA baseline, GEJ achieves an average Span-Recall improvement of 0.23 for *mistranslation*, 0.29 for *Accuracy*, and 0.23 for *Fluency*. GEJ also exhibits higher FPR_s and ER_w values than the baseline (a 39% relative increase) indicating a greater tendency to label errors in reference translations, which should be considered when interpreting downstream analyses.

Our performance impact analysis, including regression modeling, shows that *mistranslation* errors have the largest and most frequent negative effect on LLM accuracy ($P_{\text{sig}} = 0.86$, $\bar{\beta} = 0.35$), followed by *Fluency* errors ($P_{\text{sig}} = 0.68$, $\bar{\beta} = 0.30$), while *Accuracy* errors are rarely significant ($P_{\text{sig}} = 0.30$, $\bar{\beta} = 0.26$). These findings highlight the importance of fine-grained translation error analysis for understanding and improving LLM performance in multilingual benchmarks.

6 Future Work

For future work, we identify several directions: First, extending error analysis to additional MQM error categories and linguistic phenomena. Second, contributing to more diverse gold-standard datasets for meta-evaluating TQE methods, which are essential for reliable progress. Third, advancing LLM-based methods to automate and support benchmark creation and quality control, enabling more scalable and robust multilingual evaluation.

7 Limitations

Despite its merits, certain limitations should be acknowledged in our study. First, high-quality, span-annotated reference translations are scarce, especially for less common languages and MQM error categories. Second, no widely accepted gold standard exists for span-level MQM annotation in most languages investigated, and expert annotation remains costly and hard to scale. Third, our statistical analyses are correlational in nature and may be affected by sample size, model assumptions, and the rarity of certain errors. Fourth, annotation quality depends on the LLM and prompt; some combinations, like DeepSeek with GEMBA-ESA, may fail for certain tasks. Finally, our results reflect a specific selection of benchmarks, languages, and models, and applicability to different settings may be limited.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Irene Baucells, Javier Aula-Blasco, Iria de Dios-Flores, Silvia Paniagua Suárez, Naiara Perez, Anna Salles, Susana Sotelo Docio, Júlia Falcão, Jose Javier Saiz, Robiert Sepulveda Torres, Jeremy Barnes, Pablo Gamallo, Aitor Gonzalez-Agirre, German Rigau, and Marta Villegas. 2025. IberoBench: A benchmark for LLM evaluation in Iberian languages. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10491–10519, Abu Dhabi, UAE. Association for Computational Linguistics.

Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan, and André Martins. 2023. Findings of the WMT 2023 shared task on quality estimation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653, Singapore. Association for Computational Linguistics.

Rochelle Choenni, Sara Rajaee, Christof Monz, and Ekaterina Shutova. 2024. On the evaluation practices in multilingual nlp: Can machine translation offer an alternative to human translations? *Preprint*, arXiv:2406.14267.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *Preprint*, arXiv:1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. *Preprint*, arXiv:2110.14168.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Martin Fajcik, Martin Docekal, Jan Dolezal, Karel Ondrej, Karel Beneš, Jan Kapsa, Pavel Smrz, Alexander Polok, Michal Hradis, Zuzana Neverilova, Ales Horak, Radoslav Sabol, Michal Stefanik, Adam Jirkovsky, David Adamczyk, Petr Hyner, Jan Hula, and Hynek Kydlicek. 2025. Benczechmark : A czech-centric multitask and multimetric benchmark for large language models with duel scoring mechanism. *Preprint*, arXiv:2412.17933.

Manuel Faysse, Patrick Fernandes, Nuno M. Guerreiro, António Loison, Duarte M. Alves, Caio Corro, Nicolas Boizard, João Alves, Ricardo Rei, Pedro H. Martins, Antoni Bigata Casademunt, François Yvon, André F. T. Martins, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2025. Croissantllm: A truly bilingual french-english language model. *Preprint*, arXiv:2402.00786.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Xu Huang, Zhirui Zhang, Xiang Geng, Yichao Du, Jiajun Chen, and Shujian Huang. 2024. Lost in the source language: How large language models evaluate the quality of machine translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3546–3562, Bangkok, Thailand. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: Detecting Translation Quality Error Spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error Span Annotation: A Balanced Approach for Human Evaluation of Machine Translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned Large Language Models in Multiple Languages with Reinforcement Learning from Human Feedback. In

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 318–327, Singapore. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Arle Lommel, Serge Gladkoff, Alan Melby, Sue Ellen Wright, Ingemar Strandvik, Katerina Gasova, Angelika Vaasa, Andy Benzo, Romina Marazzato Sparano, Monica Foresi, Johani Innis, Lifeng Han, and Goran Nenadic. 2024. The multi-range theory of translation quality measurement: MQM scoring models and statistical quality control. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: Presentations)*, pages 75–94, Chicago, USA. Association for Machine Translation in the Americas.

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2013. Multidimensional quality metrics: A flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Qingyu Lu, Liang Ding, Kanjian Zhang, Jinxia Zhang, and Dacheng Tao. 2025. MQM-APE: Toward high-quality error annotation predictors with automatic post-editing in LLM translation evaluators. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5570–5587, Abu Dhabi, UAE. Association for Computational Linguistics.

Nikita Moghe, Arnisa Fazla, Chantal Amrhein, Tom Kocmi, Mark Steedman, Alexandra Birch, Rico Sennrich, and Liane Guillou. 2025. Machine Translation Meta Evaluation through Translation Accuracy Challenge Sets. *Computational Linguistics*, pages 1–65.

Dan Nielsen. 2023. ScandEval: A benchmark for Scandinavian natural language processing. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 185–201, Tórshavn, Faroe Islands. University of Tartu Library.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

ChaeHun Park, Koanho Lee, Hyesu Lim, Jaeseok Kim, Junmo Park, Yu-Jung Heo, Du-Seong Chang, and Jaegul Choo. 2024. Translation deserves better: Analyzing translation artifacts in cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5193–5221, Bangkok, Thailand. Association for Computational Linguistics.

Jan Pfister and Andreas Hotho. 2024. SuperGLEBer: German language understanding evaluation benchmark. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7904–7923, Mexico City, Mexico. Association for Computational Linguistics.

Irene Plaza, Nina Melero, Cristina Pozo, Javier Conde, Pedro Reviriego, Marina Mayor-Rocher, and María Grandury. 2024. Spanish and llm benchmarks: is mmlu lost in translation? *Preprint*, arXiv:2406.17789.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Sebastian Ruder, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, and 5 others. 2025. Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation. *Preprint*, arXiv:2412.03304v2.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, and Mehdi Ali. 2024. Towards Multilingual LLM Evaluation for European Languages. *Preprint*, arXiv:2410.08928.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can

a Machine Really Finish Your Sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Chrysoula Zerva, Frederic Blain, José G. C. De Souza, Diptesh Kanojia, Sourabh Deoghare, Nuno M. Guerreiro, Giuseppe Attanasio, Ricardo Rei, Constantin Orasan, Matteo Negri, Marco Turchi, Rajen Chatterjee, Pushpak Bhattacharyya, Markus Freitag, and André Martins. 2024. Findings of the Quality Estimation Shared Task at WMT 2024: Are LLMs Closing the Gap in QE? In *Proceedings of the Ninth Conference on Machine Translation*, pages 82–109, Miami, Florida, USA. Association for Computational Linguistics.

Haofei Zhao, Yilun Liu, Shimin Tao, Weibin Meng, Yimeng Chen, Xiang Geng, Chang Su, Min Zhang, and Hao Yang. 2024. From handcrafted features to llms: A brief survey for machine translation quality estimation. In *2024 International Joint Conference on Neural Networks (IJCNN)*, page 1–10. IEEE.

A Appendix: GEJ, MQM and Span-ACES

A.1 GEJ-MQM Error Categories

Table 6 presents a hierarchical mapping of the MQM error taxonomy used in our evaluation. It aligns Level-1 MQM error types (e.g., *addition*, *mistranslation*, *omission*) with ACES labels such as `real-world-knowledge-entailment` or `xnli-addition-neutral`.

As an illustration, Table 7 provides brief definitions and multilingual examples for each Level-1 error type. For instance, the *addition* category is illustrated with an English to Dutch example in which extraneous information ("to buy bread") is added in the translation. Similarly, *mistranslation* is exemplified via a Greek translation that changes the meaning of "He was murdered" to "He was attacked." These examples were curated to span both general-purpose and structured text, used as input in our evaluation setup.

Table 6: Hierarchical overview of MQM error types and their alignment with Span-ACES error labels.

MQM Type	Level-0	Level-1	Span-ACES Label
Addition	accuracy	addition	addition xnli-addition-contradiction xnli-addition-neutral
Ambiguous src content	accuracy	mistranslation	coreference-based-on-commonsense
Ambiguous tgt content	accuracy	mistranslation	ambiguous-*-since-causal ambiguous-*-since-temporal ambiguous-*-while-contrast ambiguous-*-while-temporal ambiguous-*-female-anti ambiguous-*-female-pro ambiguous-*-male-anti ambiguous-*-male-pro
Do not translate	accuracy	no-translate	do-not-translate
Grammar	fluency	grammar	anaphoric_group_it-they:subst. anaphoric_intra_non-subject_it:subst. anaphoric_intra_subject_it:subst. anaphoric_intra_they:subst. anaphoric_singular_they:subst.
Mistranslation	accuracy	mistranslation	modal_verb:subst. pleonastic_it:subst. real-world-know.-entailment real-world-know.-hypernym-vs-distr. real-world-know.-syn.-vs-antonym
MT hallucination	accuracy	mistranslation	hallucination-date-time
Omission	accuracy	omission	anaphoric_group_it-they:deletion anaphoric_intra_non-subject_it:deletion anaphoric_intra_subject_it:deletion anaphoric_intra_they:deletion anaphoric_singular_they:deletion modal_verb:deletion omission pleonastic_it:deletion xnli-omission-contradiction xnli-omission-neutral

(Continued on next page)

Table 6: Hierarchical overview of MQM error types and their alignment with Span-ACES error labels. *(Continued from previous page)*

MQM Type	Level-0	Level-1	Span-ACES Label
Addition	accuracy	addition	addition xnli-addition-contradiction xnli-addition-neutral
Overly literal	accuracy	mistranslation	overly-literal-vs-explanation overly-literal-vs-ref-word overly-literal-vs-synonym
Overtranslation	accuracy	over-translation	hyponym-replacement
Punctuation	fluency	punctuation	punctuation:deletion_all punctuation:deletion_commas punctuation:deletion_quotes punctuation:statement-to-question
Undertranslation	accuracy	under-translation	hypernym-replacement
Unintelligible	fluency	unintelligible	nonsense
Untranslated	accuracy	untranslated	copy-source untranslated-vs-ref-word untranslated-vs-synonym
Word order	accuracy	mistranslation	ordering-mismatch
Wrong language	accuracy	wrong-language	similar-language-high similar-language-low

Table 7: Overview of MQM error categories with definitions and examples.

Level-1 Category	Description and Example
`addition`	Adds extra content not in source. EN: "She might go to the store." → NL: "Ze gaat misschien naar de winkel om brood te kopen." Added: "to buy bread"

Continued on next page

Level-1 Category	Description and Example
`over-translation`	Adds unnecessary detail or specificity. EN: ”room” → HU: ”konyha” Should be: ”szoba”. ”konyha” means kitchen.
`omission`	Important information from the source is missing. EN: ”She bought a red dress.” → CS: ”Koupila si šaty.” Should be: ”Koupila červené šaty.”. ”red” is omitted
`under-translation`	Translation is too general. EN: ”colonel” → SK: ”vojak” Should be: ”plukovník”. ”vojak” means soldier
`mistranslation`	Incorrect meaning due to word or grammar. EN: ”He was murdered.” → EL: ”Τον επιτέθηκαν.” ”He was attacked” wrong meaning.
`reordering`	Word order change that alters meaning. EN: ”She only loves him.” → PL: ”Tylko ona go kocha.” ”Only she loves him”
`untranslated`	Source (or part of it) left untranslated. EN: ”He is a teacher.” → DE: ”He ist ein Lehrer.” ”He” should be ”Er”
`wrong-language`	Wrong or related language used. EN: ”Danish colleague” → DE: ”dänischen kollega” ”kollega” is Danish; should be ”Kollege”
`do-not-translate`	Elements marked non-translatable are translated. EN: ”Apple released a new update.” → BG: ”Ябълка пусна нов ъпдейт.” ”Apple” → fruit
`grammar`	Syntax or agreement error. EN: ”She go to school...” → DA: ”Hun går til skole...” Should be: ”i skole”

Continued on next page

Level-1 Category	Description and Example
`spelling`	Misspelled words affect readability or meaning. EN: ”definately” → SV: ”definetivt” Should be:”definitivt”
`punctuation`	Incorrect or missing punctuation. EN: ”Let’s eat, Grandma!” → SV: ”Låt oss äta mormor!” Missing comma
`inconsistent`	Inconsistent terminology or style. EN: ”Prime Minister”/”Premier” → PL: ”Premier”/”Prezes Rady Ministrów” Should be consistent
`awkward`	Grammatically correct but unnatural. EN: ”He made a photo.” → EL: ”Έκανε μια φωτογραφία.” Should be: ”Τράβηξε μια φωτογραφία”
`unintelligible`	Nonsensical translation. EN: ”The cat sat on the mat.” → SL: ”Mačka je stol, ki hodi po ulici.” Should be: ”The cat is a chair that walks on the street”

A.2 GEJ Prompts

Figure 4 details the system prompt used in our GEMBA-ESA adaptation, which guides the LLM to identify span-level errors based on MQM categories, while enforcing strict output formatting and semantic fidelity over stylistic preferences. It emphasizes translation adequacy, fluency, and structured error explanation, with instructions for multiple-choice and general-purpose content.

To enhance annotation quality, we created a multilingual prompt containing structured few-shot examples for both continuous and structured text (Figure 5). These examples include a diverse set of error types – such as subject-verb agreement errors in Portuguese (*grammar*), overgeneralization in German (*undertranslation*), and missing specificity in medical contexts (*mistranslation*). Translations were generated using ChatGPT-4o, with backtranslations via DeepL. We verified correctness through backtranslation analysis and cross-referenced definitions using multilingual resources such as dict.cc[12], LEO[13], monolingual dictionaries, conjugation tools, and DeepL Write[14].

Instead of adding the instruction to every few-shot example within the prompt, we include it in the system prompt, which is sent to the model only once for each translation to be evaluated. Additionally, we provide a separate input field that briefly describes the text type, alongside the original and target segments, to help the model better assess the severity of translation errors.

[12] https://www.dict.cc
[13] https://dict.leo.org
[14] https://www.deepl.com/write

Figure 4: GEMBA-ESA System Prompt (Error Type Definitions omitted)

"**role**": "**system**",
"**content**": "Your task is to assess the quality of machine translations and identify translation errors.

Given the source and target segments (enclosed in triple backticks), identify error spans in the translation and classify them using the MQM-style (Multidimensional Quality Metrics) categories: accuracy (addition, over-translation, omission, under-translation, mistranslation, reordering, untranslated, wrong-language, do-not-translate), fluency/style (grammar, spelling, punctuation, inconsistent, awkward, unintelligible), and other.

Error Type Definitions (with one example each): …

Severity Guidelines:

- Label an error as major if it significantly alters the meaning, causes confusion, or omits important information.
- Label an error as minor if it causes a slight loss in precision or naturalness, but the meaning is still clear.
- Do not flag translations as errors if they preserve meaning and naturalness, even if they are not literal.

Guidance for Output Format and Error Spans:

- Respond with a single, plain, well-formed JSON object, without any markup, containing:
  - `target_seg_backtranslation`: literal backtranslation of target segment to source language, preserving any errors
  - `evaluation`: an array of error entries, each with:
    - `span_target`: text span with the found error in the translation (if present, else use an empty string)
    - `span_original`: corresponding text in the source
    - `error_type`: the error category
    - `severity`: "major" or "minor"
    - `explanation`: reason the translation is incorrect (or correct) in one sentence
    - `correction`: suggested correction of the target span
- For omission or untranslated content, also use an empty string as the error span.
- It is important to use only Guillemets «...» and no regular double quotes "" when quoting parts of the source or target in the explanation.
  Make sure that all JSON strings are in double quotes, as single quotes are not valid JSON, e.g. "severity": "minor" and not 'severity': 'minor'.
- If no errors are found, leave evaluation as an empty array.

Translation Quality Evaluation Guidelines:
Your task is to evaluate the translation quality, not to solve or interpret the content.
Focus on whether the translation is faithful to the original meaning, and appropriate for the text type. Avoid nitpicking stylistic preferences unless they impact meaning or clarity. Types of text that can occur are: structured or multiple-choice (e.g., science questions, math problems), general-purpose or continuous text.

For structured or multiple-choice text:

- Evaluate only whether each part — the question and all answer options — is faithfully translated.
- Do not assess factual correctness of any answer options.
- Do not solve the question or make assumptions about the correct answer.
- Grammaticality and fluency are not required for short fragments (e.g., single words or sentence parts in answer options), as long as their meaning is translated.
- Ensure that all options from the source are preserved; omissions or additions are critical."

Figure 5: GEMBA-ESA-JSON prompt with illustrative subset of structured few-shot examples (without system prompt).

```
[  {
    "role": "user",
    "content": "Text type: general-purpose or continuous text
     English source: My friends went to the cinema.
     Portuguese translation: Os meus amigos foi ao cinema."
  }, {
    "role": "assistant",
    "content": "{
    'target_seg_backtranslation': 'My friends was to the cinema.',
    'evaluation': [{
    'span_target': 'foi',
    'span_original': 'went',
    'error_type': 'grammar',
    'severity': 'major',
    'explanation': 'Wrong verb agreement; «amigos» is plural, so verb should be «foram».
                    This is a source-side error and should not be corrected in the target.',
    'correction': 'Os meus amigos foram ao cinema.'}]
    }"
  }, {
    "role": "user",
    "content": "Text type: general-purpose or continuous text
     English source: She is an experienced cardiac surgeon.
     German translation: Sie ist eine Ärztin."
  }, {
    "role": "assistant",
    "content": "{
    'target_seg_backtranslation': 'She is a doctor.',
    'evaluation': [{
    'span_target': 'Ärztin',
    'span_original': 'cardiac surgeon',
    'error_type': 'undertranslation',
    'severity': 'major',
    'explanation': 'The translation omits the specificity of the profession and experience;
                    «cardiac surgeon» is reduced to «doctor».',
    'correction': 'Sie ist eine erfahrene Herzchirurgin.'}]
    }"
  }, {
    "role": "user",
    "content": "Text type: general-purpose or continuous text
    English source:
      A 44-year-old female presents to the office for evaluation of
      a lump on her neck that she noted 1 week ago.
      She denies any tenderness, fever, weight loss, or fatigue.
      Physical examination reveals a 2cm freely movable mass in the lower left lobe of the thyroid.
      In addition to thyroid-stimulating hormone and free thyroxine levels,
      the most appropriate initial method to investigate this lesion is:
      (1) a nuclear thyroid scan
      (2) an iodine-131 scan
      (3) fine-needle aspiration
      (4) ultrasonography of the thyroid gland

    Italian translation:
      Una donna di 44 anni si presenta in ufficio per valutare un nodulo sul collo notato una settimana fa.
      La donna nega di avere dolori, febbre, perdita di peso o affaticamento.
      L'esame fisico rivela una massa di 2 cm liberamente mobile nel lobo inferiore sinistro della tiroide.
      Oltre ai livelli di ormone stimolante la tiroide e di tiroxina libera,
      il metodo iniziale più appropriato per indagare questa lesione è
      (1) un'ecografia della tiroide
      (2) un'ecografia allo iodio
      (3) un'aspirazione con ago fine
      (4) un'ecografia della tiroide"
  }
]
```

A.3 Meta-Evaluation

To assess automated span annotation quality of GEJ, we computed the Span-F1 score according to the Span-ACES reference implementation[15], which compares predicted error spans with the human-labeled spans. Empty or whitespace-only error spans in either prediction or gold standard were excluded from F1 calculation as in the reference implementation[16]. In addition to Span-F1, we compute Span-Recall, which captures how many of the human-annotated spans were successfully identified by the model, without trying to account for false positives.

Both scores are computed based on textual content overlap between predicted and reference error spans (tagged with <v>...</v>), not on character offsets or positions or multiplicity. Empty and whitespace-only spans in either prediction or gold standard were excluded from the F1 and Recall computation, following the Span-ACES reference implementation. Span-F1 balances precision and recall, while Span-Recall focuses on how many gold spans were recovered.

Empty tags are valid annotations in Span-ACES, especially for omission and punctuation, where the error involves missing or absent content. However, since Span-F1 and Span-Recall exclude empty and whitespace-only spans in a preprocessing step, omission samples lack usable reference spans, which prevents reliable scoring. We did not develop an alternative metric for omissions, as unlike for non-empty spans there is no way to distinguish between correctly and incorrectly detected omissions based on the (empty) span content. For this reason, we excluded omissions (alongside other empty-span categories) even though it had good language coverage. In contrast to the Span-ACES approach, we require the detected translation error type to match according to Table 6 for the Span-Recall.

L	N	GPT	DeepSeek	LLaMA	Mistral
		B/GEJ	B/GEJ	B/GEJ	B/GEJ
Mistrans.					
DA	23	.09/.56	-/.60	.32/.39	.60/.32
DE	982	.11/.21	0/.29	.14/.10	.22/.18
ES	28	.18/.62	0/.43	.21/.32	.35/.38
ET	14	.14/.42	-/.50	.24/.39	.47/.38
FR	214	.09/.18	0/.16	.13/.06	.15/.06
HU	32	.08/.21	-/.32	.19/.17	.48/.47
LT	2	.50/.33	-/.58	.53/.00	.17/.33
NL	23	.07/.35	0/.30	.22/.11	.31/.45
PL	7	.23/.71	-/.30	.08/.08	.10/.08
PT	20	.10/.55	-/.53	.30/.36	.45/.45
RO	33	.06/.28	-/.69	.08/.10	.47/.40
SK	8	.00/.46	-/.25	.11/.22	.40/.30
SL	18	.00/.30	-/.18	.13/.11	.41/.35
SV	20	.00/.37	-/.52	.10/.13	.39/.32
Acc.					
BG	9	.10/.11	-/.48	.41/.47	.45/.37
CS	341	.00/.01	0/.57	.01/.01	.01/.31
DA	11	.00/.17	-/.47	.19/.07	.14/.35
DE	249	.00/.02	0/.11	.01/.01	.02/.07
EL	6	.00/.25	-/.38	.20/.17	.44/.38
ES	309	.00/.00	0/.70	.01/.00	.01/.09
ET	3	.00/.00	-/.47	.00/.00	.00/.33
FI	6	.00/.25	-/.56	.00/.00	.43/.60
FR	10	.00/.25	-/.44	.21/.17	.50/.33
HU	3	.00/.00	-/.75	.22/.33	.33/.75
IT	5	.00/.50	1/.73	.33/.17	.20/.40
LT	5	.00/.00	-/.33	.00/.00	.23/.29
LV	8	.00/.17	0/.56	.17/.20	.33/.48
NL	14	.04/.00	-/.37	.19/.00	.30/.17
PL	299	.00/.01	0/.72	.00/.01	.01/.29
PT	11	.00/.09	-/.67	.17/.57	.59/.45
RO	7	.00/.25	-/.77	.15/.13	.48/.57
SK	6	.00/.00	-/.25	.00/.00	.19/.12
SL	11	.00/.00	-/.52	.07/.07	.23/.50
SV	12	.14/.12	-/.61	.10/.29	.39/.50
Fluency					
DE	678	.00/.00	.00/.26	.02/.04	.01/.43

Figure 6: Span-F1 by model and language for Baseline (B) and GEJ, shown for MQM categories: (a) Mistranslation, (b) Accuracy, and (c) Fluency.

[15] https://github.com/EdinburghNLP/ACES/blob/6912157d/span_predictions/eval_span.py

[16] https://github.com/EdinburghNLP/ACES/blob/6912157d/span_predictions/eval_span.py#L36-L41

	BG	CS	DA	DE	EL	ES	ET	FI	FR	HU	IT	LT	LV	NL	PL	PT	RO	SK	SL	SV
Accuracy																				
addition	10	10	12	11	7	8	5	6	11	4	5	7	9	15	5	11	10	8	13	13
no-translate	-	-	-	100	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
untranslated	-	-	-	210	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
wrong-language	-	352	-	-	-	329	-	-	-	-	-	-	-	-	328	-	-	-	-	-
Total	10	362	12	321	7	337	5	6	11	4	5	7	9	15	333	11	10	8	13	13
Fluency																				
grammar	-	-	-	721	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 8: Sample counts per target language in the Span-ACES dataset

Languages	N
CS, ES, SL	7265
HU	7264
IT, PL, PT, RO, SK	7263
BG, DA, EL, LT, LV, NL, SV	7262
DE, ET, FI, FR	7261
Total	145252

Table 9: Sample counts per target language in the FLORES dataset

B Appendix: Model Acc.

B.1 Datasets

To ensure the completeness and structural integrity of the multilingual evaluation datasets used in this study, a validation and cleaning pipeline was applied to EU20 and Global-MMLU datasets. The structural analysis and cleaning of the translated evaluation datasets is conducted across various subsets, splits, and languages. It ensures that in the cleaned dataset, translations are correctly assigned, all questions and answers are present, answers remain consistent across languages, and that all languages are fully represented. Each dataset is described as consisting of one or more subsets, which again consist of one or more splits, which contain a set of per-language samples. Each per-language sample in a dataset is identified based on its sample ID, subset, split, and language. If there are no IDs the original dataset, a sample counter combined with the subset and split of the dataset serves as a surrogate ID. The sample ID may be a string or tuple of strings, and is chosen to be unique across splits and subsets of the dataset while being identical for different language versions of the same sample. Doing so allows for the English original dataset to be matched with the translations. Also, when there are multiple alternative translations of a dataset (EU20 vs. Global-MMLU), care is taken to ensure that corresponding samples have the same ID, so alternative translations can be matched with each other. Each individual language version of each sample is formatted into a single text intended for quality assurance purposes, and missing questions, answers and choices are identified. For the purpose of cross-language checks, the translated samples are matched with their English originals. Checks are performed to determine whether one or more, but not all translations of a sample are missing in all target languages. Table 10 reports the results of these checks. Samples that were not translated into any target language are identified separately. Where possible, a check is performed whether translated samples have been assigned to the correct split or subset. As a sanity check for the choice of sample IDs, cases where correct answers in the translated version are inconsistent with the English original answers are identified – doing so is possible because the answers in all analyzed datasets are either the index of a multiple-choice option or, in case of GSM8K, plain numbers, and thus are identical across all languages. Furthermore, statistics on the number of uncleaned per-language samples, samples with missing content and missing translations are collected on the number of samples per subset, split, and language. In a cleaning step, all samples with missing or incomplete translations are removed. Once all inconsistencies have been identified and resolved, the cleaned dataset is saved as a JSONL file. Table 11 summarizes the final cleaned dataset statistics.

B.2 Models

Table 12 lists the models whose prediction accuracy is analyzed in the Error vs. Correction comparison and the Error-Performance Association analysis.

B.3 Error vs. Correction Comparison - Details

Table 13 shows the detailed results of the error vs. correction comparison.

B.4 Error-Performance Association Analysis - Details

Tables 14, 15 and 16 illustrate more detailed results of the error-performance association analysis.

B.5 Error Impact Quantification - Details

Dataset	Subset	Split	Missing Lang.	Missing Content
eu20_arc	challenge	test	0	11
eu20_arc	challenge	train	948	3
eu20_arc	challenge	validation	0	3
eu20_arc	easy	test	0	39
eu20_arc	easy	train	1869	5
eu20_arc	easy	validation	0	5
eu20_gsm8k	main	train	1972	0
eu20_hellaswag	main	validation	1857	327
eu20_mmlu	business	test	0	12
eu20_mmlu	humanities	test	0	99
eu20_mmlu	medical	test	0	40
eu20_mmlu	other	test	0	15
eu20_mmlu	social_sciences	test	0	99
eu20_mmlu	stem	dev	0	2
eu20_mmlu	stem	test	0	96
eu20_truthfulqa	multiple_choice	validation	0	3
globalmmlu	humanities	test	1	0
globalmmlu	stem	test	2	2

Table 10: EU20, and GlobalMMLU: Number of samples with missing translations or content.

Dataset	Task	Split	N
eu20	arc	test	3498
eu20	arc	val	861
eu20	arc	train	0
eu20	gsm8k	test	1319
eu20	gsm8k	train	0
eu20	hellaswag	val	9658
eu20	hellaswag	train	0
eu20	mmlu	dev	283
eu20	mmlu	val	0
eu20	mmlu	test	13681
eu20	truthfulqa	val	814
globalmmlu	mmlu	test	2845

Table 11: EU20 and GlobalMMLU after cleaning (i.e., removing samples with missing content or translations).

Model Name / HF Link	#Params
Aya	32.3B
Command-A	111B
Gemma	27.4B
Mistral	24B
Pharia	7.0B
Phi	5.6B
Qwen	32.8B
Salamandra	7.8B

Table 12: Model Overview

Lang.	Model	Count				Fraction				Aff.	Unaff.	N_C	NAI
		$\mathbf{E}^+$		$\mathbf{E}^-$		$\mathbf{E}^+$		$\mathbf{E}^-$					
		$\mathbf{G}^+$	$\mathbf{G}^-$	$\mathbf{G}^+$	$\mathbf{G}^-$	$\mathbf{G}^+$	$\mathbf{G}^-$	$\mathbf{G}^+$	$\mathbf{G}^-$	($\mathbf{N_A}$)			
CS	Gemma	79	8	14	31	.60	.06	.11	.23	22	110	132	−.045
	Mistral	87	9	9	27	.66	.07	.07	.20	18	114	132	+.000
	Pharia	28	15	10	79	.21	.11	.08	.60	25	107	132	+.038
	Phi	47	16	12	57	.36	.12	.09	.43	28	104	132	+.030
	Qwen	88	9	14	21	.67	.07	.11	.16	23	109	132	−.038
	Salamandra	61	4	9	58	.46	.03	.07	.44	13	119	132	−.038
	Aya	79	5	13	35	.60	.04	.10	.27	18	114	132	−.061
	Command-A	90	8	12	22	.68	.06	.09	.17	20	112	132	−.030
DE	Gemma	69	3	6	26	.66	.03	.06	.25	9	95	104	−.029
	Mistral	75	3	8	18	.72	.03	.08	.17	11	93	104	−.048
	Pharia	44	9	8	43	.42	.09	.08	.41	17	87	104	+.010
	Phi	57	5	11	31	.55	.05	.11	.30	16	88	104	−.058
	Qwen	75	3	8	18	.72	.03	.08	.17	11	93	104	−.048
	Salamandra	45	8	9	42	.43	.08	.09	.40	17	87	104	−.010
	Aya	67	4	5	28	.64	.04	.05	.27	9	95	104	−.010
	Command-A	76	2	10	16	.73	.02	.10	.15	12	92	104	−.077
ES	Gemma	61	3	9	33	.58	.03	.08	.31	12	94	106	−.057
	Mistral	68	6	9	23	.64	.06	.08	.22	15	91	106	−.028
	Pharia	38	9	7	52	.36	.08	.07	.49	16	90	106	+.019
	Phi	56	6	8	36	.53	.06	.08	.34	14	92	106	−.019
	Qwen	70	4	9	23	.66	.04	.08	.22	13	93	106	−.047
	Salamandra	45	4	7	50	.42	.04	.07	.47	11	95	106	−.028
	Aya	67	5	8	26	.63	.05	.08	.25	13	93	106	−.028
	Command-A	69	7	8	22	.65	.07	.08	.21	15	91	106	−.009
FR	Gemma	68	3	6	24	.67	.03	.06	.24	9	92	101	−.030
	Mistral	73	4	10	14	.72	.04	.10	.14	14	87	101	−.059
	Pharia	40	8	8	45	.40	.08	.08	.45	16	85	101	+.000
	Phi	50	6	10	35	.50	.06	.10	.35	16	85	101	−.040
	Qwen	66	4	9	22	.65	.04	.09	.22	13	88	101	−.050
	Salamandra	48	2	10	41	.48	.02	.10	.41	12	89	101	−.079
	Aya	69	4	5	23	.68	.04	.05	.23	9	92	101	−.010
	Command-A	76	1	9	15	.75	.01	.09	.15	10	91	101	−.079
IT	Gemma	90	13	5	38	.62	.09	.03	.26	18	128	146	+.055
	Mistral	97	10	4	35	.66	.07	.03	.24	14	132	146	+.041
	Pharia	68	12	12	54	.47	.08	.08	.37	24	122	146	+.000
	Phi	67	18	22	39	.46	.12	.15	.27	40	106	146	−.027
	Qwen	97	4	11	34	.66	.03	.08	.23	15	131	146	−.048
	Salamandra	60	5	5	76	.41	.03	.03	.52	10	136	146	+.000
	Aya	88	10	6	42	.60	.07	.04	.29	16	130	146	+.027
	Command-A	108	7	5	26	.74	.05	.03	.18	12	134	146	+.014
RO	Gemma	62	6	8	21	.64	.06	.08	.22	14	83	97	−.021
	Mistral	71	5	6	15	.73	.05	.06	.15	11	86	97	−.010
	Pharia	30	8	10	49	.31	.08	.10	.51	18	79	97	−.021
	Phi	46	8	7	36	.47	.08	.07	.37	15	82	97	+.010
	Qwen	63	5	7	22	.65	.05	.07	.23	12	85	97	−.021
	Salamandra	37	4	5	51	.38	.04	.05	.53	9	88	97	−.010
	Aya	57	4	5	31	.59	.04	.05	.32	9	88	97	−.010
	Command-A	75	3	7	12	.77	.03	.07	.12	10	87	97	−.041

Table 13: Cross-tabulation of model predictions comparing EU20-MMLU samples with detected mistranslation errors to their error-free counterparts in Global-MMLU, reported in absolute counts and relative proportions. Legend: $\mathbf{E}^+/\mathbf{E}^-$ = EU20 correct/incorrect, $\mathbf{G}^+/\mathbf{G}^-$ = Global correct/incorrect; Affected (N_A) = predictions differ between EU20 and Global-MMLU; Unaffected = predictions match; N_C = number of samples; Net Accuracy Impact (NAI) = $\mathbf{G}^+/\mathbf{E}^-$ - $\mathbf{E}^+/\mathbf{G}^-$, i.e., gain minus loss in prediction acc due to improved translations.

Language	BG	CS	DA	DE	EL	ES	ET	FI	FR	HU	IT	LT	LV	NL	PL	PT	RO	SK	SL	SV
ARC																				
acc.	0	25	38	0	12	38	38	62	38	62	12	0	25	0	12	50	12	0	12	0
f./s.	12	0	0	0	0	0	62	0	0	50	0	0	0	0	0	25	0	0	0	0
mis.	12	50	0	0	12	0	88	12	0	88	0	100	75	0	12	38	50	0	88	0
Hellaswag																				
acc.	12	0	0	0	12	12	0	0	0	0	0	0	25	0	0	0	0	0	0	0
f./s.	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
mis.	0	0	12	0	0	12	12	25	12	12	0	0	0	0	12	0	0	0	12	25
MMLU																				
acc.	0	0	12	25	0	88	0	0	12	0	12	75	88	75	0	0	50	12	50	0
f./s.	62	0	12	0	0	100	12	25	100	12	12	88	25	0	12	0	25	75	62	12
mis.	88	62	38	75	75	88	88	25	100	75	88	88	75	75	100	88	100	88	100	50

Table 14: % of LLMs with significant ($p < 0.05$) χ^2 result for LLM accuracy independence between samples without detected errors and samples with detected errors exclusively of specified error category as detected by GPT (across benchmarked LLMs, per benchmark).

Language	BG	CS	DA	DE	EL	ES	ET	FI	FR	HU	IT	LT	LV	NL	PL	PT	RO	SK	SL	SV
ARC																				
acc.	0	75	25	75	100	38	50	50	75	88	12	25	62	0	38	12	88	0	25	25
f./s.	38	62	12	75	50	100	25	50	0	62	75	75	100	0	50	100	88	62	25	25
mis.	12	88	38	25	75	62	88	75	25	75	0	88	100	0	88	88	88	25	75	88
Hellaswag																				
acc.	0	0	0	50	0	0	0	0	0	0	0	0	12	0	12	12	0	38	0	0
f./s.	0	0	0	12	0	0	0	0	0	0	0	0	0	12	0	0	0	0	0	0
mis.	0	0	12	0	0	100	0	0	12	38	0	38	0	0	12	75	75	0	25	25
MMLU																				
acc.	12	25	88	25	38	12	25	38	12	38	0	12	12	0	25	50	0	50	12	75
f./s.	50	0	0	25	0	25	38	12	0	12	12	25	0	0	0	0	0	0	0	12
mis.	75	12	12	62	12	50	12	38	100	25	38	100	88	38	50	62	25	25	62	25

Table 15: % of LLMs with significant ($p < 0.05$) χ^2 result for LLM accuracy independence between samples without detected errors and samples with detected errors exclusively of specified error category as detected by Llama (across benchmarked LLMs, per benchmark).

Language	BG	CS	DA	DE	EL	ES	ET	FI	FR	HU	IT	LT	LV	NL	PL	PT	RO	SK	SL	SV
ARC																				
acc.	12	38	12	12	0	88	0	62	75	75	38	0	25	12	75	38	62	25	75	0
f./s.	25	0	50	62	0	0	12	0	50	12	0	12	62	0	0	75	0	0	0	0
mis.	75	88	88	88	25	88	62	88	100	75	100	75	75	100	100	100	75	88	100	75
Hellaswag																				
acc.	0	0	25	0	0	0	0	12	12	0	0	0	0	0	0	0	12	0	12	0
f./s.	0	0	0	0	38	0	0	25	0	0	0	0	0	12	12	12	0	0	0	0
mis.	0	0	12	50	38	0	0	50	25	62	0	0	12	0	62	0	38	25	0	12
MMLU																				
acc.	38	12	38	0	75	0	0	12	25	25	12	12	50	0	12	12	0	0	0	12
f./s.	12	88	12	25	0	100	75	12	88	25	100	0	38	88	38	12	38	88	38	62
mis.	88	88	100	100	38	88	75	88	100	88	100	88	88	100	88	100	100	100	100	100

Table 16: % of LLMs with significant ($p < 0.05$) χ^2 result for LLM accuracy independence between samples without detected errors and samples with detected errors exclusively of specified error category as detected by Mistral (across benchmarked LLMs, per benchmark).

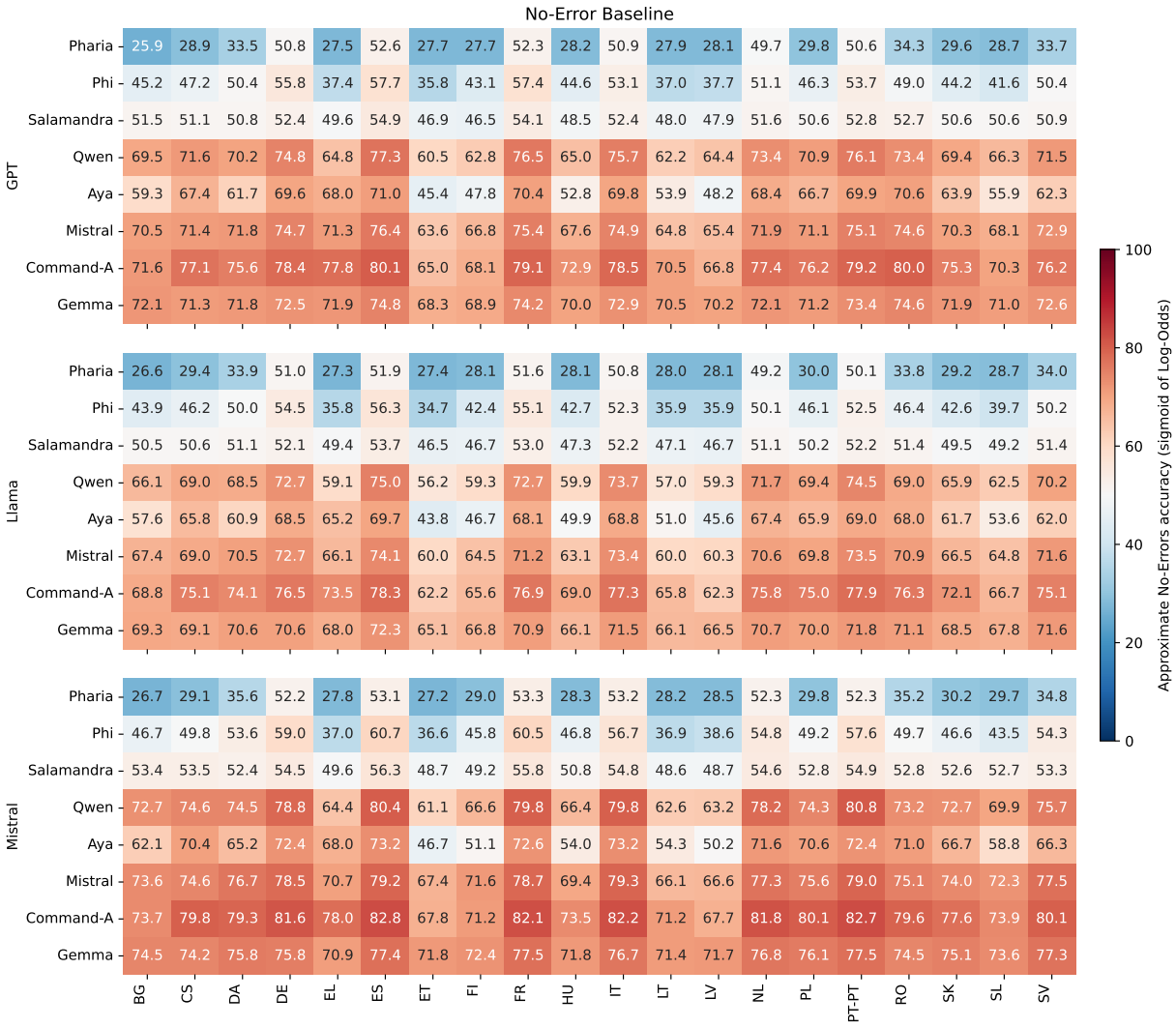

Figure 7: This figure shows the approximate baseline model accuracies one heatmap per annotator LLM for examples with no detected errors. Each heatmap shows, for each model (rows) and language (columns), the approximate probability of a correct LLM response when no translation errors are present, computed as the sigmoid of the intercept term (β_0) of the logistic regression model.