Quantifying the Impact of Translation Errors on Multilingual LLM Evaluation

Anonymous ACL submission

Abstract

Machine-translated benchmarks are widely used to assess the multilingual capabilities of large language models (LLMs), yet translation errors in these benchmarks remain underexplored, raising concerns about the reliability and comparability of multilingual evaluation. We address two practical gaps: (i) how well LLM-produced MQM-style error spans match expert human span annotations on real benchmark translations, and (ii) how strongly translation errors (as opposed to source-side issues in the English original) explain accuracy drops on translated benchmarks. We find that span agreement is non-trivial on naturally occurring benchmark translations, and that target-side translation errors are consistently associated with measurable, percentage-point drops in translated accuracy even after controlling for English correctness and source-side anomalies.

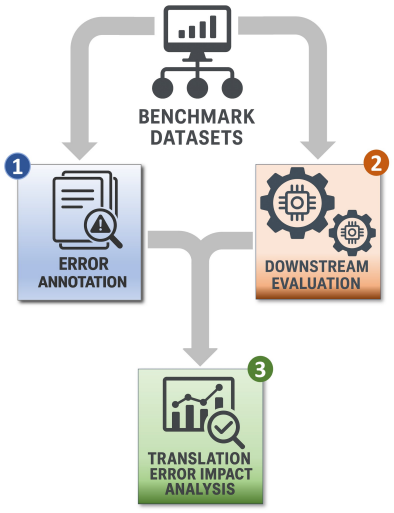

Figure 1: LLM-based TQE pipeline: (1) automatic span-level MQM error annotation, (2) model evaluation, (3) estimating the accuracy impact of translation errors.

1 Introduction

In multilingual evaluation, machine-translated datasets are widely used as reference data, yet translation quality is often overlooked, undermining reliability and comparability (Choenni et al., 2024; Artetxe et al., 2020; Plaza et al., 2024). Human protocols such as MQM (Lommel et al., 2013, 2024) and Error Span Annotation (ESA) (Kocmi et al., 2024) provide increasingly diagnostic assessments (Freitag et al., 2021).

More recently, researchers have treated LLMs themselves as translation judges ("LLM-as-a-judge"; Kocmi and Federmann, 2023), using zero- or few-shot prompting to tag MQM-style error spans. This trend is exemplified by GEMBA and GEMBA-ESA (Freitag 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)).

Two practical questions remain under-addressed: (i) how well do LLM-produced MQM-style spans match expert human spans on real benchmark translations, and (ii) how strongly do translation errors (vs. source-side issues) explain accuracy drops on translated benchmarks?

We study these questions on EU20 (Thellmann et al., 2024) and annotate target-side translation errors with four LLM annotators using an MQM-inspired prompt (§3).

We make three contributions:

1. Human reference and LLM span agreement on EU20 (§3). We release a professional span-level MQM reference for an EU20 subset (225 items, nine target languages) and benchmark four LLM annotators with strict character-offset span matching; GPT-5.2 achieves the highest agreement with humans (mean span-level F1 **0.55** under position-overlap matching).

2. Meta-evaluation with Span-ACES$_{\text{Ref}}$ (§4). We evaluate span localization on Span-ACES$_{\text{Ref}}$ (1,407 items), a cleaned projection of SPAN-ACES (Moghe et al., 2025), and validate the transformation on 178 records where manual review and GPT-5.2 agree on 165/178 (0.93), enabling controlled classic vs. tolerant span metrics.

3. Performance impact with source controls (§5). Using the pipeline in Figure 1, we estimate the impact of target-side translation errors (T) and source-side issues (S) on translated accuracy via logistic regression (fixed effects; English controls; bootstrap CIs): across annotators, T is associated with drops of about **6–8 pp** (full model) and **6–11 pp** among English-solvable items.

The annotated datasets, Span-ACES$_{\text{Ref}}$ resources, and our prompts are included in the supplementary material.[1]

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 Human and LLM MQM Annotation

In this section, we present our MQM-based annotation setup for capturing translation errors and assess the span-level annotation accuracy of LLM-based annotators. We compare LLM-generated MQM annotations against a professionally produced human reference on an EU20 subset and additionally evaluate LLM annotators on Span-ACES$_{\text{Ref}}$.

3.1 Methodology

Human Annotation. Using EU20 (Thellmann et al., 2024) (DeepL translations[2] of 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) into 20 official EU languages), we create a span-level human reference for translation error analysis by manually annotating nine target languages spanning Germanic (DE, DA), Romance (FR, IT, RO), Uralic (ET, HU), Slavic (SL), and Baltic (LT), covering a range of high- to low-resource settings.

For each language, we randomly selected 25 segments (225 total) using a fixed per-dataset quota to balance task types (HellaSwag: 9, ARC: 4, TruthfulQA: 4, GSM8K: 2, MMLU: 6 per language). To avoid trivially short instances, we applied a minimum-length filter during sampling. Dataset statistics for the sampled EU20 subset are shown in Table 1.

Annotation was carried out by professional translators/linguists with one annotator per language using an MQM-inspired, span-based pro-

[1] Code will be released upon publication.

[2] developers.deepl.com/docs

Dataset	k/lang	Ex.	Tok.	Tok. min–max	Sent.
ARC	4	36	56.6	31–98	5.0
GSM8K	2	18	129.5	93–191	8.3
HellaSwag	9	81	166.2	112–253	12.3
MMLU	6	54	153.5	62–402	8.7
TruthfulQA	4	36	99.3	66–201	10.5

Table 1: Aggregated statistics for the manually annotated EU20 subset across nine languages. **k/lang** is the sampling quota per language, yielding **Ex.** total instances per dataset across all languages. **Tok.** and **Sent.** report the mean number of tokens and sentences per instance, respectively. **Tok. min–max** gives the observed token-length range within each dataset.

tocol implemented in a custom Argilla interface (Appendix A.1). For segments with errors, annotators marked erroneous spans and assigned an MQM label and severity (Major/Minor), distinguishing *Accuracy* errors (faithfulness/meaning transfer; Table 8) from *Fluency/Style* errors (grammaticality and naturalness; Table 9). They highlighted target-side error spans and, where applicable, the corresponding source spans (Figure 2), and then submitted a minimally post-edited corrected translation and optional clarification comment (Figure 3).

After collection, we performed systematic consistency checks (e.g., presence of error labels/severity, and source–target span alignment where applicable) and investigated irregular cases such as source–target span mismatches by consulting annotator comments and the original segment context. Where necessary, we corrected format inconsistencies and obvious annotation-entry issues to obtain a clean, machine-readable reference. In addition, we conducted a targeted manual cross-check by comparing human-annotated spans with LLM-proposed annotations as a complementary sanity check.

LLM-based Annotation. To generate automatic, span-level MQM annotations, we use four instruction-tuned LLMs: GPT-5.2[3], GPT-4o-mini[4], Llama-4[5], and Mistral[6]. We adopt a variant of the GEMBA-ESA prompting approach[7] and extend it 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. To ensure comparability with the reference, the prompt restricts labels to the same MQM inventory (Accuracy vs. Fluency/Style, see Table 8 and 9 from Appendix A.1). Each model is prompted to produce a JSON-structured response containing a list of annotated error spans with MQM label and severity.

In addition, we annotate source-side anomalies on the same EU20 subset using GPT-5.2 to account for cases where the source has irregularities that may affect downstream TQE. These annotations follow an MQM-inspired, span-based protocol for fluency/coherence anomalies and distinguish surface issues (e.g., *typo*, *grammar*, *punctuation*, *awkward*) from semantic oddness (e.g., *contradiction*, *broken_coreference*, *implausible_logic*), with severity (major/minor) indicating potential impact on interpretability or answerability.

Metrics for span-level Agreement. We compare span-level MQM annotations from humans and four LLM-based annotators by converting all outputs into a unified span representation and assigning character offsets whenever they can be unambiguously determined from the raw target text. We report span-level Recall and F1 computed with greedy one-to-one matching under fixed thresholds and micro-aggregation over items (TP/FP/FN summed per language). For the position-based overlap coefficient (OC), a candidate pair is matchable if $\mathrm{OC} \geq 0.8$, where OC is the overlap length divided by the minimum span length. For the string-based metric (SIM), we use raw character 3-gram Dice similarity (no normalization) with threshold $\mathrm{SIM} \geq 0.6$. SIM is computed on spans deduplicated by exact target-span text (on both sides), whereas OC operates on the full span lists and requires valid target offsets. Spans with missing offsets cannot be matched under OC and therefore contribute to FP/FN in the OC-based Span-F1 computation. The SIM-based results are provided in Appendix A.2 (Table 10).

Source-overlap. To quantify the tendency to annotate source-driven oddness as target-side errors, we compute the fraction of an annotator's target spans (with valid offsets) that overlap target regions linked to GPT-5.2 source-anomaly annotations (OC threshold 0.8, each annotator span counted at most once).

3.2 Results and Discussion

Span-level Agreement. Table 2 shows that GPT-5.2 yields the highest agreement with the human reference (mean OC F1 $= 0.55$, range 0.34–0.78

[3] platform.openai.com/docs/models/gpt-5.2
[4] platform.openai.com/docs/models/gpt-4o-mini
[5] HF: meta-llama/Llama-4-Scout-17B-16E-Instruct
[6] HF: mistralai/Mistral-Large-Instruct-2411
[7] github.com/MicrosoftTranslator/GEMBA

Comparison	OC Recall	OC F1
Human vs GPT-5.2	.44	.55 (.34–.78)
Human vs Mistral	.17	.23 (.19–.32)
Human vs GPT-4o	.09	.15 (.08–.25)
Human vs LLaMA-4	.08	.13 (.03–.23)
GPT-5.2 vs Mistral	.24	.25 (.18–.36)
GPT-5.2 vs GPT-4o	.15	.21 (.09–.30)
GPT-5.2 vs LLaMA-4	.12	.17 (.05–.26)

Table 2: Span-level agreement based on position overlap (OC) averaged over nine languages (25 items each). We report mean Span-Recall and Span-F1 across languages. Parentheses denote the min–max Span-F1 over languages. OC uses greedy one-to-one matching on target character offsets with threshold 0.8.

Annotator	Mean overlap rate	Range
Human	.08	.04–.13
Mistral	.14	.09–.20
GPT-5.2	.06	.00–.13
GPT-4o	.10	.00–.25
LLaMA-4	.13	.06–.29

Table 3: Source-overlap: fraction of an annotator's target spans with valid offsets that overlap target regions linked to GPT-5.2 source-anomaly annotations (OC threshold 0.8). Each annotator span is counted at most once via its best overlap. Higher values indicate a stronger tendency to mark source-driven oddness as target-side errors.

across languages), outperforming the other three LLM-based annotators. Besides GPT-5.2, Mistral generally achieves higher agreement scores than GPT-4o-mini and LLaMA-4, with LLaMA-4 typically lowest. Further details of the matching procedure are provided in Appendix A.2: it reports the SIM-based agreement (Table 10) and per-language Span-F1 for OC and SIM (Table 11), illustrating that relative model rankings are stable but absolute agreement can differ markedly across language batches. The low agreement between the LLM annotators and the human reference indicates that span boundaries are often chosen differently and that annotators may disagree on what should count as a translation error in the target text.

Source-overlap (source-bias) indicator. As shown in Table 3, Mistral and LLaMA-4 exhibit the highest source-overlap rates, while GPT-5.2 is lowest, consistent with a stricter target-only annotation tendency. Human annotations are non-zero, reflecting that some source issues are preserved in the translation. Per-language source-overlap rates are reported in Appendix A.2 (Table 12), which helps contextualize language-dependent differences in span agreement. The source-overlap analysis helps explain part of this gap: annotators that frequently mark regions tied to source-side irregularities can disagree with systems that focus more strictly on target-only errors.

Qualitative inspection. To contextualize the agreement scores, we performed a manual spot-check of a small number of items across DE/DA/IT/RO. While not exhaustive, the inspection revealed recurring divergence patterns that align with the quantitative results. First, GPT-4o-mini, LLaMA-4 and Mistral sometimes label issues in answer options as *Mistranslation* even when the translation itself is plausible, especially in multiple-choice settings. Such cases are less frequent in the human reference and GPT-5.2. Second, both human and LLM annotators occasionally mark source-side ill-formedness (e.g., ungrammatical or unintelligible source segments) as target-side translation errors, which is consistent with non-zero source-overlap rates (Appendix A.2, Table 12). Third, we observed occasional data irregularities (e.g., source–target mismatches) that can depress span agreement independently of annotation quality. Finally, we found model-specific tendencies: GPT-5.2 sometimes reverts acceptable human post-edits and may miss subtle meaning nuances (e.g., false friends), whereas GPT-4o-mini can produce comparatively broad spans, making the exact error locus less clear. Overall, these observations suggest that disagreement is driven not only by span boundary variation but also by differing judgments about what should count as a target-side translation error, especially in the presence of source anomalies. These findings motivate treating source anomalies as a confounder and reporting agreement both with and without source-linked regions in future work.

4 SpanACES-Ref

To complement the EU20 study with an external benchmark that provides gold target-side error spans, we evaluate the same LLM annotators on Span-ACES$_{\text{Ref}}$, a cleaned projection of Span-ACES.

4.1 Methodology

Span-ACES$_{\text{Ref}}$ construction. A key limitation of SPAN-ACES is that it annotates only the introduced error span per item: any additional errors present in the translation remains unlabeled. As a result, span-based annotators that correctly flag

such extra errors are penalized as false positives, biasing span-level precision (and thus Span-F1) downward. Each item contains a human reference, a good translation, and an incorrect translation that introduces exactly one targeted phenomenon. We construct Span-ACES$_{\text{Ref}}$ by taking the single contentful diff between good and incorrect (case-sensitive token diff), projecting it into the human reference, and discarding items with multiple diffs or ambiguous matches. This projection reduces unlabeled noise and yields more reliable span-level precision/Span-F1.

Mapping to MQM categories. For consistency with our EU20 annotation setup, we map Span-ACES phenomena to MQM types using the ACES authors' released mapping[8] and collapse the resulting MQM types into two coarse categories: *Accuracy* and *Fluency/Style*. In total, Span-ACES$_{\text{Ref}}$ contains 1,155 *Accuracy* instances spanning the 20 EU20 target languages, plus 252 *Fluency/Style* instances (German-only). Dataset composition and the mapping breakdown are reported in Table 13.

Dataset validation (human vs. GPT-5.2). To sanity-check the transformation pipeline and catch projection errors, we manually reviewed a subset of 178 Span-ACES$_{\text{Ref}}$ records obtained by stratified sampling using a fixed checklist: (i) does the projected edit match the original good→incorrect change, (ii) is it still the intended error type in the reference context, and (iii) does it avoid introducing new issues elsewhere? We applied the same checklist with GPT-5.2 and compared final verdicts (`pass`/`fail`). The two assessments agree on 165/178 items (0.93 agreement rate), indicating that most cases are straightforward under these criteria and that GPT-5.2 can serve as a useful auxiliary signal for scalable dataset validation. The remaining 13 disagreements are mostly borderline cases (e.g., type drift after projection or subtle side effects).

Span metrics (classic vs. tolerant). We report classic span metrics (Span-F1 $F1$ and Span-Recall R) and a tolerant variant ($F1_t$, R_t) that accounts for common boundary near-misses (a few extra tokens) without using overlap thresholds. Classic matching require exact equality between a predicted span and the gold span. Tolerant matching is gold-centered: a prediction is counted correct if it contains the gold span and (optionally) up to k tokens on either side (left/right) of the gold span. We first average metrics within each phenomenon and then aggregate per MQM category by two weighting schemes: (i) **mean (N)** weights each phenomenon by its sample count, and (ii) **mean (cap)** caps each phenomenon's weight at $C{=}25$ to avoid domination by a few large phenomena.

Baseline vs. our prompt. We evaluate two prompting setups on Span-ACES$_{\text{Ref}}$: (i) the initial GEMBA-ESA baseline prompt, and (ii) our updated MQM-style prompting (Section 3.1, paragraph "LLM-based Annotation"). Unless stated otherwise, tables report each metric as *Baseline/Ours*.

4.2 Results and Discussion

Accuracy. Table 4 reports span localization performance on the *Accuracy* subset. Mistral is strongest overall, and tolerant matching yields large gains for all models, indicating frequent boundary near-misses even when the correct error region is detected. Under mean (N), our updated prompt improves classic F1 for GPT-4o-mini (**.10**→**.21**) and Mistral (**.24**→**.34**), while LLaMA-4 drops slightly (**.13**→**.10**). The classic-to-tolerant gap is particularly large for GPT-4o-mini and LLaMA-4 (e.g., GPT-4o-mini $F1_t$: **.33**→**.48**), consistent with spans that are wider or shifted relative to gold boundaries. We interpret the LLaMA-4 drop as a plausible prompt–model sensitivity effect (rather than a causal claim), given that it occurs under both classic and tolerant matching.

Fluency/Style. Table 5 shows results on *Fluency/Style*. The absolute scores are higher than for *Accuracy*, but this subset is German-only and limited to a small set of phenomena. We therefore interpret it primarily as a controlled diagnostic rather than a cross-lingual benchmark result. Compared to the baseline, our prompt yields a large improvement for GPT-4o-mini on this controlled *Fluency/Style* set (classic F1 **.05**→**.32**, mean (N)), while Mistral (**.45**→**.46**) and LLaMA-4 (**.15**→**.16**) improve only slightly. This pattern mirrors the *Accuracy* results: prompt refinements help some models substantially, but do not transfer uniformly across architectures.

Takeaways. Across models and categories, tolerant scores ($F1_t$, R_t) are consistently higher than classic scores, indicating frequent boundary near-misses rather than complete detection failures. This matches a common pattern in LLM annotations:

[8] github.com/EdinburghNLP/ACES

	GPT-4o-mini				LLaMA-4				Mistral			
Aggregation	**F1**	**R**	**F1$_t$**	**R$_t$**	**F1**	**R**	**F1$_t$**	**R$_t$**	**F1**	**R**	**F1$_t$**	**R$_t$**
mean (N)	.10/.21	.17/.23	.33/.48	.57/.55	.13/.10	.17/.12	.34/.31	.43/.36	.24/.34	.36/.42	.47/.60	.70/.71
mean (cap)	.17/.28	.25/.31	.31/.46	.49/.51	.21/.20	.26/.22	.34/.35	.43/.40	.28/.39	.39/.45	.41/.57	.59/.65

Table 4: Span-ACES$_{\text{Ref}}$ span localization on the *Accuracy* subset. Classic matching (Span-F1 $F1$, Span-Recall R) requires exact span equality; tolerant matching ($F1_t$, R_t) allows gold-centered boundary slack up to k=3 tokens per side. Each cell reports Baseline/Ours (initial GEMBA-ESA vs. our updated prompt).

	GPT-4o-mini				LLaMA-4				Mistral			
Aggregation	**F1**	**R**	**F1$_t$**	**R$_t$**	**F1**	**R**	**F1$_t$**	**R$_t$**	**F1**	**R**	**F1$_t$**	**R$_t$**
mean (N)	.05/.32	.09/.37	.41/.57	.68/.67	.15/.16	.19/.17	.42/.41	.59/.47	.45/.46	.62/.54	.55/.56	.82/.67
mean (cap)	.04/.32	.08/.36	.47/.61	.74/.71	.12/.14	.16/.15	.40/.41	.58/.49	.47/.48	.62/.56	.58/.59	.83/.70

Table 5: Span-ACES$_{\text{Ref}}$ span localization on the *Fluency/Style* subset (German-only). Matching and aggregation follow Table 4. Each cell reports Baseline/Ours.

different systems often point to the same error region but choose slightly different span boundaries (e.g., broader or shifted spans)

5 Performance Impact Analysis

Following the TQE pipeline in Figure 1, we link span-level error annotations to whether models answer translated instances correctly and analyze these associations with logistic regression to assess how translation quality relates to downstream LLM performance.

5.1 Methodology

Data and annotations. For our performance analysis, we use MMLU, ARC, and GSM8K from the EU20 benchmark suite and annotate target-side translation errors with an MQM-inspired prompting scheme (Section 3) using four LLM-based annotators: GPT-5.2, GPT-4o-mini, Llama-4, and Mistral. To control for issues already present in the original English items (e.g., inconsistencies or ambiguities that may affect model accuracy independent of translation), we additionally annotate the English source for a paired subset of 651 translated instances (item–language pairs) using GPT-5.2. Together with the corresponding target-side annotations, this paired subset enables a direct comparison of source-side issues (S) versus translation-induced errors (T). For the performance-impact regressions, we exclude HellaSwag and TruthfulQA: in HellaSwag, many "odd" endings are intentionally constructed distractors (task design rather than source noise), which would blur the interpretation of S, and for TruthfulQA we could not reliably derive binary correctness labels from the released evaluation logs. Dataset statistics are provided in Table 6.

Statistic	**MMLU**	**GSM8K**	**ARC**	**Total**
N_{pair}	237	215	199	651
N_{obs}	1896	1505	1592	4993
y=1	1127	847	966	2940
y=0	769	658	626	2053
y^{EN}=1	1373	1017	1255	3645
y^{EN}=0	523	488	337	1348
SET=∅	814	673	722	2209
SET≠∅	134	187	74	395
TET=∅	472	347	478	1297
TET≠∅	476	513	318	1307

Table 6: Dataset statistics for the regression dataset (HellaSwag and TruthfulQA excluded). N_{pair} counts unique (sample, target-language) pairs and N_{obs} translated instances across evaluation models. y and y^{EN} are translated vs. English correctness (counts over N_{obs}). SET/TET indicate empty vs. non-empty annotated error lists. These counts are over (pair, annotator) records (i.e., $4 \times N_{\text{pair}}$).

Model evaluation. We evaluate eight instruction-tuned multilingual LLMs (Appendix C.1, Table 14) using a modified version of EleutherAI's LM Evaluation Harness[9] and record a binary outcome (correct/incorrect) for each item in English and in each translated variant.

Analysis dataset. We merge per-instance correctness with the corresponding span-level error annotations to construct an analysis dataset that links (i) the presence of source-side issues and/or translation errors to (ii) each evaluated LLM's correctness

[9] github.com/EleutherAI/lm-evaluation-harness

on the English original and its translated variants. We analyze this dataset with logistic regression, focusing on the paired-annotation subset when comparing translation errors against source-side issues.

Regression setup. We use logistic regression to estimate how the presence of translation errors (T) and source-side issues (S) predicts correctness on translated items, while controlling for whether the model solves the English original and for systematic differences across languages, datasets, and evaluation models. A key property of our data is that for each underlying item and evaluation model we observe both (i) performance on the English original and (ii) performance on its EU20 translations, enabling a clean separation between English-level ability and translation-associated failures.

Instances and variables. Our unit of analysis is a translated instance: we record whether evaluation model m answers an item in target language ℓ correctly ($y \in \{0, 1\}$; English excluded). For the same item and model, we record whether the English original is answered correctly, denoted $y^{EN} \in \{0, 1\}$. From the span-level annotations, we derive two binary indicators: $T = 1$ if the translation contains at least one annotated target-side error (annotator-specific), and $S = 1$ if at least one issue was annotated in the English source. Since S refers to the English original, it is shared across all translations of the same item.

Specifications. We fit logistic regressions with fixed effects for target language, dataset, and evaluation model, denoted by C.

(A) Full model with English control. This specification asks: *How strongly are translation errors associated with lower accuracy after controlling for whether the model solves the English original?*

$$\Pr(y = 1) = \sigma\big(\beta_T T + \beta_S S + \beta_{EN} y^{EN} + C\big). \quad (1)$$

(B) Only items solved in English. Estimated on the subset with $y^{EN} = 1$, this specification asks: *Among items the model answers correctly in English, how are translation errors (T) and source issues (S) associated with accuracy in translation?*

$$\Pr(y = 1) = \sigma\big(\beta_T T + \beta_S S + C\big). \quad (2)$$

Ablations (omit S). To assess sensitivity to omitted-variable bias, we refit (A) and (B) without S. If translation errors correlate with source-side issues that also depress accuracy, omitting S can make the estimated translation-error association more negative.

	Spec.	AME(T)	AME(S)
GPT-4o	A	-6.74 [-10.44, -3.17]	-3.50 [-7.46, 0.61]
	$A_{\neg S}$	-6.99 [-10.62, -3.22]	–
	B	-9.64 [-14.05, -5.49]	-4.34 [-8.75, 0.19]
	$B_{\neg S}$	-10.01 [-14.52, -5.86]	–
GPT-5.2	A	-6.76 [-10.61, -3.11]	-3.54 [-7.47, 0.57]
	$A_{\neg S}$	-7.01 [-10.80, -3.26]	–
	B	-8.98 [-13.35, -4.93]	-4.64 [-9.19, -0.10]
	$B_{\neg S}$	-9.26 [-13.46, -5.03]	–
LLaMA-4	A	-7.51 [-12.47, -2.70]	-3.47 [-7.56, 0.62]
	$A_{\neg S}$	-7.83 [-12.50, -3.33]	–
	B	-10.71 [-16.56, -4.91]	-4.15 [-8.66, 0.42]
	$B_{\neg S}$	-11.23 [-17.10, -5.57]	–
Mistral	A	-5.59 [-9.50, -1.69]	-3.69 [-7.66, 0.44]
	$A_{\neg S}$	-5.82 [-9.47, -2.04]	–
	B	-6.12 [-10.30, -1.86]	-4.74 [-9.29, -0.20]
	$B_{\neg S}$	-6.48 [-10.74, -2.18]	–

Table 7: Average marginal effects (AMEs; probability points) of target-side translation errors T and source-side issues S on correctness in translation. Spec. A: full model with English correctness control y^{EN}. Spec. B: subset with y^{EN}=1. $A_{\neg S}$ and $B_{\neg S}$ omit S. All models include fixed effects for target language, dataset, and evaluation model. 95% CIs from block bootstrap over items (A: 4993 translated instances / 626 items; B: 3645 / 598).

Effect reporting and uncertainty. For each target annotator and specification, we report average marginal effects (AMEs) of T and S in probability points (more interpretable than log-odds). We compute confidence intervals using a block bootstrap over underlying items: we resample items with replacement, include all associated rows (across target languages and evaluation models) for each resampled item, refit the model, and take percentile intervals from the bootstrap distribution.

5.2 Results and Discussion

Table 7 answers our main questions about how translation quality relates to downstream accuracy. Across all four annotators, target-side translation errors (T) are consistently associated with lower correctness in translation, even when controlling for source-side issues (S), English correctness, and fixed effects. In the full model with an English control (Spec. A), AME(T) ranges from **-5.59 to -7.51** pp and all 95% bootstrap CIs lie entirely below zero. Conditioning on items solved in English (Spec. B; y^{EN}=1) yields larger drops, with AME(T) between **-6.12 and -10.71** pp, again with CIs fully below zero. Together, A and B provide complementary evidence that translation errors are associated with accuracy losses and that these losses persist

even when the underlying item is solvable in English.

Source-side issues (S) are directionally negative but weaker and less stable. In Spec. A, AME(S) is around -3.5 pp, yet CIs generally overlap zero, consistent with y^{EN} absorbing much of the item difficulty signal. In Spec. B, AME(S) is larger (-4.1 to -4.7 pp) and becomes significant for some annotators (e.g., GPT-5.2, Mistral), suggesting that source anomalies can additionally depress translated performance once we restrict to "English-solvable" items. Omitting S changes AME(T) only marginally (typically about -0.2 to -0.5 pp, e.g., GPT-5.2: -6.76 to -7.01 in A, -8.98 to -9.26 in B). Thus, T is not simply acting as a proxy for S. Logit coefficients and odds ratios mirror these patterns (Appendix Tables 16 and 15), with OR(T) consistently below 1 (roughly .65–.72 in A and .54–.69 in B).

Practical impact. Table 7 reports effects in probability points (pp), which are directly interpretable as accuracy drops. Across annotators, a translation with at least one annotated error (T=1) is less likely to be answered correctly by about **6-8 pp** in the full model (Spec. A) and by about **6-11 pp** when we restrict to items solved in English (Spec. B; y^{EN}=1). That is, translation errors are associated with a noticeable loss in measured accuracy, even when the underlying item is solvable in English.

To translate these per-item drops into an overall impact, we can scale by how often T=1 occurs. A simple approximation is: *overall loss* $\approx \Pr(T{=}1) \times |\text{AME}(T)|$. With GPT-5.2 annotations, T=1 occurs in **63%** of cases. Combined with AME(T)=**-6.76** pp (Spec. A), this implies roughly **4.3 pp** lower overall accuracy attributable to translation errors. With LLaMA-4 annotations, T=1 occurs in **24%** of cases. With AME(T)=**-7.51** pp (Spec. A), the implied loss is about **1.8 pp**. In the English-correct subset (Spec. B), the same calculation yields larger impacts (e.g., GPT-5.2: **62%**×**8.98** pp ≈ **5.6 pp**).

For intuition, every additional **10%** of items that contain a translation error corresponds to about $|\text{AME}(T)|/10$ points of accuracy loss. In our results, this is roughly **0.6–0.8 pp per 10%** in Spec. A and **0.6–1.1 pp per 10%** in Spec. B. These calculations are meant as a readable approximation. They ignore overlap between T and source-side issues S, beyond what is already accounted for by the regression controls.

6 Conclusion

We analyze how translation noise in machine-translated benchmarks affects multilingual LLM evaluation by combining span-level MQM-style error annotation with performance modeling. On an EU20 subset with a professional human reference (225 items, nine target languages), GPT-5.2 shows the highest agreement with humans (mean span-level F1 $= 0.55$ under position-overlap matching), but disagreement remains common due to span-boundary variation and differing views of what counts as a target-side error. We further evaluate span localization on Span-ACES$_{\text{Ref}}$ (1,407 items), which provides gold target-side error spans for controlled comparisons: tolerant span metrics are much higher than classic ones, suggesting many residual errors are boundary near-misses. Finally, in regressions with English controls and explicit source-side issue indicators, target-side translation errors are robustly associated with lower translated accuracy (about 6–8 pp in the full model and 6–11 pp when restricting to items solved in English), while source-side issues have smaller and less stable effects. Overall, translation-aware audits and source controls are needed to make multilingual benchmark scores more interpretable and comparable.

7 Future Work

We see several directions for future work. First, expanding beyond coarse MQM groupings to finer-grained error types and multi-span phenomena could improve diagnostic power and support more targeted benchmark cleanup. Second, developing and standardizing boundary-robust span evaluation (and prompting strategies that encourage minimal, well-aligned spans) may reduce apparent disagreement that is driven mainly by span placement. Third, scaling dataset quality control by combining human checks with reliable LLM-assisted validation, and extending source-side anomaly annotation beyond small subsets, would enable translation-aware benchmark releases with clearer guarantees about both source and target quality.

Limitations

First, high-quality span-annotated references remain scarce, particularly for lower-resource languages and for fine-grained MQM error inventories; our EU20 human reference covers nine target languages and a limited sample size (225 items), and

Span-ACES$_{\text{Ref}}$ covers controlled phenomena rather than naturally occurring error distributions. Second, span-level MQM annotation has no widely accepted cross-lingual gold standard, and expert annotation is costly; while we validate parts of our setup (human reference checks and a Span-ACES$_{\text{Ref}}$ validation subset), residual subjectivity in span boundaries and in what counts as a target-side error is unavoidable. Third, our performance-impact analysis is correlational: although we control for English correctness, fixed effects, and source-side issues, unobserved confounders and model/specification choices can still affect effect sizes, and some signals (especially source-side issues) are sparse and therefore estimated with greater uncertainty. Fourth, automatic annotation quality depends on the chosen LLMs and prompts; we mitigate this by comparing multiple LLM annotators and reporting agreement/robustness analyses, but conclusions about "best" annotators may not transfer to other models, prompting styles, or annotation interfaces. Finally, our results reflect a specific selection of benchmarks (EU20 tasks), MT system, target languages, and evaluated LLMs; translation artifacts and their downstream impact may differ in other domains, genres, or translation pipelines.

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.

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 Bene, 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.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs Breaking MT Metrics? Results of the WMT24 Metrics Shared Task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

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. 2023. 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 Popovi, 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.

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.

A Appendix: MQM Annotation

A.1 Annotation Guide and Interface

This subsection summarizes the manual protocol used to create our human reference and documents the annotation interface. Our guidelines are based on the Multidimensional Quality Metrics (MQM) framework,[10] using a reduced, task-focused typology. Concretely, we group error labels into two high-level dimensions—*Accuracy* (semantic faithfulness) and *Fluency/Style* (well-formedness and naturalness)—and annotate errors as spans. For each error, annotators highlight the erroneous span in the target translation and, where applicable, the corresponding source span to support span-level alignment analyses. We treat *Addition* and *Omission* as asymmetric cases: *Addition* is annotated only on the target side (no source span), whereas *Omission* is annotated only on the source side (no target span).

Label	Description
Addition	Adds content not supported by source
Omission	Drops source content
Mistranslation	Meaning is changed or incorrect
Under-translation	Meaning is too vague; nuance is lost
Over-translation	Adds unwarranted specificity or detail
Reordering	Changes attachment or meaning
Untranslated	Leaves a source fragment untranslated
Wrong language	Uses tokens from unintended language
Do-not-translate	Content that should remain unchanged

Table 8: Reduced MQM label set for *Accuracy* errors used in manual annotation.

Label	Description
Grammar	Grammatical error (agreement, tense, case)
Spelling	Spelling or character error
Punctuation	Punctuation error affecting readability
Inconsistent	Inconsistent terminology or naming
Awkward	Unnatural or non-idiomatic phrasing
Unintelligible	Output is not reliably interpretable

Table 9: Reduced MQM label set for *Fluency/Style* errors used in manual annotation.

Annotators assign one of two severity levels to each annotated span. *Major* errors change meaning, can mislead, or substantially harm understanding/faithfulness, whereas *Minor* errors largely preserve meaning and primarily affect fluency/style or constitute a limited local defect. If no label fits or annotators are uncertain, they mark the span as *Other/Unknown* (used sparingly).

[10] https://themqm.org/the-mqm-full-typology/

Figure 2 and Figure 3 illustrate the Argilla-based interface and the annotation workflow. Concretely, annotators proceeded as follows:

1. Decide whether the translation is error-free (*Yes/No/Unsure*).
2. If errors are present, mark erroneous target spans and assign an MQM label and severity.
3. Mark corresponding source spans where applicable; *Addition* is annotated only on the target side, whereas *Omission* is annotated only on the source side.
4. Provide a minimally post-edited corrected translation.
5. Answer a control question indicating whether all important errors were captured.
6. Optionally add comments for clarification.

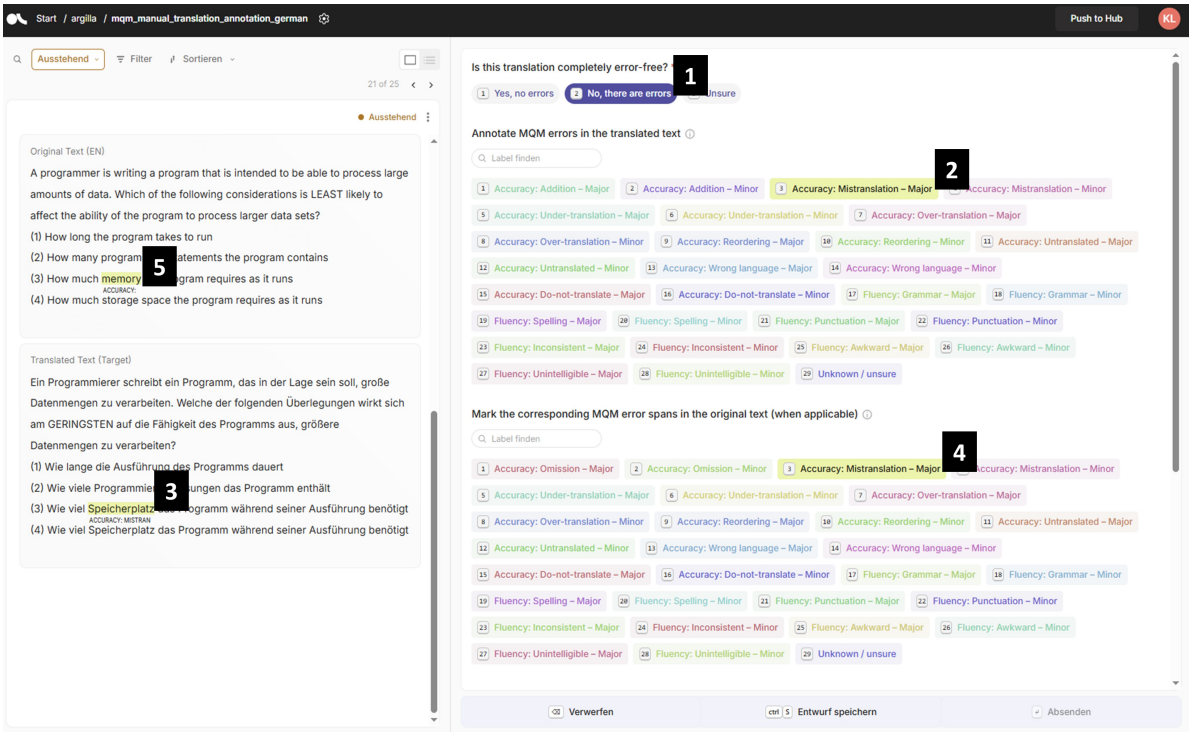

Figure 2: Argilla interface for span-based MQM annotation: target span highlighting, label selection, severity assignment, and source–target span alignment.

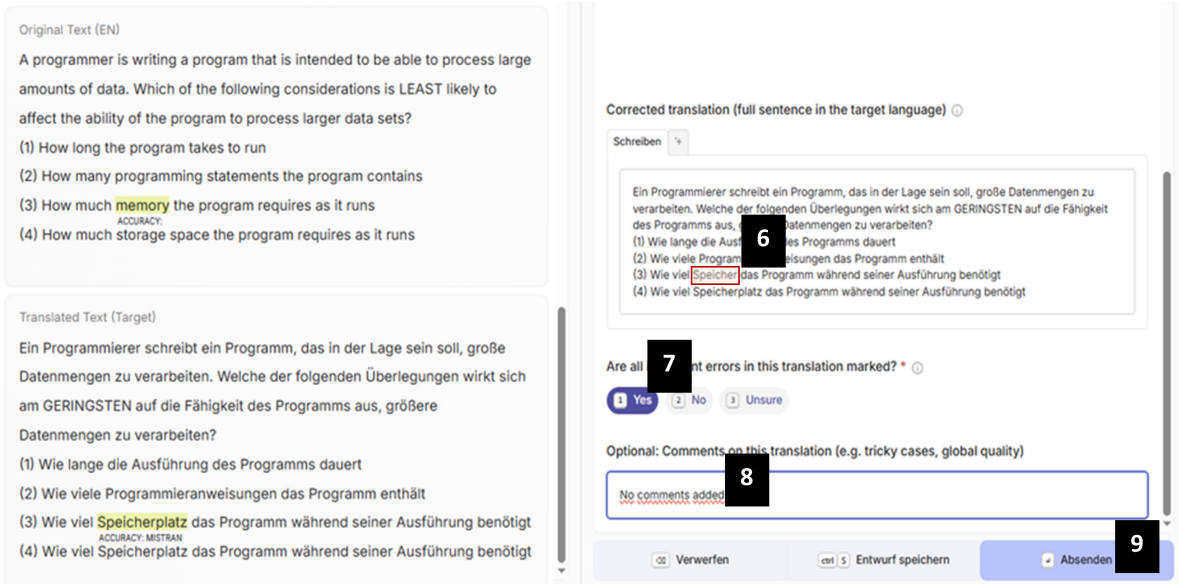

Figure 3: Argilla interface for post-editing and metadata: corrected translation field, completeness control question, and optional global comments.

A.2 Metrics and Matching

We compare span sets using (i) a position-based overlap coefficient (OC) computed from target character offsets, and (ii) a string-based similarity metric (SIM) defined as raw character 3-gram Dice similarity, without text normalization. Offsets are validated against the raw target string. If missing or inconsistent, we attempt to recover them via exact substring search and leave offsets missing in ambiguous cases.

For OC-based matching, a gold span and a predicted span are matchable if $\mathrm{OC} \geq 0.8$, where OC is the intersection length divided by the minimum span length. OC matching requires valid target offsets. Spans without offsets cannot be matched and contribute to FP/FN. For SIM-based matching, we compute raw 3-gram Dice and apply threshold $\mathrm{SIM} \geq 0.6$ on spans deduplicated by exact target-span text on both sides.

To compute precision/recall/F1, we perform greedy one-to-one matching: all matchable span pairs are sorted by score (descending) and selected if neither span has been matched before. For each comparison, we aggregate TP/FP/FN across samples to obtain micro-precision, micro-recall, and micro-F1.

Comparison	SIM Recall	SIM F1
Human vs GPT-5.2	.38	.45 (.30–.78)
Human vs Mistral	.14	.17 (.10–.26)
Human vs GPT-4o-mini	.11	.15 (.09–.25)
Human vs LLaMA-4	.06	.10 (.02–.19)
GPT-5.2 vs Mistral	.16	.17 (.09–.24)
GPT-5.2 vs GPT-4o-mini	.15	.21 (.11–.30)
GPT-5.2 vsLLaMA-4	.09	.12 (.03–.19)

Table 10: Span-level agreement based on string similarity (SIM) averaged over nine languages (25 items each). We report mean Span-Recall and Span-F1 across languages. Parentheses denote the min–max Span-F1 over languages. SIM uses raw character 3-gram Dice with threshold 0.6 and greedy one-to-one matching on spans deduplicated by exact target-span text.

Lang	Metric	GPT-5.2 vs			Human vs			
		G4o	L4	Mis	G4o	L4	Mis	GPT-5.2
DA	OC	.23	.17	.25	.12	.13	.19	.48
DE	OC	.16	.14	.19	.16	.13	.20	.61
ET	OC	.20	.20	.24	.14	.14	.20	.57
FR	OC	.17	.12	.24	.10	.09	.19	.34
HU	OC	.28	.23	.28	.15	.12	.22	.44
IT	OC	.21	.18	.36	.16	.14	.29	.53
LT	OC	.23	.05	.18	.16	.04	.25	.49
RO	OC	.09	.18	.25	.08	.14	.21	.68
SL	OC	.30	.26	.31	.25	.23	.32	.78
DA	SIM	.23	.08	.10	.09	.08	.15	.30
DE	SIM	.14	.12	.19	.16	.10	.14	.39
ET	SIM	.20	.13	.13	.19	.09	.15	.44
FR	SIM	.18	.13	.19	.09	.06	.10	.34
HU	SIM	.16	.14	.20	.10	.10	.18	.49
IT	SIM	.20	.12	.24	.18	.10	.25	.51
LT	SIM	.23	.03	.11	.16	.02	.26	.39
RO	SIM	.11	.17	.12	.09	.09	.12	.42
SL	SIM	.30	.19	.24	.25	.19	.22	.78

Table 11: Span-F1 by language for OC and SIM. G4o = GPT-4o-mini, L4 = LLaMA-4, Mis = Mistral.

Lang	G4o	L4	Mis	GPT-5.2	Human
DA	.00	.07	.10	.00	.04
DE	.17	.09	.20	.06	.13
ET	.09	.09	.18	.04	.05
FR	.10	.17	.09	.07	.07
HU	.25	.15	.14	.09	.04
IT	.15	.22	.18	.13	.13
LT	.00	.06	.15	.04	.09
RO	.04	.07	.12	.03	.07
SL	.14	.29	.09	.12	.11

Table 12: Source-overlap rate by language (Span-Recall-style micro aggregation): fraction of target error spans (with valid offsets) that overlap target regions linked to GPT-5.2 source-anomaly annotations (OC threshold 0.8).

B Appendix: RefSpanACES

Samples	Lang.	MQM-Error-Type	Span-ACES-Phenomena
Category: Accuracy			
470	DE	mistranslation	pleonastic_it:substitution
180	DE	mistranslation	coreference-based-on-commonsense
55	DE	mistranslation	overly-literal-vs-ref-word
42	FR	mistranslation	coreference-based-on-commonsense
32	DE	mistranslation	overly-literal-vs-explanation
25	RO	mistranslation	hallucination-date-time
18	DA	mistranslation	hallucination-date-time
18	ES	mistranslation	hallucination-date-time
17	DE	mistranslation	overly-literal-vs-synonym
16	DE	mistranslation	hallucination-date-time
16	FR	mistranslation	hallucination-date-time
16	HU	mistranslation	hallucination-date-time
16	SV	mistranslation	hallucination-date-time
15	PT	mistranslation	hallucination-date-time
13	SL	mistranslation	hallucination-date-time
11	DE	mistranslation	real-world-knowledge-hypernym-vs-distractor
10	DE	mistranslation	real-world-knowledge-entailment
10	NL	mistranslation	hallucination-date-time
9	DE	mistranslation	real-world-knowledge-synonym-vs-antonym
9	ET	mistranslation	hallucination-date-time
8	SK	mistranslation	hallucination-date-time
5	PL	mistranslation	hallucination-date-time
1	DE	mistranslation	ordering-mismatch
1	LT	mistranslation	hallucination-date-time
105	DE	untranslated	untranslated-vs-ref-word
71	DE	no-translate	do-not-translate
31	DE	untranslated	untranslated-vs-synonym
7	CS	addition	addition
7	FR	addition	addition
7	NL	addition	addition
6	DE	addition	addition
6	PT	addition	addition
5	SK	addition	addition
5	SL	addition	addition
4	BG	addition	addition
4	DA	addition	addition
4	RO	addition	addition
4	SV	addition	addition
3	LT	addition	addition
3	LV	addition	addition
2	EL	addition	addition
2	ES	addition	addition
2	HU	addition	addition
2	IT	addition	addition
1	ET	addition	addition
1	FI	addition	addition
1641	**20**	**4**	**14**
Category: Fluency			
123	DE	grammar	anaphoric_intra_non-subject_it:substitution
103	DE	grammar	anaphoric_intra_subject_it:substitution
85	DE	grammar	anaphoric_intra_they:substitution
35	DE	grammar	anaphoric_group_it-they:substitution
346	**1**	**1**	**4**

Table 13: Span-ACES$_{\text{Ref}}$ statistics grouped by MQM error categories (Accuracy and Fluency) comprising unfiltered **1641** annotated instances. Each row lists per-language sample counts and the corresponding Span-ACES phenomenon. For each block, *Totals* report: total samples in the block, number of distinct languages represented, number of distinct MQM types in the block, and number of distinct Span-ACES phenomena.

C Appendix: Performance Analysis

C.1 Models

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 14: Model Overview.

C.2 Regression Analysis

Annotator	**Spec.**	**coef(T)**	**coef(S)**
GPT-4o	A	-.40 [-.62, -.19]	-.21 [-.45, .04]
	$A_{\neg S}$	-.41 [-.64, -.20]	–
	B	-.58 [-.84, -.34]	-.27 [-.55, .01]
	$B_{\neg S}$	-.60 [-.86, -.36]	–
GPT-5.2	A	-.41 [-.64, -.19]	-.21 [-.45, .03]
	$A_{\neg S}$	-.42 [-.66, -.20]	–
	B	-.56 [-.86, -.31]	-.28 [-.57, -.01]
	$B_{\neg S}$	-.57 [-.86, -.31]	–
LLaMA-4	A	-.44 [-.72, -.16]	-.21 [-.45, .04]
	$A_{\neg S}$	-.45 [-.73, -.20]	–
	B	-.61 [-.95, -.29]	-.25 [-.54, .03]
	$B_{\neg S}$	-.64 [-.97, -.34]	–
Mistral	A	-.33 [-.58, -.10]	-.22 [-.47, .03]
	$A_{\neg S}$	-.35 [-.58, -.12]	–
	B	-.38 [-.66, -.12]	-.29 [-.58, -.01]
	$B_{\neg S}$	-.40 [-.69, -.13]	–

Table 15: Logit coefficients (log-odds) for translation errors T and source-side issues S with 95% bootstrap CIs. Spec. A includes English correctness y^{EN}. Spec. B restricts to y^{EN}=1. $A_{\neg S}$/$B_{\neg S}$ omit S. All models include fixed effects for target language, dataset, and evaluation model.

Annotator	**Spec.**	**OR(T)**	**OR(S)**
GPT-4o	A	.67 [.54, .83]	.81 [.64, 1.04]
	$A_{\neg S}$	.66 [.53, .82]	–
	B	.56 [.43, .71]	.77 [.58, 1.01]
	$B_{\neg S}$	.55 [.42, .70]	–
GPT-5.2	A	.67 [.53, .83]	.81 [.64, 1.03]
	$A_{\neg S}$	.66 [.52, .82]	–
	B	.57 [.43, .74]	.75 [.56, .99]
	$B_{\neg S}$	.56 [.43, .73]	–
LLaMA-4	A	.65 [.49, .85]	.81 [.63, 1.04]
	$A_{\neg S}$	.63 [.48, .82]	–
	B	.54 [.39, .75]	.78 [.58, 1.03]
	$B_{\neg S}$	.53 [.38, .71]	–
Mistral	A	.72 [.56, .90]	.80 [.63, 1.03]
	$A_{\neg S}$	.71 [.56, .88]	–
	B	.69 [.52, .89]	.75 [.56, .99]
	$B_{\neg S}$	.67 [.50, .88]	–

Table 16: Odds ratios (OR) for translation errors T and source-side issues S with 95% bootstrap CIs. Spec. A includes English correctness y^{EN}. Spec. B restricts to y^{EN}=1. $A_{\neg S}$/$B_{\neg S}$ omit S. All models include fixed effects for target language, dataset, and evaluation model.