Diagnosing Translated Benchmarks: An Automated Quality Assurance Study of the EU20 Benchmark Suite

Klaudia Thellmann[1†], **Bernhard Stadler**[1,2†], **Michael Färber**[1]
[1]TU Dresden and ScaDS.AI, [2]InfAI e.V.
{klaudia-doris.thellmann, bernhard.stadler, michael.faerber}@tu-dresden.de
[†] Main authors

Abstract

Machine-translated benchmark datasets reduce costs and offer scale, but noise, loss of structure, and uneven quality weaken confidence. What matters is not merely whether we can translate, but also whether we can measure and verify translation reliability at scale. We study translation quality in the EU20 benchmark suite, which comprises five established benchmarks translated into 20 languages, via a three-step automated quality assurance approach: (i) a structural corpus audit with targeted fixes; (ii) quality profiling using a neural metric (COMET, reference-free and reference-based) with translation service comparisons (DeepL / ChatGPT / Google); and (iii) an LLM-based span-level translation error landscape. Trends are consistent: datasets with lower COMET scores exhibit a higher share of accuracy/mistranslation errors at span level (notably HellaSwag; ARC is comparatively clean). Reference-based COMET on MMLU against human-edited samples points in the same direction. We release cleaned/corrected versions of the EU20 datasets, and code for reproducibility. In sum, automated quality assurance offers practical, scalable indicators that help prioritize review – complementing, not replacing, human gold standards.

1. Introduction

Large Language Models (LLMs) have transformed NLP, yet rigorous multilingual evaluation remains challenging beyond high-resource settings. Across Europe, language- and region-specific suites have emerged, covering e.g. Scandinavian languages (Nielsen, 2023), Norwegian (Samuel et al., 2023), German (Pfister and Hotho, 2024), Italian (Magnini et al., 2025), and Iberian languages (Baucells et al., 2025), Czech (Fajcik et al., 2025), Polish (Jassem et al., 2025), Greek (Peng et al., 2025), and French (Faysse et al., 2025). These native resources improve quality and task relevance, but heterogeneity in scope, construction protocol, and task mix limits parallelism and cross-language comparability at scale (Ott et al., 2022; Srivastava et al., 2023; Yang et al., 2019)

Translating existing benchmarks automatically is a pragmatic alternative that scales, but concerns about translation noise, loss of structure, and uneven quality limit trust in such evaluations (Plaza et al., 2024; Meng et al., 2022; NLLB Team et al., 2022). As a result, the question is not merely whether we can translate benchmarks, but whether machine-translated benchmarks meet quantifiable reliability and diagnostic criteria to guide LLM development and cross-language comparison of LLMs at scale.

We ground our study in EU20 (Thellmann et al., 2024), which translates five established English benchmarks into 20 European languages using DeepL. While the EU20 benchmark suite offers scale and coverage, comprehensive quality assurance, whether human-based or automated, was not the primary focus of the initial release. We take a first step toward scalable validation for EU20 by combining structural diagnostics with two complementary, automated translation quality estimation (TQE) methods (i) neural quality estimation based on COMET scores (Rei et al., 2020, 2023; Guerreiro et al., 2024) and (ii) an LLM-as-a-judge procedure (Kocmi and Federmann, 2023b). Automated TQE does not replace expert human review, but helps prioritize where to look first (e.g., accuracy vs. fluency issues) and provides indicators of translation quality under budget constraints.

In this paper, we operationalize translation quality validation on EU20 along three dimensions:

1. **Structural integrity**: a corpus-level audit of field completeness, split/subset consistency, and cross-language coverage (Section 3).
2. **Item-level quality profiling**: a quality landscape, by task and language, for translated benchmark entries using reference-free and reference-based xCOMET-XXL (Guerreiro et al., 2024), including paired comparisons across distinct translation services – EU20/DeepL[1], Okapi/ChatGPT[2] (Lai et al., 2023), and Global-MMLU/Google Translate[3] (Singh et al., 2025) (Section 4).
3. **Span-level diagnostic validity**: an interpretable error landscape from an LLM-as-a-judge TQE setup, quantifying error categories (Accuracy/Mistranslation, Fluency, Other) and severities across languages and

[1] https://www.deepl.com
[2] https://openai.com/research/gpt-4
[3] https://translate.google.com

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9030–9043
11–16 May 2026.

tasks, and testing convergence with xCOMET-XXL (Section 5).

To make this operationalization actionable and reproducible, we provide the following artifacts:

- A cleaned and corrected version of EU20 with documented fixes from our structural audit[4] (Section 3).
- Code for the structure-preserving cleaning and correction used in the audit, enabling reproducible maintenance of the EU20 benchmark suite[5].
- Our LLM-as-a-judge TQE setup (prompts, few-shot exemplars, and scripts) for span-level error annotation[6].

2. Related Work

Prior work on European LLM evaluation spans three strands. First, gold-standard resources are created directly in the target language or via human translation/editing, yielding high in-language validity but slower coverage growth and limited cross-task parallelism. Examples include ScandEval (Scandinavian) (Nielsen, 2023), NorBench (Norwegian) (Samuel et al., 2023), SuperGLEBer (German) (Pfister and Hotho, 2024), Evalita-LLM (Italian) (Magnini et al., 2025), IberoBench (Iberian) (Baucells et al., 2025), as well as BenCzechMark (Czech) (Fajcik et al., 2025), LLMzSzŁ (Polish) (Jassem et al., 2025), Plutus (Greek finance) (Peng et al., 2025), and CroissantLLM (French) (Faysse et al., 2025). Task-specific multilingual benchmarks such as MLQA (Lewis et al., 2020), XNLI (Conneau et al., 2018), XCOPA (Ponti et al., 2020), XQuAD (Artetxe et al., 2020) and TyDi QA (Clark et al., 2020) offer deep, domain-focused insights but cover a narrow slice of capabilities, which limits cross-task comparability and pan-European parallelism.

Second, machine-translated with human quality assurance (QA) suites broaden language reach while retaining human curation on a subset. Global-MMLU (Singh et al., 2025) provides MT variants (Google Translate) and human-edited subsets for selected languages, enabling reference-based checks.

Third, there are machine-translated with limited or no documented QA resources that emphasize scalability. EU20 (Thellmann et al., 2024) translates five complementary task families – knowledge-heavy QA (ARC; Clark et al., 2018), mathematical reasoning (GSM8K; Cobbe et al., 2021), commonsense (HellaSwag; Zellers et al., 2019), NLI/knowledge probing (MMLU; Hendrycks et al., 2020), and truthfulness (TruthfulQA; Lin et al., 2022) – into 20 European languages using DeepL, maximizing parallelism across tasks and languages. Okapi (Lai et al., 2023) provides ChatGPT-based translations of ARC, HellaSwag, MMLU, and TruthfulQA into 31 languages, spanning both EU and non-EU languages. In both cases, a systematic QA pass is not the primary focus of the initial releases.

Against this landscape, our contribution is an automated QA layer for EU20 that is cost-efficient, scalable, and complementary to expert review. We focus on EU20 because its five task families and 20 languages offer the parallelism required for cross-lingual evaluation at scale. Our automated QA increases the reliability of this testbed without claiming to replace human-curated gold data.

With respect to other automated translation quality assessment tasks, the WMT translation quality estimation and metrics shared tasks (Lavie et al., 2025; Zerva et al., 2024; Freitag et al., 2024; etc.) consider translations of "flat" text passages. In comparison, LLM evaluation examples have a structure which has to be preserved during translation. Furthermore, in our work, we also consider consistency across multiple (target) languages as a dataset-level quality criterion.

3. Structural Diagnostics and Dataset Maintenance for EU20

In this section we follow Thellmann et al. (2024), who introduced EU20 by translating five established English benchmarks into 20 European languages, and provide a structural assessment of the released corpora (Section 3.1), applying targeted updates and completions where needed (Section 3.2).

3.1. Structural Analysis

We verify the integrity of all translated samples along four criteria:

A) **Answer-index alignment (for multiple-choice tasks):** Language-independent indices of the correct choice(s) match the English original.

B) **Field completeness:** Essential fields are non-empty ("question"; "choices" for multiple choice; "answer" for generative tasks).

C) **Split/subset consistency:** For uniquely identifiable samples, split and subset mirror the English version.

[4] https://hf.co/eu20-cleaned/datasets

[5] https://github.com/eu20-cleaned/lang-integrity

[6] https://github.com/eu20-cleaned/translation-quality-analysis

D) **Cross-language coverage:** The sample exists across all 20 translations for the evaluated splits.

Criterion A) is satisfied for all datasets. Table 1 summarizes the remaining criteria: N_C is the number of samples violating criterion B), N_T is the number of samples fulfilling criterion C) across the 20 target languages, and N_L is the number of samples violating criterion D) after removing samples that violate the other criteria.

Regarding B), we observe missing content primarily in the HellaSwag validation split: among 10,042 English originals, 327 samples (3.26%) have at least one translation with empty answer options; 257 of these (78%) affect two or more target languages. Manual inspection suggests that DeepL can be confused by the context-continuation format used in LM-Eval-Harness when answer options are fragments rather than full sentences, which complicates finding a common prefix across languages.

For criterion C), we did not expect any mismatches, but found two problems that are most likely due to operating errors during the translation:

Firstly, the train splits of our ARC translations, from which the few-shot samples are drawn, consist of a mix of samples from the `easy` and `challenge` subsets. This applies to each of the translated languages, but not to the English version, where the few-shot samples are drawn from the same subset as the sample under test, so comparability of k-shot accuracies ($k \geq 1$) between English and non-English languages might be limited.

Secondly, in HellaSwag the per-language `train` split is not a subset of `train` but a 99-item subset of `validation`, which can leak answers into few-shot contexts. Because the 10 few-shot examples per query are sampled from this small set (≈ 100) that itself belongs to the validation set ($\approx 10{,}000$), the probability that the evaluated item appears in the context is $\sim 0.1\%$. The resulting measurement error is proportional to this chance and decreases as the true model accuracy increases. Even in the worst case (four answer options, true accuracy $\approx 25\%$), the overestimation is only about 0.08 percentage points (e.g., $25.00\% \rightarrow 25.08\%$). Given this upper bound, we consider the expected fraction of leaks (and thus their impact on reported accuracies) too small to warrant re-evaluating all models on this relatively large and resource-intensive benchmark.

Criterion D) holds for all evaluation splits except HellaSwag and for MMLU-dev. It is not met for the `train` splits of ARC, GSM8K, and HellaSwag because sub-splits were selected independently per target language. Among the HellaSwag validation splits, DE is missing 63 items, and ES, FR and IT 4 items each.

Dataset	Split	N_{en}	N_T	N_C	N_L
ARC	train	3,370	6,420*	8	6,420
ARC	val	869	17,380	17	0
ARC	test	3,548	70,960	109	0
GSM8K	train	7,473	2,288	0	2,288
GSM8K	test	1,319	26,380	0	0
He.Sw.	train	39,905	1,980**	10	1,980
He.Sw.	val	10,042	200,765	1,039	1,185
MMLU	test	14,042	280,840	678	0
MMLU	dev	285	5,700	3	0
Tr.QA	val	817	16,340	3	0

Table 1: N_{en}: #English samples. N_T: #Non-English samples. N_C: #Non-English samples with missing content. N_L: #Non-English samples with other Non-English version(s) missing. *3,060 samples from the respective other subset. **All samples present translated from validation (val) split.

3.2. Corrections and Completion

Based on the structural findings, we repair defective entries and complete missing ones via targeted, sample-level re-translation:

- **Scope selection.** We operate either on full split/subset combinations or on JSONL manifests that enumerate individual sample IDs (with associated split/subset and target language) for re-translation. This enables surgical fixes without reprocessing entire corpora.
- **Provenance guarantees.** For each sample, we ensure that the record contains sufficient ID fields so its lineage to the English original is unambiguous before and after modification.
- **Update policy.** Only fields flagged as defective (missing/empty or structurally inconsistent) are overwritten. Intact fields remain unchanged. All edits are logged in an extended diagnostics format alongside the EU20-conform outputs to support auditability and regression checks.

We reuse a minimal translation processor primarily for correction/augmentation rather than bulk creation. The design allows alternative engines, but we currently use DeepL with formatting equivalent to the original setup except for XML escaping.

For each dataset we (i) extract translatable fragments, (ii) ensure/normalize ID/key fields, and (iii) write translated fragments back. Fragments are serialized into a single XML string (`Frag_1<x>SEP</x> ... <x>SEP</x>Frag_n`) with XML-escaping. The DeepL API is configured to ignore `<x>`, and the response is de-serialized to a fragment list.

We support API-level batching and cache formatted inputs/outputs (key: formatted source;

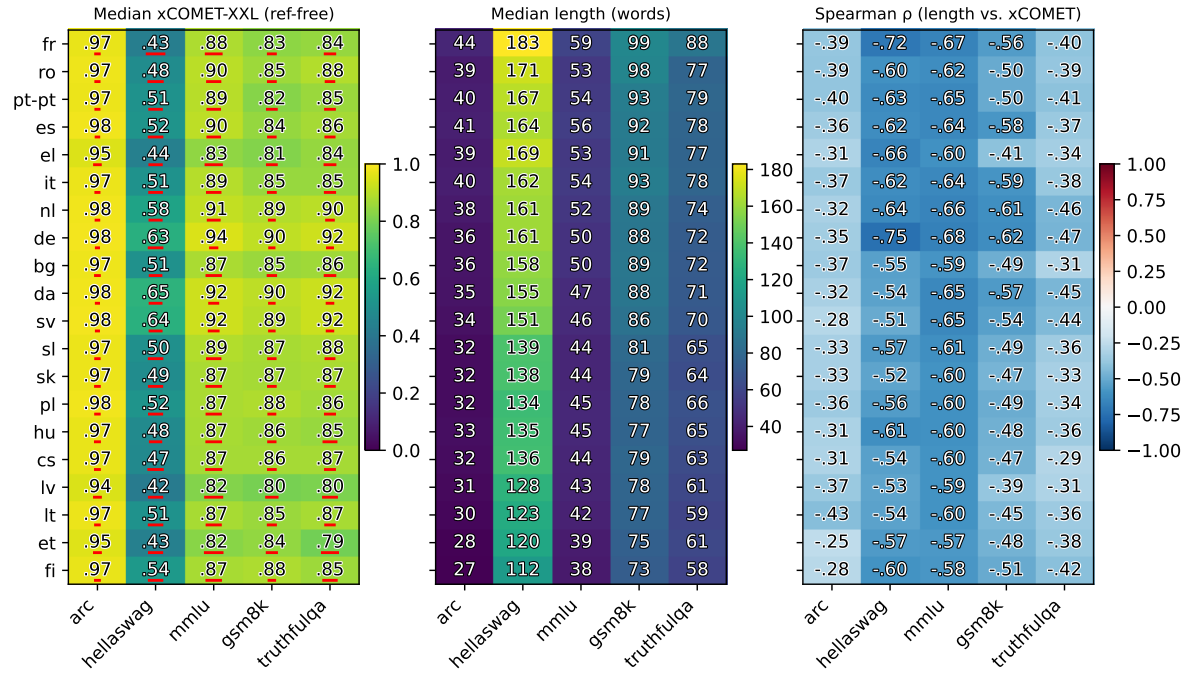

Figure 1: EU20 reference-free quality landscape. Left: median xCOMET-XXL per language×dataset on a unified $[0, 1]$ scale; short in-cell tick encodes IQR (Q_3-Q_1). Middle: median target-side sentence length (words). Right: Spearman correlation (ρ) between length and score (negative ρ indicates lower scores for longer outputs). Rows are aligned across panels and sorted by the language-wise median across datasets.

value: formatted target) to avoid duplicate calls during iterative fixes. Outputs are emitted as (a) EU20-conform JSON and (b) an extended diagnostics format for quality control and diffing.

For HellaSwag-style continuations, we optionally reformat options (e.g., prefix repeated context) to reduce empty-field failures, validate completeness post-translation, and route failed/ambiguous cases to manual inspection queues recorded in the diagnostics.

4. EU20 Translation Quality Profiling with a Neural QE Metric

In this section, we profile translation quality with xCOMET-XXL, a neural QE metric (reference-free and reference-based), in three steps across tasks and systems: First, we build a reference-free quality landscape over the EU20 benchmark suite, complemented by a length profile and a length–quality correlation analysis (Section 4.1). Second, we run paired, reference-free comparisons of EU20 versus Okapi across three suites and ten languages, reporting median gaps and win-rates with paired-bootstrap confidence intervals (Section 4.2). Third, on MMLU we compare EU20, Okapi, and the human-edited Global-MMLU via average-rank testing, and we also perform a reference-based comparison against the human-edited references to validate the trends (Section 4.3 and 4.4).

4.1. Quality Landscape

Method. In this comparison, we present a quality landscape of the EU20 translations across five widely used benchmarks: ARC, HellaSwag, MMLU, GSM8K, and TruthfulQA, using the reference-free xCOMET-XXL quality estimator. To better interpret the results we posit two hypotheses: (i) longer outputs tend to receive lower xCOMET-XXL scores; (ii) scores follow a pattern driven by dataset design, with standardized QA (e.g., ARC, MMLU) yielding higher medians than open-ended continuations (HellaSwag).

We test (i) by computing target-side word-count medians for each language-dataset combination, then estimating Spearman's ρ between length and score. Hypothesis (ii) is treated as a design-based rationale consistent with the observed pattern rather than a tested causal claim. The results are shown in Figure 1 (left heatmap) as a 20×5 matrix (languages $\times$ datasets) on a unified $[0, 1]$ scale. Each cell reports the median score, a short in-cell tick encodes the interquartile range (IQR, Q_3-Q_1), and rows are sorted by the language-wise median across datasets. The same figure also presents median sentence-length statistics (middle heatmap) and Spearman correlations be-

tween length and score (right heatmap).

Results and discussion. Figure 1 (left) shows the median xCOMET-XXL scores as a 20×5 matrix. ARC is highest overall (.,97-.,98), HellaSwag is lowest (.,42-.,65), and MMLU (.,83-.,94), GSM8K (.,80-.,90), and TruthfulQA (.,79-.,92) lie in between. This pattern is consistent with Hypothesis (ii): standardized QA prompts (ARC, MMLU) tend to yield more semantically aligned translations, whereas open-ended continuations (HellaSwag) induce greater lexical and structural variability. Across languages, DE/DA/SV lead (median across datasets ≈.,87-.,88), followed by NL (.,85) and a mid-cluster (PL/FI/SL/ES/RO, ≈.,82); EL/ET (≈.,77) and LV (.,75) trail. In addition to median levels, longer IQR ticks on HellaSwag and MMLU indicate greater within-dataset variability: HellaSwag combines lower medians with high dispersion (uneven outputs), while MMLU shows higher medians yet similarly wide spread (heterogeneous subjects, formats, and length effects).

Figure 1 (middle) reports median output lengths, supporting Hypothesis (i)'s premise about verbosity: HellaSwag is longest (medians ∼112-183 words), followed by GSM8K (73-99), TruthfulQA (58-88), and MMLU (38-59); ARC is shortest (27-44). This aligns with task design, where open-ended reasoning naturally produce longer outputs than fixed-format multiple-choice questions.

Figure 1 (right) shows Spearman correlations between length and xCOMET-XXL. Correlations are predominantly negative, indicating that longer translations tend to receive lower quality estimates. Effect sizes vary by dataset: ARC weak-moderate ($\rho \approx -0.25$ to -0.43), TruthfulQA moderate ($\rho \approx -0.29$ to -0.47), GSM8K and MMLU stronger ($\rho \approx -0.39$ to -0.68), and HellaSwag strongest (often < -0.60; e.g., de $\rho \approx -0.75$), with $p \ll .001$ in most cells. Overall, the dataset pattern remains the dominant signal, while length effects help explain within-cell dispersion and some cross-dataset differences.

4.2. EU20 vs. Okapi

Method. To directly compare the quality of our translations against Okapi across tasks, we compute reference-free xCOMET-XXL scores on the shared subset of items for three representative benchmarks ARC, HellaSwag, and MMLU, and ten overlapping languages (CORE-10: DA, DE, ES, FR, HU, IT, NL, RO, SK, SV). Working on the paired item overlap (identical (*language, subset, split, id*)) ensures a fair comparison, as both translation systems are evaluated on exactly the same segments. For each (language, dataset) pair, we compute the median score difference $\Delta = \text{median}(EU20) - \text{median}(Okapi)$, where positive Δ indicates higher predicted adequacy/fluency for EU20.

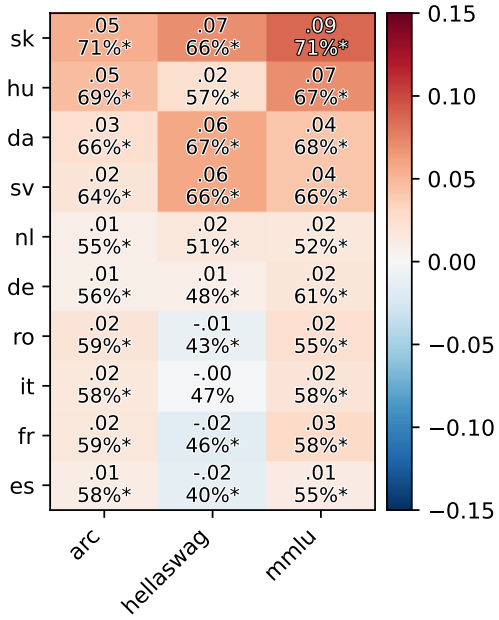

Figure 2: EU20 vs. OKAPI xCOMET-XXL reference-free quality comparison per language×dataset. Cells report the median difference $\Delta = \text{median}(EU20) - \text{median}(Okapi)$ on the paired overlap and the win-rate (% items where EU20 > Okapi). Positive Δ favors EU20.

To visualize these results, we use a heatmap with a diverging, zero-centered color scale fixed to a symmetric range, enabling comparison across datasets and languages. Each cell also reports the win-rate $P(EU20 > Okapi)$, i.e., the proportion of samples for which EU20 receives a higher xCOMET-XXL score than Okapi. This provides a complementary, distribution-sensitive indicator of relative system quality beyond the median.

We assess the statistical reliability of Δ using a paired bootstrap, a standard distribution-free method in MT/NLP for system-level score comparisons (Koehn, 2004; Dror et al., 2018; Efron and Tibshirani, 1994). We resample paired segment scores $B \approx 5000$ times, recompute the median difference for each replicate, and form a $(1-\alpha)$ confidence interval from the empirical quantiles of the bootstrap distribution. Figure 2 summarizes the paired, reference-free comparison of EU20 vs. Okapi on ARC/HellaSwag/MMLU across the CORE-10 languages. Given the large per-cell sample sizes (e.g., ∼1.4k for ARC, ∼7.3k for HellaSwag, ∼11.4k for MMLU), the intervals are typically narrow and differences often reach statistical significance. A star marks cells where the 95% CI does not include zero, indicating a statistically significant difference.

Results and discussion. On ARC and MMLU, EU20 is ahead in every language: the median score advantage is typically $\Delta \approx .009$-$.054$ on ARC and $.012$-$.086$ on MMLU. These gains are small in absolute terms but statistically reliable: bootstrap confidence intervals for (Δ) do not include 0, and win-rates are mostly $[0.55, 0.71]$ with intervals above 0.5 (i.e., EU20 wins on a clear majority of segments). Because scores are bounded in $[0, 1]$, differences between strong systems are compressed, so improvements of a few hundredths (together with win-rates > 0.5 and CIs excluding 0) are therefore meaningful when consistent.

HellaSwag is mixed: EU20 leads in SK/SV/DA/HU/NL/DE, while Okapi (Google Translate) edges out in ES/FR/RO; IT is at parity (small, non-significant Δ). One plausible interpretation is a better stylistic fit of Okapi on colloquial continuations in some Romance languages. We treat this as an observation consistent with the data rather than a causal claim.

Two signals support the overall conclusion that in the majority of language-task combinations, the EU20 translations have higher xCOMET-XXL scores on average than the Okapi translations: (i) win-rates (> 0.5) show that EU20 wins on a majority of segments, and (ii) large per-cell sample sizes ($\sim 1.4k$ on ARC, $\sim 7.3k$ on HellaSwag, $\sim 11.4k$ on MMLU) yield narrow bootstrap intervals, so many gaps are statistically significant.

Finally, this pattern is consistent with the reference-based MMLU analysis (Section 4.4): the ref-free advantage of EU20 over Okapi persists when evaluated against human-edited references and is significant in 4/5 languages. This triangulation suggests the effect is systematic rather than specific to a single evaluation mode.

4.3. MMLU Ranks: EU20, Okapi, Global

Method. We compare three translation sources EU20 (DeepL), Okapi (ChatGPT), and Global-MMLU (Google Translate, human-edited) on MMLU in five languages {DE, ES, FR, IT, RO}, using the triple item overlap (segments with identical language, subset, split, and id). On this shared set of items, we compute per-system median xCOMET-XXL scores, then convert the three medians to ranks (1 = highest median; ties share average rank).

To determine whether observed rank gaps are more than descriptive, we use the standard ML/NLP workflow – Friedman's omnibus test (blocks = languages) followed by the Nemenyi all-pairs test on mean ranks, to determine which systems differ across datasets (Demšar, 2006; García and Herrera, 2008). The resulting critical difference (CD) is the minimum gap between average ranks that must be exceeded for a pair to be considered significantly different at $\alpha = 0.05$ (Demšar, 2006). Graphically, the CD is shown as Nemenyi intervals (avg$\pm$CD/2) where overlap indicates "not significantly different" and non-overlap indicates a significant difference. Figure 3 visualizes average ranks with Nemenyi intervals and bridges indicating pairs that are not significantly different.

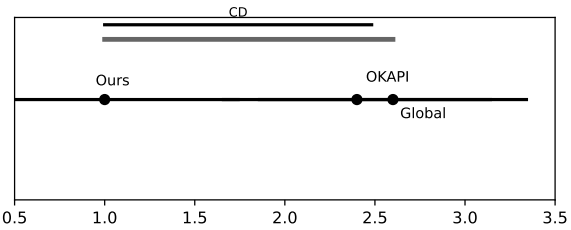

Figure 3: Critical-difference (CD) diagram on MMLU (ref-free). Points are systems' average ranks across five languages (lower is better). Thin bars show Nemenyi intervals (avg$\pm$CD/2; α=0.05, k=3, N=5). A grey bridge links systems that are not significantly different (no bridge = significant).

lang	m_{EU20}	m_{Okapi}	m_{Global}	items	r_{EU20}	r_{Okapi}	r_{Global}
de	.96	.94	.95	2378	1	3	2
es	.93	.92	.91	2378	1	2	3
fr	.91	.89	.88	2378	1	2	3
it	.92	.90	.91	2378	1	3	2
ro	.93	.91	.88	2378	1	2	3

Table 2: Per-language medians and ranks on the triple overlap (MMLU, ref-free). For each language, we report median xCOMET-XXL for EU20, Okapi, and Global-MMLU, the number of common items, and per-language ranks (1 = best).

Results and discussion. On the triple overlap (2,378 items per language), our translation attains rank 1 in all five languages (average rank = 1.00), while Okapi and Global-MMLU occupy ranks 2-3 (see Table 2 for medians and ranks). With k=3 systems and N=5 languages, the Nemenyi CD is $\approx$ 1.48. The EU20-Global average-rank gap is 1.60 > CD $\Rightarrow$ significant, whereas EU20-OKAPI (1.40 $\leq$ CD) and OKAPI-Global (0.20 $\leq$ CD) are not significant (see Figure 3). The per-language medians in Table 2 show the same ordering. We hypothesize that minor differences in paraphrasing/verbosity and morphology handling on EN$\rightarrow${de, es, fr, it, ro} make EU20/DeepL align slightly better with COMET's adequacy/fluency signals. This finding is consistent with the reference-based results (see Section 4.4).

lang	$m^{\text{ref}}_{\text{EU20}}$	$m^{\text{ref}}_{\text{Okapi}}$	Δ_{ref}	CI[low,high]	items
it	.92	.89	.029	[.0195, .0398]	2342
de	.96	.93	.026	[.0218, .0320]	2342
fr	.90	.88	.025	[.0150, .0344]	2342
ro	.92	.91	.015	[.0070, .0232]	2342
es	.91	.91	-.004	[-.0237, .0158]	516

Table 3: Per-language reference-based medians and deltas on MMLU. For each language, we report median($EU20_{\text{ref}}$), median($Okapi_{\text{ref}}$), Δ_{ref}, 95% CI, and n of common items.

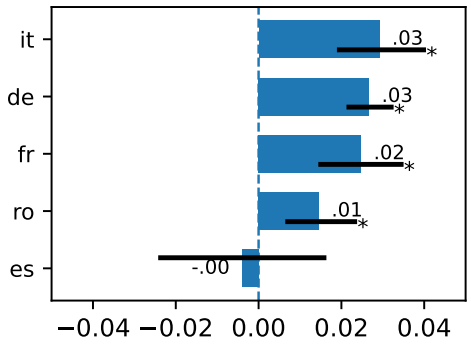

Figure 4: Δ_{ref} (EU20−Okapi) of reference-based xCOMET-XXL on MMLU (reference = Global-MMLU). Bars show Δ_{ref} per language with 95% paired bootstrap CIs. Zero line indicates parity. Sorted by Δ_{ref}. Common items only.

4.4. Ref-Based MMLU: EU20 vs. Okapi

Method. We compare two MMLU translation variants in a reference-based setting: EU20 (DeepL-derived) vs. Okapi (ChatGPT-derived), using Global-MMLU as the human-edited reference. To ensure strict comparability, we restrict both systems to the paired item overlap, i.e., the common set of segments with identical language, subset, split and id, present in both sources, for the five languages with vetted references {DE, ES, FR, IT, RO}. On this shared set, we summarize the effect per language as $\Delta_{\text{ref}} = \text{median}(EU20_{\text{ref}}) - \text{median}(Okapi_{\text{ref}})$, where larger values favor EU20. Uncertainty is quantified with a 95% paired bootstrap CI on Δ_{ref} (resampling the same indices in both systems; $B \approx 5000$) to avoid distributional assumptions (Efron and Tibshirani, 1994). Figure 4 displays horizontal bars per language with CIs and a vertical zero line indicating parity ($\Delta_{\text{ref}}=0$), using fixed axes for visual comparability and sorting languages by Δ_{ref}. The corresponding per-language medians, deltas, CIs, and item counts are reported in Table 3.

Results and discussion. Across the five languages, EU20 is significantly better than Okapi in 4/5 cases (see Figure 4 and Table 3): IT ($\Delta_{\text{ref}}=.029$, 95% CI $[.020 - .040]$), DE ($.027[.022 - .032]$), FR ($.025[.015 - .034]$), RO ($.015[.007 - .023]$) ES shows no significant difference ($\Delta_{\text{ref}}=-.004$, CI $[-.024$, $.016]$). Effect sizes are small-to-moderate (hundredths) but consistent over large item counts (n).

The wider CI in ES reflects the much smaller paired overlap ($n=516$ vs. 2,342 elsewhere). These findings complement the ref-free comparison in Section 4.3: although EU20 differs significantly from Global-MMLU in the reference-free setting (reflecting stylistic or phrasing divergences that xCOMET-XXL judges more favorably), it still aligns more closely with the human-edited Global-MMLU translations than Okapi does when those translations are used as explicit references. In other words, EU20 is not identical to Global-MMLU, but it captures key aspects of phrasing and morphology that bring it significantly closer to the reference than Okapi across four of five languages.

5. Translation Error Landscape for EU20 from Span-Level Judgments

As a complement to the sentence-level quality profiling with xCOMET-XXL – a neural, dual-mode QE metric (reference-free/-based; Section 4) – we conduct a span-level error profiling of EU20 translations. We adopt an LLM-as-a-judge TQE setup based on GEMBA-ESA (Kocmi and Federmann, 2023a; Kocmi et al., 2024) and the MQM taxonomy (Lommel et al., 2013) to annotate translation errors with category and severity. The two perspectives are complementary: xCOMET-XXL provides scalar sentence-level quality signals, whereas GEMBA-ESA exposes an interpretable error structure (what went wrong, and how severe) at span level.

Method. We adapt GEMBA-ESA with a structured JSON output and multilingual few-shot prompts that instruct the model to detect span-level errors and label them with MQM categories and a severity (major if meaning is impaired, minor if generally understandable). We use three independent LLM annotators – GPT-4o-mini[7], Llama-4 Scout[8], and Mistral-Large-Instruct-2411[9] – all prompted with our adapted GEMBA-ESA prompt. For comparability across languages and tasks, we aggregate the MQM span labels produced by the LLM annotators into the high-level categories: `A+M` (accuracy+mistranslation), `F` (fluency/style), and `O` (other).

For each item (one translated benchmark entry) and category (`A+M`, `F`, `O`), we set $\text{maj}(i, T, S) = 1$ iff

[7] platform.openai.com/docs/models/gpt-4o-mini
[8] HF: meta-llama/Llama-4-Scout-17B-16E-Instruct
[9] HF: mistralai/Mistral-Large-Instruct-2411

Figure 5: EU20 error overview per language × dataset. Each cell shows four horizontal bars: A+M, F, O, and Clean. Error rates per 1,000 items.

at least two annotators flagged any span of type T with severity S ($S \in \{\text{minor}, \text{major}\}$; if both severities would apply, we keep major to avoid double counting). For Figure 5, we collapse severities and mark a category as present if either minor or major reached majority. CLEAN holds if a majority judged "no error" (items without error spans count as no error for that annotator). We report rates per 1,000 items for each category and CLEAN for every language and dataset.

Although Figure 5 collapses severities for space, Table 4 reports the share of major/minor within each category per dataset. We pool counts across all 20 languages: for each dataset and category, the major percentage is the share of agreed major errors out of all agreed errors in that category (major+minor), and minor percentage is the complementary share.

Results and discussion. A consistent pattern emerges across languages: (i) HellaSwag shows the highest A+M rates (e.g., LV: 744/1k, RO: 691/1k, BG: 693/1k), with F smaller (typically ∼120-290/1k) and O low (single digits to ∼20/1k). Clean is correspondingly low (often ∼120-285/1k). (ii) GSM8K and MMLU sit mid-range in A+M (∼170-347/1k and ∼150-322/1k, respectively) with CLEAN mostly ∼500-750/1k. (iii) ARC is comparatively clean (A+M ∼90-250/1k; CLEAN ∼700-850/1k). (iv) TruthfulQA shows moderate A+M (146-303/1k) and high CLEAN (535 - 709/1k). CLEAN is generally higher in Germanic/Romance languages (e.g., DA/NL/SV/DE/FR) than in Baltic/Balkan languages. However, these cross-language gaps are modest compared to the dataset effect – differences between datasets (e.g. HellaSwag vs. ARC) are substantially larger than differences between languages within a dataset. Overall, A+M dominates the error mass, F is secondary, and O marginal.

	AM	F	O	CLEAN
	maj/min	maj/min	maj/min	-/-
arc	73.9/26.1	3.7/96.3	0.9/99.1	72.6
gsm8k	83.4/16.6	6.1/93.9	2.2/97.8	56.7
hellaswag	87.6/12.4	2.0/98.0	4.1/95.9	19.3
mmlu	81.9/18.1	5.1/94.9	9.4/90.6	66.3
truthfulqa	75.4/24.6	7.3/92.7	1.4/98.6	65.2

Table 4: Share of major/minor (maj/min) severities (in %) among majority-agreed errors per category (A+M F, O, CLEAN)

As shown in Table 4, A+M errors are predominantly major across datasets – highest on HellaSwag (87.6% maj), followed by GSM8K (83.4%) and MMLU (81.9%). ARC and TruthfulQA are lower yet still predominantly major (73.9% and 75.4%). By contrast, F and O are mostly minor in all datasets (e.g., HellaSwag F 2.0/98.0, O 4.1/95.9). CLEAN is lowest on HellaSwag (19.3%) and highest on ARC (72.6%)

The translation error profile/landscape from span-level judgements aligns closely with the sentence-level xCOMET-XXL quality landscape (Section 4): datasets with lower xCOMET-XXL medians (notably HellaSwag) are precisely those with high A+M rates here. This supports our earlier hypotheses: task design drives difficulty (open-ended continuations induce adequacy/mistranslation pressure), and length effects exacerbate it, producing lower sentence-level quality scores and more adequacy-span agreements. Conversely, ARC (short, standardized QA) shows high CLEAN and low A+M, matching its high xCOMET-XXL medians. Taken together, xCOMET-XXL provides scalar, task-sensitive sentence-level quality signals, while GEMBA-ESA reveals where the quality is lost: predominantly accuracy/mistranslation (often major), rather than fluency or style issues. This suggests that targeted clean-up of high-A+M clusters (especially on HellaSwag) would yield the largest quality gains for EU20.

6. Conclusion

We examined the reliability of machine-translated benchmark evaluation at pan-European scale by diagnosing translation quality in EU20. Our automated QA stack combines (i) a corpus audit and targeted repairs, (ii) COMET-based sentence-level profiling (reference-free and reference-based) with comparisons across translation services, and (iii) span-level LLM-as-a-judge (MQM). Results converge: HellaSwag has the highest accuracy/mistranslation error mass and lowest COMET; ARC is cleanest; longer outputs correlate with lower quality. On MMLU, reference-based COMET against human-edited Global-MMLU samples supports ref-free rankings. Our take-away is pragmatic: automated QA provides quantifiable, diagnostic evidence that helps decide where scarce human QA is most valuable, strengthening cross-lingual comparisons at scale.

7. Limitations

Our study is centered on EU20 (five tasks and 20 European languages) so conclusions may not transfer to other domains or non-European languages. We rely on automated quality assurance as a proxy for human review: COMET and LLM-as-a-judge can introduce metric and judge biases, and span-level MQM may suffer from boundary ambiguity, over-fragmentation of single errors, and unstable severity calibration under prompt/model changes. Human verification is limited to a reference-based check on a subset of MMLU (human-edited Global-MMLU). We do not provide large-scale human adjudication across all tasks and languages (out of scope).

Stability remains a concern: results can vary with prompts, seeds, and evolving MT/LLM/COMET versions despite reporting confidence intervals. Comparisons across translation services (DeepL, ChatGPT, Google) are observational and tied to specific pipelines and time windows. Small deltas should be interpreted cautiously rather than causally. Finally, structural audits ensure format integrity (fields, splits, coverage), and TQE-based evaluations cannot guarantee full semantic faithfulness or cultural appropriateness of translations.

8. Ethical & Broader Impact

The ability to evaluate large language models (LLMs) across a wide range of European languages, particularly underrepresented ones, is a critical step toward enhancing inclusivity and accessibility in natural language processing (NLP). By ensuring that LLMs can perform well in languages beyond English or other high-resource languages, we contribute to a more equitable digital landscape where speakers of less widely spoken languages have equal access to advanced language technologies. However, this inclusivity brings unique challenges, particularly in achieving benchmarks that are comparable across diverse linguistic and cultural contexts.

Acknowledgments

This work was funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) through the project OpenGPT-X (project no. 68GX21007D). The authors acknowledge the financial support by the Federal Ministry of Research, Technology and Space of Germany (BMFTR) and by Sächsische Staatsministerium für Wissenschaft, Kultur und Tourismus in the programme Center of Excellence for AI-research „Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig“, project identification number: ScaDS.AI, and by the BMFTR under grant number 01IS24077A. The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for funding this project by providing computing time on the GCS Supercomputer JUWELS at Jülich Supercomputing Centre (JSC) as well as the Center for Information Services and High Performance Computing [Zentrum für Informationsdienste und Hochleistungsrechnen (ZIH)] at TU Dresden for providing its facilities for automatic evaluation computations.

9. Bibliographical References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, 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.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

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. *CoRR*, abs/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. *arXiv 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.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Bradley Efron and R.J. Tibshirani. 1994. *An Introduction to the Bootstrap*, 0 edition. Chapman and Hall/CRC.

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.

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.

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.

Salvador García and Francisco Herrera. 2008. An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. *Journal of Machine Learning Research*, 9:2677–2694.

Nuno M. Guerreiro, Ricardo Rei, Daan Van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xCOMET: Transparent Machine Translation Evaluation through Fine-grained Error Detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Krzysztof Jassem, Michał Ciesiółka, Filip Graliński, Piotr Jabłoński, Jakub Pokrywka, Marek Kubis, Monika Jabłońska, and Ryszard Staruch. 2025. Llmzszł: a comprehensive llm benchmark for polish.

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.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. 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.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483, Suzhou, China. Association for Computational Linguistics.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. 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 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.

Bernardo Magnini, Roberto Zanoli, Michele Resta, Martin Cimmino, Paolo Albano, Marco Madeddu, and Viviana Patti. 2025. Evalita-llm: Benchmarking large language models on italian.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. *Advances in Neural Information Processing Systems*, 35:462–477.

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.

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, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Simon Ott, Adriano Barbosa-Silva, Kathrin Blagec, Jan Brauner, and Matthias Samwald. 2022. Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *Nature Communications*, 13(1):6793.

Xueqing Peng, Triantafillos Papadopoulos, Efstathia Soufleri, Polydoros Giannouris, Ruoyu Xiang, Yan Wang, Lingfei Qian, Jimin Huang, Qianqian Xie, and Sophia Ananiadou. 2025. Plutus: Benchmarking large language models in low-resource greek finance.

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 del Pozo, Javier Conde, Pedro Reviriego, Marina Mayor-Rocher,

and María Grandury. 2024. Spanish and llm benchmarks: is mmlu lost in translation?

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

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.

David Samuel, Andrey Kutuzov, Samia Touileb, Erik Velldal, Lilja Øvrelid, Egil Rønningstad, Elina Sigdel, and Anna Palatkina. 2023. NorBench – a benchmark for Norwegian language models. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 618–633, Tórshavn, Faroe Islands. University of Tartu Library.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cesar Ferri, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Christopher Waites, Christian Voigt, Christopher D Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, C. Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-Lopez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Francis Anthony Shevlin, Hinrich Schuetze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fer-

nández Fisac, James B Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros-Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje Ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramirez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael Andrew Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Andrew Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Russ Salakhutdinov, Ryan Andrew Chi, Seungjae Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel Stern Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Shammie Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven Piantadosi, Stuart Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, vinay uday prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

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.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference*

on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *ACL (1)*, pages 4791–4800. 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.