Gold vs. Translation: Are We Measuring Language, Knowledge, or Locality in Translated Benchmarks?

Klaudia-Doris Thellmann[1], Bernhard Stadler[1]
Dan Saattrup Nielsen[2], Michael Färber[1], Jens Lehmann[1]
[1]firstname.lastname@tu-dresden.de [2]dan.smart@alexandra.dk

[1]TUD Dresden University of Technology and ScaDS.AI,
[2]Alexandra Institute

Abstract

Translated versions of English-origin benchmarks such as MMLU are widely used to evaluate large language models in non-English languages, yet it remains unclear whether they measure the same constructs as natively created benchmarks. We frame this as a *proxy validity* question and compare DeepL-translated MMLU proxies against six native benchmarks in seven pairings across five Nordic and Baltic languages, evaluating 29 models on each pairing. Beyond ranking agreement, we test *incremental* proxy validity: does the translated score add predictive value for native performance after controlling for model size and language proficiency? Proxy validity varies substantially across benchmark pairings. Applying the same robustness criteria to all seven pairings, we find four robust effects, two marginal effects, and one null result. The strongest evidence occurs for curriculum-overlap benchmarks such as Skolprov ($\Delta R^2 = .128$, 95% CI $[.038, .253]$), whereas translated MMLU provides no significant incremental value for Sweden-specific factual knowledge in SwedishFacts ($p = .242$). Because SwedishFacts is also the only open-ended benchmark, we cannot fully separate the effects of local content and response format. Overall, the results indicate a clear pattern: translated benchmarks provide a better proxy when their source content overlaps with the native evaluation target, while locally anchored knowledge is less reliably captured.

1 Introduction

Multilingual evaluation of large language models (LLMs) frequently relies on translated versions of English-origin benchmarks. Datasets such as MMLU (Hendrycks et al., 2021) are routinely machine-translated into dozens of languages and used to assess model performance as if they were native evaluation instruments in the target language. Yet a growing body of evidence suggests that this practice is problematic. Translated benchmarks carry cultural and linguistic biases from the source language (Singh et al., 2025), exhibit translation artifacts that can inflate scores and alter system rankings (Zhang and Toral, 2019; Wang et al., 2023), and correlate less with human judgments than natively developed benchmarks (Wu et al., 2025).

These findings raise a fundamental question: when we evaluate a model on MMLU translated into Estonian or Swedish, what do we actually measure – the model's language proficiency, its general knowledge, or its familiarity with locally relevant content?

Most prior work addresses the *channel question*: is machine translation a good enough substitute for human translation (Rajaee et al., 2025; Bijl de Vroe et al., 2025; Agrawal et al., 2024)? A few efforts go further: Wu et al. (2025) find that localized benchmarks correlate more with human preferences than translated ones, and Hupkes and Bogoychev (2025) quantify a substantial Locality Effect. However, these studies remain descriptive. They document that gaps exist but do not control for the confound that strong models score well on any benchmark, nor formalize when a translated benchmark is a *valid proxy* for a native one.

We address this gap by framing the problem as a **proxy validity** question: does a translated benchmark lead to the same evaluation conclusions as a native one, once we account for general model strength? We compare translated MMLU proxies against six native benchmark sources, yielding seven native–proxy pairings across five Nordic and Baltic languages and evaluating 29 LLMs. Beyond ranking agreement, we test *incremental* proxy validity—whether the proxy score adds predictive value for native performance after controlling for model size and language proficiency—and examine content differences through culturally sensitive (CS) vs. agnostic (CA) stratification and required-knowledge annotations.

Our analysis reveals that proxy validity varies:

ranking agreement ranges from $\tau_b = 0.89$ for curriculum-overlap benchmarks to 0.51 for Sweden-specific factual knowledge. Under the same evidence definition for all seven pairings, four show robust incremental proxy validity, two are marginal, and one is null. The strongest evidence occurs for Skolprov ($\Delta R^2 = .128$, 95% CI $[.038, .253]$), whereas translated MMLU provides no significant incremental value for SwedishFacts ($p = .242$), potentially due to its Sweden-centric content and/or its open-ended generation format. Adding English MMLU as a control supports the same overall pattern, indicating that general MMLU-content familiarity alone does not explain the proxy-validity differences. Taken together, these results indicate that the usefulness of a translated benchmark depends on how closely its content matches the knowledge targeted by the native benchmark: broadly shared curriculum content transfers well, whereas locally specific knowledge is less reliably captured.

Our contributions are:

(1) an *incremental proxy validity framework* that goes beyond rank correlation by controlling for confounds and testing whether translated benchmarks carry predictive signal beyond background factors;

(2) a multilingual cultural-sensitivity and required-knowledge annotation pipeline validated against three-annotator majority references for 500 items across five languages;

(3) empirical evidence across seven native-vs-proxy pairings showing that proxy validity varies with content and locality overlap, offering practical guidance on when translated benchmarks approximate native evaluation signal and when native benchmarks remain necessary.

We release the annotations, annotation pipeline, and analysis code to support reproducibility and extension to other languages and translation systems.[1]

2 Related Work

Translation methods. Studies comparing MT and HT benchmarks find high rank correlations on average, but also reveal systematic overestimation for commonly benchmarked low-resource languages (Rajaee et al., 2025), disproportionate effects on low-resource languages (Agrawal et al., 2024), sensitivity to model type (Bijl de Vroe et al., 2025), and substantial cross-lingual gaps (Ahuja et al., 2023; Lai et al., 2023). Post-editing (Huang et al., 2025) and adapted translate-test (Artetxe et al., 2023) mitigate but do not eliminate these effects. All of these compare translation *methods* on English-origin content; we ask whether *any* translated version measures the same construct as a native L2 benchmark.

Native versus translated benchmarks. Wu et al. (2025) find that localized benchmarks correlate more with human preferences than translated ones; Singh et al. (2025) show that rankings shift on culturally sensitive items in GlobalMMLU; Keleg and Magdy (2023) show culturally patterned factual knowledge in pretrained LMs; and Hupkes and Bogoychev (2025) quantify a substantial Locality Effect across 31 languages. However, none of these formalize proxy validity with statistical controls. We build on these findings by testing whether the divergence persists after controlling for confounds.

Translationese. Translation artifacts inflate scores, alter rankings (Zhang and Toral, 2019; Wang et al., 2023), and create spurious cross-lingual performance gains (Park et al., 2024; Plaza Ortiz et al., 2025); we ask whether the net effect undermines proxy validity.

Validity frameworks. Recent work has called for more rigorous evaluation methodology: Salaudeen et al. (2025) propose a validity-centered framework distinguishing different types of validity threats in AI evaluation, Bean et al. (2025) find widespread construct validity threats across 445 LLM benchmarks, and Alaa et al. (2025) show that high benchmark scores do not predict real-world performance. More broadly, Wallach et al. (2025) and Weidinger et al. (2025) argue that evaluating GenAI requires social science measurement rigor. On the empirical side, LLM capabilities are non-monolithic with a strong general ability factor (Burnell et al., 2023; Ilić and Gignac, 2024; Kim et al., 2025) – motivating our use of controls to separate general strength from benchmark-specific signal. Our notion of proxy validity is distinct from construct validity as discussed in these works: we ask whether one *observable* benchmark (translated) can substitute for another *observable* benchmark (native), rather than whether a benchmark measures a latent construct (see Appendix A for extended discussion).

[1] Annotations and merged analysis data: https://github.com/gold-vs-translation/cs-annotation-data; analysis scripts and annotation pipeline: https://github.com/gold-vs-translation/annotation-and-analysis-scripts.

3 Datasets and Comparison Setup

In this section we introduce the native benchmark collection used in our study and present the pairings that match each native benchmark against a translated MMLU proxy in the same language.

Native (Gold)	n	MMLU-Proxy	n
Exam-ET	1,614	ET	2,800
INCLUDE-ET	224	ET	2,800
Skolprov	545	SV	2,800
SwedishFacts	1,292	SV	2,800
DanishCitizenTest	668	DA	2,800
NRK-Quiz	3,600	NB	2,701
INCLUDE-LT	534	LT	2,800

Table 1: Dataset overview and native-vs-proxy comparison pairings. Each native benchmark is paired with the DeepL-translated MMLU in the same language (G-MMLUX subset). Proxies sourced from EU20-MMLU for ET, LT, SV, DA and from m_mmlu for NB.

Our study compares *translated* MMLU proxies against *natively created* benchmarks across five target languages: Estonian (ET), Swedish (SV), Danish (DA), Norwegian Bokmål (NB), and Lithuanian (LT). Table 1 summarizes all datasets used in our study, organized into *translated global benchmarks* derived from MMLU and *native local benchmarks* created independently in the target language.

3.1 Translated Benchmarks (Proxies)

Our MMLU proxies are drawn from EU20-MMLU (Thellmann et al., 2024) for ET, LT, SV, and DA, and from m_mmlu (Alexandra Institute, 2024b) for NB—both DeepL translations of the full MMLU (Hendrycks et al., 2021). We additionally construct **G-MMLUX**, a unified subset mapping the culturally annotated items from GlobalMMLU (Singh et al., 2025) to their translations in each target language, including the English originals. After agreement filtering to exclude items with low annotator consensus, G-MMLUX contains 16,701 items across six languages (Appendix B).

3.2 Native Benchmarks

We include six natively created benchmark sources across five languages, yielding seven native sets through the independent Estonian and Lithuanian INCLUDE subsets:

Exam-ET (Lillepalu and Alumäe, 2026) contains 1,614 MC questions from official Estonian secondary- and high-school exams across seven subjects, one of which we subdivide.

Skolprov (Ekgren, 2025) comprises 545 MC items from the Swedish university entrance exam and professional licensing exams for physicians, dentists, audiologists, and pharmacists.

SwedishFacts (Kunz, 2026) is a generative factual QA benchmark of 1,292 items targeting Sweden-specific personalities and events, and is the only non-MC dataset in our study.

DanishCitizenTest (Alexandra Institute, 2024a; Saattrup Nielsen, 2023) consists of 668 MC items from the official Danish citizenship and residence permit exams.

NRK-Quiz (Mikhailov et al., 2025) provides 3,600 MC quiz questions from NRK, Norway's public broadcaster, covering Norwegian culture, geography, and current affairs.

INCLUDE (Romanou et al., 2025) contributes the Estonian (224 items) and Lithuanian (534 items) subsets of natively sourced exam questions.

All native sets were created in the target language and are free of translation artifacts by construction, but they differ markedly in locality and content specificity. The proportion of culturally sensitive (CS) items ranges from 14.3% (Skolprov) to 97.8% (SwedishFacts), compared to 27.8% in GlobalMMLU (Table 12).

3.3 Subject Mapping

To assess whether the native and translated benchmarks are thematically comparable, we mapped each native subject to the most appropriate MMLU category (STEM, HUMANITIES, SOCIAL SCIENCES, BUSINESS, MEDICAL, OTHER).

A direct mapping between native subjects and individual MMLU subjects is unreliable because the underlying taxonomies differ in granularity and scope and often lack one-to-one equivalents. We therefore conduct cross-benchmark analyses at the broader category level, which provides a more robust basis for comparison. Across native-exam and GlobalMMLU-Lite cross-validation items, GPT-5.2 category labels agree with reference categories on 85.1% of 3,044 items. Disagreements mainly reflect MMLU taxonomy ambiguities rather than clear mapping errors. The full mapping and examples of ambiguous cases are provided in Appendix B.2.

The category mapping also reveals how much thematic ground the native benchmarks share with

MMLU. Table 1 lists our seven comparison pairings. Exam-ET and Skolprov are curriculum-based exam benchmarks whose items can be mapped to MMLU categories at *approximate* confidence for 70% (1,130/1,614) and 85% (465/545) of items, respectively—primarily in STEM, HUMANITIES, and MEDICAL. DanishCitizenTest maps entirely to SOCIAL SCIENCES at approximate confidence, but its 95.4% culturally sensitive content has no meaningful MMLU counterpart. In our analysis, INCLUDE (ET, LT), NRK-Quiz, and SwedishFacts are classified entirely via item-level annotation at the category level, instead of using native subject labels (Appendix B.3).

4 Annotation and Meta-Evaluation

This section describes how we obtain culturally sensitive (CS) versus culturally agnostic (CA) labels for benchmark items and how we validate the automatic annotation pipeline. The distinction is central to our proxy validity analysis (§5), as it allows us to test whether translated benchmarks fail specifically on items requiring local or culture-specific knowledge. We pursue a two-pronged approach: **(1)** a native-language human reference with three independent annotators per item for 500 items across five languages, and **(2)** a GPT-5.2 annotation pipeline calibrated on GlobalMMLU-Lite.

4.1 Human Annotation

We adapted the GlobalMMLU cultural-sensitivity schema (Singh et al., 2025) for our native benchmark items. For each item, annotators record a primary `required_knowledge` type (*none*, *cultural*, *regional*, or *dialect*), a binary `time_sensitive` flag, an explicit `cultural_sensitivity_label` (CS/CA), a `subject_category`, and optional metadata including trigger spans and geographic/-cultural identifiers.

The CS/CA label captures whether annotators judge an item to require culture-specific, regional, or dialect knowledge, but it is recorded as a separate binary judgment rather than mechanically derived from `required_knowledge`. For 10 of the 500 items (2%), the explicit CS/CA majority label differs from the label one would obtain by mechanically collapsing `required_knowledge` into `none` versus non-`none`. For each target language, we sampled 100 native-benchmark items and obtained three independent professional annotations per item, yielding 1,500 raw ratings over 500 items. Annotators were recruited based on native-speaker status and professional language expertise. For the reliability and meta-evaluation analyses, we use the original raw/pre-QA human annotations as the reference labels, before any GPT-assisted consistency review. The full schema is given in Appendix C.1. Sampling, recruitment, and raw/pre-QA reconstruction details are documented in Appendix C.3.

Table 2 reports the resulting three-annotator majority reference. Explicit CS/CA has a unique majority for all 500 items. For four-way `required_knowledge`, 477 items have a unique majority. The remaining 23 items (4.6%) are unresolved 1–1–1 ties, most frequently in Lithuanian (13 of 100).

	Required knowledge					Explicit CS/CA	
L	`none`	`reg.`	`cult.`	`dial.`	**ties**	**CS**	**CA**
DA	4	85	6	0	5	97	3
ET	77	19	3	1	0	22	78
LT	34	41	12	0	13	66	34
NB	23	30	41	2	4	75	25
SV	45	2	52	0	1	59	41
All	**183**	**177**	**114**	**3**	**23**	**319**	**181**

Table 2: Three-annotator majority reference for 100 native items per language (500 total). Required knowledge and explicit CS/CA (culturally sensitive / culturally agnostic) are independently annotated. Required-knowledge counts exclude 23 unresolved 1–1–1 ties. Explicit CS/CA has a unique majority for all 500 items.

Inter-annotator agreement on the explicit CS/CA field is generally high but varies across languages. Unanimous 3-of-3 agreement ranges from 65% in Norwegian Bokmål to 91% in Estonian. Fleiss' κ ranges from .25 in Danish to .83 in Estonian, and Gwet's AC1 from .60 in Norwegian Bokmål to .91 in Estonian. The Danish case illustrates the prevalence effect: although Fleiss' κ is low (.25, 95% CI $[-.05, .51]$), annotators still reach 84% unanimous agreement and AC1=.88 because almost all raw ratings are CS. Full multilingual IAA statistics, required-knowledge agreement, and bootstrap CIs are reported in Appendix C.4; annotation-provenance details in C.3.

4.2 Automatic Annotation Pipeline

To scale annotation to the full native benchmark collection (Table 1), we use GPT-5.2 with structured JSON output. The pipeline assigns the same GlobalMMLU-style labels as the human annotators; Appendix C.2 reproduces the full

prompt in Figure 3. We calibrated the prompt on GlobalMMLU-Lite, which provides 3–8 independent annotations per item (median: 4). We aggregate `required_knowledge` by majority vote, resolving simple ties between *none* and one non-none label in favor of the non-none category and discarding complex ties with multiple non-none labels. This yields 391 usable items (of 400; 226 *none*, 48 *cultural*, 114 *regional*, 3 *dialect*).

For GlobalMMLU-Lite only, the CS/CA reference follows its documented rule: CA if strictly more than half of annotators assigned *none*, and CS otherwise, resulting in 191 CS and 200 CA items. From the 391 usable items, we created a stratified dev/test split (174/217 items). Iterative refinement on the dev split addressed five recurring error patterns, including under-detection of Western philosophical terminology and implicit US-centric regional cues. Details are provided in Appendix C.5.2.

4.3 Meta-Evaluation

We evaluate the annotation pipeline against two reference sets: **(1)** the held-out GlobalMMLU-Lite test split ($n = 217$), and **(2)** the Native-500 reference described above. Both evaluations cover binary CS/CA classification and four-way `required_knowledge` classification. For Native-500, the binary CS/CA evaluation uses the explicit three-annotator majority label for all 500 items, whereas the four-way `required_knowledge` evaluation uses the 477 items with a unique majority. The prompt was calibrated only on GlobalMMLU-Lite. Native-500 was not used for prompt tuning, and GPT annotations for the native items were joined to the human-majority reference only for this meta-evaluation (Appendix C.5).

	n	Acc.	CS		CA	
			P	R	P	R
GlobalMMLU-Lite	217	73.3	79.5	61.7	69.4	84.5
Native (5 langs)	500	90.6	89.8	96.2	92.4	80.7

Table 3: Binary CS/CA agreement between GPT-5.2 and reference labels (%). CS/CA denotes culturally sensitive / culturally agnostic. For GlobalMMLU-Lite, the CS/CA reference follows the documented rule over human `required_knowledge` votes. For Native-500, the reference is the majority of three independent explicit CS/CA annotations per item.

CS/CA agreement. The pipeline identifies binary CS/CA labels more reliably on Native-500 than on the held-out GlobalMMLU-Lite test split. Table 3 reports binary CS/CA performance. On GlobalMMLU-Lite, GPT-5.2 reaches 73.3% accuracy, with lower recall for CS items (61.7%) than for CA items (84.5%), indicating that implicit cultural cues in English-origin items remain difficult. Against the explicit three-annotator native majority, accuracy rises to **90.6%** (macro-F1 89.5%). The native error profile is asymmetric: CS recall is high (96.2%), whereas CA recall is lower (80.7%), so remaining errors more often over-predict cultural sensitivity than miss it. As sensitivity checks, deriving GPT CS/CA from predicted `required_knowledge` yields 92.8% accuracy against the explicit native majority; deriving CS/CA from `required_knowledge` on both sides yields 94.4% (Appendix C.5).

Set	Class	P	R	F1
GlobalMMLU ($n = 217$)	*cultural*	59.5	75.8	66.7
	regional	82.5	55.9	66.7
	none	76.1	82.9	79.4
	dialect	100	50.0	66.7
Native ($n = 477$)	*cultural*	78.6	57.9	66.7
	regional	76.4	93.2	84.0
	none	96.0	90.7	93.3
	dialect	75.0	100	85.7

Table 4: Per-class precision (P), recall (R), and F1 (%) for `required_knowledge` predictions by GPT-5.2. Native metrics are computed on the 477 items with a unique three-annotator majority. Confusion matrices and tie handling are in Appendix C.5.

Required knowledge. Table 4 reports per-class precision, recall, and F1 for the four-way task. Although harder than binary CS/CA classification, performance remains strong for the main classes. On GlobalMMLU-Lite, *regional* recall is comparatively low (55.9%), reflecting confusion between implicit regional cues and *none*. On Native-500, the primary evaluation uses the 477 items with a unique human majority. Accuracy is **83.9%** and macro-F1 is 82.4%. If the 23 unresolved human ties are conservatively counted as errors, accuracy is 80.0% (400/500). The dominant native error is confusion from *cultural* to *regional*, while *none* and *regional* achieve F1 scores above 83%. Dialect scores are based on only three resolved native items and should be treated with caution.

5 Proxy Validity Evaluation

In this section we test whether translated MMLU proxies produce the same model rankings as native benchmarks (§5.2), then ask the stronger validity question: whether translated scores add predictive signal beyond model size and language proficiency (§5.3). We quantify uncertainty by bootstrapping models and test model-sample stability with leave-one-family-out (LOFO) refits and an English-MMLU control sensitivity (§5.4).

5.1 Evaluation Setup

We evaluate 29 instruction-tuned LLMs spanning four size categories: M ($\geq$70B, 7 models), S (15–32B, 8 models), XS (7–9B, 7 models), and XXS ($\leq$1.7B, 7 models), covering model families including Llama, Qwen, Gemma, Mistral, Falcon, and several European-focused models (Appendix D.1). All evaluations use the LM Evaluation Harness (Sutawika et al., 2025) with custom task implementations for the native benchmarks. We compute accuracy scores on all native benchmarks and their MMLU proxies (Table 1), as well as on Belebele (Bandarkar et al., 2024) in all five target languages as a language-proficiency control. Per-model outputs are merged with the CS/CA annotations from §4 for the stratified secondary analyses (Appendix D.2).

5.2 Ranking Analysis

We assess ranking consistency between native benchmarks and their MMLU proxies using Kendall's τ_b, reporting 95% bootstrap confidence intervals and top-5 overlap (Appendix D.4). As a preliminary check, CA and CS model rankings within the translated MMLU proxies are highly stable across languages ($\tau_b = .91$–$.94$; Appendix D.3).

Native	L	τ_b [95% CI]	T5	τ_{CA}	τ_{CS}
Skolprov	SV	.89 [.78, .96]	5/5	.88	.77
Exam-ET	ET	.88 [.78, .96]	5/5	.83	.88
INCLUDE	LT	.87 [.79, .93]	5/5	.82	.85
NRK-Quiz	NB	.85 [.76, .92]	4/5	.82	.88
DanishCT	DA	.82 [.73, .90]	3/5	.66	.84
INCLUDE	ET	.72 [.58, .83]	3/5	.68	.72
SwedishFacts	SV	.51 [.25, .74]	3/5	.50	.52

Table 5: Native-vs-proxy ranking consistency. τ_b: Kendall's τ_b on overall accuracy across 29 models. **T5**: top-5 model overlap. τ_{CA}/τ_{CS}: τ_b restricted to culturally agnostic / sensitive items. All $p < .001$. Full statistics in Appendix D.4.

Table 5 presents the native-vs-proxy ranking results. Three patterns emerge.

First, benchmarks with broad curriculum or content overlap show the strongest ranking agreement. Skolprov ($\tau_b = .89$), Exam-ET (.88), and INCLUDE-LT (.87) all achieve top-5 overlaps of 5/5, meaning that the five best-performing models are identical on the native benchmark and its same-language MMLU proxy.

Second, high-CS benchmarks can still show substantial ranking agreement. NRK-Quiz ($\tau_b = .85$) and DanishCitizenTest (.82) correlate strongly with their translated proxies despite substantial local or culturally specific content. This already suggests that CS/CA status alone does not determine ranking agreement.

Third, SwedishFacts shows the weakest ranking agreement. Its correlation with the Swedish MMLU proxy is much lower ($\tau_b = .51$), suggesting that Sweden-specific factual knowledge is less well captured by the translated benchmark. INCLUDE-ET also shows weaker agreement (.72) than the other multiple-choice pairings.

It is the smallest multiple-choice native benchmark ($n = 224$, versus 534–3,600 for the other MC benchmarks), which plausibly contributes to its wider rank uncertainty. We treat this as an explanation to test rather than an identified mechanism.

Most importantly, ranking correlation alone cannot establish proxy validity. Strong models tend to perform well on many benchmarks (Ilić and Gignac, 2024), so high τ_b values may reflect general model strength rather than genuine proxy signal. We therefore next test whether translated MMLU adds predictive value beyond model size and target-language proficiency.

5.3 Incremental Proxy Validity

To test whether the translated proxy carries information *beyond* what background factors already explain, we fit two nested OLS models for each pairing, where y is the native score, $s = \log(\text{total parameters})$, b is the Belebele score, and x is the proxy score, all in language L:

$$\textbf{M1 (baseline):} \quad y_L \sim s + b_L$$
$$\textbf{M2 (extended):} \quad y_L \sim s + b_L + x_L$$

If Model 2 improves over Model 1 in the nested F-test, the proxy carries predictive information beyond the controls. These regressions use total benchmark scores rather than individual CS/CA

labels. The annotations enter only in the content analysis and exploratory CS/CA stratification below. Because $n = 29$ models is small for regression, we pair the nested F-test with 5,000 model-level bootstrap resamples and LOFO stability checks (Appendix D.6). We classify the strength of evidence using the same criteria for all seven pairings. A pairing is *robust* if the full-sample incremental F-test is significant, the 95% bootstrap CI for β_{proxy} excludes zero, and β_{proxy} remains positive with nominal $p < .05$ in all 17 LOFO refits. A *marginal* pairing is full-sample significant but fails at least one robustness criterion. A *null* pairing has a non-significant full-sample F-test and a bootstrap β_{proxy} CI containing zero. These labels describe evidence strength, not intrinsic benchmark quality. We do not use the ΔR^2 CI as a null test, because in-sample ΔR^2 is non-negative by construction. Near-zero lower bounds instead indicate uncertainty about the size of the incremental effect.

Native	ΔR^2 [95% CI]	β_{proxy} [95% CI]	Class
Skolprov	.128 [.038,.253]	1.06 [.80,1.19]	robust
INCLUDE-LT	.076 [.021,.148]	0.94 [.67,1.13]	robust
NRK-Quiz	.041 [.009,.087]	0.46 [.29,.59]	robust
Exam-ET	.024 [.010,.044]	0.52 [.34,.85]	robust
INCLUDE-ET	.059 [.000,.199]	0.84 [−.02,1.30]	marginal
DanishCT	.009 [.000,.027]	0.31 [.01,.57]	marginal
SwedishFacts	.035 [.000,.163]	0.14 [−.10,.33]	null

Table 6: Incremental proxy validity on all items. Brackets show 95% model-bootstrap CIs. Classes follow the fixed evidence definition in the text. Full details are in Appendix D.6. The ΔR^2 CI is not used as a null test because in-sample $\Delta R^2 \geq 0$ by construction.

Table 6 shows the main result: **four pairings are robust, two are marginal, and one is null**. Skolprov has the strongest robust incremental signal ($\Delta R^2 = .128$, 95% CI $[.038, .253]$). INCLUDE-LT and NRK-Quiz are also robust despite substantial culturally sensitive content, showing that high CS content alone does not imply weak proxy validity. Exam-ET shows a smaller but robust increment ($\Delta R^2 = .024$), partly because the baseline is already very strong ($R^2_{\text{base}} = .954$). DanishCitizenTest is positive in the full sample but marginal under LOFO. INCLUDE-ET is likewise marginal because its β_{proxy} CI crosses zero and one family exclusion loses significance. For SwedishFacts, the proxy adds no significant incremental signal ($p = .242$), and its bootstrap coefficient CI contains zero. An exploratory CS/CA stratification is reported in Appendix D.5.

5.4 Robustness

Leave-one-family-out sensitivity. To check whether the incremental effects depend on a single model family, we exclude each of 17 model families in turn and refit the main model, leaving 25–28 of the 29 models per refit. All four robust pairings remain positive and nominally significant in every exclusion. The two marginal pairings are less stable: DanishCitizenTest loses nominal significance when Cohere, Llama, or Qwen are excluded ($p = .110, .072, .096$), and INCLUDE-ET when Qwen is excluded ($p = .102$). SwedishFacts remains non-significant in all 17 refits. We use the LOFO $p < .05$ rule only as a conservative stability criterion for the *robust* label, not as 17 multiplicity-corrected confirmatory tests (Appendix D.7).

Native	β_{proxy} [95% CI]	p_F
Skolprov	1.17 [.88,1.28]	$< 10^{-12}$
INCLUDE-LT	1.08 [.84,1.30]	$< 10^{-9}$
INCLUDE-ET	1.06 [.36,1.42]	$< 10^{-5}$
NRK-Quiz	0.54 [.34,.68]	$< 10^{-6}$
Exam-ET	0.59 [.39,.95]	$< 10^{-7}$
DanishCT	0.57 [.34,.92]	$< 10^{-3}$
SwedishFacts	0.25 [.06,.48]	.092

Table 7: Sensitivity analysis adding English G-MMLUX as a third control. Brackets are 95% percentile model-bootstrap CIs. The main evidence classes remain those in Table 6. Full ΔR^2, regression, and bootstrap statistics are in Appendix D.8.

Sensitivity: controlling for English MMLU performance. To test whether the translated-proxy association is merely an echo of general MMLU-content familiarity, we add each model's English G-MMLUX score as a third control. Six pairings remain significant by the nested F-test. SwedishFacts remains non-significant by the F-test ($p = .092$), but its bootstrap CI for β_{proxy} is positive ([.06,.48]), giving weak mixed evidence rather than a clean null in this alternative specification. This does not change its *null* classification under the fixed main-specification evidence definition. The English-control analysis preserves the broad ordering but does not identify a purely language-specific causal effect. In particular, the stronger DanishCitizenTest sensitivity result does not override its marginal classification in the main specification.

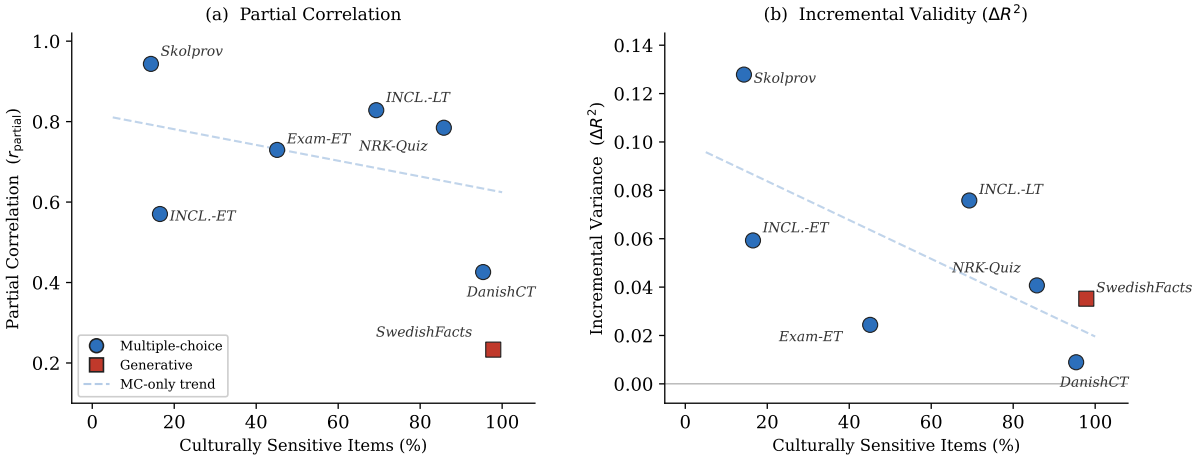


Figure 1: Proxy validity as a function of cultural sensitivity. Each point represents one native-vs-proxy pairing. **(a)** Partial correlation after residualizing on controls. **(b)** Incremental variance (ΔR^2) from the nested F-test. The dashed line shows the MC-only linear trend (6 pairings; descriptive and not statistically significant at $n = 6$). DanishCitizenTest (95% CS, MC format) already shows weak incremental validity. SwedishFacts lies close to the descriptive MC extrapolation in panel (b), which is consistent with a locality/content-overlap interpretation, but the design does not isolate locality from its open-ended response format.

6 Discussion

6.1 Interpreting Proxy Validity

What proxy validity does and does not show. Our incremental test shows where translated scores predict native-benchmark performance beyond model size and target-language proficiency. This is stronger evidence than ranking correlation alone, but it does not prove that translated and native benchmarks measure an identical latent construct or are interchangeable for every practical decision. The robust/marginal/null labels therefore describe evidence strength for incremental association within the evaluated model set, not intrinsic benchmark quality. Practitioners should read the results as evidence about when translated benchmarks approximate native evaluation signal, not as a blanket endorsement of translated benchmarks.

A content/locality-overlap gradient. The main pattern is a gradient: proxy validity is strongest when the native benchmark shares broad curriculum or content coverage with MMLU, and weaker or less stable as the native target becomes more locally anchored. Skolprov and Exam-ET are robust curriculum-overlap cases. INCLUDE-LT and NRK-Quiz also remain robust, showing that high cultural load alone does not imply poor proxy validity. DanishCitizenTest and INCLUDE-ET are marginal, while SwedishFacts is null in the main specification. This pattern appears across ranking analyses, incremental regressions, bootstrap uncertainty, LOFO stability, and the English-MMLU sensitivity analysis, but it is better understood as a content/locality-overlap gradient than as a deterministic function of the percentage of CS items.

Locality and response format. SwedishFacts is the clearest limitation case, but it is also the only open-ended benchmark in our study. Its null result is consistent with a locality/content-overlap interpretation, especially because the multiple-choice pairings already vary substantially in incremental validity and DanishCitizenTest shows only a small and fragile proxy signal despite its MC format. SwedishFacts also lies near the descriptive MC-only trend for ΔR^2 in Figure 1. However, the current design cannot separate Sweden-specific content from open-ended response format. The open-ended format may amplify the mismatch, and a controlled comparison would require additional generative benchmarks with low local content or MC benchmarks with comparable Sweden-specific factual content.

6.2 Annotation Reliability

The Native-500 reference provides a stronger validation basis because every item has three independent annotations in its target language. For explicit CS/CA, unanimous agreement ranges from 65–91% across languages, with Gwet's AC1=.60–.91. Danish illustrates the prevalence sensitivity of Fleiss' κ: despite 84% unanimous agreement, κ is only .25 because 277/300 raw ratings are CS. Norwegian Bok-

mål is the weakest language-level explicit-CS/CA reliability case, with 65% unanimous agreement, Fleiss' $\kappa = .44$, Gwet's AC1=.60, and GPT-5.2 CS/CA accuracy of 86%. We therefore interpret NB-stratified content analyses with particular caution. Overall, GPT-5.2 reaches 90.6% CS/CA accuracy against the explicit three-annotator majority. Four-way `required_knowledge` is harder: 23/500 items have unresolved 1–1–1 human ties, and GPT-5.2 reaches 83.9% accuracy on the 477 resolved items, or 80.0% if all ties are counted as errors. We therefore treat required-knowledge distinctions as noisier than the binary CS/CA variable. Importantly, the main incremental-validity analysis uses total benchmark scores and is independent of individual annotation labels. Annotation uncertainty affects the stratified analyses and descriptive content characterization rather than the core regression result.

6.3 Generalizability

Our evidence is restricted to five Nordic and Baltic target languages and DeepL-based MMLU proxies. The exact τ_b, ΔR^2, and coefficient magnitudes should not be assumed to transfer one-to-one to human or LLM translations, because translation quality can change item difficulty and model rankings. At the same time, translation cannot introduce source content that is absent from MMLU. The observed content-overlap pattern therefore motivates a broader hypothesis: translated benchmarks should be most useful when the source benchmark covers knowledge similar to the native evaluation target, and less reliable when the native target is strongly local. The language sample spans Germanic (SV, DA, NB), Finnic (ET), and Baltic (LT) languages, but remains geographically and typologically narrow. Whether the proxy-validity gradient generalizes to typologically distant and lower-resource languages remains an open question. Native benchmarks in other language settings can support such tests in future work.

7 Conclusion

We framed the question of whether translated benchmarks can stand in for native ones as a proxy-validity problem and tested whether translated MMLU scores carry incremental predictive value for native benchmark performance after controlling for model size and target-language proficiency. Across 29 models and seven pairings in five Nordic and Baltic languages, we find a clear gradient rather than a binary answer. Proxy validity is strongest when native benchmarks share broad subject or curriculum coverage with MMLU, as in Skolprov ($\Delta R^2 = .128$, 95% CI $[.038, .253]$). It remains robust for some culturally loaded benchmarks such as INCLUDE-LT and NRK-Quiz, becomes marginal for DanishCitizenTest and INCLUDE-ET, and is null in the main specification for SwedishFacts. A sensitivity analysis controlling for English MMLU performance supports the same overall pattern, indicating that general MMLU-content familiarity alone does not explain the observed proxy-validity differences. These results offer practical guidance: translated benchmarks are most defensible when their source content aligns with the native evaluation target, but locally grounded competences may require native benchmarks rather than translated substitutes.

Limitations

Our regression analyses cover 29 models, which is substantial for an evaluation study but still small for regression. The effect sizes and bootstrap intervals should therefore be interpreted as uncertainty over the evaluated model sample rather than over a population of all LLMs. Similarly, LOFO results are used as a conservative stability check. The nominal $p < .05$ criterion across 17 refits is not multiplicity-corrected.

All translated proxies in this study use DeepL. Other translation approaches, including human translation or LLM-based translation, may change absolute scores, rankings, and effect magnitudes. A multi-translator design would be needed to separate content effects from translation-quality effects more directly.

SwedishFacts is the only generative benchmark in the study, so locality/content and response format remain entangled for this pairing. The descriptive MC-only trend is useful for interpretation, but it is not a controlled format comparison.

The human reference covers 500 items, with 100 per language. Although three annotations per item support multilingual IAA, 23 required-knowledge items have unresolved 1–1–1 ties, showing that fine-grained cultural versus regional distinctions can be ambiguous.

Benchmark-level CS proportions used in Figure 1 are estimated by the automatic annotation pipeline rather than directly human-labeled. On

Native-500, CA recall is 80.7%, with 35 CA→CS versus 12 CS→CA errors, suggesting that estimated CS shares may be slightly upward biased. This may shift points horizontally in the descriptive gradient but does not affect the total-score regression outcomes.

Finally, the study covers five European target languages. Evaluation on typologically distant and lower-resource languages is needed before treating the observed content/locality gradient as broadly language-general.

Ethics Statement

This study involves human annotation of cultural-sensitivity and required-knowledge labels for benchmark items. Professional native-language annotators were recruited through ProZ.com and compensated on individually agreed terms.

The final Native-500 reference contains three independent annotations per item in each language, corresponding to 15 pseudonymous annotator IDs in total, and no personally identifying information is reported. Annotators were informed about the purpose of the study and intended use of their annotations.

GPT-5.2 is used for automatic annotation and for a separate model-assisted QA workflow. To avoid using model-assisted corrections as the human gold standard, the IAA analyses and Native-500 meta-evaluation use reconstructed raw/pre-QA human submissions and majority consensus, before GPT-assisted consistency review. Model-generated cultural-sensitivity labels may themselves reflect biases in the model's training data. We therefore report multilingual human agreement, confusion patterns, and tie rates transparently (§4.3; Appendix C).

All benchmark datasets used in this study are publicly available under their respective licenses. We release our annotations, the annotation pipeline, and the analysis code to support reproducibility.

Acknowledgments

The authors acknowledge the financial support by the Federal Ministry of Research, Technology and Space of Germany and by Sächsisches 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. The authors gratefully acknowledge the computing time made available to them on the high-performance computer at the NHR Center of TU Dresden. This center is jointly supported by the Federal Ministry of Research, Technology and Space of Germany and the state governments participating in the NHR (www.nhr-verein.de/unsere-partner). Evaluations were performed with computing resources granted by WestAI under project `westai0018` and by ZIH at TU Dresden for infrastructure used in automated evaluation computations.

References

Ashish Agrawal, Barah Fazili, and Preethi Jyothi. 2024. Translation Errors Significantly Impact Low-Resource Languages in Cross-Lingual Learning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 319–329, St. Julian's, Malta. Association for Computational Linguistics.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual Evaluation of Generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Ahmed Alaa, Thomas Hartvigsen, Niloufar Golchini, Shiladitya Dutta, Frances Dean, Inioluwa Deborah Raji, and Travis Zack. 2025. Position: Medical large language model benchmarks should prioritize construct validity. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 80991–81004. PMLR.

Alexandra Institute. 2024a. Danish-Citizen-Tests. Rev. `74158d9` (2024-02-15).

Alexandra Institute. 2024b. Multilingual MMLU. Rev. `18e6c8e` (2024-03-11).

Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. 2023. Revisiting Machine Translation for Cross-lingual Classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6489–6499, Singapore. Association for Computational Linguistics.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The Belebele Benchmark: A Parallel Reading Comprehension Dataset in 122 Language Variants. In *Proceedings of the 62nd Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan Eghlidi, Chris Schmitz, Karolina Korgul, Hunar Batra, Oishi Deb, Emma Beharry, Cornelius Emde, Thomas Foster, Anna Gausen, María Grandury, Sophia Han, Valentin Hofmann, Lujain Ibrahim, and 23 others. 2025. Measuring what Matters: Construct Validity in Large Language Model Benchmarks. In *Advances in Neural Information Processing Systems 38*, pages 19868–19949, San Diego, California, USA and Mexico City, Mexico. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Sander Bijl de Vroe, George Stampoulidis, Kai Hakala, Aku Rouhe, Mark van Heeswijk, and Jussi Karlgren. 2025. Comparing human and machine translations of generative language model evaluation datasets. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 80–85, Tallinn, Estonia. University of Tartu Library.

Ryan Burnell, Han Hao, Andrew R. A. Conway, and Jose Hernandez Orallo. 2023. Revealing the structure of language model capabilities. *Preprint*, arXiv:2306.10062.

Ariel Ekgren. 2025. Ekgren/swedish_skolprov. Rev. `a5218aa` (2025-12-15).

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, Wenhao Zhu, Hanxu Hu, Conghui He, Lei Li, Shujian Huang, and Fei Yuan. 2025. BenchMAX: A Comprehensive Multilingual Evaluation Suite for Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16751–16774, Suzhou, China. Association for Computational Linguistics.

Dieuwke Hupkes and Nikolay Bogoychev. 2025. MultiLoKo: A multilingual local knowledge benchmark for LLMs spanning 31 languages. *Preprint*, arXiv:2504.10356.

David Ilić and Gilles E. Gignac. 2024. Evidence of interrelated cognitive-like capabilities in large language models: Indications of artificial general intelligence or achievement? *Intelligence*, 106:101858.

Amr Keleg and Walid Magdy. 2023. DLAMA: A Framework for Curating Culturally Diverse Facts for Probing the Knowledge of Pretrained Language Models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6245–6266, Toronto, Canada. Association for Computational Linguistics.

Dongjun Kim, Gyuho Shim, Yongchan Chun, Minhyuk Kim, Chanjun Park, and Heuiseok Lim. 2025. Benchmark Profiling: Mechanistic Diagnosis of LLM Benchmarks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, page 15635–15650, Suzhou, China. Association for Computational Linguistics.

Jenny Kunz. 2026. A diagnostic benchmark for Sweden-related factual knowledge. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pages 5328–5334, Palma de Mallorca, Spain. ELRA Language Resource Association.

Viet Dac Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Helena Grete Lillepalu and Tanel Alumäe. 2026. Estonian native large language model benchmark. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pages 4257–4267, Palma de Mallorca, Spain. ELRA Language Resource Association.

Vladislav Mikhailov, Petter Mæhlum, Victoria Ovedie Chruickshank Langø, Erik Velldal, and Lilja Øvrelid. 2025. A Collection of Question Answering Datasets for Norwegian. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 397–407, Tallinn, Estonia. University of Tartu Library.

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.

Irene Plaza Ortiz, Nina Melero, Cristina Del Pozo, Javier Conde, Pedro Reviriego, Marina Mayor Rocher, and María Grandury. 2025. Spanish and LLM Benchmarks: Is MMLU Lost in Translation? In *2025 2nd International Generative AI and Computational Language Modelling Conference (GACLM)*, pages 104–108, Valencia, Spain. IEEE.

Sara Rajaee, Rochelle Choenni, Ekaterina Shutova, and Christof Monz. 2025. An Empirical Analysis of Machine Translation for Expanding Multilingual Benchmarks. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1–30, Suzhou, China. Association for Computational Linguistics.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Zeming Chen, Mohamed Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya K Dalmia, Abraham Diress, Sharad Duwal, and 38 others. 2025. INCLUDE: Evaluating multilingual language understanding with regional knowledge. In *International Conference on Learning Representations*.

Dan Saattrup 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.

Olawale Elijah Salaudeen, Anka Reuel, Ahmed M Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Benjamin W. Domingue, Angelina Wang, and Sanmi Koyejo. 2025. Measurement to Meaning: A Validity-Centered Framework for AI Evaluation. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*.

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, and 4 others. 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.

Lintang Sutawika, Hailey Schoelkopf, Leo Gao, Baber Abbasi, Stella Biderman, Jonathan Tow, ben fattori, Charles Lovering, farzanehnakhaee70, Jason Phang, Anish Thite, Fazz, Aflah, Niklas, Thomas Wang, sdtblck, nopperl, gakada, tttyuntian, and 11 others. 2025. EleutherAI/lm-evaluation-harness. Zenodo. Rev. `d021bf84` (2025-08-04).

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.

Hanna Wallach, Meera Desai, A. Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Nicholas J Pangakis, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs. 2025. Position: Evaluating Generative AI Systems Is a Social Science Measurement Challenge. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 82232–82251. PMLR.

Jiaan Wang, Fandong Meng, Yunlong Liang, Tingyi Zhang, Jiarong Xu, Zhixu Li, and Jie Zhou. 2023. Understanding Translationese in Cross-Lingual Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3837–3849, Singapore. Association for Computational Linguistics.

Laura Weidinger, Inioluwa Deborah Raji, Hanna Wallach, Margaret Mitchell, Angelina Wang, Olawale Salaudeen, Rishi Bommasani, Deep Ganguli, Sanmi Koyejo, and William Isaac. 2025. Toward an Evaluation Science for Generative AI Systems. *Preprint*, arXiv:2503.05336.

Minghao Wu, Weixuan Wang, Sinuo Liu, Huifeng Yin, Xintong Wang, Yu Zhao, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. The Bitter Lesson Learned from 2,000+ Multilingual Benchmarks. *Preprint*, arXiv:2504.15521.

Mike Zhang and Antonio Toral. 2019. The Effect of Translationese in Machine Translation Test Sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics.

A Extended Related Work

This appendix provides extended discussion of the related work summarized in §2, with more detailed descriptions of individual studies.

A.1 Translation in Multilingual Evaluation

Rajaee et al. (2025) translate four NLP tasks: XNLI, PAWS-X (paraphrase identification), XCOPA (causal commonsense reasoning) and XStoryCloze (story completion) into 198 languages via NLLB. They find that MT-based evaluations broadly track human-translated evaluations, although the strength of agreement varies across tasks and model settings. They further show that commonly benchmarked low-resource languages can lead to systematic overestimation of model performance relative to a broader set of low-resource languages.

Bijl de Vroe et al. (2025) compare English, Finnish-HT, and Finnish-MT versions of ARC Challenge and observe a nuanced picture: while model-size rankings are broadly preserved, base and instruction-tuned models swap places depending on the translation variant, suggesting that translation effects interact with model type.

Agrawal et al. (2024) provide a fine-grained analysis showing that translation inconsistencies in XNLI disproportionately impact low-resource languages, with accuracy gaps reaching 10 points. At scale, Huang et al. (2025) implement an MT→post-editing pipeline across 16 languages and 10 tasks, demonstrating that post-editing can mitigate MT artifacts.

Artetxe et al. (2023) demonstrate that carefully adapted translate-test can be competitive with cross-lingual transfer for classification, though the evaluation remains within translated English-origin content.

Ahuja et al. (2023) evaluate generative LLMs across a wide set of NLP tasks in multiple languages (MEGA) and find that performance gaps are strongly language-dependent, with some languages showing near-English performance while others lag far behind.

Lai et al. (2023) evaluate ChatGPT on non-English benchmarks and document substantial cross-lingual performance variation, with large gaps for some languages with low CommonCrawl representation.

A.2 Native versus Translated Benchmarks

Wu et al. (2025) conduct a meta-analysis of over 2,000 multilingual benchmarks and report that localized benchmarks correlate more strongly with human preferences than translated ones. However, the analysis does not control for general model strength, leaving open the possibility that the correlation difference is partly driven by the "strong models are strong everywhere" confound.

Singh et al. (2025) build Global MMLU with improved translations and cultural-sensitivity annotations (CS/CA labels) across 42 languages. They demonstrate that model rankings shift when evaluation is restricted to culturally sensitive items, a key finding that motivates our own CS/CA-stratified analysis, but they do not compare against independently created native exams.

Keleg and Magdy (2023) introduce DLAMA, a culturally diverse probing benchmark, and show that LLMs underperform on factual knowledge associated with non-Western cultures, providing direct evidence that knowledge gaps in LLMs are culturally patterned rather than uniform.

Closest to our work, Hupkes and Bogoychev (2025) introduce MultiLoKo with locally sourced questions for 31 languages alongside human and machine translations, and quantify a substantial Locality Effect. The reported performance gap between locally sourced and translated data exceeds 20 accuracy points for some model-language combinations. However, MultiLoKo focuses on benchmark construction and descriptive comparisons. It does not formalize proxy validity with statistical controls or attempt to disentangle the factors driving the observed gaps. Our work extends this line by providing such controls and by analyzing whether proxy validity differs between culturally agnostic and culturally sensitive subsets.

A.3 Translationese and Translation Artifacts

Zhang and Toral (2019) demonstrate that translated MT test sets inflate human evaluation scores for machine translation systems and can alter system rankings, because evaluators find translated references more predictable.

Wang et al. (2023) show that translationese in cross-lingual summarization data creates systematic discrepancies between automatic metrics and human evaluation, calling into question the reliability of evaluation on translated data.

Park et al. (2024) provide evidence that models

trained on translated data adapt to translation artifacts rather than learning genuine cross-lingual representations, inflating their apparent cross-lingual performance on benchmarks that share these artifacts.

Plaza Ortiz et al. (2025) conduct an item-level analysis of MMLU translated into Spanish and confirm that a significant share of model errors trace back to translation problems including ambiguous answer choices, mistranslated domain terminology, and loss of pragmatic cues rather than to genuine knowledge gaps.

These findings collectively establish that translation effects are not mere noise but systematic biases that can fundamentally alter evaluation conclusions.

A.4 Validity Frameworks for LLM Evaluation

Salaudeen et al. (2025) propose a validity-centered framework for AI evaluation that emphasizes claim-first reasoning and distinguishes criterion validity from proxy validity. Their use of "proxy validity" refers to whether a benchmark score (observable) validly represents a latent construct (e.g., reasoning ability). Our use of the term is narrower and more practical. We ask whether one observable benchmark (translated) can substitute for another observable benchmark (native) in producing equivalent evaluation conclusions.

Bean et al. (2025) conduct a systematic review of 445 LLM benchmarks from leading ML and NLP conferences and find widespread threats to construct validity, including vague phenomenon definitions, insufficient sampling, and missing reliability assessments, concerns that apply directly to translated benchmarks used as stand-ins for native tests.

Alaa et al. (2025) focus specifically on the medical domain, drawing parallels between LLM benchmarks and psychological tests. They demonstrate empirically that high scores on medical exam benchmarks (e.g., MedQA) do not predict performance on real clinical records, exposing a fundamental gap in predictive validity that resonates with our concern about translated benchmarks as proxies.

More broadly, Wallach et al. (2025) argue that evaluating GenAI systems is fundamentally a social science measurement challenge and present a four-level framework grounded in measurement theory that highlights how current benchmarking practices often skip the crucial systematization step.

Weidinger et al. (2025) call for maturing an evaluation science for generative AI, drawing lessons from transportation, aerospace, and pharmaceutical engineering where evaluation metrics must be applicable to real-world performance and iteratively refined.

On the empirical side, Burnell et al. (2023) apply factor analysis to a large battery of LLM evaluations and find that capabilities are not monolithic but explained by multiple latent factors (reasoning, comprehension, language modeling).

Complementing this, Ilić and Gignac (2024) analyze 591 LLMs and identify a strong general ability factor analogous to the g-factor in psychometrics. Both findings motivate our use of statistical controls: without accounting for general model strength, any two benchmarks will correlate simply because strong models tend to perform well everywhere.

Kim et al. (2025) further show through benchmark profiling that individual benchmarks typically measure mixtures of abilities rather than single constructs, reinforcing that a translated benchmark may tap into a different ability mixture than a native exam even when both nominally test "knowledge in language L."

B Appendix: Datasets

B.1 G-MMLUX Construction

G-MMLUX is a unified subset that maps the culturally annotated items from GlobalMMLU (Singh et al., 2025) to their DeepL translations in each target language, including the English originals. Before filtering, G-MMLUX contains 16,996 items (EN/DA/ET/LT/SV = 2,850 each; NB = 2,746). We apply agreement filtering to exclude items with low annotator consensus in GlobalMMLU, removing 50 items per language for EN/DA/ET/LT/SV and 45 for NB, yielding 16,701 items in the final G-MMLUX (EN/DA/ET/LT/SV = 2,800; NB = 2,701). All downstream analyses use this filtered version.

GlobalMMLU-Lite consists of 391 items after label aggregation (174 dev + 217 test); we use the 217-item test split in all evaluations. DanishCitizenTest was deduplicated from the original 720 items to 668.

B.2 Subject and Category Mapping

This appendix provides the complete subject-to-MMLU mapping, native subject translations, dataset statistics, and cross-validation details referenced in §3.

Mapping confidence levels. To capture the ambiguity inherent in cross-benchmark subject alignment, we assigned each mapping a *confidence level*: **direct** (1:1 MMLU subject match), **approximate** (plausible but not exact), or **no-equivalent** (category-level only). Table 8 shows the distribution across datasets.

Notably, all native exam items fall into the *approximate* or *no-equivalent* categories, illustrating why subject-level comparison with MMLU is inherently imprecise.

Mapping challenges: detailed examples. Three types of mismatches complicate direct subject-level alignment between native benchmarks and MMLU.

Granularity mismatches arise when a native exam covers multiple difficulty levels that MMLU separates. Estonian *matemaatika* in Exam-ET spans content from basic algebra to university-level calculus, which MMLU distributes across `elementary_mathematics`, `high_school_mathematics`, and `college_mathematics`. Similarly, *füüsika* maps to `high_school_physics`, `college_physics`, and `conceptual_physics`. Without item-level difficulty annotation, this ambiguity cannot be fully resolved.

Dataset	Direct	Approx.	No-equiv.
GlobalMMLU	2,850	—	—
GlobalMMLU-Lite	217	—	—
Skolprov	—	465	80
Exam-ET	—	1,130	484
DanishCitizenTest	—	668	—
SwedishFacts	—	—	1,292
INCLUDE-base (LT)	—	—	534
INCLUDE-base (ET)	—	—	224
NRK-Quiz-QA	—	—	3,600

Table 8: Items per mapping confidence level. *Direct*: 1:1 MMLU subject match. *Approx.*: plausible subject mapping with granularity or scope mismatch. *No-equiv.*: no MMLU subject equivalent; mapped at category level only. Datasets without native subject labels (bottom block) are classified entirely via item-level annotation.

Scope mismatches occur when a single native subject conflates content from distinct MMLU subjects. The Swedish *matematik_och_fysikprovet* combines mathematics and physics in a single test, spanning three MMLU subjects. Estonian *ühiskonnaõpetus* (Social Studies) covers both government/politics and sociology. The Danish citizenship and residence tests similarly span civics, history, and cultural knowledge that MMLU distributes across multiple subjects.

Native Subject	English
Skolprov (SV)	
kunskapsprov_läkare	Prof. Knowl.: Physician
kunskapsprov_tandläkare	Prof. Knowl.: Dentist
kunskapsprov_audionom	Prof. Knowl.: Audiologist
kunskapsprov_apotekare	Prof. Knowl.: Pharmacist
matematik_och_fysikprovet	Math. & Physics Test
högskoleprovet	Scholastic Apt. Test
Exam-ET (ET)	
ajalugu	History
bioloogia	Biology
füüsika	Physics
geograafia	Geography
keemia	Chemistry
ühiskonnaõpetus	Social Studies
eesti_keel_teise_k._güm.	Estonian as L2 (Gymn.)
eesti_keel_teise_k._põhi.	Estonian as L2 (Basic)
DanishCitizenTest (DA)	
citizenship	Citizenship Test
residence	Residence Permit Test

Table 9: Native exam subject labels and English translations.

No-equivalent subjects have no MMLU counterpart. Language proficiency tests—such as *eesti*

keel teise keelena (Estonian as a Second Language) at gymnasium and basic school levels—assess L2 competence absent from MMLU.

Dataset	L	Items	Subj.	M-subj.	M-cat.	Ann.
GlobalMMLU	EN	2,850	57	57	6	—
G-MMLU-Lite	EN	217	43	43	6	6
Skolprov	SV	545	6	9	3	5
Exam-ET	ET	1,614	8	13	3	5
DanishCT	DA	668	2	2	1	6
SwedishFacts	SV	1,292	—	—	—	6
INCLUDE-LT	LT	534	—	—	—	6
INCLUDE-ET	ET	224	—	—	—	5
NRK-Quiz-QA	NB	3,600	—	—	—	6

Table 10: Dataset overview. **Subj.**: number of native subject labels. **M-subj.**: number of distinct MMLU subjects covered by the mapping. **M-cat.**: number of MMLU categories covered. **Ann.**: number of distinct `subject_category` labels assigned by GPT-5.2 annotation. Datasets without native subject labels rely on item-level annotation for category assignment.

Vocabulary comprehension (ORD) and reading comprehension (LÄS) in the Swedish Högskoleprov test linguistic aptitude rather than domain knowledge.

Professional licensing exams in Skolprov—*kunskapsprov läkare* (physician), *kunskapsprov tandläkare* (dentist), *kunskapsprov audionom* (audiologist), and *kunskapsprov apotekare* (pharmacist)—test specialized clinical knowledge that MMLU approximates at best through `professional_medicine` or `clinical_knowledge`, but without the vocational specificity of the Swedish exams.

These mismatches motivate our three-level confidence system (*direct*, *approximate*, *no-equivalent*) and our recommendation that category-level rather than subject-level comparison is more appropriate for cross-benchmark analysis.

For detailed numbers, please refer to Table 11.

Cross-validation with automated annotations. To validate category assignments, we compared them against the `subject_category` labels produced by our GPT-5.2 annotation pipeline (§4). Native exam items use our manual category mappings; GlobalMMLU-Lite items use their direct MMLU categories. Across 3,044 cross-validated items, agreement reaches **85.1%** (Table 13). The original MMLU taxonomy (Hendrycks et al., 2021) contains 57 subjects across four broad categories expanded to six by Singh et al. (2025): STEM (18), Humanities (13), Social Sciences (12), Medical (7), Business (5), and Other (2).

The 10 subjects with disagreements are attributable to inherent MMLU taxonomy ambiguities (law subjects classified as HUMANITIES vs. SOCIAL SCIENCES), heterogeneous catch-all categories, or low sample sizes ($n \leq 4$). We found no clear cases where disagreements indicated a manual mapping error. This supports category-level comparison as the more appropriate basis for cross-benchmark analysis; subject-level alignment is too noisy for reliable interpretation.

B.3 Comparison Pairing Rationale

Each pairing in Table 1 matches a translated MMLU proxy with a native benchmark in the same language, designed to test proxy validity across a range of scenarios.

Curriculum-overlap pairings. Exam-ET vs. MMLU_ET: Both are MC knowledge/exam QA. Exam-ET draws from the Estonian national curriculum across eight subjects; MMLU_ET covers 57 subjects from a US-centric perspective. Our category mapping shows that 70% of Exam-ET items (1,130 of 1,614) map at *approximate* confidence, primarily in STEM (biology, chemistry, physics), HUMANITIES (history), and SOCIAL SCIENCES (social studies, geography). The remaining 30% (484 items) are language proficiency tests with no MMLU equivalent.

Skolprov vs. MMLU_SV: Skolprov is rooted in Swedish educational standards and professional licensing, while MMLU_SV is a DeepL translation of US-centric content. Of Skolprov's 545 items, 85% (465) map at approximate confidence—primarily MEDICAL (340 items from the four professional licensing exams) and STEM (125 items from the mathematics and physics test). The remaining 80 items from the Högskoleprov are a mixed aptitude test with no clean MMLU equivalent.

High-cultural-load pairings. DanishCitizenTest vs. MMLU_DA: The Danish citizenship and residence permit tests specifically target knowledge about Danish society, culture, history, and democratic values—content that translated English-origin benchmarks are unlikely to capture well. With 95.4% culturally sensitive items and the majority requiring regional knowledge (558 of 668 items annotated as `regional`), this pairing provides a strong test of whether translated proxies retain signal for locally anchored civic content.

Dataset	Native Subject	MMLU Subject(s)	Cat.	Conf.	n
Skolprov (SV)					
	kunskapsprov_läkare	professional_medicine, clinical_knowledge, college_medicine	MED	appr.	132
	matematik_och_fysikprovet	high_school_mathematics, high_school_physics, college_physics	STEM	appr.	125
	kunskapsprov_audionom	clinical_knowledge, anatomy	MED	appr.	85
	högskoleprovet	miscellaneous	OTH	no-eq.	80
	kunskapsprov_tandläkare	professional_medicine, clinical_knowledge, anatomy	MED	appr.	70
	kunskapsprov_apotekare	clinical_knowledge, nutrition	MED	appr.	53
Exam-ET (ET)					
	ühiskonnaõpetus	hs_gov_and_politics, sociology	SOC	appr.	310
	ajalugu	hs_european_history, hs_world_history	HUM	appr.	261
	eesti_keel_teise_keelena_güm.	miscellaneous	HUM	no-eq.	246
	eesti_keel_teise_keelena_põhi.	miscellaneous	HUM	no-eq.	238
	bioloogia	high_school_biology, college_biology	STEM	appr.	207
	keemia	high_school_chemistry, college_chemistry	STEM	appr.	133
	geograafia	high_school_geography	SOC	appr.	111
	füüsika	hs_physics, college_physics, conceptual_physics	STEM	appr.	108
DanishCitizenTest (DA)					
	citizenship	hs_gov_and_politics, sociology	SOC	appr.	336
	residence	hs_gov_and_politics, sociology	SOC	appr.	332
Datasets without native subjects					
SwedishFacts (SV)	*No native subjects; category via item-level annotation*		mixed	no-eq.	1,292
INCLUDE-base (LT)			mixed	no-eq.	534
INCLUDE-base (ET)			mixed	no-eq.	224
NRK-Quiz-QA (NB)			mixed	no-eq.	3,600

Table 11: Native subject to MMLU mapping. **Cat.**: mapped MMLU category (STEM, HUM=Humanities, SOC=Social Sciences, MED=Medical, BUS=Business, OTH=Other). **Conf.**: mapping confidence—*appr.*=approximate, *no-eq.*=no equivalent. GlobalMMLU and GlobalMMLU-Lite are omitted as their subjects map directly (1:1) to MMLU. n: number of items.

Dataset	Lang.	Items	CA	CS	Required Knowledge	Time-sens.
GlobalMMLU	EN	2,850	2,058	792	none=9505; regional=1809; cultural=1253; dialect=84	Yes=428
GlobalMMLU-Lite	EN	217	121	96	none=134; cultural=42; regional=40; dialect=1	Yes=32
Skolprov	SV	545	467	78	none=529; regional=14; dialect=2	Yes=66
Exam-ET	ET	1,614	886	728	none=1120; regional=485; cultural=9	Yes=229
DanishCitizenTest	DA	668	31	637	regional=558; cultural=77; none=33	Yes=127
SwedishFacts	SV	1,292	28	1,264	regional=641; cultural=622; none=27; dialect=2	Yes=185
INCLUDE-base (LT)	LT	534	164	370	regional=294; none=187; cultural=53	Yes=119
INCLUDE-base (ET)	ET	224	187	37	none=199; regional=20; dialect=3; cultural=2	Yes=4
NRK-Quiz-QA	NB	3,600	512	3,088	cultural=1483; regional=1441; none=578; dialect=98	Yes=1023

Table 12: Cultural annotation statistics. **CA**: culturally agnostic items. **CS**: culturally sensitive items. **Required Knowledge**: distribution of annotated knowledge types (note: GlobalMMLU counts reflect individual annotator votes, not aggregated labels). **Time-sens.**: items with time-sensitive content. The annotation-derived counts are used only to characterize the benchmark collections. The annotation procedure, reliability analysis, and meta-evaluation are documented separately in Appendix C.

Dataset	Native Subject	Our Cat.	GPT-5.2 Majority	n	Agree %	Agreement
Native exam datasets (approximate / no-equivalent mappings)						
Skolprov	matematik_och_fysikprovet	STEM	STEM	125	100.0	✓
Skolprov	kunskapsprov_läkare	MED	MED	132	85.6	✓
Skolprov	kunskapsprov_tandläkare	MED	MED	70	100.0	✓
Skolprov	kunskapsprov_audionom	MED	MED	85	82.4	✓
Skolprov	kunskapsprov_apotekare	MED	MED	53	71.7	✓
Skolprov	högskoleprovet	OTH	HUM	80	8.8	×
Exam-ET	ühiskonnaõpetus	SOC	SOC	310	100.0	✓
Exam-ET	ajalugu	HUM	HUM	261	95.0	✓
Exam-ET	bioloogia	STEM	STEM	207	98.1	✓
Exam-ET	keemia	STEM	STEM	133	99.2	✓
Exam-ET	füüsika	STEM	STEM	108	100.0	✓
Exam-ET	geograafia	SOC	SOC	111	57.7	✓
Exam-ET	eesti_keel_teise_keelena_güm.	HUM	HUM	246	93.1	✓
Exam-ET	eesti_keel_teise_keelena_põhi.	HUM	HUM	238	99.2	✓
DCT	citizenship	SOC	SOC	336	70.5	✓
DCT	residence	SOC	SOC	332	76.2	✓
Selected MMLU-Lite subjects with disagreements (direct mappings)						
MMLU-Lite	international_law	HUM	SOC	5	0.0	×
MMLU-Lite	jurisprudence	HUM	SOC	5	0.0	×
MMLU-Lite	professional_law	HUM	SOC	5	0.0	×
MMLU-Lite	human_aging	MED	SOC	10	10.0	×
MMLU-Lite	global_facts	OTH	SOC	9	0.0	×
MMLU-Lite	hs_us_history	HUM	SOC	5	0.0	×
MMLU-Lite	miscellaneous	OTH	HUM	29	27.6	×

Table 13: Cross-validation of our category mapping against GPT-5.2 `subject_category` annotations. **Our Cat.**: category assigned by our mapping. **GPT-5.2 Majority**: most frequent category in the independent annotation. **Agree %**: proportion of items where our category matches the annotation. Overall agreement across 3,044 items: **85.1%**. The bottom block shows disagreements originating from the MMLU taxonomy itself (law → humanities vs. social sciences), not from our mapping procedure. Category abbreviations: STEM, HUM=Humanities, SOC=Social Sciences, MED=Medical, BUS=Business, OTH=Other.

NRK-Quiz-QA vs. MMLU_NB: NRK-Quiz-QA covers Norwegian-specific general knowledge and cultural topics from public broadcasting quizzes, with 85.8% culturally sensitive items. Unlike DanishCitizenTest (which tests civic/institutional knowledge), NRK-Quiz-QA spans broader cultural trivia including geography, language, and current affairs.

Middle-ground pairings. INCLUDE (ET, LT) vs. MMLU_ET/LT: INCLUDE provides natively sourced exam questions from an independent multilingual benchmark effort, offering an additional native comparison point outside the national-exam collections. The Estonian subset (224 items) and Lithuanian subset (534 items) test region-specific academic and professional knowledge.

SwedishFacts vs. MMLU_SV: SwedishFacts complements Skolprov by probing Sweden-specific factual knowledge in a generative (non-MC) format, with 97.8% culturally sensitive items—the highest CS proportion in our study. Because it is the only non-MC benchmark in our study, we interpret it separately from the MC-based comparisons.

C Appendix: Annotation

C.1 Annotation Schema and Interface

This appendix provides the annotation schema, interface example, automatic-annotation prompt, human-reference reconstruction, inter-annotator agreement, and meta-evaluation details referenced in §4.

Full annotation schema. Each item is a short benchmark item in one of five target languages. For each item, annotators provide the following labels:

- **Required knowledge type (Q1).** Whether answering correctly requires geographic, cultural, dialect-specific, or no specialized knowledge (exactly one of: `regional`, `cultural`, `dialect`, `none`). The key criterion is *necessity*: a knowledge type is selected only if it is required to answer correctly, not merely mentioned in passing. A tie-break rule (dialect > geographic > cultural > none) resolves cases where multiple types appear equally necessary.
- **Time sensitivity (Q2).** Whether the item contains explicit temporal or trend-based framing (e.g., "currently", "in recent years"). Annotators mark `Yes` only when such cues are explicitly present in the text.
- **Trigger spans (Q3).** Minimal, non-overlapping text spans that justify the Q1 decision, each labeled as `Geographic`, `Cultural`, or `Dialect`. If the required knowledge is implicit, spans may be left empty without changing the Q1 decision.
- **Conditional metadata (Q4–Q6).** If Q1 = `cultural`, annotators select one of seven culture groups (e.g., Western, Eastern Asian, Middle Eastern). If Q1 = `regional`, they select a world region and optionally a specific country in ISO 3166-1 format. These fields are mutually exclusive by design to enforce schema consistency.
- **Cultural sensitivity label (CS/CA).** A binary label indicating whether the item is *culturally sensitive* (CS) or *culturally agnostic* (CA). CA is assigned only when the item is free of any cultural, geographic, or dialect cues. This label is not strictly determined by Q1: an item may be labeled CS even if Q1 = `none`, in borderline cases where cultural cues are present but not strictly necessary for answering.
- **Subject category and explanation.** One of six categories (STEM, Humanities, Social Sciences, Medical, Business, Other), based on the knowledge domain required to answer correctly, accompanied by a short English justification ($\leq$ 250 characters).

The annotation interface was deployed using Argilla[2] hosted on Hugging Face Spaces and configured with the full schema described above. Figure 2 shows a completed example, including the item panel, form fields, metadata fields, and submission controls.

C.2 Annotation Prompt

Figure 3 reproduces the system prompt used for automatic annotation. The prompt asks GPT-5.2 to return the same schema as the human annotation interface in structured JSON format.

API and reproducibility details. To support reproducibility, all automatic benchmark annotations were generated in February 2026 using the pinned model snapshot `gpt-5.2-2025-12-11` via the OpenAI API, with `reasoning.effort=high` and `max_output_tokens=8000`. Each item was annotated in a single zero-shot call with the full system prompt (Figure 3). The human-reference reconstruction and the separate model-assisted QA workflow are described below; importantly, the majority gold and IAA use raw/pre-QA human submissions rather than model-assisted corrections.

C.3 Human Annotation Details

C.3.1 Sampling and Recruitment

For each target language (DA, ET, LT, NB, SV), we sampled 100 native-benchmark items. Sampling was stratified using automatic metadata, including required knowledge, explicit CS/CA, time sensitivity, and subject category. These automatic labels were used only for sampling and were removed before human annotation. When a language contained multiple native datasets, we allocated the sample across datasets to preserve coverage.

Professional native-language annotators were recruited through ProZ.com. The final Native-500 reference contains three independent annotations per item in each language, corresponding to 15 pseudonymous annotator IDs in total. Annotators independently recorded `required_knowledge`,

[2] Argilla: https://github.com/argilla-io/argilla

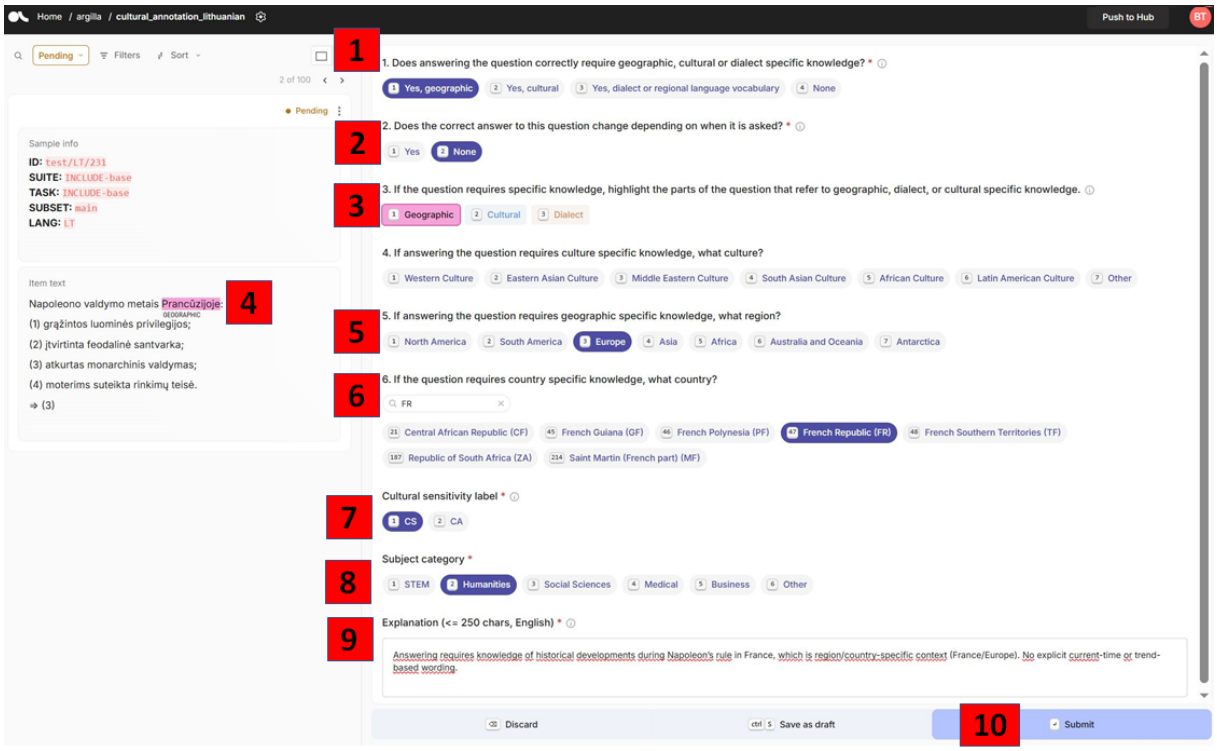


Figure 2: Argilla annotation interface shown through a completed Lithuanian item about Napoleonic rule in France. The interface includes the item text in the target language, guideline and navigation controls, required-knowledge selection, time-sensitivity flag, trigger-span highlighting, conditional culture/region/country metadata, explicit CS/CA label, subject category, explanation field, and submission controls. In this example, the annotator selected geographic/regional knowledge, marked no time sensitivity, highlighted the geographic trigger span “Prancūzijoje” (France), selected region *Europe* and country *France (FR)*, assigned the *CS* label, chose *Humanities* as the subject category, and provided a short explanation before submitting.

System Prompt for Cultural Sensitivity Annotation

You are an annotator for cultural sensitivity in exam-style multiple-choice questions.
You will receive ONE item as plain text in language: {LANG}.
The text may include the answer key marker like "=> (1)". Do NOT use the answer key to decide spans; you may ignore it.

Goal: Create a GlobalMMLU-style annotation that captures whether answering correctly requires: 1) Cultural knowledge 2) Geographic (regional) knowledge 3) Dialect or regional language vocabulary knowledge 4) Temporal (time-sensitive) knowledge

Language requirement: All annotation fields and the explanation must be in English. Only the "span" strings in references/other_references must be copied verbatim from the provided TEXT (and may be in {LANG}).

Definitions:
Cultural knowledge: Culture-specific facts, norms, values, customs, religion, arts/literature, pop culture, historically rooted traditions tied to a cultural group. Label as cultural if there is a clear cultural anchor such as: religion or religious law/ritual (e.g., Sharia, Halakha, canon law) when relevant for answering; a named canonical philosopher/scholar/text/school/tradition (esp. in philosophy/psychology/political theory); culturally specific norms or traditions. Treat named canonical figures/schools as a cultural anchor (Western Culture), especially when the question asks about their theories/views (e.g., Freud, Kant, Aristotle, Mill, Rawls, Bentham, Hobbes, Locke, Marx, Weber, Durkheim). Do NOT treat "According to Author & Author (Year)" as cultural; this is domain-specific academic recall. Also treat standard terminology from Western moral philosophy, logic, and philosophy of mind (e.g., consequentialism, virtue ethics, deontology, argument from ignorance, substance dualism) as a cultural anchor with culture="Western Culture". Religion note: Do NOT map religions mechanically to a region; set culture="Other" if no clear region is anchored.
Geographic/regional knowledge: Knowledge tied to specific regions/countries. Important cues: Imperial/US customary units (inch, mile, feet, Fahrenheit, pounds, mph) -> regional, region="North America". Currency "$" is ambiguous alone; only regional with additional US anchor. US state names or US-specific events (e.g., World Series) -> regional. Named country-specific regulations (e.g., EMAS, GDPR, Sarbanes-Oxley) -> regional. Place-in-scenario: if a real country/region is part of the scenario setup, label regional even if the principle is general, UNLESS mentioned only as historical background. Do NOT over-label: country mentions in historical passages are not automatically regional.
Dialect knowledge: Region/community-specific language variations, slang, idioms, or dialect forms necessary to answer. Treat common idioms (e.g., "Mom & Pop", "blue-collar", "red tape", "y'all") as dialect and include the idiom as a span.
Temporal knowledge: Mark Yes if the item contains explicit time references or trend/change language (e.g., "in recent years", "currently", "has increased"), even if the fact might still be true.

Subject category (always required): Assign exactly one of STEM, Humanities, Social Sciences, Medical, Business, Other.

Decision rules: Only mark a knowledge type if it is necessary to answer correctly. Choose exactly ONE required_knowledge. Tie-break: dialect > regional > cultural > none. Set cultural_sensitivity_label="CA" ONLY if the item is clearly free of cultural/geographic/dialect cues; otherwise "CS", even if required_knowledge="none" (borderline sensitivity).

Span rules: references must contain only spans justifying the chosen required_knowledge. Offsets are 0-indexed, end exclusive, matching TEXT exactly. If knowledge is implicit, references can be [].

Output: Return ONLY JSON: {"culture_ann": {"required_knowledge": "none"|"cultural"| "regional"|"dialect", "time_sensitive": "Yes"|"No", "cultural_sensitivity_label": "CS"|"CA", "subject_category": "STEM"|"Humanities"|"Social Sciences"|"Medical"|"Business"|"Other", "culture": str|null, "region": str|null, "country": str|null, "references": [{"start": int, "end": int, "span": str, "label": "Geographic"|"Cultural"|"Dialect"}], "other_references": [...], "explanation": str (max 250 chars)}}

Figure 3: Full system prompt for automatic cultural sensitivity annotation via GPT-5.2. The placeholder {LANG} is replaced with the target language code. All output fields are in English; only span extractions appear in the original language. The prompt is also available in the accompanying code repository.

explicit `cultural_sensitivity_label`, `time_sensitive`, `subject_category`, and the associated schema metadata. Explicit CS/CA was a separate annotation field and was not mechanically derived from `required_knowledge`.

C.3.2 Raw/Pre-QA Reference Reconstruction

The annotation campaign produced multiple archived exports and workflow snapshots. For the reliability and meta-evaluation analyses, we reconstruct the earliest usable *raw/pre-QA* submission for each annotator and item. We use these raw/pre-QA labels, rather than labels after GPT-assisted quality review, to avoid defining the human gold standard through the same model family used for automatic annotation.

The reconstruction starts from archived per-user raw snapshots, restores missing initial items from pre-validation "finished" snapshots where necessary, and uses the first stable completed but unvalidated submission when an earlier workflow record was incomplete. Some archived records retain a workflow status such as `pending` despite populated annotation fields. Inclusion is therefore based on annotation coverage, not the Argilla status flag. The resulting reference contains complete values for all five evaluation fields on all 100 items for each of the 15 annotators. The reconstruction reports and final item-level consensus files are retained with the accompanying analysis artifacts.

For each item, explicit CS/CA gold is the simple majority of the three independently annotated `cultural_sensitivity_label` values. All 500 items have a unique explicit-CS/CA majority. Required knowledge is aggregated independently: 477 items have a unique majority and 23 are unresolved 1–1–1 ties, which are excluded from the primary four-way meta-evaluation. No model output is used to break these human ties.

C.4 Inter-Annotator Agreement

We report IAA on the raw/pre-QA labels, before model-assisted review. Confidence intervals are 95% percentile intervals from 5,000 item-level bootstrap resamples (seed 20260501). Because the explicit CS/CA label is the binary construct used in downstream stratification, it is the primary reliability result.

Table 14 shows generally high explicit CS/CA agreement, but also illustrates why multiple agreement measures are useful. Danish has the clearest prevalence effect: 277 of 300 raw ratings are CS.

L	3/3 agree [95% CI]	Pairwise [95% CI]
DA	.84 [.77,.91]	.89 [.85,.94]
ET	.91 [.85,.96]	.94 [.90,.97]
LT	.80 [.72,.88]	.87 [.81,.92]
NB	.65 [.56,.74]	.77 [.71,.83]
SV	.83 [.76,.90]	.89 [.84,.93]

L	Fleiss κ [95% CI]	Gwet AC1 [95% CI]
DA	.25 [−.05,.51]	.88 [.81,.93]
ET	.83 [.71,.92]	.91 [.84,.96]
LT	.72 [.59,.82]	.75 [.65,.85]
NB	.44 [.28,.57]	.60 [.47,.72]
SV	.77 [.66,.86]	.78 [.68,.87]

Table 14: Raw/pre-QA inter-annotator agreement for the independently annotated explicit CS/CA field. Each language has three annotators and 100 items. Pairwise is mean pairwise agreement over the three annotator pairs. Brackets show 95% percentile bootstrap CIs from 5,000 item resamples.

Consequently, Fleiss' $\kappa = .25$ has a wide CI crossing zero even though unanimous agreement is 84%, mean pairwise agreement is .89, and prevalence-robust Gwet's AC1 is .88.

Required knowledge is more difficult than the binary CS/CA decision because it permits disagreement among four categories. Table 15 reports the corresponding consensus structure.

The largest ambiguity occurs in LT and NB, where unanimous agreement is 36% and 37%, respectively, and LT contributes 13 of the 23 unresolved ties. This motivates reporting both the resolved-majority accuracy ($n = 477$) and a conservative lower bound that counts all ties as errors.

Relationship to the model-assisted QA workflow. A later GPT-5.2-assisted review flagged schema consistency, missing fields, metadata constraints, and potential content-level issues for annotator review. That workflow remains part of the annotation provenance and is useful for producing cleaned release annotations, but it is *not* the source of the reported IAA or Native-500 majority gold. This separation prevents the QA model from mechanically increasing its own measured agreement with the human reference.

C.5 Meta-Evaluation Details

This section provides the full split statistics, confusion matrices, construction checks, and per-

L	3/3	2/1	1/1/1	Unique
DA	63	32	5	95
ET	81	19	0	100
LT	36	51	13	87
NB	37	59	4	96
SV	52	47	1	99
All	**269**	**208**	**23**	**477**

L	Fleiss κ [95% CI]
DA	.20 [.06,.34]
ET	.67 [.58,.76]
LT	.33 [.23,.42]
NB	.37 [.26,.47]
SV	.48 [.38,.56]

Table 15: Raw/pre-QA agreement structure for four-way `required_knowledge`. A 1/1/1 row is an unresolved three-way tie among the three annotators. These 23 items are excluded from the primary four-way meta-evaluation.

language error breakdowns summarized in §4.3.

C.5.1 GlobalMMLU-Lite Label Aggregation and Split

GlobalMMLU-Lite provides annotations from 3–8 annotators per item (median: 4; 69.2% of items have exactly 4 annotators). We aggregate `required_knowledge` via majority vote. For ties, we distinguish two cases:

(a) simple ties between *none* and exactly one non-none label are resolved in favor of the non-none label;

(b) complex ties involving multiple non-none labels are removed.

Of the 400 items, 144 (36.0%) are unanimous, 185 (46.3%) have a clear majority, 62 (15.5%) are simple ties resolved via rule (a), and 9 (2.3%) are complex ties removed via rule (b), yielding 391 items with resolved human labels. The average majority share across all items is 77.0%.

The provided `cultural_sensitivity_label` in GlobalMMLU follows a strict majority rule over the raw annotator votes: CA if strictly more than half of the annotators assigned `required_knowledge` = *none*, CS otherwise. We verified this rule against the provided labels (100% agreement on 391 items) and use these provided CS/CA labels directly in all GlobalMMLU-Lite evaluations.

Split	n	*none*	*cult.*	*reg.*	*dial.*	CA	CS
Dev	174	103	15	55	1	90	84
Test	217	123	33	59	2	110	107
Total	391	226	48	114	3	200	191

Table 16: GlobalMMLU-Lite dev/test split after human label aggregation and tie resolution. Columns show `required_knowledge` counts after majority-vote aggregation and provided CS/CA labels.

From the 391 resolved items, we create a stratified split (∼30% dev, seed = 13), yielding 174 dev and 217 test items. The dev split is used for iterative prompt refinement; the test split is held out for final evaluation. Table 16 reports the resulting label distribution.

C.5.2 Prompt Calibration Error Patterns

Iterative refinement on the dev split identified five recurring misclassification patterns, each addressed through targeted prompt modifications:

(A) under-detection of Western philosophical terminology as cultural anchors (e.g., consequentialism, virtue ethics);

(B) under-detection of implicit US-centric regional cues such as imperial units, dollar references, and place-in-scenario setups;

(C) under-detection of region-specific idioms as dialect cues (e.g., “Mom & Pop”, “blue-collar”);

(D) over-triggering of cultural labels on generic academic citations (“According to Author & Author (Year)”);

and (E) over-labeling of regional for incidental country mentions in historical passages. All prompt modifications are motivated exclusively by dev-split findings.

The test split was never inspected during development.

C.5.3 GlobalMMLU-Lite Test Evaluation

Table 17 shows the four-way `required_knowledge` confusion matrix on the held-out GlobalMMLU-Lite test split. The main errors are *regional*→*none* and *none*→*cultural*, reflecting the difficulty of detecting implicit regional and cultural cues in English-origin items.

C.5.4 Native-500 as External Meta-Evaluation

The Native-500 sample was not used to tune the annotation prompt. The item-level GPT annotations

H↓\G→	*cult.*	*dial.*	*none*	*reg.*	Σ
cultural	**25**	0	8	0	33
dialect	0	**1**	1	0	2
none	14	0	**102**	7	123
regional	3	0	23	**33**	59
Σ	42	1	134	40	217

Table 17: Required-knowledge confusion matrix on the GlobalMMLU-Lite test split ($n = 217$). H=human, G=GPT. Overall accuracy: 74.2%.

used for the native benchmarks were produced before the human-majority meta-evaluation and are joined to the reconstructed human reference only afterward. Thus Native-500 is held out with respect to GlobalMMLU-Lite prompt calibration, while sharing the same annotation schema.

Tables 18 and 19 report the two Native-500 confusion matrices. The CS/CA matrix uses the explicit three-annotator majority for all 500 items. The required-knowledge matrix uses the 477 items with a unique three-annotator majority.

Human ↓ / GPT →	CS	CA	Σ
CS	**307**	12	319
CA	35	**146**	181
Σ	342	158	500

Table 18: Explicit CS/CA confusion matrix on Native-500. Human gold is the majority of the three independently annotated explicit CS/CA labels. GPT uses its explicit `cultural_sensitivity_label`. Accuracy is 90.6%; macro-F1 is 89.5%.

H↓\G→	*cult.*	*dial.*	*none*	*reg.*	Σ
cultural	**66**	1	3	44	114
dialect	0	**3**	0	0	3
none	10	0	**166**	7	183
regional	8	0	4	**165**	177
Σ	84	4	173	216	477

Table 19: Required-knowledge confusion matrix on the 477 Native-500 items with a unique three-annotator majority. Overall accuracy: 83.9%; macro-F1: 82.4%. Counting all 23 unresolved human ties as errors gives a conservative lower bound of 400/500=80.0%.

The dominant native four-way error is *cultural→regional* (44 items), reflecting the fuzzy boundary between cultural and geographically anchored local knowledge. Because explicit CS/CA is independently annotated, we do not collapse these four-way errors into the primary binary gold construction.

C.5.5 CS/CA Construction Sensitivity

The explicit human CS/CA field and a CS/CA label mechanically derived from majority `required_knowledge` differ on 10/500 items (2%). This confirms that the two human fields are correlated but non-redundant. Table 20 therefore distinguishes the primary CS/CA evaluation from two sensitivity constructions.

Construction	Acc.	Role
Human explicit vs. GPT explicit	90.6%	primary; downstream-aligned
Human explicit vs. GPT RK-derived	92.8%	sensitivity
Human RK-derived vs. GPT RK-derived	94.4%	legacy diagnostic

Table 20: Native-500 CS/CA meta-evaluation under different constructions. "RK-derived" maps `none`→CA and non-`none`→CS. The primary result compares the independently annotated explicit human field with GPT's explicit field.

C.5.6 Native Error Distribution by Language

Table 21 reports explicit-CS/CA mismatches by language. Errors are asymmetric: the model more often predicts CS for human-CA items than the reverse, especially in Norwegian Bokmål and Estonian.

Lang.	Acc.	Errors	CS→CA	CA→CS
DA	92%	8	5	3
ET	90%	10	0	10
LT	92%	8	2	6
NB	86%	14	2	12
SV	93%	7	3	4

Table 21: Explicit CS/CA performance by language on Native-500. CS→CA denotes a human-CS/GPT-CA miss. CA→CS denotes a human-CA/GPT-CS false alarm. NB is the weakest language-level case (86% accuracy; 12/14 errors are CA→CS).

Table 22 reports four-way required-knowledge accuracy by language, using only items with a unique three-annotator majority. The table also shows how many items per language are excluded as unresolved ties.

L	Resolved n	Correct	Accuracy	Ties
DA	95	83	87.4%	5
ET	100	96	96.0%	0
LT	87	77	88.5%	13
NB	96	72	75.0%	4
SV	99	72	72.7%	1
All	477	400	83.9%	23

Table 22: Four-way `required_knowledge` accuracy by language using only items with a unique three-annotator majority.

The required-knowledge results are strongest for Estonian and weakest for Swedish and Norwegian Bokmål. These differences support the interpretation in the main text that fine-grained knowledge-type labels are noisier than binary CS/CA labels.

D Proxy Validity: Supplementary Material

D.1 Model Overview

Table 23 lists all 29 models evaluated in our study, grouped by size category.

D.2 Data Foundation

Table 24 summarizes the item-level data used for the ranking and incremental-validity analyses. After merging benchmark outputs, CS/CA annotations, and per-model LM Evaluation Harness results, the final item-level analysis file contains 886,762 rows across 29 models.[3] This total decomposes as follows: G-MMLUX (16,701 items across 6 languages), native benchmarks excluding SwedishFacts (7,185 items), SwedishFacts (1,292 accuracy rows), and Belebele (5,400 items across 5 languages), giving $30{,}578 \times 29 = 886{,}762$ rows. These balanced subsets are used only for the ranking analyses. The main incremental-validity regressions use total benchmark scores rather than balanced CA/CS subsets.

Evaluation infrastructure. All evaluations use our custom fork of the LM Evaluation Harness[4] (Sutawika et al., 2025). Inference runs on NVIDIA H100 GPUs (94 GiB HBM2e). Models in categories S, XS, and XXS use a single GPU, while M-category models use 4-way tensor parallelism.

Balanced CA/CS subsets. For the ranking analysis in §5.2, we construct balanced CA/CS subsets: within each benchmark group (MMLU-only and non-MMLU native) and language, we sample $\min(n_{\text{CA}}, n_{\text{CS}})$ items per class to ensure that accuracy differences between CA and CS rankings are not driven by sample size imbalance.

D.3 Block A: CA vs. CS Ranking Consistency (Details)

Table 25 provides the full Block A results. The Danish native CA/CS comparison is shown for completeness but not interpreted in the main text because only 31 CA items are available.

[3]The merged analysis data and data-processing pipeline are available at https://github.com/gold-vs-translation/data-processing.

[4]LM Evaluation Harness: https://github.com/EleutherAI/lm-evaluation-harness; our fork with native benchmark task implementations: https://github.com/gold-vs-translation/lm-evaluation-harness.

Cat.	Model	B	BA
M (≥70B)			
	c4ai-command-a-03-2025-111B	111	111
	Apertus-70B-Instruct-2509	70	70
	Llama-4-Scout-17B-16E	109	17
	Mistral-Large-2411	123	123
	Llama-3.1-70B-Instruct	70	70
	Qwen2.5-72B-Instruct	72	72
	Xwen-72B-Chat	72	72
S (15–32B)			
	aya-expanse-32b	32	32
	Phi-4-15B-Instruct	15	15
	gpt-oss-20b	20	3.6
	Nemotron-3-Nano-30B-A3B	30	3
	Qwen3-30B-A3B-2507	30	3
	EuroLLM-22B-Instruct-Prev.	22	22
	Mistral-Small-24B-2501	24	24
	gemma-3-27b-it	27	27
XS (7–9B)			
	salamandra-7b-instruct	7	7
	Teuken-7B-instruct-v0.6	7	7
	Llama-3.1-8B-Instruct	8	8
	Qwen3-8B	8	8
	Falcon3-7B-Instruct	7	7
	EuroLLM-9B-Instruct	9	9
	aya-expanse-8b	8	8
XXS (≤1.7B)			
	SmolLM2-360M-Instruct	0.4	0.4
	Qwen3-0.6B	0.6	0.6
	LFM2-700M	0.7	0.7
	gemma-3-1b-it	1	1
	Falcon3-1B-Instruct	1.7	1.7
	Llama-3.2-1B-Instruct	1.2	1.2
	bloomz-560m	0.6	0.6

Table 23: Models evaluated in our study. **B**: total parameters (billions). **BA**: active parameters (billions; differs from B for mixture-of-experts models). All evaluated checkpoints are instruction-tuned. Model labels are shortened, with exact identifiers in the released harness configuration. Regressions use B (total parameters). Inference uses NVIDIA H100 GPUs (94 GiB HBM2e). M-category models use 4-way tensor parallelism, others use a single GPU.

Group	Dataset	Lang.	Items	CA	CS	Balanced	Used in
Translated proxy benchmarks (MMLU)							
MMLU-only	G-MMLUX	DA	2,800	2,059	741	741+741	A, B
MMLU-only	G-MMLUX	EN	2,800	2,059	741	741+741	A
MMLU-only	G-MMLUX	ET	2,800	2,059	741	741+741	A, B
MMLU-only	G-MMLUX	LT	2,800	2,059	741	741+741	A, B
MMLU-only	G-MMLUX	NB	2,701	2,008	693	693+693	A, B
MMLU-only	G-MMLUX	SV	2,800	2,059	741	741+741	A, B
Native benchmarks							
non-MMLU	DanishCitizenTest	DA	668	31	637	31+31†	B
non-MMLU	Exam-ET	ET	1,614	886	728	765+765	A, B
non-MMLU	INCLUDE-base	ET	224	187	37		A, B
non-MMLU	INCLUDE-base	LT	534	164	370	164+164	A, B
non-MMLU	NRK-Quiz-QA	NB	3,600	512	3,088	512+512	A, B
non-MMLU	Skolprov	SV	545	467	78	495+495	A, B
non-MMLU	SwedishFacts	SV	1,292	28	1,264		A, B
Language proficiency							
—	Belebele	5 langs	5,400	(not annotated)		—	Incr.

Table 24: Data foundation for the proxy validity analysis. **Items**: unique annotated items. **CA/CS**: culturally agnostic/sensitive item counts. **Balanced**: items per class in the balanced CA/CS subsets used for Block A ($\min(n_{\text{CA}}, n_{\text{CS}})$ per group $\times$ language). **Used in**: A = Block A (CA vs. CS ranking), B = Block B (native vs. proxy ranking), Incr. = incremental proxy validity (§5.3). †DA non-MMLU excluded from Block A interpretation due to insufficient CA items ($n = 31$). For ET and SV, multiple native datasets are pooled before balancing.

Group	L	n_{CA}	n_{CS}	τ_b	[95% CI]	p
MMLU-only						
	DA	741	741	0.91	[.82, .97]	$<10^{-11}$
	EN	741	741	0.91	[.85, .97]	$<10^{-11}$
	ET	741	741	0.91	[.84, .97]	$<10^{-11}$
	LT	741	741	0.94	[.86, .98]	$<10^{-12}$
	NB	693	693	0.91	[.82, .97]	$<10^{-11}$
	SV	741	741	0.91	[.82, .97]	$<10^{-11}$
non-MMLU (native benchmarks)						
	DA†	31	31	0.66	[.48, .80]	$<10^{-5}$
	ET	765	765	0.87	[.77, .94]	$<10^{-10}$
	LT	164	164	0.79	[.66, .90]	$<10^{-8}$
	NB	512	512	0.88	[.79, .96]	$<10^{-10}$
	SV	495	495	0.63	[.43, .77]	$<10^{-5}$

Table 25: Block A: CA vs. CS ranking consistency (Kendall's τ_b) within MMLU-only and non-MMLU native benchmarks, using balanced CA/CS subsets. †DA non-MMLU: excluded from main-text interpretation due to insufficient CA items ($n_{CA} = 31$; ≈1 item per model).

Methodological note: balanced subsets. To ensure that CA vs. CS ranking differences are not confounded by unequal sample sizes, we construct balanced subsets by sampling $\min(n_{CA}, n_{CS})$ items per class within each benchmark group and language. For MMLU-only, this yields 741 items per class per language, except for Norwegian Bokmål, where 693 items per class are available because fewer translations are present. For native benchmarks, the balancing is constrained by the minority class: ET and SV pool multiple datasets before balancing; DA is limited by only 31 CA items in DanishCitizenTest. Sampling is reproducible via a fixed random seed ($s = 42$).

Interpretation of the SV non-MMLU result ($\tau_b = 0.63$). The low Swedish consistency is a compositional effect rather than a substantive finding about cultural sensitivity. Skolprov contributes 467 of 495 CA items (86% CA, primarily STEM and medical), while SwedishFacts contributes 1,264 of 1,342 CS items (98% CS, primarily cultural trivia). The CA vs. CS comparison therefore effectively contrasts two distinct benchmarks with different knowledge requirements, rather than two facets of the same construct. This is reflected in the Block B results (§5.2), where Skolprov alone achieves $\tau_b = 0.89$ against its MMLU proxy while SwedishFacts achieves only $\tau_b = 0.51$.

D.4 Block B: Native vs. Proxy Ranking (Details)

Table 26 extends Table 5 with the full set of statistics including item counts and top-3 overlap.

Note on the DanishCT CA result ($\tau_{CA} = 0.66$, T3 = 3/3, T5 = 5/5). The seemingly contradictory pattern—low τ_b but perfect top-k overlap—arises because only 31 CA items are available in DanishCitizenTest, yielding noisy accuracy estimates. The top models are sufficiently strong to maintain their relative order despite this noise, but mid-range and lower-ranked models swap positions frequently, pulling down the global correlation.

Methodology: Kendall's τ_b. We use Kendall's τ_b (the variant correcting for ties) rather than Spearman's ρ for two reasons: (1) with $n = 29$ models, Kendall's τ_b is less sensitive to large rank displacements because it counts pairwise concordances rather than squared rank differences; and (2) τ_b explicitly corrects for tied ranks, which are common when models achieve identical accuracy on small item sets.

Given $n = 29$ and the presence of tied ranks, we interpret the reported p-values together with the bootstrap CIs rather than as standalone evidence. Bootstrap CIs use 2,000 resamples over models; each resample draws 29 models with replacement and recomputes τ_b.

Methodology: top-k overlap. Top-k overlap counts the number of models appearing in the top-k positions of *both* the native and proxy rankings (using dense ranks, where ties share the same rank). We report $k \in \{3, 5\}$ as these correspond to the practical question of whether the "best few" models are the same regardless of which benchmark is used.

D.5 CS/CA-Stratified Incremental Proxy Validity

Table 28 reports the CS/CA-stratified incremental proxy validity results referenced in §5.3.

D.6 Incremental Proxy Validity (Details)

The main analysis fits nested OLS models on the 29 evaluated models. For each pairing, the baseline model uses log(total parameters) and same-language Belebele. The extended model adds the translated MMLU proxy score. We quantify uncertainty by resampling the 29 models with replacement 5,000 times and refitting both models in every

Native	L	Sub.	n_{nat}	n_{prx}	τ_b	[95% CI]	p	T3	T5
Skolprov	SV	ALL	545	2800	.89	[.78, .96]	$<10^{-11}$	1/3	5/5
		CA	467	2059	.88	[.78, .95]	$<10^{-10}$	2/3	4/5
		CS	78	741	.77	[.66, .86]	$<10^{-8}$	1/3	4/5
Exam-ET	ET	ALL	1614	2800	.88	[.78, .96]	$<10^{-10}$	2/3	5/5
		CA	886	2059	.83	[.72, .91]	$<10^{-9}$	1/3	4/5
		CS	728	741	.88	[.78, .95]	$<10^{-10}$	2/3	5/5
INCLUDE	LT	ALL	534	2800	.87	[.79, .93]	$<10^{-10}$	1/3	5/5
		CA	164	2059	.82	[.73, .90]	$<10^{-9}$	2/3	5/5
		CS	370	741	.85	[.74, .93]	$<10^{-10}$	2/3	5/5
NRK-Quiz	NB	ALL	3600	2701	.85	[.76, .92]	$<10^{-10}$	1/3	4/5
		CA	512	2008	.82	[.73, .90]	$<10^{-9}$	1/3	3/5
		CS	3088	693	.88	[.79, .94]	$<10^{-10}$	2/3	5/5
DanishCT	DA	ALL	668	2800	.82	[.73, .90]	$<10^{-13}$	1/3	3/5
		CA	31	2059	.66	[.53, .79]	$<10^{-5}$	3/3	5/5
		CS	637	741	.84	[.75, .90]	$<10^{-9}$	1/3	4/5
INCLUDE	ET	ALL	224	2800	.72	[.58, .83]	$<10^{-7}$	2/3	3/5
		CA	187	2059	.68	[.54, .80]	$<10^{-6}$	1/3	3/5
		CS	37	741	.72	[.54, .85]	$<10^{-7}$	0/3	4/5
SwedishFacts	SV	ALL	1292	2800	.51	[.25, .74]	$<10^{-3}$	1/3	3/5
		CA	28	2059	.50	[.23, .71]	$<10^{-3}$	1/3	3/5
		CS	1264	741	.52	[.26, .74]	$<10^{-4}$	1/3	2/5

Table 26: Block B: full native-vs-proxy ranking consistency. **Sub.**: subset (ALL, CA-only, CS-only). n_{nat}/n_{prx}: number of items on native/proxy side. **T3/T5**: top-3/top-5 model overlap. The proxy is always G-MMLUX in the corresponding language. All comparisons use 29 models.

resample. We report 95% percentile bootstrap CIs for R^2, ΔR^2, and coefficients. The final reported outputs are reproduced with bootstrap seed 42; this seed is distinct from the Native-500 IAA bootstrap seed reported in Appendix C.4.

A pairing is classified as *robust* if (i) the full-sample nested F-test is significant, (ii) the 95% model-bootstrap CI for β_{proxy} excludes zero, and (iii) β_{proxy} is positive with nominal $p < .05$ in all 17 LOFO refits. A *marginal* pairing is full-sample significant but fails at least one robustness criterion. A *null* pairing has a non-significant full-sample F-test and a bootstrap β_{proxy} CI containing zero. The labels describe evidence strength, not benchmark quality. Because in-sample ΔR^2 is non-negative by construction, its percentile-bootstrap CI cannot serve as a null test; near-zero lower bounds describe uncertainty in effect magnitude but do not independently determine the class.

The CS/CA-stratified regressions in Table 28 are exploratory secondary analyses and do not determine the robust/marginal/null class. For Exam-ET, the small $\Delta R^2 = .024$ occurs against a high baseline ($R^2 = .954$), so the translated proxy adds a modest but robust increment after controls. For SwedishFacts, the low baseline ($R^2 = .353$) indicates that model size and Belebele explain substantially less variance than for the exam-style pairings; the data do not by themselves identify this residual variance as a uniquely local construct.

OLS assumptions and collinearity. With $n = 29$ and three predictors in the extended main model, the regression has limited degrees of freedom, motivating the model-level bootstrap and LOFO analyses. Variance Inflation Factors indicate substantial overlap among model strength, language proficiency, and proxy performance; this inflates coefficient uncertainty and reinforces interpreting β_{proxy} as approximate within the evaluated model set. The nested F-test and ΔR^2 quantify whether the added proxy improves the fitted model, while the bootstrap and LOFO analyses assess how stable that conclu-

Native	L	R^2_{base}	R^2_{ext}	ΔR^2 [95% CI]	β_{prx} [95% CI]	p_F	LOFO	Class
Skolprov	SV	.856	.984	.128 [.038,.253]	1.056 [.798,1.187]	$< 10^{-13}$	17/17	robust
INCLUDE	LT	.890	.965	.076 [.021,.148]	0.938 [.667,1.126]	$< 10^{-7}$	17/17	robust
NRK-Quiz	NB	.934	.975	.041 [.009,.087]	0.463 [.289,.592]	$< 10^{-6}$	17/17	robust
Exam-ET	ET	.954	.979	.024 [.010,.044]	0.515 [.340,.852]	$< 10^{-5}$	17/17	robust
INCLUDE	ET	.818	.877	.059 [.000,.199]	0.842 [−.021,1.300]	.002	16/17	marginal
DanishCT	DA	.951	.960	.009 [.000,.027]	0.309 [.006,.575]	.027	14/17	marginal
SwedishFacts	SV	.353	.388	.035 [.000,.163]	0.140 [−.096,.329]	.242	0/17	null

Table 27: Full all-item incremental proxy-validity results with model-bootstrap uncertainty. The baseline model predicts native performance from log(total parameters) and Belebele. The extended model adds the same-language translated MMLU proxy. **LOFO** reports how many of the 17 leave-one-family-out refits retain positive β_{proxy} with nominal $p < .05$. The evidence class follows the fixed definition in the main text.

		CA items		CS items	
Native	L	ΔR^2	p	ΔR^2	p
Skolprov	SV	.140	$<10^{-13}$	.060	$<10^{-3}$
INCLUDE	LT	.095	$<10^{-5}$	.071	$<10^{-7}$
INCLUDE	ET	.059	.003	.046	.010
NRK-Quiz	NB	.048	$<10^{-4}$	.044	$<10^{-7}$
Exam-ET	ET	.031	$<10^{-3}$	.015	$<10^{-4}$
DanishCT	DA	.005	.333	.014	.005
SwedishFacts	SV	.007	.573	.046	.180

Table 28: CS/CA-stratified incremental proxy validity. ΔR^2 and p from the nested F-test, computed separately on CA and CS items of each native benchmark. Proxy scores use the corresponding CA/CS split of G-MMLUX.

sion is to model sampling. In particular, because in-sample $\Delta R^2 \geq 0$ by construction, a percentile-bootstrap interval whose lower endpoint rounds to .000 (or is slightly above zero) should not be interpreted as a separate rejection of the null.

D.7 Robustness (Details)

Partial correlations. Table 29 reports the partial correlation between native and proxy performance after residualizing both on the main controls (log(total parameters)+ Belebele). The partial-correlation pattern mirrors the main incremental results: robust pairings retain strong residual native–proxy associations, whereas SwedishFacts remains weak and non-significant after the same controls.

Leave-one-family-out sensitivity. LOFO removes every model family in turn, including singleton families, and refits the ALL-items main model. There are 17 excluded families and

Native	L	Sub.	r_{partial}	p
Skolprov	SV	ALL	.943	$< 10^{-14}$
		CA	.945	$< 10^{-14}$
		CS	.670	$< 10^{-4}$
INCLUDE	LT	ALL	.828	$< 10^{-7}$
		CA	.747	$< 10^{-5}$
		CS	.825	$< 10^{-7}$
NRK-Quiz	NB	ALL	.785	$< 10^{-6}$
		CA	.716	$< 10^{-4}$
		CS	.823	$< 10^{-7}$
Exam-ET	ET	ALL	.730	$< 10^{-5}$
		CA	.661	$< 10^{-4}$
		CS	.718	$< 10^{-5}$
INCLUDE	ET	ALL	.570	.001
		CA	.545	.002
		CS	.485	.008
DanishCT	DA	ALL	.426	.021
		CA	−.194	.314
		CS	.520	.004
SwedishFacts	SV	ALL	.233	.223
		CA	.113	.558
		CS	.266	.163

Table 29: Partial correlations between native and proxy scores after residualizing on the main controls. $n = 29$ for all entries.

Native	L	R^2_{base}	R^2_{ext}	ΔR^2 [95% CI]	β_{prx} [95% CI]	p_F
Skolprov	SV	.882	.987	.105 [.018,.224]	1.166 [.883,1.285]	$< 10^{-12}$
INCLUDE	LT	.890	.977	.087 [.024,.160]	1.082 [.843,1.298]	$< 10^{-9}$
INCLUDE	ET	.846	.934	.087 [.006,.233]	1.057 [.365,1.418]	$< 10^{-5}$
NRK-Quiz	NB	.940	.977	.037 [.006,.080]	0.537 [.342,.679]	$< 10^{-6}$
Exam-ET	ET	.956	.986	.030 [.013,.051]	0.587 [.393,.947]	$< 10^{-7}$
DanishCT	DA	.952	.971	.020 [.003,.045]	0.572 [.340,.916]	$< 10^{-3}$
SwedishFacts	SV	.356	.430	.073 [.004,.187]	0.247 [.059,.476]	.092

Table 30: Sensitivity analysis with English G-MMLUX as an additional control. Brackets show 95% percentile model-bootstrap CIs. This specification is a robustness check against general MMLU-content familiarity and is not used to define the main robust/marginal/null evidence classes.

119 total refits (7 pairings $\times$ 17 families); each refit retains 25–28 of the 29 models. Table 31 summarizes all 17 exclusions per pairing.

Pairing	Sig.	Drops	Class
Skolprov	17/17	none	robust
INCLUDE-LT	17/17	none	robust
NRK-Quiz-QA	17/17	none	robust
Exam-ET	17/17	none	robust
INCLUDE-ET	16/17	Qwen (.102) Cohere (.110);	marginal
DanishCT	14/17	Llama (.072); Qwen (.096)	marginal
SwedishFacts	0/17	all 17	null

Table 31: LOFO stability across all 17 model-family exclusions. **Sig.** reports how many refits retain a positive β_{proxy} with nominal $p < .05$. **Drops** lists the family exclusions for which this criterion is not met. These nominal tests are not multiplicity-corrected and are used only as a conservative stability requirement for the *robust* label.

The four robust pairings preserve their direction and nominal significance under every family exclusion. DanishCT and INCLUDE-ET remain positive in the full sample but are sensitive to changes in the evaluated model set. SwedishFacts remains non-significant throughout, supporting the stability of the null classification within this model sample.

D.8 Sensitivity: English MMLU Control (Details)

The sensitivity analysis adds English G-MMLUX performance as a third baseline control. This checks whether the translated-proxy association mainly reflects general familiarity with MMLU content. Let C_L denote the three controls log(total parameters), Belebele$_L$, and GMMLUX-EN. We compare native$_L \sim C_L$ with native$_L \sim C_L + \text{proxy}_L$. As in the main analysis, confidence intervals use 5,000 model-level bootstrap resamples with seed 42.

Six pairings remain significant by the nested F-test. SwedishFacts remains non-significant by the F-test ($p = .092$), but the bootstrap CI for β_{prx} is positive ([.059,.476]). We therefore treat SwedishFacts as weak mixed positive evidence in this alternative specification, not as a clean null. This does not change its *null* classification under the fixed main-specification evidence definition.

The sensitivity analysis preserves the broad ordering after controlling for English MMLU performance. DanishCT shows a stronger signal in this specification, but its main evidence class remains *marginal* because the primary classification is based on the main specification and LOFO/bootstrap criteria. The sensitivity should not be read as identifying a purely language-specific causal effect, since English MMLU, Belebele, and the translated proxy are correlated measures of overlapping model abilities.