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

Anonymous ACL submission

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? For curriculum-overlap benchmarks (Skolprov, Exam-ET), the proxy is strongly valid ($\tau_b \geq 0.88$; up to $\Delta R^2 = .128$ beyond controls, $p < 10^{-13}$). For culturally loaded benchmarks (DanishCitizenTest, NRK-Quiz-QA), rankings remain moderately strong but incremental validity is smaller. For Sweden-specific factual knowledge (SwedishFacts), the proxy provides no significant incremental value ($p = .24$), consistent with a locality gradient that is already visible within MC-format benchmarks. Our results indicate when translated benchmarks approximate native evaluation signal and when native benchmarks remain necessary.

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 MT a good enough substitute for HT (Choenni et al., 2024; 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 benchmarks in seven pairings across five Nordic and Baltic languages, 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 analyze mechanisms via culturally sensitive (CS) vs. agnostic (CA) stratification.

Our analysis reveals that proxy validity varies: ranking agreement ranges from $\tau_b = 0.89$ (curriculum-overlap benchmarks) to 0.51 (Sweden-specific factual knowledge). For six of seven pairings, the proxy significantly improves prediction beyond controls, with the strongest effect for Skolprov ($\Delta R^2 = .128$, $p < 10^{-13}$). The sole excep-

tion is SwedishFacts ($p = .24$), suggesting a dimension not well captured by translated multiple-choice benchmarks, potentially due to its Sweden-centric content and/or its open-ended generation format. A sensitivity analysis adding English MMLU performance as an additional control confirms that these effects reflect language-specific proxy signal, not general MMLU-content familiarity.

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 cultural sensitivity annotation pipeline validated across five languages with 500 human-annotated items;

(3) empirical evidence across seven native-vs-proxy pairings, offering practical guidance on when translated benchmarks approximate native evaluation signal and when native benchmarks remain necessary.

2 Related Work

Translation methods. Studies comparing MT and HT benchmarks find that MT approximates HT with high rank correlations on average, but with systematic biases for low-resource languages (Choenni et al., 2024; 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 Global-MMLU; Keleg and Magdy (2023) demonstrate systematic LLM underperformance on non-Western knowledge; 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 (Appendix A).

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 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).

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-QA	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), Dan-

ish (DA), Norwegian Bokmål (NB), and Lithuanian (LT). Table 1 summarises all datasets used in our study, organised into *translated global benchmarks* derived from MMLU and *native local benchmarks* created independently in the target language.

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, 2024) 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 Global-MMLU (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).

Native benchmarks. We include seven natively created benchmarks spanning five languages: **Exam_ET** (Lillepalu and Alumäe, 2025) contains 1,614 MC questions from official Estonian secondary and high-school exams across eight subjects; **Skolprov** (Kurfalı et al., 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 – the only non-MC dataset in our study; **DanishCitizenTest** (Nielsen, 2024, 2023) consists of 668 MC items from the official Danish citizenship and residence permit exams; **NRK-Quiz-QA** (Mikhailov et al., 2025) provides 3,600 MC quiz questions from NRK, Norway's public broadcaster, covering Norwegian culture, geography, and current affairs; and **INCLUDE** (Romanou et al., 2025), from which we use the Estonian (224 items) and Lithuanian (534 items) subsets of natively sourced exam questions. All native benchmarks are strongly locally anchored and free of translation artefacts by construction. The proportion of culturally sensitive (CS) items ranges from 14.3% (Skolprov) to 97.8% (SwedishFacts), compared to 27.8% in Global-MMLU (Table 12).

Subject Mapping and Comparison Pairings. 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).

Three types of mismatches—granularity, scope, and missing equivalents—make direct subject-level alignment unreliable (see Appendix B.2 for the full mapping and detailed examples). We therefore conduct all cross-benchmark analyses at **category level**, which provides a more robust basis for comparison. Cross-validation of our manual category assignments against independent GPT-5.2 annotations confirms this choice, with 85.1% agreement across 3,044 items at category level; disagreements trace to inherent MMLU taxonomy ambiguities rather than mapping errors (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-QA, 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) vs. culturally agnostic (CA) labels for all benchmark items—a distinction 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) human reference annotations on a stratified sample of 500 native benchmark items across five languages, and (2) an automatic annotation pipeline using GPT-5.2, calibrated on GlobalMMLU-Lite and validated against the human references.

4.1 Human Annotation

We adapted the GlobalMMLU cultural sensitivity schema (Singh et al., 2025) for our native benchmark items. For each item, annotators assign a primary `required_knowledge` type (*none*, *cultural*, *regional*, or *dialect*), a binary `time_sensitive` flag, a `cultural_sensitivity_label` (CS/CA),

a `subject_category`, and optional metadata including trigger spans and geographic/cultural identifiers. CS is assigned whenever the item requires culture-specific, regional, or dialect knowledge; CA is assigned only when the item is free of such cues (full schema in Appendix C.1).

		Required Knowledge				CS/CA	
Lang	**n**	`none`	`reg.`	`cult.`	`dial.`	CS	CA
DA	100	3	77	20	0	97	3
ET	100	77	9	14	0	20	80
LT	100	43	50	6	1	55	45
NB	100	22	33	43	2	78	22
SV	100	45	40	15	0	55	45
All	500	190	209	98	3	305	195

Table 2: Label distribution of 500 human-annotated native benchmark items (100 per language, stratified sample). **Required Knowledge**: knowledge type needed to answer correctly. **CS/CA**: culturally sensitive vs. culturally agnostic.

For each language, we drew a stratified sample of 100 items across the native benchmarks described in Table 1, yielding 500 items total. Seven professional annotators were recruited via ProZ.com based on native-speaker status and professional language expertise: three independent annotators for Danish and one each for Estonian, Lithuanian, Norwegian Bokmål, and Swedish.

For Danish—the only language with multiple annotators—inter-annotator agreement on CS/CA reaches 84% raw agreement across 75 common items, though chance-corrected κ is low (0.083) due to the extreme CS prevalence in the Danish data (only 3 of 100 items are CA; Appendix C.4).

Annotation followed a two-round protocol: an initial calibration round with individualized feedback on the first 10 items, followed by full-batch annotation and an automated consistency review using GPT-5.2 to flag schema-level errors (Appendix C.3).

Table 2 reports the resulting label distributions. The data reveal substantial cross-lingual variation mirroring benchmark composition: Danish items are predominantly regional (77%), Estonian overwhelmingly CA (80%), and Norwegian shows high cultural load (43%). The `dialect` category is used very rarely (3 items total).

4.2 Automatic Annotation Pipeline

To scale CS/CA labeling to the full native benchmark collection (Table 1), we developed an automatic annotation pipeline using GPT-5.2 with structured JSON output. The pipeline assigns the same GlobalMMLU-style labels as the human annotators: `required_knowledge`, `time_sensitive`, CS/CA, `subject_category`, and associated metadata. The full prompt is reproduced in Appendix C.2 (Figure 5).

We calibrated the prompt on GlobalMMLU-Lite, which provides 3–8 independent annotations per item (median: 4). We aggregated `required_knowledge` labels via majority vote, resolving simple ties between *none* and a single 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*). The GlobalMMLU-Lite CS/CA labeling follows the rule: CA if strictly more than half of annotators assigned *none*, CS otherwise—resulting in 191 CS and 200 CA items. From these, we created a stratified dev/test split (174/217 items). Iterative refinement on the dev split addressed five error patterns—including under-detection of Western philosophical terminology as cultural anchors and of implicit US-centric regional cues—before evaluating on the held-out test split. Further details on prompt calibration are provided in Appendix C.5.2.

4.3 Meta-Evaluation

We assess the reliability of the automatic pipeline in two independent evaluations: (1) on the GlobalMMLU-Lite test split (English, multi-annotator reference, $n = 217$), and (2) on the 500 human-annotated native benchmark items (5 languages, single-annotator reference).

CS/CA agreement. Table 3 (top) summarizes the binary CS/CA classification. On GlobalMMLU-Lite, the pipeline achieves 73.3% accuracy with an asymmetric error profile: CS recall (61.7%) lags behind CA recall (84.5%), indicating a tendency to under-predict cultural sensitivity for English items with implicit cultural cues.

On native benchmarks, accuracy rises to 92.0%, with high recall for both classes. The higher native agreement reflects that native exam items tend to carry unambiguous regional or cultural anchors, whereas GlobalMMLU-Lite contains more borderline cases.

Required knowledge classification. Table 4 (bottom) reports per-class precision, recall, and F1 for the four-way `required_knowledge` classification.

	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	92.0	91.0	96.4	93.8	85.1

Table 3: Binary CS/CA agreement between GPT-5.2 and human references (%). The higher native benchmark accuracy reflects more explicit cultural cues in exam-style items compared to the implicit anchors in GlobalMMLU-Lite.

On GlobalMMLU-Lite, *regional* recall is notably low (55.9%), driven by confusion with *none* – the model misses items where geographic cues are implicit.

On native benchmarks, all classes except *dialect* (too few items) reach F1 above 64%, with *none* and *regional* exceeding 83%. The dominant error on native benchmarks is confusion between *cultural* and *regional* (53 items), reflecting genuine ambiguity when items involve country-specific traditions that are simultaneously place-bound and culture-bound. Crucially, this confusion does not affect the binary CS/CA decision, since both labels map to CS.

	GlobalMMLU-Lite			Native (5 langs)		
	P	R	F1	P	R	F1
cultural	59.5	75.8	66.7	68.2	61.2	64.5
regional	82.5	55.9	66.7	79.2	87.6	83.2
none	76.1	82.9	79.4	94.4	87.9	91.0
dialect	100	50.0	66.7	50.0	66.7	57.1

Table 4: Per-class precision, recall, and F1 (%) for `required_knowledge` (GPT-5.2 vs. human reference).

5 Proxy Validity Evaluation

In this section we test whether translated MMLU proxies produce the same model rankings as native benchmarks (§5.2), whether they carry predictive value beyond model size and language proficiency (§5.3), and verify robustness via leave-one-family-out analysis and a sensitivity check controlling for English MMLU performance (§5.4).

5.1 Evaluation Setup

We evaluate 29 instruction-tuned LLMs spanning four size categories: M (≥70B, 7 models), S (15–32B, 8 models), XS (7–9B, 7 models), and XXS (≤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 (Gao et al., 2024) 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 downstream 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-k overlap (methodological details in Appendix D.4). As a preliminary check, we confirm that CA and CS model rankings within MMLU translations are highly stable across all languages ($\tau_b = 0.91$–0.94; 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.

Curriculum-overlap benchmarks show strong proxy agreement: Skolprov ($\tau_b = 0.89$), Exam-ET (0.88), and INCLUDE-LT (0.87) all achieve top-5 overlaps of 5/5, meaning the five best-performing models are identical on native and proxy benchmarks.

Culturally loaded benchmarks maintain moderate-to-strong agreement: DanishCitizenTest ($\tau_b = 0.82$; 95% CS) and NRK-Quiz-QA (0.85; 86% CS) still correlate substantially with their proxies, despite having content with no meaningful MMLU counterpart.

SwedishFacts is a clear outlier ($\tau_b = 0.51$), suggesting a knowledge dimension not well captured by translated benchmarks. However, high ranking

correlations do not resolve a key confound: strong models tend to score well on *any* benchmark (Ilić and Gignac, 2024), so high τ_b values may reflect general model strength rather than genuine proxy validity. We address this next.

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 (y = native score, s = log(params), b = Belebele score, x = 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 significantly improves over Model 1 (nested F-test), the proxy carries incremental predictive value—our operational definition of proxy validity. We keep the control set small ($n = 29$; two baseline predictors leave ∼10 observations per parameter); see Appendix D.6 for design rationale and full regression statistics.

Table 7 presents the main results. Six of seven pairings show significant incremental proxy validity ($p < .05$). The strongest effect is **Skolprov** ($\Delta R^2 = .128$, $p < 10^{-13}$): after controlling for model size and Swedish language proficiency, the MMLU proxy explains an additional 12.8 percentage points of variance in native performance.

INCLUDE-LT ($\Delta R^2 = .076$) and **NRK-Quiz-QA** ($\Delta R^2 = .041$) also show substantial effects, demonstrating that the proxy carries meaningful signal even for benchmarks with high culturally sensitive content (69% and 86% CS, respectively).

For **Exam-ET** ($\Delta R^2 = .024$), the modest increment reflects the already high baseline ($R^2_{\text{base}} = .954$): model size and language proficiency alone explain 95% of the variance. **DanishCitizenTest** ($\Delta R^2 = .009$, $p = .027$) is only marginally significant and fragile under robustness checks (§5.4).

The exception is **SwedishFacts** ($p = .242$): the proxy fails to add predictive value, and even the baseline explains only 35% of the variance—suggesting a knowledge dimension that neither general ability nor translated proxies capture. An exploratory CS/CA stratification (Appendix D.5) shows that for most pairings, both CA and CS items contribute to the proxy effect.

5.4 Robustness

Leave-one-family-out sensitivity. To verify that incremental effects are not driven by a single model family, we iteratively exclude each multi-model family and re-run Model 2. Skolprov remains highly significant in all seven leave-out conditions ($\beta_{\text{proxy}} > 0.98$, $p < 10^{-9}$); NRK-Quiz-QA and INCLUDE-LT likewise remain significant throughout. SwedishFacts remains non-significant in every condition. Two pairings show sensitivity: DanishCitizenTest loses significance when Cohere, Llama, or Qwen are excluded ($p = .110, .072, .096$), and INCLUDE-ET becomes non-significant when Qwen is excluded ($p = .102$; Appendix D.7).

Native	L	R^2_{base}	R^2_{ext}	ΔR^2	β_{proxy}	p
Skolprov	SV	.882	.987	.105	1.17	$<10^{-12}$
INCLUDE	LT	.890	.977	.087	1.08	$<10^{-9}$
INCLUDE	ET	.846	.934	.087	1.06	$<10^{-5}$
NRK-Quiz	NB	.940	.977	.037	0.54	$<10^{-6}$
Exam-ET	ET	.956	.986	.030	0.59	$<10^{-7}$
DanishCT	DA	.952	.971	.020	0.57	$<10^{-3}$
SwedishFacts	SV	.356	.430	.073	0.25	.092

Table 6: Sensitivity: incremental proxy validity with GMMLUX-EN as third control. All six previously significant pairings remain significant ($p < .001$); SwedishFacts remains non-significant. Full statistics in Appendix D.8.

Sensitivity: controlling for English MMLU performance. A potential concern is that the proxy effect reflects a "strong models are strong" confound rather than language-specific signal. To test this, we add each model's G-MMLUX score in English as a third control.

Native	L	R^2_{base}	R^2_{ext}	ΔR^2	β_{proxy}	p
Skolprov	SV	.856	.984	.128	1.06	$<10^{-13}$
INCLUDE	LT	.890	.965	.076	0.94	$<10^{-7}$
INCLUDE	ET	.818	.877	.059	0.84	.002
NRK-Quiz	NB	.934	.975	.041	0.46	$<10^{-6}$
Exam-ET	ET	.954	.979	.024	0.52	$<10^{-5}$
DanishCT	DA	.951	.960	.009	0.31	.027
SwedishFacts	SV	.353	.388	.035	0.14	.242

Table 7: Incremental proxy validity. R^2_{base}: baseline (log params + Belebele). R^2_{ext}: extended (+ proxy). ΔR^2: variance explained by the proxy beyond controls. p: nested F-test ($n = 29$). Six pairings significant at $p < .05$; DanishCT only marginally. SwedishFacts non-significant.

Table 6 shows that all six previously significant pairings remain significant ($p < .001$). Notably, DanishCitizenTest becomes robustly significant ($p < 10^{-3}$): once English MMLU per-

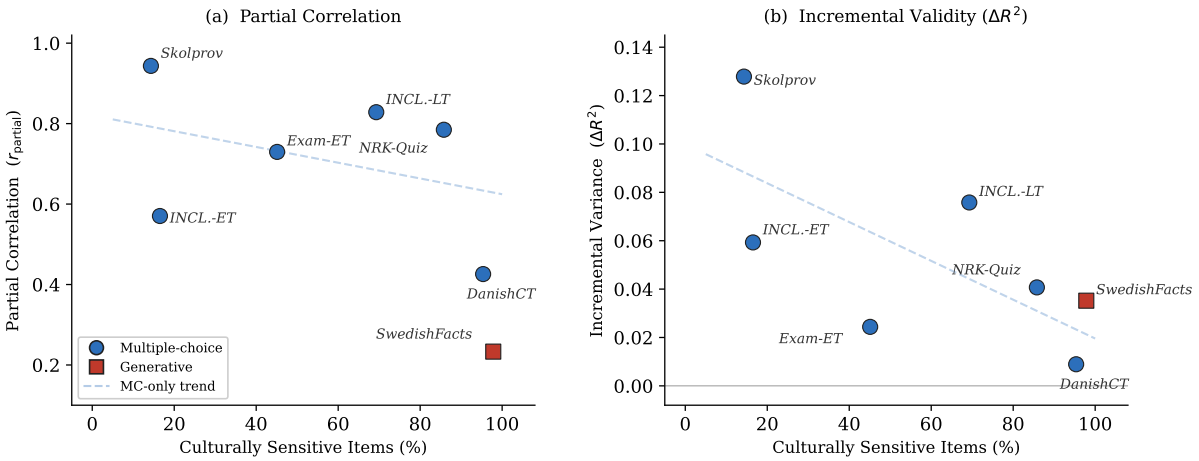

Figure 1: Proxy validity as a function of cultural sensitivity. Each point represents one native-vs-proxy pairing. **(a)** Partial correlation after residualizing on controls. **(b)** Incremental variance (ΔR^2) from the nested F-test. The dashed line shows the MC-only linear trend (6 pairings; indicative, not statistically significant at $n = 6$). DanishCitizenTest (95% CS, MC format) already shows weak proxy validity ($r_{\text{partial}} = .43$, $\Delta R^2 = .009$). In panel (b), SwedishFacts falls near the MC-trend extrapolation; removing Skolprov (the highest-leverage point) brings the predicted ΔR^2 even closer to the observed value (.030 vs. .035), confirming that this pattern is not driven by a single outlier.

formance absorbs general-ability variance, the language-specific proxy signal emerges more clearly. SwedishFacts remains non-significant ($p = .092$). These results suggest that the proxy's predictive value reflects language-specific signal rather than general MMLU-content familiarity.

6 Discussion

6.1 Interpreting Proxy Validity

What proxy validity does and does not show. Our incremental validity test shows that translated proxy scores are *predictive* of native benchmark performance beyond model size and language proficiency. This is a necessary condition for using translated benchmarks as stand-ins, but not a sufficient one: predictive association does not entail that the proxy measures the same underlying construct, nor that ranking differences are small enough for practical purposes. Practitioners should therefore interpret our results as evidence about *relative* proxy quality across benchmark types, not as blanket endorsement of translated benchmarks.

The proxy validity gradient. Our findings reveal a consistent gradient tied to content overlap between native and translated benchmarks. When native benchmarks test knowledge with international equivalents (curriculum-overlap: Skolprov, Exam-ET), proxy validity is strong and robust. When they test culturally loaded content with no direct MMLU counterpart (DanishCitizenTest, NRK-Quiz-QA), ranking agreement remains substantial but incremental validity is weaker—suggesting that much of the apparent agreement is driven by general model strength rather than content-specific signal. When they test locally anchored factual knowledge (SwedishFacts), the proxy provides no significant predictive value beyond controls. This gradient aligns with the locality effects reported descriptively by Hupkes and Bogoychev (2025), but our analysis additionally shows that the effect persists after controlling for confounds. An unexpected pattern in the ranking data reinforces this interpretation: for most pairings, $\tau_{\text{CS}} \geq \tau_{\text{CA}}$ (Table 5), likely because both native and MMLU CS items reward broad world knowledge that correlates with general capability, while CA items introduce more curriculum-specific ranking variation.

The SwedishFacts confound. SwedishFacts is the only generative benchmark in our study, leaving format and locality as entangled explanations for its low proxy validity. However, the proxy validity gradient is already visible within MC-format benchmarks alone (Figure 1). DanishCitizenTest—pure MC format, 95% culturally sensitive—shows the weakest partial correlation ($r = .43$) and the smallest ΔR^2 (.009) of all significant MC pairings, with fragile significance under LOFO (§5.4). This demonstrates that locally anchored content, not response format, is sufficient to substantially degrade

proxy validity. In the incremental validity analysis (Figure 1b), SwedishFacts falls near the MC-only trend extrapolation, suggesting that its ΔR^2 is consistent with what the locality gradient predicts at 98% cultural sensitivity. The partial correlation (Figure 1a) drops further below the MC trend, leaving room for an additional format contribution. We therefore interpret SwedishFacts as primarily locality-driven, with format as a potential amplifying factor rather than the primary cause. Disentangling format from locality conclusively would require additional generative benchmarks with low cultural load, or MC benchmarks with comparable Sweden-specific factual content; we leave this for future work.

6.2 Annotation Reliability

Our automatic CS/CA pipeline is substantially more accurate on native exam items (92%) than on GlobalMMLU-Lite (73%). This gap largely reflects differences in reference quality rather than pipeline failure: GlobalMMLU-Lite's human annotators agree with only 77% average majority share, and 62 of 391 items required tie resolution—indicating genuinely ambiguous boundary cases where the reference labels themselves are uncertain.

Native benchmark items, by contrast, tend to carry explicit regional or cultural anchors (e.g., national institutions, civic procedures), making the CS/CA distinction clearer for both human annotators and the model. The Danish inter-annotator data supports this interpretation: disagreement concentrates on the *regional/cultural* boundary (consistent with Singh et al., 2025), which is irrelevant for the binary CS/CA variable since both map to CS—hence the much higher CS/CA agreement (84%) compared to four-way `required_knowledge` agreement (58%).

The residual 8% error rate on native benchmarks introduces noise but should not systematically bias the proxy validity results, as misclassifications near the CS/CA boundary affect both proxy and native scores similarly. Moreover, the main incremental validity results use total accuracy scores and are therefore independent of individual CS/CA labels; annotation noise affects only the stratified subanalyses and the x-axis of Figure 1, where ± 8 pp shifts would not alter the benchmark ordering.

6.3 Generalizability

Our focus on five Nordic and Baltic languages reflects the rare combination of independently developed native benchmarks with sufficient item counts, existing DeepL-translated MMLU versions, and Belebele coverage for language-proficiency controls. While typologically limited to European languages, the selection spans two language families (Germanic: SV, DA, NB; Finno-Ugric: ET; Baltic: LT) and a range of resource levels, from moderate (LT, ET) to well-resourced (SV, DA, NB). The proxy validity gradient we observe—strong for curriculum-overlap, weaker for culturally loaded content—is plausibly language-independent, as it is driven by content overlap rather than language-specific properties. However, the absolute strength of the proxy effect may differ for typologically distant or low-resource languages where translation quality degrades and training data coverage is thinner. Native MMLU-style benchmarks now exist for Chinese (C-Eval, CMMLU), Korean (KMMLU), and Arabic (Native Arabic MMLU), which would enable testing this directly—an important direction for 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 language proficiency. Across 29 models and seven pairings in five languages, we find a clear gradient: proxy validity is strong when native benchmarks share broad subject coverage with MMLU (ΔR^2 up to .128, $p < 10^{-13}$), moderate for culturally loaded benchmarks where ranking agreement persists but incremental signal weakens, and absent for locally anchored factual knowledge (SwedishFacts). A sensitivity analysis controlling for English MMLU performance confirms that this signal is language-specific rather than an echo of general MMLU-content familiarity. These results offer practical guidance: translated benchmarks are reasonable proxies for international academic knowledge, but native benchmarks remain necessary for evaluating culturally grounded or locally anchored competences. We release all annotations, the cultural sensitivity pipeline, and analysis code to support reproducibility and extension to other languages.[1]

[1] Annotations and merged analysis data: https://anonymous.4open.science/r/cs-annotation-data-0231; analysis scripts and annotation pipeline: https://anonymous.4open.science/

8 Limitations

Our sample of 29 models is small from a regression perspective; incremental validity results indicate effect directions and approximate magnitudes rather than precise estimates.

The sensitivity analysis with three control variables (log params, Belebele, GMMLUX-EN) reduces the ratio to approximately 7 observations per parameter, which remains adequate for OLS but leaves less margin for detecting small effects.

Our proxies use a single translation method (DeepL); results may differ for human-translated or LLM-translated variants.

SwedishFacts is the only generative benchmark in our study; proxy validity findings for non-MC formats require confirmation with additional datasets.

Finally, extending the analysis to typologically distant or low-resource languages—where translation quality and training data coverage may differ substantially—is needed to test whether the proxy validity gradient we observe generalizes beyond European languages (§6.3).

Ethics Statement

This study involves human annotation of cultural sensitivity labels for benchmark items. Seven professional annotators were recruited through ProZ.com, a platform for freelance translators and linguists, and compensated on an individually agreed hourly basis (median approximately €30/hour). Annotators were informed about the purpose of the study and the intended use of their annotations before beginning work. No personally identifiable information was collected beyond the pseudonymized annotator IDs used in the inter-annotator agreement analysis.

Our use of GPT-5.2 for automatic annotation and quality review raises two ethical considerations. First, model-generated cultural sensitivity labels may reflect biases present in the model's training data; we mitigate this by validating against human references and reporting error rates transparently (§4.3). Second, the model-assisted review of human annotations (§4.1) could introduce circularity if similar models are later evaluated; we address this concern in Appendix C.3.

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

r/annotation-and-analysis-scripts-0A89.

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 Jain, Divyanshu Goyal, Saif Kumar, Sarvesh Meetei, Varun Kaul, Anoop Gupta, 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. 2024. m_mmlu: Machine-translated MMLU. Norwegian Bokmål translated via DeepL.

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, Chris Schmitz, Karolina Korgul, Hunar Batra, Oishi Deb, Emma Beharry, Cornelius Emde, Thomas Foster, Anna Gausen, María Grandury, Simeng Han, Valentin Hofmann, Lujain Ibrahim, and 23 others. 2025. Measuring what matters: Construct validity in large language model benchmarks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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.

Rochelle Choenni, Sara Rajaee, Christof Monz, and Ekaterina Shutova. 2024. On the Evaluation Practices in Multilingual NLP: Can Machine Translation Offer an Alternative to Human Translations? *Preprint*, arXiv:2406.14267.

Alvan R. Feinstein and Domenic V. Cicchetti. 1990. High Agreement but Low Kappa: I. The Problems of Two Paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The Language Model Evaluation Harness.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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*, pages 15646–15661, Suzhou, China. Association for Computational Linguistics.

Jenny Kunz. 2026. A Diagnostic Benchmark for Sweden-Related Factual Knowledge. *Preprint*, arXiv:2510.21360.

Murathan Kurfalı, Shorouq Zahra, Evangelia Gogoulou, Luise Dürlich, Fredrik Carlsson, and Joakim Nivre. 2025. SweSAT-1.0: The Swedish university entrance exam as a benchmark for large language models. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 331–339, Tallinn, Estonia. University of Tartu Library.

Viet Dac Lai, Nghia Trung 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.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Helena Grete Lillepalu and Tanel Alumäe. 2025. Estonian Native Large Language Model Benchmark. *Preprint*, arXiv:2510.21193.

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.

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.

Dan Saattrup Nielsen. 2024. Danish citizen test. ScandEval benchmark, `https://scandeval.com`.

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, 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.

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*, volume 2025, pages 83291–83322.

Olawale Salaudeen, Anka Reuel, Ahmed Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Ben Domingue, Angelina Wang, and Sanmi Koyejo. 2025. Measurement to Meaning: A Validity-Centered Framework for AI Evaluation. *Preprint*, arXiv:2505.10573.

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.

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 summarised in §2, with more detailed descriptions of individual studies.

Translation methods in multilingual evaluation. Choenni et al. (2024) translate four NLP tasks (NLI, QA, sentiment, commonsense) into 198 languages via NLLB and find that MT approximates HT with high rank correlations on average, though these vary considerably across languages; low-resource languages are systematically overestimated due to translation quality issues. 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 12 tasks, demonstrating that post-editing can substantially reduce 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) systematically evaluate ChatGPT on non-English benchmarks and document substantial cross-lingual performance variation, particularly for languages with limited web presence.

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 systematically 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 performance gap between locally sourced and translated data, which can exceed 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.

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 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.

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—background concept, systematized concept, measurement instruments, and measurements—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). Ilić and Gignac (2024) extend this with an analysis of 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 Global-MMLU (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 Global-MMLU, 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, confirming that subject-level comparison with MMLU is inherently imprecise.

Dataset	Direct	Approx.	No-equiv.
Global-MMLU	2,850	—	—
Global-MMLU-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.

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.

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.

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.

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 specialised 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 9.

Cross-validation with automated annotations. To validate our manual category assignments, we compared them against the `subject_category` labels produced by our GPT-5.2 annotation pipeline (§4). For each native subject, we computed the majority category assigned by the annotator and checked whether it matches our mapping. Across 3,044 cross-validated items, agreement reaches **85.1%** (Table 13).

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 ($n \leq 4$). None stem from mapping errors. This confirms that category-level comparison provides a robust basis for cross-benchmark analysis; subject-level alignment is too noisy to be informative.

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 benchmarks cannot capture. With 95.4% culturally sensitive items and the majority requiring regional knowledge (558 of 668 items annotated as `regional`), this pairing provides the strongest test of where proxies are expected to fail.

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 a less biased comparison point. The Estonian subset (224 items) and

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 9: 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. Global-MMLU and Global-MMLU-Lite are omitted as their subjects map directly (1:1) to MMLU. n: number of items.

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 10: Native exam subject labels and English translations.

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. Due to its non-MC format, we evaluate it separately from the MC-based comparisons.

C Appendix: Annotation

C.1 Annotation Schema and Interface

This section provides the complete annotation schema, interface screenshots, and prompt details referenced in §4.

Full annotation schema. Each item is a short, exam-style multiple-choice question 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, configured with the full schema described above. Figures 2–4 show the interface layout, form fields, and a worked example.

C.2 Annotation Prompt

API and reproducibility details. To ensure traceability and facilitate reruns, all automatic annotations were generated in February 2026 using the pinned model snapshot `gpt-5.2-2025-12-11` via the OpenAI

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

Dataset	Lang.	Items	Subjects	MMLU Subj.	MMLU Cat.	Ann. Cat.
Global-MMLU	EN	2,850	57	57	6	—
Global-MMLU-Lite	EN	217	43	43	6	6
Skolprov	SV	545	6	9	3	5
Exam-ET	ET	1,614	8	13	3	5
DanishCitizenTest	DA	668	2	2	1	6
SwedishFacts	SV	1,292	—	—	—	6
INCLUDE-base	LT	534	—	—	—	6
INCLUDE-base	ET	224	—	—	—	5
NRK-Quiz-QA	NB	3,600	—	—	—	6

Table 11: Dataset overview. **Subjects**: number of native subject labels. **MMLU Subj.**: number of distinct MMLU subjects covered by the mapping (one native subject may map to multiple). **MMLU Cat.**: number of MMLU categories covered. **Ann. Cat.**: number of distinct `subject_category` labels assigned by GPT-5.2 annotation. Datasets without native subjects (bottom block) rely entirely on item-level annotation for category assignment.

Dataset	Lang.	Items	CA	CS	Required Knowledge	Time-sens.
Global-MMLU	EN	2,850	2,058	792	none=9505; regional=1809; cultural=1253; dialect=84	Yes=428
Global-MMLU-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: Global-MMLU counts reflect individual annotator votes, not aggregated labels). **Time-sens.**: items with time-sensitive content.

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.

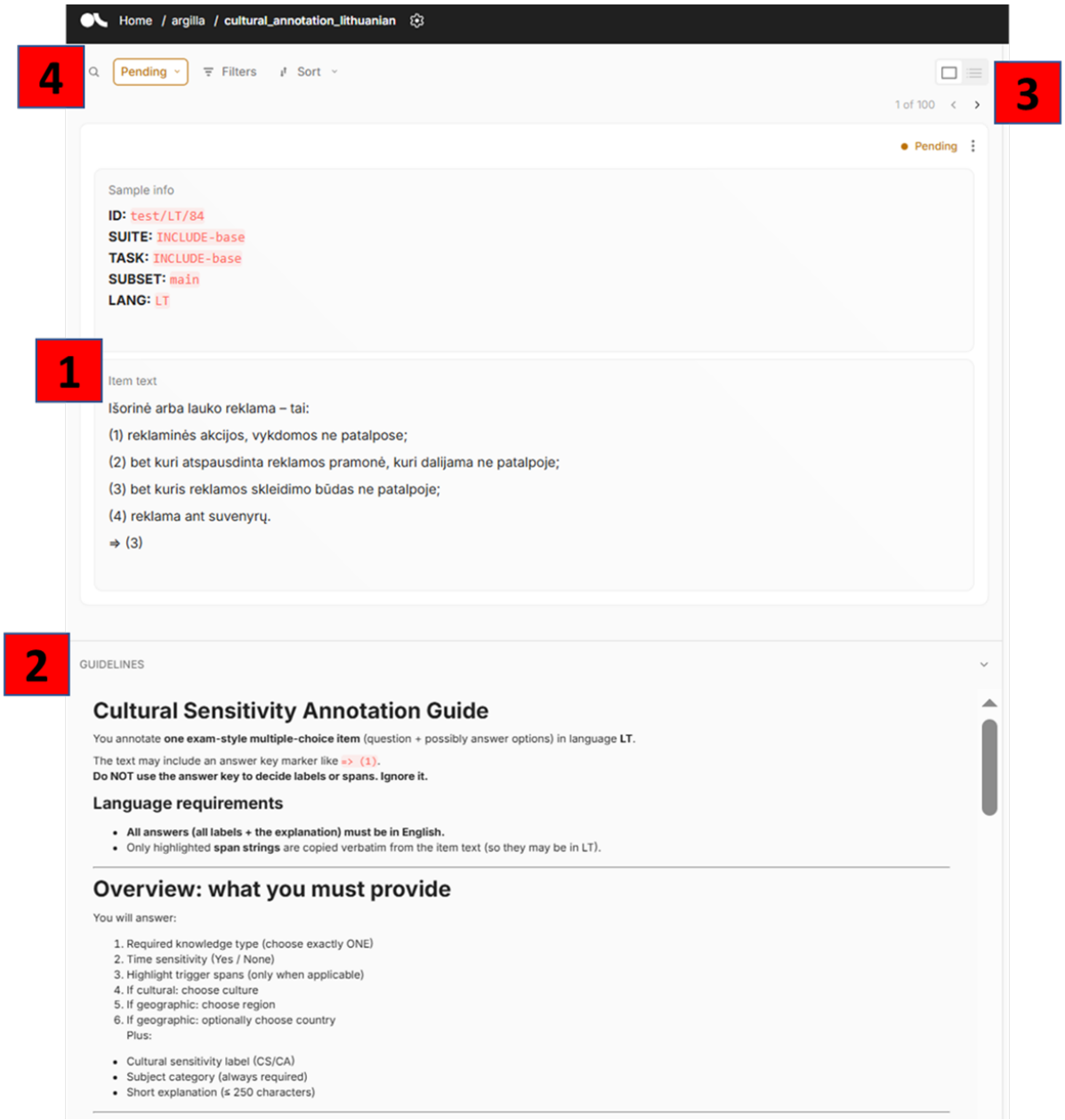

Figure 2: Overview of the Argilla annotation interface. **(1)** Item text panel displaying the multiple-choice question in the target language. **(2)** Expandable guidelines panel with the full annotation guide. **(3)** Navigation controls showing item progress (e.g., 1 of 100). **(4)** Status filter for selecting items by annotation state (Pending, Draft, Submitted, Discarded).

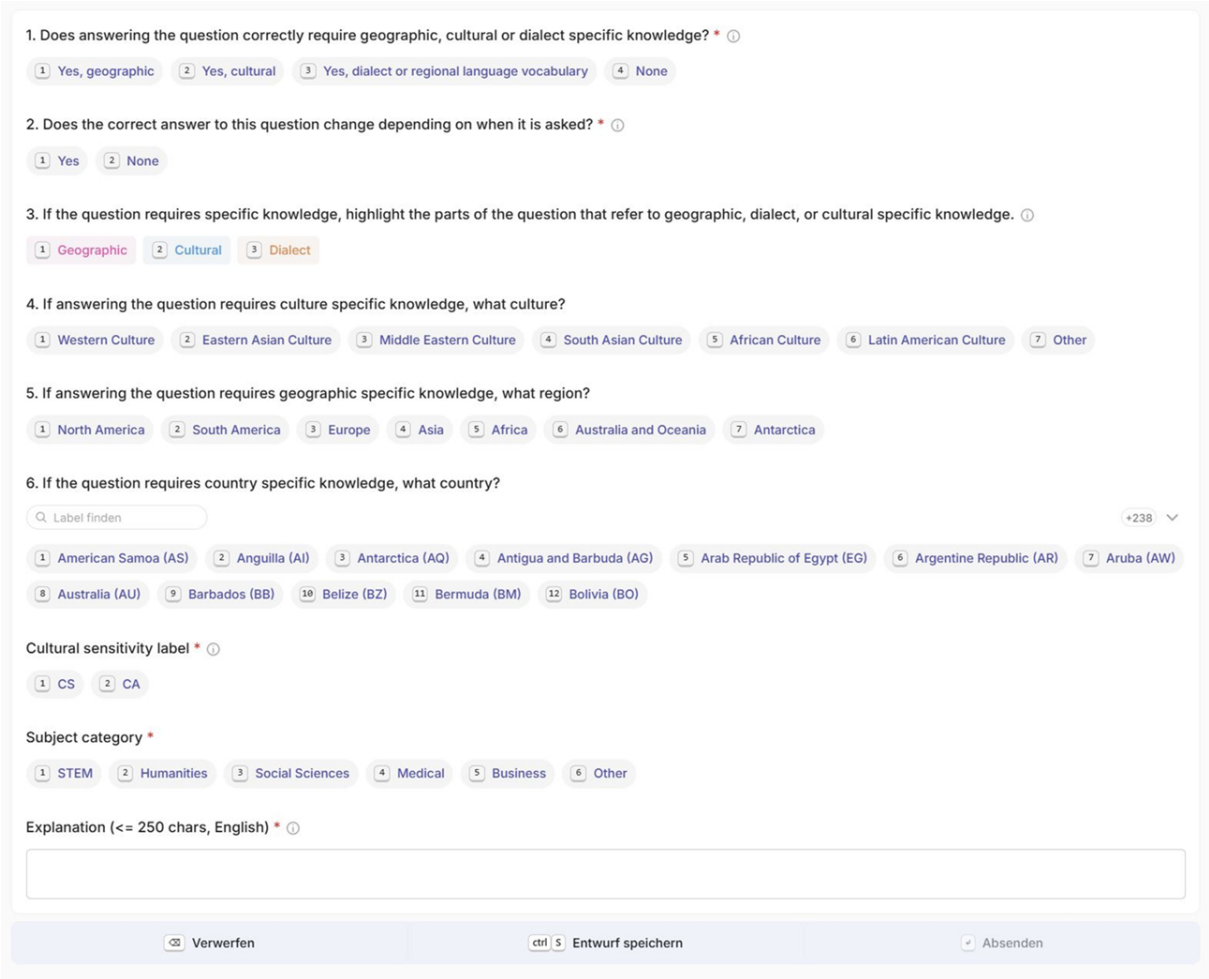

Figure 3: Annotation form fields in the Argilla interface. Annotators provide: (Q1) required knowledge type, (Q2) time sensitivity, (Q3) trigger span highlighting with geographic/cultural/dialect labels, (Q4–Q6) conditional culture, region, and country metadata, the CS/CA cultural sensitivity label, a subject category, and a short English explanation.

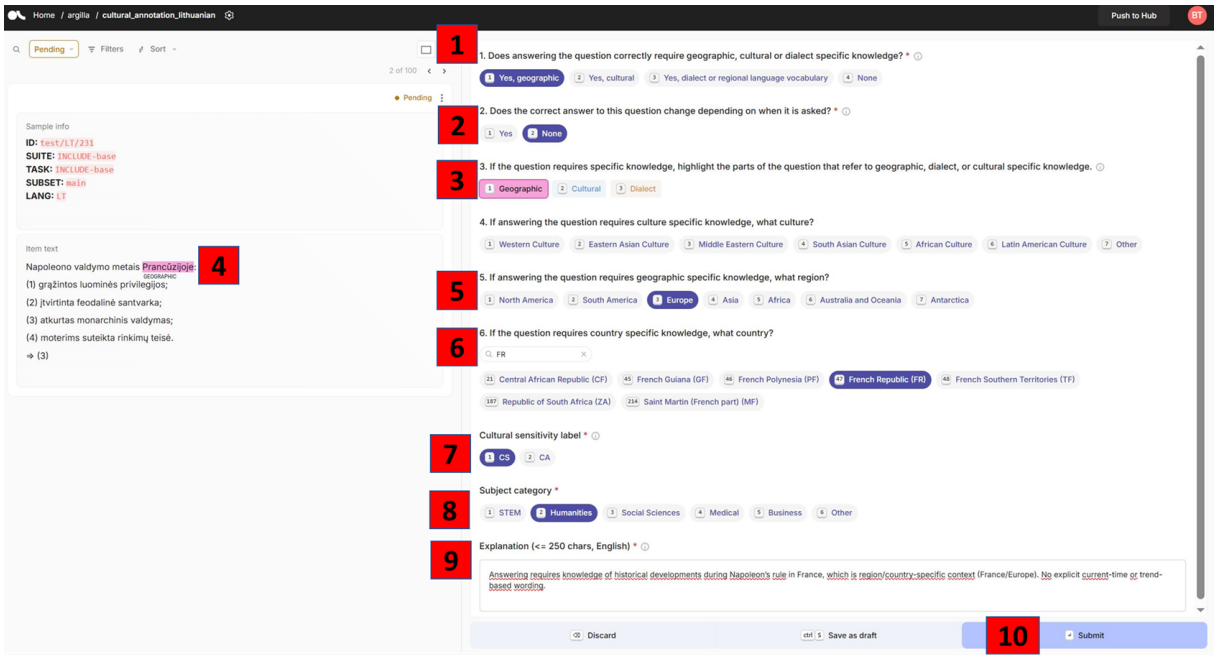

Figure 4: Worked example of a completed annotation for a Lithuanian item about Napoleonic rule in France. The annotator selected *Yes, geographic* (1), *None* for time sensitivity (2), highlighted the geographic trigger span "Prancūzijoje" (France) (3–4), chose region *Europe* (5) and country *France (FR)* (6), assigned the *CS* label (7), subject category *Humanities* (8), and provided a short explanation (9) before submitting (10).

MMLU Category	Subjects
STEM	18
Humanities	13
Social Sciences	12
Medical	7
Business	5
Other	2
Total	*57*

Table 14: MMLU category composition. The full list of 57 subjects and their category assignments follows the original MMLU taxonomy (Hendrycks et al., 2021).

API. We used `reasoning.effort=high` and `max_output_tokens=8000`. Each item was annotated in a single zero-shot call with the full system prompt (Figure 5). The validation review (Figure 6) used the same model snapshot and request configuration. We release all prompts, scripts, and raw outputs to support auditability and reruns

C.3 Human Annotation Details

This section provides the sampling procedure, annotator information, calibration protocol, and quality review details summarized in §4.1.

C.3.1 Sampling Procedure

We employ stratified sampling with a backoff mechanism to ensure broad coverage across relevant dimensions. Stratification is guided by proxy metadata obtained from the automatic annotation pipeline (GPT-5.2), specifically: predicted required knowledge type, cultural sensitivity label, time sensitivity, and subject category. These proxy labels are used *exclusively* for sampling and are removed from all items before upload to the annotation interface—annotators never see them.

The sampling algorithm attempts four-way stratification first and progressively reduces complexity (three-way, two-way, one-way) if the resulting strata become too fragmented. If all stratified approaches yield excessively fine-grained strata, it falls back to simple random sampling. For languages with multiple source datasets (e.g., Estonian with both Exam_ET and INCLUDE-ET), the sample budget is allocated proportionally across datasets with a minimum allocation per dataset to ensure coverage.

C.3.2 Annotator Recruitment and Compensation

We recruited professional annotators through ProZ.com, a platform for freelance translators and linguists. We contacted 15 candidates across the

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 5: 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 our public code repository.

five target languages and selected annotators based on native-speaker status and professional language expertise. Compensation was agreed upon individually on an hourly basis (median approximately €30/hour) with renegotiatable cap, with a total effort of 4–5 hours per annotator batch of 100 items.

A practical challenge was the limited availability of qualified annotators for some target languages. While we obtained three independent annotators for Danish, we were able to recruit only one annotator per language for Estonian, Lithuanian, Norwegian Bokmål, and Swedish. Additionally, not all DA annotators completed the full item set: one annotator submitted 95 of 100 items and another submitted 80, yielding 75 items annotated by all three DA annotators.

C.3.3 Calibration and Review Protocol

The annotation followed a two-round protocol designed to ensure guideline adherence and annotation quality.

Round 1: Calibration. Each annotator received the annotation guide (available in both a detailed and a condensed version), access credentials for the Argilla interface, and a request to annotate the first 10 items in their assigned language batch. After submission, we reviewed these calibration items and provided individualized feedback, including item-by-item corrections with short comments. Com-

mon calibration issues included (i) confusion between `cultural` and `regional` categories (e.g., labeling Denmark-specific legal knowledge as "cultural" rather than "regional"), (ii) inconsistent handling of conditional metadata fields (e.g., populating both culture and region fields when only one should be set), and (iii) overly liberal use of the time-sensitivity flag. After addressing the feedback, annotators proceeded to annotate the remaining items.

Throughout the process, annotators were instructed to rely exclusively on their own expert judgment and not use LLM or AI assistance, since the human labels serve as the reference standard against which model-generated annotations are evaluated.

Round 2: Automated review and correction. After all annotations were submitted, we ran an automated consistency check consisting of three steps: (1) *Per-annotator extraction:* annotations were exported from Argilla and split into individual JSONL files per annotator. (2) *Model-assisted validation:* each annotation was independently reviewed using GPT-5.2 with structured output, prompted with the full annotation schema and guidelines (see Figure 6 for the condensed system prompt). The model was instructed to detect missing fields, contradictory labels, invalid formats, span issues, and content-level sanity flags (e.g., `likely_regional_not_cultural`, `time_sensitive_without_time_cues`). For each flagged item, the model assigned a severity level (ERROR or WARN) and proposed minimal corrections. (3) *Review summary generation:* flagged items were aggregated into a TSV file per annotator, listing the item ID, severity, detected issues, and suggested corrections.

Each annotator then received their review file with instructions to prioritize ERROR-level items. The review was explicitly framed as decision support: annotators were free to reject suggestions and were asked to provide a short justification when doing so. The research team manually inspected all review files before distributing them, adding comments and corrections where the automated suggestions were incorrect or ambiguous.

C.3.4 Quality Review Statistics

Table 15 summarizes the results of the automated quality review conducted on the pre-review annotations.

ERROR rates vary considerably, ranging from 6% (DA ann3090) to 56% (SV). The majority of ERRORS concern schema-level inconsistencies—such as contradictory metadata fields (e.g., setting a culture value when Q1 = `regional`), missing required fields, or span-label mismatches—rather than substantive disagreements about the correct knowledge type. The high WARN rate (55% overall) largely reflects minor issues such as weak explanations or reconstructable missing span text.

Annotator	Lang	n	OK	Warn	Error	Fix
ann0172	DA	95	0	72	23	53
ann3090	DA	100	2	92	6	31
ann6129	DA	80	3	49	28	38
ann7708	ET	100	60	19	21	30
ann4956	LT	100	7	45	48	65
ann8933	NB	100	14	76	10	44
ann8395	SV	100	26	18	56	63
All		675	112	371	192	324

Table 15: Automated review statistics per annotator (pre-review). ERROR: issues requiring correction; WARN: minor consistency issues; OK: no issues detected; Fix: items flagged as needing correction.

Lang	Stage	none	regional	cultural	dialect
DA	pre	3	73	24	0
	post	3	77	20	0
	Δ	0	+4	−4	0
ET	pre	77	9	14	0
	post	77	9	14	0
	Δ	0	0	0	0
LT	pre	46	4	9	41
	post	43	50	6	1
	Δ	−3	+46	−3	−40
NB	pre	22	34	42	2
	post	22	33	43	2
	Δ	0	−1	+1	0
SV	pre	43	6	51	0
	post	45	40	15	0
	Δ	+2	+34	−36	0

Table 16: Label distribution shift from pre-review to post-review for `required_knowledge`. Δ indicates net change after the review and correction round.

C.3.5 Pre- to Post-Review Label Shifts

Table 16 shows the label shift from pre-review to post-review for `required_knowledge`.

The most substantial corrections occurred in Lithuanian and Swedish. For Lithuanian, 40 items originally labeled `dialect` were relabeled, predominantly to `regional` (+46), indicating that the annotator initially interpreted Lithuanian-language vocabulary and expressions as dialect-specific, whereas the review identified them as requiring knowledge of Lithuanian institutions, geography, or

System prompt for model-assisted annotation validation (GPT-5.2)

```
You are a strict QA reviewer for human annotations of cultural sensitivity in short exam-style items.

You will be given ONE annotation instance (one item + one user_id), with: metadata (sample_id, suite, task, subset,
language), item_text, and annotation fields (may be missing): required_knowledge, cultural_sensitivity_label,
time_sensitive, subject_category, culture, region, country, references, explanation.

Your job: (1) Validate each annotation against the annotation schema and guideline rules. (2) Detect missing fields,
contradictory labels, invalid formats, and suspicious entries. (3) Propose concrete corrections and short improvement
suggestions. (4) Output ONLY a JSON object matching the provided schema.
```

```
Core rules: (A) required_knowledge must be one of: none | cultural | regional | dialect. (B)
cultural_sensitivity_label: CA if required_knowledge == none; CS otherwise (default rule; deviations are flagged as
csl_vs_required_knowledge for review but are permitted by the annotation schema, see §C.1). (C) Metadata constraints:
if cultural → culture set, region/country empty; if regional → region/country may be set, culture empty; if dialect
or none → all empty. (D) time_sensitive: Yes only if item text contains explicit time/trend cues (``currently'', ``in
recent years'', ``has increased'', etc.); not just because a fact could change. (E) subject_category: one of STEM |
Humanities | Social Sciences | Medical | Business | Other. (F) references: list of span objects (start, end, label);
labels must match required_knowledge type; non-overlapping, sorted by start.

Label-choice sanity checks: regional → country/region-specific institutions, legal regimes, government bodies, local
policies, local history tied to a place. cultural → culture-/religion-/tradition-/arts-/canon-knowledge tied to a
```

```
cultural group. dialect → idioms, slang, dialectal words/phrases necessary to answer. none → general domain knowledge
not tied to a specific culture/region/dialect.

Output: One JSON object per instance with fields: missing_fields, contradictions, format_issues, span_issues,
content_sanity_flags (pipe-separated codes), needs_fix (YES/NO), severity (OK/WARN/ERROR), suggested corrections
per field, suggest_span_fix, and a 1--2 sentence comment.

Issue codes (excerpt): csl_vs_required_knowledge, cultural_has_region_or_country, regional_has_culture,
likely_regional_not_cultural, likely_none_not_cultural, time_sensitive_without_time_cues, weak_explanation,
span_label_mismatch.

Do not grade the annotator; just flag issues and suggest minimal, guideline-based corrections.
```

Figure 6: Condensed system prompt used for model-assisted validation of human annotations (§4.1, Round 2). The full prompt additionally specifies the complete JSON output schema, all issue code vocabularies, and detailed metadata constraint rules.

legal context. For Swedish, 36 items shifted from `cultural` to `regional` (+34), reflecting a systematic correction where items requiring knowledge of Swedish-specific institutions or policies were more appropriately classified as regional rather than cultural. Estonian and Norwegian Bokmål showed minimal or no shifts, suggesting that these annotators' initial labels were already largely consistent with the schema.

C.3.6 Limitations of the Human Annotation

Single-annotator coverage. For four of five languages, annotations were obtained from a single annotator per item. This prevents computation of inter-annotator agreement for these languages and limits our ability to quantify annotation uncertainty beyond Danish. The two-round protocol (calibration + automated review) partially mitigates this by surfacing schema-level inconsistencies, but cannot detect substantive disagreements that would only emerge with multiple annotators.

Boundary between regional and cultural. The inter-annotator analysis for Danish (§C.4) reveals that the distinction between `regional` and `cultural` knowledge is the primary source of disagreement. This is consistent with findings in the original GlobalMMLU study, where this boundary was identified as inherently challenging. Items involving country-specific traditions, legal customs, or historically rooted practices often straddle both categories. The tie-break rule (geographic > cultural) provides a default resolution, but annotators applied it with varying consistency.

Model-in-the-loop review. Our use of GPT-5.2 for automated quality review introduces a potential circularity concern: if similar models are later used in the evaluation comparison, the review step could inadvertently align human annotations toward model behavior. We mitigated this risk in several ways: (i) the review focused on schema consistency and format compliance rather than substantive label choices; (ii) annotators retained full authority to reject model suggestions; (iii) the research team manually inspected all review files before distribution. The pre-to-post comparison (Table 16) shows that the most substantial corrections were systematic category reassignments (`dialect`→`regional` in LT, `cultural`→`regional` in SV) rather than item-by-item alignment with model predictions.

Label	κ_F	3/3	2/3	0/3
Req. knowl.	0.206	43/74 (58%)	28/74 (38%)	3/74 (4%)
CS/CA	0.083	51/61 (84%)	10/61 (16%)	0/61 (0%)
Subject cat.	0.444	—	—	—
Time sens.	−0.009	—	—	—

Table 17: Inter-annotator agreement for Danish (3 annotators, 75 common items, pre-review). κ_F: Fleiss' κ. Agreement breakdown shows unanimous (3/3), majority (2/3), and complete disagreement (0/3) counts.

Label	Pair	κ	Agreement
Req. knowledge	0172–3090	0.281	51/75 (68%)
	0172–6129	0.249	55/74 (74%)
	3090–6129	0.149	51/74 (69%)
CS/CA	0172–3090	0.307	64/75 (85%)
	0172–6129	0.349	52/61 (85%)
	3090–6129	0.792	60/61 (98%)
Subject cat.	0172–3090	0.446	51/75 (68%)
	0172–6129	0.402	50/75 (67%)
	3090–6129	0.521	58/75 (77%)

Table 18: Pairwise Cohen's κ between Danish annotators (pre-review).

C.4 Inter-Annotator Agreement

For Danish, three annotators independently annotated the same set of items, providing an opportunity to assess inter-annotator agreement. Of the 100 items in the DA batch, 75 were annotated by all three annotators. Within these 75 items, annotator 6129 left 1 `required_knowledge` label and 14 `cultural_sensitivity_label` fields empty (pre-review), reducing the effective denominator to $n = 74$ for `required_knowledge` and $n = 61$ for CS/CA in all pairwise comparisons involving this annotator.

Tables 17 and 18 report agreement statistics on these common items, computed on the *pre-review* annotations to ensure that the statistics reflect genuine independent annotator concordance.

For the required knowledge type (Q1), Fleiss' $\kappa = 0.206$ indicates fair agreement according to the scale of Landis and Koch (1977). Full three-way agreement is reached on 58% of items, with a two-thirds majority on an additional 38%. Only 4% show complete disagreement. The primary source of disagreement is the boundary between `regional` and `cultural`, which is inherently fuzzy for items involving Denmark-

specific traditions or historical events that are simultaneously place-bound and culture-bound.

For CS/CA, the apparent discrepancy between high raw agreement (84% full, no full disagreements) and low Fleiss' κ (0.083) is an instance of the well-known prevalence effect (Feinstein and Cicchetti, 1990): all three annotators assign CS to 80–97% of items, leaving very little room for chance-corrected agreement to differentiate from expected agreement. The pairwise Cohen's κ between ann3090 and ann6129 (0.792) reflects near-identical CS/CA assignments (98.4%), while ann0172 assigns CS somewhat more conservatively (80% vs. 95–97%).

Subject category shows moderate agreement (Fleiss' $\kappa = 0.444$, pairwise Cohen's κ 0.402–0.521), reflecting genuine ambiguity in categorizing items that span multiple domains (e.g., legal history items classifiable as either Humanities or Social Sciences).

Time sensitivity is not meaningfully evaluated via κ because nearly all DA items were unanimously labeled `No` (97–100%), resulting in degenerate κ values.

C.5 Meta-Evaluation Details

This section provides the full confusion matrices, per-language error breakdowns, and calibration details 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 Global-MMLU 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 downstream analyses.

From the 391 resolved items, we create a stratified split (∼30% dev, seed = 13), yielding 174 dev and 217 test items (Table 19). The dev split is used for iterative prompt refinement; the test split is held out for final evaluation.

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 19: 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.

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 Required Knowledge Confusion Matrices

Tables 20 and 21 show the full four-way `required_knowledge` confusion matrices for both meta-evaluation settings.

On GlobalMMLU-Lite, the dominant errors are *regional*→*none* (23 items) and *none*→*cultural* (14 items). Both patterns directly affect CS/CA classification, as they shift items across the CS/CA boundary.

On native benchmarks, the dominant confusion is between *cultural* and *regional* (35+18 = 53 items). This reflects a genuine ambiguity in native-language exam items: when a question concerns, e.g., Esto-

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

H↓\G→	*cult.*	*dial.*	*none*	*reg.*	Σ
cultural	**60**	1	2	35	98
dialect	0	**2**	0	1	3
none	10	1	**167**	12	190
regional	18	0	8	**183**	209
Σ	88	4	177	231	500

Table 21: Required knowledge confusion matrix on the native benchmark sample ($n = 500$, 5 languages). Overall accuracy: 82.4%.

nian history or Danish civic institutions, the boundary between cultural and geographic knowledge is inherently unclear. Crucially, this confusion has no impact on CS/CA classification, since both labels map to CS. The CS/CA-relevant errors are comparatively few: 22 false alarms (*none*→*non-none*) versus 10 misses (*non-none*→*none*).

C.5.4 CS/CA Error Distribution by Language

Table 22 shows the CS/CA error distribution by language on the native benchmark sample. Lithuanian and Norwegian exhibit the highest error rates, predominantly in the direction CA→CS, suggesting that the model detects cultural or regional cues that the human annotator did not consider relevant. Swedish shows near-perfect agreement.

Lang.	Errors	CS→CA	CA→CS
DA	6	4	2
ET	6	1	5
LT	16	2	14
SV	1	1	0
NB	11	3	8

Table 22: CS/CA mismatches by language on the native benchmark sample ($n = 500$). CS→CA = human CS, GPT CA (miss); CA→CS = human CA, GPT CS (false alarm).

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 summarises the evaluation data underlying the proxy validity analysis. We merge the CS/CA annotation tables from §4 with per-model LM Evaluation Harness outputs (Gao et al., 2024), yielding 924,230 rows in the merged item-level analysis file 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 items × 2 rows per item, as both accuracy and F1 are scored), and Belebele (5,400 items across 5 languages), giving $31{,}870 \times 29 = 924{,}230$ rows.

Evaluation infrastructure. All evaluations use our custom fork of the LM Evaluation Harness[4] (Gao et al., 2024). 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, including DA non-MMLU which is excluded from interpretation in the main text.

[3] The merged analysis data and data-processing pipeline are available at https://anonymous.4open.science/r/data-processing-3D19.

[4] LM Evaluation Harness: https://github.com/EleutherAI/lm-evaluation-harness; our fork with native benchmark task implementations: https://anonymous.4open.science/r/lm-evaluation-harness-5E5A.

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 models are instruction-tuned. Inference uses NVIDIA H100 GPUs (94 GiB HBM2e); M-category models use 4-way tensor parallelism, others use a single GPU.

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 $\min(n_{\text{CA}}, n_{\text{CS}})$ items per class within each benchmark group and language. For MMLU-only, this yields 741 items per class per language (693 for NB due to missing translations in m_mmlu). 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 ($\tau_{\text{CA}} = 0.66$, 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 three reasons: (1) with $n = 29$ models, Kendall is more robust to individual outliers because it counts pairwise concordances rather than squaring rank differences; (2) τ_b explicitly corrects for tied ranks, which are common when models achieve identical accuracy on small item sets; and (3) p-values for τ_b do not rely on large-sample

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 ($n_{\text{CA}} = 31$; $\approx$1 item per model).

normal approximations that become unreliable at $n < 30$. Bootstrap CIs use 2,000 resamples over models, stratified by the model pairing (each resample draws n 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 27 reports the CS/CA-stratified incremental proxy validity results referenced in §5.3.

D.6 Incremental Proxy Validity (Details)

Table 29 extends Table 7 with the full set of regression statistics for all subsets (ALL, CA, CS), including adjusted R^2, AIC/BIC, and all coefficient estimates.

Interpretation of baseline R^2 values. The baseline R^2 varies substantially across pairings (from .353 for SwedishFacts to .954 for Exam-ET). High baselines indicate that model size and language proficiency already explain most of the variance in native performance, leaving a small ceiling for the proxy to improve upon. The fact that the proxy still achieves significant ΔR^2 even against high baselines (e.g., Exam-ET: $\Delta R^2 = .024$ against $R^2_{\text{base}} = .954$) is noteworthy. Conversely, the low baseline for SwedishFacts ($R^2 = .353$) indicates that this benchmark measures something that neither model size, language proficiency, *nor* the translated proxy captures—a genuine "locality" dimension of knowledge.

Methodology: nested F-test and OLS assumptions. We use standard OLS regression. The nested F-test compares Model 1 (2 predictors) against Model 2 (3 predictors), with 1 and $n - 4 = 25$ degrees of freedom. With $n = 29$ models and 3 predictors in the extended model, we maintain approximately 10 observations per parameter, which is adequate for stable OLS estimates. Variance Inflation Factors for the extended model range from 3.9 (log params) to 12.5 (Belebele), reflecting substantial collinearity between the proxy and Belebele scores. While VIF values above 10 indicate non-trivial collinearity, the primary consequence is inflated standard errors for individual coefficients; the nested F-test and ΔR^2 remain valid because they test overall model improvement rather than individual predictor contributions. We therefore interpret ΔR^2 and p-values as reliable, but caution that the magnitude of β_{proxy} should be treated as approximate rather than precise.

D.7 Robustness (Details)

Partial correlations. Table 31 reports the partial correlation between native and proxy performance after residualizing both on the controls (log params + Belebele).

Leave-one-family-out sensitivity. Table 32 summarizes the leave-one-family-out analysis for the ALL-items regression, restricted to multi-model families ($n_{\text{excluded}} \geq 2$). For each family, we remove all its members and re-run Model 2.

The LOFO analysis confirms that the incremental proxy validity findings are robust for most pairings. For **Skolprov**, the proxy coefficient remains remarkably stable ($\beta = 0.98$–1.07) across all leave-out conditions; the largest drop occurs when Qwen models are excluded (ΔR^2 drops from .128 to .078), but the effect remains highly significant ($p < 10^{-9}$). For **NRK-Quiz-QA** (Table 34), the proxy effect remains significant in all conditions ($\beta > 0.42$, $p < 10^{-4}$); likewise for **INCLUDE-LT** ($\beta > 0.90$, $p < 10^{-4}$). For **Exam-ET** (Table 33), all conditions remain significant ($p < 10^{-3}$). For **SwedishFacts**, the null result is equally stable: no family

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). $n_{\text{nat}}/n_{\text{prx}}$: 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.

		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 27: 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.

exclusion produces a significant effect.

Two pairings show LOFO sensitivity. **DanishCitizenTest** (Table 33): excluding Cohere ($n = 3$), Llama ($n = 4$), or Qwen ($n = 4$) each renders the effect non-significant ($p = .110$, .072, .096, respectively). This is consistent with the marginally significant overall result ($p = .027$) and confirms that the DanishCitizenTest proxy validity finding should be interpreted with caution. **INCLUDE-ET** (Table 30): excluding Qwen ($n = 4$) raises p from .002 to .102, consistent with the moderate overall $\Delta R^2 = .059$. All other leave-out conditions remain significant for INCLUDE-ET.

D.8 Sensitivity: English MMLU Control (Details)

Table 28 extends Table 6 with the full regression statistics for the sensitivity analysis that adds GMMLUX-EN (each model's G-MMLUX score in English) as a third baseline control.

Rationale and interpretation. The English G-MMLUX score serves as a content-level control: it captures how well a model handles MMLU-style questions *in English*, factoring out language-specific transfer. If the translated proxy in language L merely reflected a model's general familiarity with MMLU content rather than language-specific signal, its coefficient should drop to near zero once GMMLUX-EN is included.

The results show that this is not the case. All six previously significant pairings remain significant, most with p-values several orders of magnitude below .05. Several pairings show *larger* ΔR^2 values than in the main analysis (e.g., INCLUDE-ET: .087 vs. .059; DanishCT: .020 vs. .009), indicating that the language-specific proxy signal is partially masked by general-ability variance in the main specification. The improvement for DanishCitizenTest is particularly noteworthy: the main analysis yields only marginal significance ($p = .027$) that is fragile under LOFO, but the sensitivity analysis produces robust significance ($p < 10^{-3}$), suggesting that the proxy does carry meaningful signal for Danish civic knowledge once general MMLU ability is partialled out.

Collinearity considerations. Adding GMMLUX-EN as a third control increases collinearity: GMMLUX-EN correlates substantially with Belebele (both capture general model ability) and with the proxy score (both are MMLU-derived). Variance Inflation Factors increase substantially with four predictors (max VIF $= 10.6$–14.5 across languages), reflecting the strong correlations among Belebele, the proxy, and GMMLUX-EN (all capture overlapping dimensions of model ability). OLS coefficient estimates remain unbiased under collinearity, but individual standard errors are inflated, making significant results *harder* to achieve. The fact that all six previously significant pairings remain highly significant ($p < 10^{-3}$) despite this collinearity is therefore a conservative test. However, the individual β_{proxy} values should be interpreted as directional rather than precisely estimated; the ΔR^2 and p-values from the nested F-test are the primary inferential quantities.

Native	L	Sub.	n	R^2_{base}	$R^2_{\text{adj,base}}$	R^2_{ext}	$R^2_{\text{adj,ext}}$	ΔR^2	β_{proxy}	SE	p
Skolprov	SV	ALL	29	.882	.868	.987	.985	.105	1.166	0.084	$<10^{-12}$
INCLUDE	LT	ALL	29	.890	.876	.977	.973	.087	1.082	0.114	$<10^{-9}$
INCLUDE	ET	ALL	29	.846	.828	.934	.923	.087	1.057	0.188	$<10^{-5}$
NRK-Quiz	NB	ALL	29	.940	.933	.977	.973	.037	0.537	0.087	$<10^{-6}$
Exam-ET	ET	ALL	29	.956	.951	.986	.983	.030	0.587	0.084	$<10^{-7}$
DanishCT	DA	ALL	29	.952	.946	.971	.966	.020	0.572	0.142	$<10^{-3}$
SwedishFacts	SV	ALL	29	.356	.279	.430	.334	.073	0.247	0.141	.092

Table 28: Sensitivity analysis: full regression statistics with GMMLUX-EN as additional control. Baseline: $\text{native}_L \sim \log(\text{params}) + \text{Belebele}_L + \text{GMMLUX-EN}$. Extended: $+\text{proxy}_L$. Only ALL-items results are shown; the sensitivity analysis was not run on CA/CS subsets.

Native	L	Sub.	n	R^2_{base}	$R^2_{\text{adj,base}}$	R^2_{ext}	$R^2_{\text{adj,ext}}$	ΔR^2	β_{proxy}	SE	p
Curriculum-overlap benchmarks											
Skolprov	SV	ALL	29	.856	.845	.984	.982	.128	1.056	0.074	$<10^{-13}$
		CA	29	.843	.831	.983	.981	.140	1.079	0.075	$<10^{-13}$
		CS	29	.867	.857	.927	.918	.060	0.814	0.180	$<10^{-3}$
Exam-ET	ET	ALL	29	.954	.951	.979	.976	.024	0.515	0.097	$<10^{-5}$
		CA	29	.930	.924	.960	.956	.031	0.600	0.136	$<10^{-3}$
		CS	29	.971	.969	.986	.984	.015	0.382	0.074	$<10^{-4}$
Independent multilingual benchmarks											
INCLUDE	LT	ALL	29	.890	.881	.965	.961	.076	0.938	0.127	$<10^{-7}$
		CA	29	.829	.816	.925	.915	.095	0.999	0.178	$<10^{-5}$
		CS	29	.896	.888	.967	.963	.071	0.941	0.129	$<10^{-7}$
INCLUDE	ET	ALL	29	.818	.804	.877	.862	.059	0.842	0.243	.002
		CA	29	.802	.787	.861	.844	.059	0.831	0.255	.003
		CS	29	.803	.788	.850	.832	.046	0.751	0.271	.010
Culturally loaded benchmarks											
NRK-Quiz	NB	ALL	29	.934	.929	.975	.972	.041	0.463	0.073	$<10^{-6}$
		CA	29	.907	.899	.954	.949	.048	0.556	0.108	$<10^{-4}$
		CS	29	.936	.931	.979	.977	.044	0.445	0.061	$<10^{-7}$
DanishCT	DA	ALL	29	.951	.947	.960	.955	.009	0.309	0.131	.027
		CA	29	.864	.853	.869	.853	.005	−0.211	0.214	.333
		CS	29	.949	.945	.963	.958	.014	0.367	0.121	.005
Locally anchored factual knowledge											
SwedishFacts	SV	ALL	29	.353	.303	.388	.314	.035	0.140	0.117	.242
		CA	29	.427	.383	.434	.367	.007	0.108	0.190	.573
		CS	29	.348	.298	.394	.322	.046	0.156	0.113	.180

Table 29: Full incremental proxy validity results. R^2_{adj}: adjusted R^2. **Sub.**: subset (ALL, CA-only, CS-only). All models use $n = 29$. Baseline: native $\sim \log(\text{params}) + \text{Belebele}$. Extended: +proxy score. Grouped by proxy-validity strength.

		INCLUDE-ET (ET)				
Excl. family	n_{ex}	n	β_{prx}	p	ΔR^2	**Sig.**
Llama	4	25	0.751	.014	.050	✓
Qwen	4	25	0.554	.102	.015	×
Cohere	3	26	0.697	.010	.040	✓
Mistral	2	27	0.829	.003	.061	✓
Google	2	27	0.806	.004	.057	✓
EuroLLM	2	27	0.897	$<10^{-3}$	.071	✓
TII	2	27	0.837	.003	.063	✓

Table 30: Leave-one-family-out sensitivity for INCLUDE-ET. The effect remains significant in six of seven conditions. Excluding Qwen ($n = 4$) raises p from .002 to .102, making the effect non-significant—consistent with the moderate overall effect ($\Delta R^2 = .059$).

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 31: Partial correlations between native and proxy scores after residualizing on controls (log params + Belebele). $n = 29$ for all entries.

		Skolprov (SV)					SwedishFacts (SV)				
Excl. family	n_{ex}	n	β_{prx}	p	ΔR^2	**Sig.**	n	β_{prx}	p	ΔR^2	**Sig.**
Llama	4	25	1.059	$<10^{-11}$	.132	✓	25	0.088	.431	.020	×
Qwen	4	25	0.982	$<10^{-9}$	.078	✓	25	0.153	.340	.027	×
Cohere	3	26	1.039	$<10^{-11}$	.115	✓	26	0.118	.365	.025	×
Mistral	2	27	1.042	$<10^{-12}$	.137	✓	27	0.151	.207	.042	×
Google	2	27	1.056	$<10^{-12}$	.134	✓	27	0.124	.321	.029	×
EuroLLM	2	27	1.066	$<10^{-12}$	.128	✓	27	0.164	.176	.051	×
TII	2	27	1.039	$<10^{-12}$	.126	✓	27	0.125	.320	.029	×

Table 32: Leave-one-family-out sensitivity for Skolprov (strongest effect) and SwedishFacts (null result). n_{ex}: models excluded. **Sig.**: significant at $\alpha = .05$. Skolprov maintains $\beta_{\text{proxy}} > 0.98$ and $p < 10^{-9}$ across all conditions. SwedishFacts remains non-significant in all conditions. Tables 33–30 provide the remaining pairings.

		DanishCitizenTest (DA)					Exam-ET (ET)				
Excl. family	n_{ex}	n	β_{prx}	p	ΔR^2	**Sig.**	n	β_{prx}	p	ΔR^2	**Sig.**
Llama	4	25	0.281	.072	.008	×	25	0.413	$<10^{-3}$	.017	✓
Qwen	4	25	0.331	.096	.007	×	25	0.726	$<10^{-4}$	.029	✓
Cohere	3	26	0.235	.110	.005	×	26	0.448	$<10^{-3}$	.017	✓
Mistral	2	27	0.298	.039	.009	✓	27	0.514	$<10^{-4}$	.026	✓
Google	2	27	0.305	.037	.009	✓	27	0.508	$<10^{-4}$	.025	✓
EuroLLM	2	27	0.342	.010	.011	✓	27	0.525	$<10^{-4}$	.026	✓
TII	2	27	0.281	.025	.008	✓	27	0.514	$<10^{-4}$	.026	✓

Table 33: Leave-one-family-out sensitivity for DanishCitizenTest (fragile result) and Exam-ET (robust result). n_{ex}: models excluded. **Sig.**: significant at $\alpha = .05$. DanishCT loses significance when Cohere ($p = .110$), Llama ($p = .072$), or Qwen ($p = .096$) are excluded, confirming the fragility of this pairing. Exam-ET remains significant in all conditions ($p < 10^{-3}$).

		NRK-Quiz-QA (NB)					INCLUDE-LT (LT)				
Excl. family	n_{ex}	n	β_{prx}	p	ΔR^2	**Sig.**	n	β_{prx}	p	ΔR^2	**Sig.**
Llama	4	25	0.466	$<10^{-4}$	.041	✓	25	0.904	$<10^{-5}$	.072	✓
Qwen	4	25	0.459	$<10^{-4}$	.028	✓	25	0.922	$<10^{-4}$	.050	✓
Cohere	3	26	0.423	$<10^{-4}$	.032	✓	26	0.901	$<10^{-5}$	.066	✓
Mistral	2	27	0.448	$<10^{-5}$	.043	✓	27	0.946	$<10^{-6}$	.084	✓
Google	2	27	0.457	$<10^{-5}$	.042	✓	27	0.957	$<10^{-6}$	.081	✓
EuroLLM	2	27	0.470	$<10^{-5}$	.041	✓	27	0.955	$<10^{-7}$	.081	✓
TII	2	27	0.442	$<10^{-5}$	.039	✓	27	0.919	$<10^{-7}$	.079	✓

Table 34: Leave-one-family-out sensitivity for NRK-Quiz-QA and INCLUDE-LT. Both pairings remain significant in all leave-out conditions. NRK-Quiz-QA maintains $\beta_{\text{proxy}} > 0.42$ and $p < 10^{-4}$ throughout. INCLUDE-LT maintains $\beta_{\text{proxy}} > 0.90$ and $p < 10^{-4}$ throughout.