

Beyond Language Resource Tier: A Survey Study of Culturally Situated Resource Scarcity in LLM-Generated Cultural Translation of Math Word Problems

Parisa Suchdev^{1,2,3}, Juniper Lovato^{1,2,3,4}

¹Department of Computer Science, University of Vermont

²Vermont Complex Systems Institute, University of Vermont

³Computational Ethics Lab, University of Vermont

⁴Complexity Science Hub

jlovato@uvm.edu

Abstract

Large Language Models (LLMs) are being used to adapt math word problems for personalized learning at scale, and whether those adaptations actually serve students is an open question. Personalized learning, when conducted poorly, can introduce extraneous cognitive load for students, which makes it important to understand when and why LLMs fail at this task. To account for these failures, we present culturally situated resource scarcity, a mismatch between language-level data availability and coverage of the specific cultural and educational context a task requires. A language can appear high-resource overall while still lacking data for the community, region, script, domain, register, or educational genre needed for a particular task. In this preregistered study, we surveyed 552 native speakers across seven languages and three regions to assess the effectiveness of three frontier models in culturally translating math word problems. Our confirmatory results show that a low-resource language tier does impact linguistic quality and cultural appropriateness scores, but does not alter the mathematical reasoning of the translations. Our exploratory analysis shows, however, that region is a better predictor of culturally adapted translation quality than language resource tier. For example, Italian is a high-resource language yet receives lower cultural appropriateness ratings than all three Indian languages that were assessed, including low-resource Punjabi. A single-axis resource classification is insufficient. We argue that contributions from multiple data types should be assessed when reviewing complex translation tasks to fully understand where translation failures are occurring. We find that Likert ratings, checklist responses, and open-ended coding reveal different failure modes, with some coded errors not reflected in holistic ratings.

Code, Data, and Extended Version —

https://github.com/juniperlovato/MWP_Survey_Paper

Introduction

Math word problems are more effective when their themes and contexts feel familiar to students (Koncel-Kedziorski et al. 2016; Terrell and DeBay 2020). A familiar scenario reduces cognitive load, rather than spending working memory interpreting an unfamiliar world, the student can focus on the reasoning the problem is designed to teach (Sweller 1988;

Cummins et al. 1988; Kintsch and Greeno 1985; Chandler and Sweller 1991). An unfamiliar context adds a second task on top of the math. The problem compounds when a context appears to belong to a student’s culture but misrepresents it, a problem designed to put a student at ease instead signals distance, introducing an identity threat (Walton and Cohen 2007; Good, Rattan, and Dweck 2012).

Educators localize math word problems to avoid this. They replace unfamiliar referents with locally meaningful ones, preserve the mathematical structure, and remove barriers to comprehension without altering what the problem is designed to teach. LLMs are increasingly used to do this work at scale (OECD 2026; Walkington 2025). A single prompt can ask a model to translate a problem into a target language and adapt it for students in a specified country or region. We call this *culturally adapted translation* to distinguish it from literal translation, which changes language but not cultural surface, and from problem generation, which creates problems from scratch. The model must transform language and culture while preserving the underlying mathematical reasoning, this task means balancing three competing demands simultaneously, where optimizing for one can compromise the others.

LLMs make cultural adaptation easy to scale, but scaled failures require scrutiny because they may affect some populations more than others. A translated problem can appear locally appropriate, with adapted names, currency, and settings, while still carrying assumptions that make it less useful for students familiar with that culture. In prior work (Suchdev and Lovato 2026), we analyzed math word problem translations at the entity level, tracking how LLMs transformed, substituted, or omitted names, foods, currencies, and settings across languages. That study showed what models do to cultural entities at the corpus level, but did not ask whether full translations were linguistically natural, culturally appropriate, or mathematically faithful to native speakers. A substitution that looks reasonable in isolation can fail in context, replacing “Easter” with “Eid” is plausible, but retaining an egg-collection activity for Eid is not.

This paper moves from entity-level corpus characterization to translation-level human evaluation. We recruited 552 native speakers across seven languages, Hindi, Bengali, and Punjabi (India); Urdu and Sindhi (Pakistan); and Italian and Sicilian (Italy), to evaluate 60 math word problems

adapted by Claude Opus 4, GPT-4.1, and Gemini 2.5 Pro. The language set spans three resource tiers and three regional contexts, with pairs that allow within-region, within-tier comparison (resulting in a total of 1,260 unique culturally adapted translations). We collected three types of evaluation data, which included holistic Likert ratings of linguistic quality (LQ), cultural appropriateness (CA), and reasoning preservation (RP); structured checklist responses identifying specific cultural mismatch categories; and open-ended comments, which we coded into a 51-label error taxonomy. We find that these survey instruments are not interchangeable, each makes different failures visible.

We preregistered five hypotheses prior to data collection (Lovato and Suchdev 2026). H1: LLMs achieve lower translation quality in low-resource languages. H2: Cultural mismatches are more frequent in low-resource languages. H3: Mathematical reasoning is less reliably preserved in low-resource languages. H4: Gemini 2.5 Pro outperforms the other models on cultural fidelity and reasoning preservation. H5: Pronoun ambiguity compounds quality deficits in low-resource languages.

Our results show that low-resource languages (Punjabi, Sicilian, Sindhi) do receive worse ratings for linguistic quality and cultural appropriateness, but mid-resource languages (Bengali, Urdu) are rated as well as high-resource ones (Hindi, Italian), and reasoning preservation is unaffected at any tier. We find that no model outperforms the others and pronoun ambiguity makes no detectable difference. The more consequential finding is what the language resource tier classification misses. Region fits more than tier across all three evaluation dimensions. Italian is a high-resource language, yet Italian translations receive lower cultural appropriateness ratings than translations into Indian languages at every resource tier. Bengali is classified as mid-resource, yet models assigned Bangladeshi taka to Indian Bengali students. Urdu is classified as mid-resource, but available resources are not equally reliable across scripts (Kreutzer et al. 2022). A language can be data-rich globally and culturally thin in the specific community, domain, and script being tested (Nigatu et al. 2024). We call this *culturally situated resource scarcity*: the idea that translation quality depends not only on how much text a model has seen in a language, but on how well that text represents the specific cultural and educational context being requested.

We collected three types of evaluation data, Likert ratings, a structured cultural mismatch checklist, and open-ended responses coded into a 51-label error taxonomy. These survey instruments do not agree on which translations are wrong and where they fail. Grammar errors are most prevalent in the open-ended taxonomy and are consistently reflected in lower Likert ratings. Cultural errors are frequent in the taxonomy but inconsistently penalized in Likert scores. Participants' lower Likert scores for wrong names and implausible scenarios, but do not for currency mismatches or register failures, even when those errors appear in open-ended responses. Math errors are rare in the taxonomy but cause the largest Likert drops when they occur. We call this the *detection gap*. The failures that matter most for educational equity do not always surface in the instruments designed to catch

them and we find that different tools catch different errors.

This paper makes three contributions. We provide a large-scale preregistered native-speaker evaluation of LLM-generated culturally adapted math word problems across seven languages, three regional contexts, and three language resource tiers. We show that the standard framing of translation quality as a function of language tier is insufficient, region predicts quality better than tier, and a language can be globally high-resource while remaining culturally thin in the specific community, domain, and register being tested. We also develop an error taxonomy from participant open-ended feedback and show a detection gap analysis to quantify which coded error types are not reflected in holistic ratings, showing that culturally important failures can be difficult to capture with a single evaluation instrument.

Related Work

This paper connects three literatures: language resource tier in Natural Language Processing (NLP), cross-cultural adaptation, and measurement validity. We argue that culturally adapted math word problem translation should be evaluated not by language alone, but by how well a model fits the target community, region, domain, and task. We also argue for measurement instruments that capture these multiple axes.

Language Resources to Culturally Situated Scarcity

In NLP, resource has generally meant the computational materials that allow systems to work for a language, though the proxy has shifted over time. Early statistical machine translation emphasized bitext and sentence-aligned corpora (Resnik 1999; Resnik and Smith 2003; Nakov and Ng 2009); supervised NLP emphasized labeled data for tasks such as tagging, parsing, and classification (Yarowsky, Ngai, and Wicentowski 2001; Duong et al. 2014); and language engineering broadened the concept to include corpora, lexicons, grammars, tools, and benchmarks (Krauwier 2003; Probst et al. 2001; Besacier et al. 2014; ELRA 2026). Across these stages, the logic remained stable: a language was treated as more or less resourced according to whether the materials required by a given NLP system existed.

In the LLM era, high, mid, and low-resource labels are in common use but remain unsettled. Hämmäläinen (2021) argues the term has been applied so loosely that almost any language other than English can be called low-resource. The cited attempt at precision is Joshi et al. (2020), who bin languages into resource classes using indicators such as labeled and unlabeled corpora. The framework makes global NLP inequality visible, but treats resource as a property of entire languages. Nigatu et al. (2024) extend the critique, showing that resource conflates data, community, language, and agency, which are not separable in practice.

This data-centric framing scales easily but becomes harder to interpret as the scale grows. Token counts function as a proxy for exposure under finite model capacity, and recent work shows performance depends not only on total data size but on how training tokens are allocated across languages (He et al. 2025; Longpre et al. 2026). The proxy

only works if the tokens are representative in form and quality. Kreutzer et al. (2022) make this gap explicit through audits of web-crawled multilingual data, showing that apparent language coverage often contains mislabeled or non-linguistic content. Their Urdu finding is central to our argument: in CCAlied, the standard Urdu subset was highly usable, while the romanized Urdu subset contained very little usable Urdu. The same issue appears beyond script. Faisal et al. (2024) show that NLP performance varies across dialects, varieties, and closely related languages, and parallel work on Arabic finds that systems built on Modern Standard Arabic underperform on the dialects people use in daily life (Naous et al. 2024; Mousi et al. 2025). The question is therefore not only whether data exists for a named language, but whether the available data represents the forms in which people actually use it. Bird (2022) and Hershcovich et al. (2022) argue that low-resourcedness is not a single property of a language but a relation between variety, task, data source, model, and evaluation setting.

Recent cultural benchmarks show that language coverage does not guarantee cultural coverage. Myung et al. (2024) show that LLM performance on everyday cultural knowledge varies across countries even within the same language, with models performing better for cultures more represented online. Chiu et al. (2025) show uneven performance across 45 countries and regions. Together, these benchmarks show that cultural performance depends not only on language support, but on whether models capture the everyday contexts, regions, and practices associated with that language. Style adds another axis. Havaladar et al. (2025a) show that culturally appropriate conversational style depends on situation, relationship, and cultural context, while Havaladar et al. (2025b) show that LLM translations often fail to preserve intended style across politeness, intimacy, and formality, especially for several non-Western target languages. Together, these studies show that literal meaning can be preserved while culturally appropriate style is lost, a concern directly relevant to translated math word problems that may preserve reasoning while still sounding unnatural or pragmatically inappropriate. Taken together, this work shows that a language is not a uniform pool of tokens and a culture is not a uniform property of a language community. Both are internally varied along scripts, regions, registers, and genres – variation the standard resource framing cannot see. We refer to the joint condition as *culturally situated resource scarcity*: a model trained on a language has not thereby been trained on every community that speaks it, every register that community uses, or every genre in which that community communicates with its children. For our task, the relevant unit is the fit between a model’s training distribution and the target educational situation: an Urdu or Sindhi problem for students in Pakistan, a Bengali problem for students in India, or a Sicilian-language problem for students in Italy. With that said, the instruments used to measure this multidimensionality must themselves be capable of capturing it.

Evaluation, Construct Validity, and Detection Gap

Automatic MT metrics have long shaped translation evaluation: BLEU enabled large-scale comparison through a

fast, inexpensive, language-independent metric (Papineni et al. 2002), while later approaches such as BERTScore and COMET moved beyond lexical overlap by using contextual embeddings and neural models trained on human quality judgments (Zhang et al. 2020; Rei et al. 2020). Yet automatic scores often miss distinctions that matter to human evaluators (Callison-Burch, Osborne, and Koehn 2006). Freitag et al. (2021) argue for expert human evaluation using Multidimensional Quality Metrics (MQM), showing that inadequate evaluation procedures can lead to erroneous conclusions. More recent work extends this concern: Error Span Annotation asks evaluators to mark where serious translation errors occur, bridging scalar ratings and MQM-style error analysis (Kocmi et al. 2024).

Psychometrics provides the vocabulary for this problem. Messick (1995, 1994) defines validity as an evaluative judgment about whether evidence supports the interpretation and use of scores, and identifies two threats that matter here: construct underrepresentation, when an instrument omits parts of the construct, and construct-irrelevant variance, when scores are shaped by factors outside it. Survey methodology adds a third concern: ceiling effects, where responses cluster at the high end and lose discriminative power (Chyung, Hutchinson, and Shamsy 2020).

Applied to cultural evaluation, these threats are central. A single Likert rating of cultural appropriateness underrepresents a construct that spans entity choice, register, regional fit, and genre. Checklists can identify specific errors but miss dimensions not anticipated by the instrument, while free-text coding can surface those dimensions at the cost of greater rater-dependent variance. No single instrument is psychometrically sufficient for a construct with this many axes.

We call this mismatch the detection gap: each instrument is selectively sensitive to a different slice of the construct. Holistic ratings can indicate that something feels wrong but cannot identify whether the issue is a name, currency, regional convention, or register mismatch. Checklists capture predefined errors but miss diffuse failures of tone or genre. Open-ended text coding recovers those details, but a coded error does not necessarily mean the rater lowered their scalar judgment. In the analysis below, we operationalize one form of this gap by comparing coded open-ended labels with the corresponding Likert ratings. The fourth instrument, Corpus-level analysis (Suchdev and Lovato 2026) captures failure on a systematic level such as models converging on the same names or regional defaults across many outputs. Because each instrument is selectively sensitive to different failures, cultural adaptation studies must combine instruments rather than rely on any one.

Methodology

Dataset and Translations

The math word problem dataset used for translation is a modified subset from Karim et al. (2025), which is constructed from the GSM8K test set (Cobbe et al. 2021). Each math word problem is accompanied by annotations that identify culturally meaningful entities within that problem, such as names, foods, currencies, and places. Using these

annotations, we selected a subset of 60 problems to capture a broad range of cultural entity types and to cover different levels of complexity. Complexity refers to the number of unique annotated cultural entity types in a problem. In addition to this subset construction, we introduced a controlled modification to 16 of the 60 selected problems by replacing gendered pronouns such as “he” and “she” with the gender-neutral pronoun “they”. This modification introduces referential ambiguity and allows us to examine whether LLMs preserve clarity when translating math word problems under less explicit linguistic conditions.

We construct a multilingual translation corpus across seven target languages from three regions and across different resource tiers (HR = high, MR = mid, LR = low). For India: Hindi (HR), Bengali (MR), and Punjabi (LR); for Pakistan: Urdu (MR) and Sindhi (LR); for Italy: Italian (HR) and Sicilian (LR). Resource tiers are derived from HPLT deduplicated word counts (see Appendix for full table) and used consistently throughout (de Gibert et al. 2024).

Translations are generated using Claude Opus 4, GPT-4.1, and Gemini 2.5 Pro, in a zero-shot setting. Each model is prompted to translate and culturally adapt the problem for the specified region using a standardized prompt (Full prompt in Appendix). We set the temperature to 0.7 for all generations to encourage lexical variation while maintaining coherence (Singh et al. 2023; Moreno García and Mangiron 2024), and keep all other parameters at their default. The translation pipeline and resulting corpus are described in Suchdev and Lovato (2026); this paper focuses on native speaker evaluation of those translations.

Using this dataset, we evaluate translations along three dimensions: linguistic quality (LQ), cultural appropriateness (CA), and reasoning preservation (RP). We also examine how translation quality varies as a function of ambiguity, complexity, and the resource level of the target language.

Survey Design

Data were collected through seven parallel Qualtrics surveys, one per language, structurally identical across languages but adapted to the linguistic and regional context. Participants were recruited from India, Pakistan, and Italy through a Qualtrics panel (IRB protocol STUDY00003740).

Eligible participants were at least 18 years old, had completed high school, and self-reported fluency in both English and the target language, as well as familiarity with the relevant regional culture. After consent and screening, participants completed a language and cultural fluency self-report and a purpose-designed bilingual comprehension check. The check was tested against frontier LLMs; none answered correctly. The check required integrating information from English text, target-language text, and an accompanying image to arrive at the correct answer. Only participants who answered correctly proceeded.

The full item pool consisted of 180 LLM-generated translations (60 source problems $\times$ 3 LLMs). Each participant was randomly assigned to 5 translations. A quota was applied to ensure each translation received approximately 3 independent ratings, thereby meeting the target sample of 100 participants per language. Problems were presented with

both the original English and the LLM-generated translation displayed side by side. For each problem, participants rated three dimensions on 5-point Likert scales (LQ, CA, RP). Each Likert item was followed by a required open-ended follow-up question. CA was additionally assessed via a 15-item mismatch checklist (e.g., names, settings, social norms, units, idioms, other, none of the above), presented in randomized order. Next, participants were asked to complete demographics questions (e.g., gender, age, race/ethnicity, culture, education, home region, country of residence). Submissions flagged as ballot stuffing, duplicates, incoherent or gibberish open-ended responses, or completion times under 120 seconds were dropped before analysis (initial recruitment = 700, final N = 552).

Survey Analysis

All confirmatory analyses were preregistered, and preregistered exploratory analyses are noted where applicable. We computed descriptive statistics such as means, medians, and standard deviations for each Likert dimension (LQ, CA, RP) disaggregated by language, region, resource tier, model, and ambiguity condition. Checklist responses were summarized as raw counts and proportions per category.

We also conducted Kruskal-Wallis tests to assess the differences across resource tiers and, as a preregistered exploratory analysis, across regions. For pairwise comparisons, we conducted Mann-Whitney U tests with rank-biserial correlation (r) to measure effect size. Effect sizes are reported alongside p-values. We computed pairwise Spearman correlations across the three Likert dimensions, overall and disaggregated by language tier, region, and language. 95% Confidence intervals were created using a bootstrap with 10,000 iterations (seed = 42).

A confirmatory analysis used ordinal logistic regression to predict each Likert dimension from resource tier (reference: high), cultural complexity (1–5 entities), ambiguity condition, LLM model (reference: Gemini 2.5 Pro), and ambiguity $\times$ resource tier interaction terms. An exploratory analysis replaced the resource tier with geographic region (reference: India) using identical predictors and interaction terms. We also fit logistic regression models predicting the presence of a binary mismatch on the cultural checklist under both tier and region specifications. Model fit was assessed using Akaike Information Criterion.

Error Taxonomy Development and Analysis Open-ended responses (N = 8,303) were coded using a collaborative inductive-deductive procedure. We developed a taxonomy of 51 labels grouped into three supercategories: grammar, culture, and math errors. Administrative labels: invalid or unusable responses were excluded before analysis.

Annotator 1 first reviewed a pilot sample of 745 responses to identify preliminary categories, then coded a stratified random sample of 1,141 responses, evenly distributed across languages and representing 14% of the dataset, to refine the taxonomy. Annotator 2 reviewed the resulting codebook, and Annotators 1 and 2 revised label definitions by consensus. The finalized codebook was then applied to the full dataset by three annotators, who divided responses and

coded them independently. Annotators 1 and 2 had participated in codebook development; Annotator 3 received only the written codebook, allowing us to test whether the taxonomy could be applied by a coder not involved in its creation.

To assess reliability, we drew a stratified random 10% subsample of responses originally coded by Annotators 2 and 3 ($n = 674$), evenly distributed across languages, and had Annotator 1 independently re-code them. Percent agreement between the original coder and independent coder was 97.7%. We report percent agreement rather than Cohen's kappa because kappa can be difficult to interpret in inductive, multi-label, or prevalence-skewed qualitative coding settings (McDonald, Schoenebeck, and Forte 2019). Coding yielded 2,231 total labeled errors across all seven languages.

For each label, we computed the mean Likert rating on the relevant dimension among responses with that label (exploratory). We define the detection gap as the difference between the overall dimension mean and the label-specific mean. Positive values indicate that raters lowered scores when the error was present; values near zero or negative indicate that raters did not penalize the error.

Results

Rating Distributions by Language and Region

Throughout the results, resource tier refers to the HPLT-derived grouping (high: Hindi, Italian; mid: Bengali, Urdu; low: Punjabi, Sicilian, Sindhi), (de Gibert et al. 2024) and region to the country specified in the translation prompt (India: Hindi, Bengali, Punjabi; Pakistan: Urdu, Sindhi; Italy: Italian, Sicilian). We see that native speaker ratings are high across all seven languages and three evaluation dimensions, with means ranging from 3.94 (Urdu, cultural appropriateness (CA)) to 4.32 (Hindi, CA) on the 1-5 Likert scale (full tables in Appendix). Overall medians are 4.0 across all three dimensions, a threshold we return to in the detection gap analysis. Standard deviations are consistently larger for Pakistan languages (1.01-1.07) than for Indian or Italian ones (0.73-0.92), suggesting greater evaluator disagreement rather than worse performance. At the tier level, aggregate means decline modestly from high to low resource: Linguistic Quality (LQ) from 4.21 to 4.06, Cultural Appropriateness (CA) from 4.22 to 4.04. Reasoning Preservation (RP) shows the smallest tier gradient of the three dimensions (4.23 to 4.16), consistent with the null H3 result. Pairwise Spearman correlations between the Likert dimensions were strong and consistent across all subgroups ($\rho = 0.56-0.72$).

Figure 1 shows the within-tier structure. Within the high tier, Hindi (India, CA = 4.32) and Italian (Italy, CA = 4.11) diverge by 0.21 points on CA. Within the mid tier, Bengali (India, LQ = 4.28) scores above Hindi (LQ = 4.23), a language one tier higher. Punjabi (low resource, India, LQ = 4.16) scores above Urdu (mid resource, Pakistan, LQ = 3.95). The language tier ordering is not preserved within regions or across them. Figure 1 shows that Pakistan languages occupy the lower portion of all three panels at every tier level; Indian languages occupy the upper portion; and Italian languages sit between them. The dots cluster by color, not by resource tier position.

Resource Tier and Model Effects

We tested five preregistered hypotheses using Kruskal-Wallis tests for group comparisons and ordinal logistic regression to identify predictors of Likert ratings (reference categories: Gemini 2.5 Pro, high resource tier, standard ambiguity; Figure 2 left panels, see Appendix for full Table).

H1, that low-resource languages receive lower translation quality scores, and **H2**, that low-resource translations introduce more cultural mismatches, receive partial support. Resource tier predicts LQ (OR = 0.78, $p = 0.012$) and CA (OR = 0.72, $p = 0.001$) at the low-resource tier; mid-resource languages are indistinguishable from high-resource ones on both dimensions. **H3** that low-resource translations produce more mathematical reasoning errors is not supported: resource tier does not predict RP ratings at any tier level (low tier: OR = 0.93, $p = 0.476$). **H4** (Gemini 2.5 Pro outperforms Claude Opus 4 and GPT-4.1) and **H5** (pronoun ambiguity compounds low-resource deficits) are not supported: model and ambiguity condition do not predict ratings on any dimension, and ambiguity $\times$ tier interactions are null throughout (all $p > 0.22$).

H1 and H2 receive partial support. Resource tier predicts LQ and CA at the low-resource tier; mid-resource languages are indistinguishable from high-resource on both dimensions (Kruskal-Wallis: LQ $H = 9.58$, $p = 0.008$, $\varepsilon^2 = 0.003$; CA $H = 14.62$, $p = 0.001$, $\varepsilon^2 = 0.004$). H3 is not supported: tier does not predict RP at any level ($H = 1.53$, $p = 0.466$). H4 and H5 are not supported: model and ambiguity condition are null across all three dimensions, including the ambiguity $\times$ tier interactions (all $p > 0.05$).

Two features of this pattern are worth noting. First, the tier effect is dimension-specific: it surfaces on surface-level LQ and CA but not on RP, suggesting that resource scarcity shapes the cultural and linguistic register of translations without systematically disrupting their mathematical logic. Second, the absence of a mid-tier effect means that tier is not a continuous predictor, it flags only the lowest-resource languages, and even then with modest effect sizes.

Region as an Alternative Predictor

The within-tier spread raises a question that the confirmatory model cannot answer, if Bengali (mid resource, India) rates higher on linguistic quality than Hindi (high resource, India), and Punjabi (low resource, India) rates higher than Urdu (mid resource, Pakistan) on all three dimensions, then resource tier is not doing the explanatory work we expected. The pattern seen in Figure 1 suggests that a language's region of origin, which indexes cultural and geographic context rather than just training data quantity, may be a more precise predictor than its resource tier. We ran exploratory ordinal logistic regressions replacing resource tier with region (reference: India), using the same covariates and model specification as the confirmatory analyses. Region is not a post-hoc covariate. The prompt names the country, so every translation was generated under an explicit regional instruction; the region models test a variable the models were directly conditioned on. We see that region models fit better than tier models on all three dimensions. AIC is lower

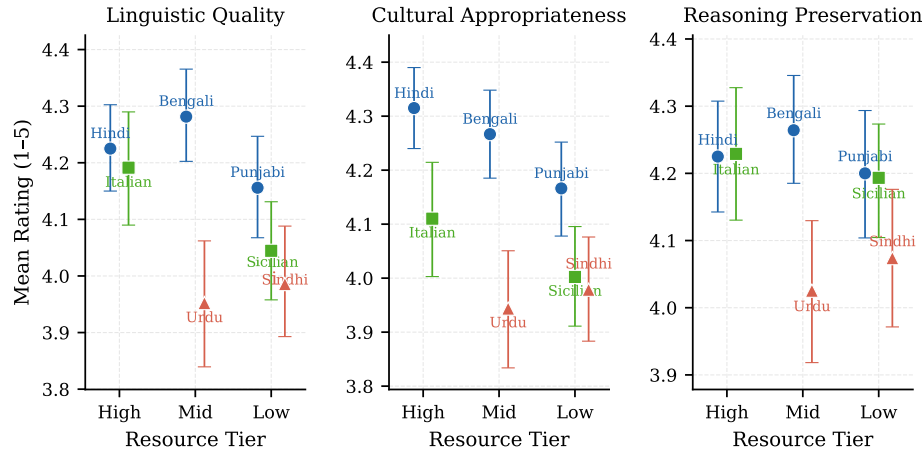


Figure 1: Mean native speaker ratings by language, grouped by language resource tier (low: Punjabi, Sicilian, Sindhi; mid: Bengali, Urdu; high: Hindi, Italian), across three evaluation dimensions. Each point represents one language; blue circles = India, green squares = Italy, red triangles = Pakistan. Error bars show 95% bootstrap confidence intervals. Within each tier, languages from the same region cluster together: India consistently above the tier center, Pakistan consistently below, while languages sharing a tier but different regions diverge substantially. Urdu scores below Punjabi on all three dimensions, and Sindhi scores below Sicilian (low-tier, Italy).

for the region specification on LQ (6800 vs. 6817, $\Delta\text{AIC} = -17.7$), CA (6802 vs. 6825, $\Delta\text{AIC} = -22.4$), and RP (6664 vs. 6674, $\Delta\text{AIC} = -10.6$), with the largest improvement on CA (Figure 2 right panels, see Appendix for full table).

Both Italy and Pakistan rate significantly lower than India on LQ and CA. For LQ, Italy shows OR = 0.81, 95% CI [0.67, 0.99], $p = 0.039$, and Pakistan shows OR = 0.73, 95% CI [0.60, 0.88], $p = 0.001$. The CA effects are larger: Italy OR = 0.74 [0.61, 0.91], $p = 0.003$; Pakistan OR = 0.62 [0.51, 0.76], $p < 0.001$. For RP, Italy is not significant (OR = 0.97, $p = 0.792$), but Pakistan is (OR = 0.82 [0.67, 1.00], $p = 0.047$), making it the only predictor in either model to reach significance on that dimension. Ambiguity condition and model effects remain null throughout.

Italian here is notable, it is a high-resource language, it is not data-scarce in any global sense, yet Italian translations receive lower CA ratings than translations into Indian languages at every resource tier. The regional contrast is visible within the high tier itself: Hindi (India) receives a mean CA rating of 4.32 while Italian (Italy) receives 4.11, a gap of 0.21 points that is larger than the full high-to-low tier aggregate difference on that dimension. This is not entirely explained by the amount of Italian text in the LLM training data alone. An alternative account involves the specific cultural domain being tested, elementary school mathematics problems embedded in Italian regional contexts, which may be underrepresented even in Italian-language training data.

Urdu is classified as mid-resource (MR) by token count, yet scores below low-resource (LR) Sicilian on all three dimensions. Kreutzer et al. (2022) found that the standard Urdu subset in CCAIaligned was highly usable, whereas the romanized Urdu subset contained very little usable Urdu text. Token count may therefore mask a script-level data quality gap rather than genuine MR coverage. Sicilian presents

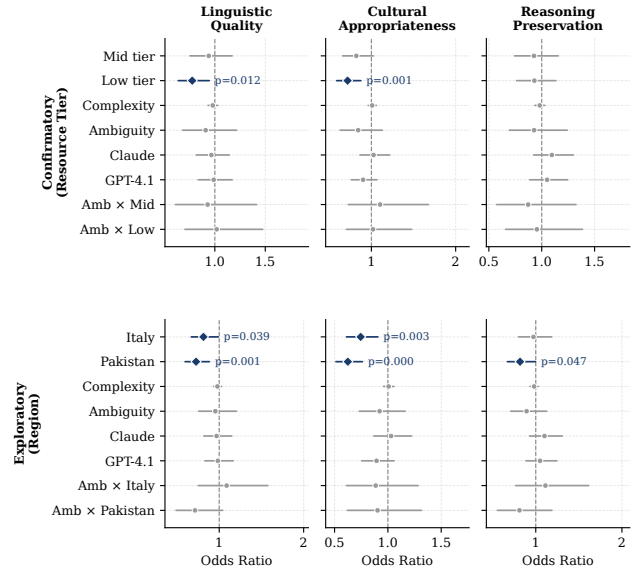


Figure 2: Odds ratios with 95% confidence intervals from ordinal logistic regressions predicting native speaker ratings. Left: confirmatory models with resource tier as primary predictor (reference: High). Right: exploratory models with region as primary predictor (reference: India). Both include complexity, ambiguity, and model (reference: Gemini 2.5 Pro) as covariates. Dark blue diamonds: $p < 0.05$; grey circles: $p \geq 0.05$.

the inverse condition. Classified as LR, it shares the Latin script, regional cultural context, and typological proximity of Italian, a high-resource (HR) language; Sicilian also lacks a standardized orthography, and contemporary written use spans multiple spontaneous spelling conventions (Gyllstad and Kupisch 2025). Italian-language data may therefore provide partial lexical, orthographic, and cultural support that Sicilian token counts alone do not reflect, consistent with evidence that cultural knowledge can transfer across typologically related languages during adaptation (Zhang, Liao, and Feng 2025). The Pakistan languages lack an equivalent HR neighbor: Urdu and Sindhi share the same script constraints with no HR Perso-Arabic language in our sample. This contrast motivates *culturally situated resource scarcity*: the degree to which training data represents a language as used by a specific community, script, domain, and register. Our findings are consistent with this framing, but we cannot measure script-level or cultural data density from model outputs alone and treat this as a hypothesis for future investigation via training-data audits and targeted probing.

Checklist Pattern Mismatch

Across all 2,827 checklist survey responses, the most frequently flagged mismatch category is names, places, or objects not reflecting the target culture (14.6% of responses), followed by awkward or unnatural phrasing (12.4%) and an unfamiliar math problem format (12.2%; full table in Appendix). The prominence of the math format item in the top three is also notable: it is a Math & Format category flag, not a cultural fit concern, appearing ahead of items like social norms (9.3%), idioms (9.2%), and humor (8.2%). Approximately 37% of responses selected “none of the above,” indicating that a substantial minority of evaluations were judged culturally appropriate across all categories.

The resource tier does not predict whether a mismatch will be flagged. In the confirmatory logistic regression predicting any mismatch presence (full table in Appendix), neither mid-resource ($OR = 0.89$, $p = 0.334$) nor low-resource tier ($OR = 1.06$, $p = 0.590$) differs from high-resource. The tier-level frequencies are consistent with this: the most-flagged items shift somewhat across tiers but the overall rates are similar and no tier shows a dramatically higher or lower flag rate. Ambiguity condition and cultural complexity are also null throughout (full table in appendix).

The region model tells a different story. In the exploratory specification, Italy shows a strong negative effect on mismatch presence: Italian evaluations are 46% less likely than Indian ones to flag any checklist item ($OR = 0.54$, 95% CI [0.44, 0.67], $p < 0.001$). Pakistan is not significant ($p = 0.165$). The AIC difference between the region and tier mismatch models is $\Delta AIC = -41.8$, considerably larger than the improvement seen in the Likert regressions, indicating that region is a better fit than resource tier.

Our checklist results show that the regional difference is one of type, not just rate (see Appendix for full figure). Indian languages produce concentrated mismatches like names, everyday practices, and Math & Format items (currency, measurement conventions). Each appears among the top four flags across all three Indian languages.

Italian raters produce a different profile. No item exceeds 7.2%, and the distribution is compressed across abstract concerns. Over-generalized representation leads, rather than concrete flags like wrong names or currency. Sicilian adds a further contrast, awkward phrasing leads all cultural fit categories, making it the only language in the dataset where a Language & Expression item tops the checklist.

Pakistan languages show another profile, they are dominated by identity and naming failures at rates higher than in any Indian or Italian checklist. Sindhi has the highest single-item rate in the dataset, with names or objects deemed culturally inappropriate at 24.3%. Urdu leads with awkward phrasing (15.5%) and names (15.0%), with math format also prominent (14.4%). These three profiles suggest that what raters notice and can articulate on a structured checklist varies by cultural context, not by resource tier. The per-language figures are available in the appendix.

Italian translations receive lower CA ratings than Indian translations, and the exploratory region model confirms that Italy is a significant negative predictor of CA ($OR = 0.74$, $p = 0.003$). Yet Italian raters flag fewer checklist items. The checklist and the rating are pointing in opposite directions. Italian raters are signaling that something is wrong with the translations, but the structured checklist categories do not capture what it is. There are two explanations: First, Italian translation failures may be more diffuse, distributed across abstract dimensions of register, tone, and cultural specificity that resist categorization into any single checklist item. Second, the checklist may have been developed with a category structure that provides better coverage of South Asian cultural mismatch patterns than European ones, an instrument-level gap rather than a translation-level one. Both explanations imply that the checklist alone is insufficient to characterize cultural mismatch in Italian. The open-ended error taxonomy, which we turn to next, provides a richer account.

Error Taxonomy

To interpret failures not captured by the checklist, we analyzed the open-ended survey responses using the 51-label error taxonomy. The taxonomy groups errors into Grammar, Culture, and Math categories and yielded 2,231 labeled errors across the seven languages (Full figure in appendix).

Grammar errors are the most frequent category overall. Fluency and awkwardness lead all labels at 9.6% of coded instances, followed by word choice and lexical mistranslation (8.2%) and unspecified grammar issues (6.7%; Figure 3). Culture errors account for the next largest share, with name form (6.1%) and unspecified cultural mismatch (6.0%) the most common labels. Math errors are the rarest supercategory: math problem confusing (5.8%) and unspecified math issue (4.6%) are the most frequent, with conversion and rate errors (2.2%) and numeric value errors appearing further down. This distribution, $grammar > culture > math$, is consistent across languages, but the magnitude of each supercategory and the specific labels that dominate vary substantially by language and region.

For Italian word choice and lexical mistranslation lead at 15.1%, highest single-label rate across languages, and the next three labels are all grammar errors. Culture errors are

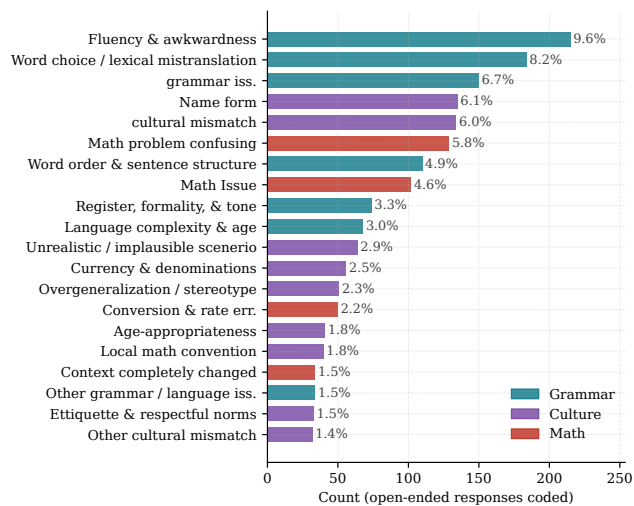


Figure 3: Top 20 error labels by frequency across all languages and models. Bars show raw counts; percentages represent the proportion of all coded observations. Colors indicate error category.

present but low-frequency; Math errors are rare. Italian also has the lowest total error count of any language (205 coded labels, vs. 454 for Sindhi and 362 for Punjabi; full figure in Appendix). Fewer errors are coded, yet Italian receives the lowest CA ratings among HR languages, a pattern consistent with errors that are felt but not easily surfaced by either structured or unstructured instruments.

Indian languages share a grammar-dominant profile, with fluency and awkwardness leading in Hindi (11.3%), Bengali (10.7%), and Punjabi (11.3%). Culture errors appear in the fourth or fifth slot in each, with the specific labels diverging by language: currency and units for Bengali, cultural mismatch and name form for Hindi and Punjabi.

Languages in Pakistan diverge most at the second-ranked label. Sindhi’s is name form (9.5%), the highest name-form rate in the dataset. Urdu’s is math problem confusing (9.9%), the only language where a Math error ranks this prominently, a pattern not visible in the confirmatory analyses (H3 was null) and not captured by the checklist. This suggests that math failures in Urdu surface more in unprompted feedback than in structured instruments. Sicilian follows a similar grammar-dominant structure to Italian but shows more cultural error diversity, including name form (8.6%) and material culture objects, placing its error profile closer to Pakistani languages on the identity dimension than to Italian.

Within Italy, Sicilian and Italian diverge at the label level. Italian’s top error is lexical precision; Sicilian’s top error is also word choice (10.9%), but name form follows closely (8.6%), a pattern absent from Italian. Sicilian also shows more cultural error diversity in the top fifteen labels than Italian does, including material culture objects and overgeneralization. The two languages share a region and a general grammar-dominance structure, but Sicilian’s error profile is closer to Pakistan languages on the identity/naming dimension than to Italian on the lexical one.

The qualitative translation data clarify what some of these coded labels mean in practice. Math word problem #568, the tooth fairy problem, illustrates the range of cultural substitution strategies for a problem where the original trigger, a child losing a tooth, had no direct cultural equivalent in most target contexts. Italian models split: Gemini 2.5 Pro correctly produced *il topolino dei denti* (the tooth mouse, Italy’s actual folk tradition), while GPT-4.1 and Claude Opus 4 used *la fatina dei denti* (the tooth fairy, an anglophone frame). South Asian models universally replaced the tooth fairy with a grandmother or elder giving money, preserving the intergenerational reward structure without a supernatural agent. Claude Urdu’s translation replaced the triggering event entirely: instead of losing teeth, the child completes sections of Quran memorization (*sipara*), with the grandmother rewarding each milestone. The math is preserved exactly. The cultural grounding is specific to Pakistani religious education practice.

Problem #131, the Santa village problem, shows a different substitution dynamic. Italian models preserved the original’s slapstick, children vomiting on workers and kicking their shins, because the Santa village context is culturally coherent in Italy. Every South Asian and Pakistani model removed this framing. Workers leave for heat, fatigue, festival obligations, or Eid crowds. The narrative register, slapstick comedy about unruly children, was systematically replaced by dignified departures across all three countries, regardless of language or resource tier. The math is unchanged, but the story has changed significantly.

These examples illustrate that translation errors in the taxonomy are not only surface-level linguistic failures. They include editorial choices about which cultural register is appropriate, which folk traditions are accessible, and what kinds of child behavior are suitable to depict in educational materials. Frequency counts capture how often each label appears; they do not capture which errors actually register in native speaker ratings. That question requires comparing the taxonomy directly with the Likert data, which is the subject of the next section.

Detection Gap Analysis

Figure 4 reports exploratory detection gap scores ($\Delta = \bar{r}_{\text{dim}} - \bar{r}_{\text{label}}$) for all taxonomy labels; positive values indicate errors that depressed ratings, values near zero indicate errors present but not penalized.

Grammar errors Nearly all grammar errors are detected. Unspecified grammar issues ($\Delta = 0.63$, $\bar{r} = 3.49$, $n = 150$), word choice and lexical mistranslation ($\Delta = 0.59$, $\bar{r} = 3.53$, $n = 184$), and other grammar errors ($\Delta = 0.59$, $\bar{r} = 3.53$, $n = 34$) show the largest gaps. Fluency, awkwardness, and redundancy, the most frequent label in the taxonomy overall, has $\Delta = 0.38$ ($\bar{r} = 3.74$, $n = 215$). Word order and sentence structure ($\Delta = 0.37$) and language complexity and age level ($\Delta = 0.32$, $n = 68$) are also penalized.

Three grammar labels are effectively invisible. Register, formality, and tone ($\Delta = 0.04$, $\bar{r} = 4.08$, $n = 74$) co-occurs with ratings at the dimension mean. Punctuation and non-math symbols ($n = 10$) and tense ($n = 16$) both sit at or

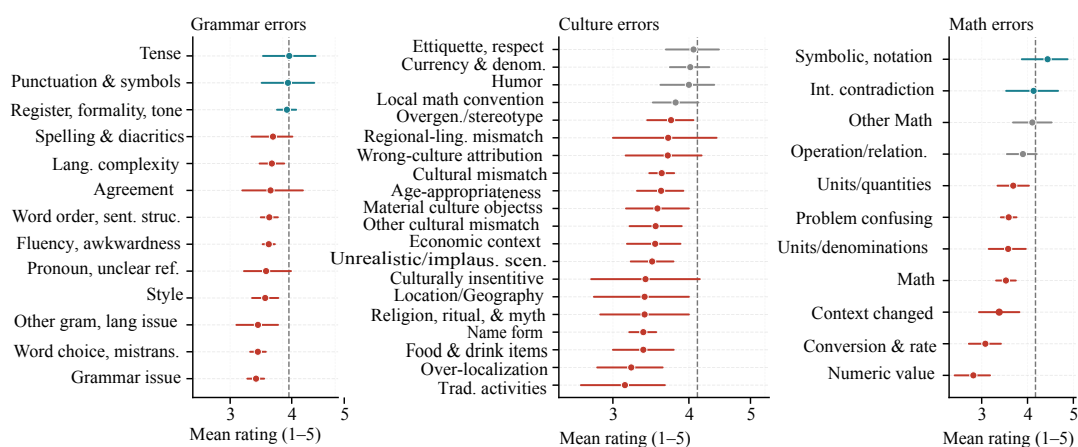


Figure 4: Detection gap by error label across Culture, Grammar, and Math supercategories, mapped to CA, LQ, and RP ratings respectively. Points show label-specific mean ratings with 95% bootstrap confidence intervals; the dashed line marks the overall dimension mean. Colors indicate no reliable gap, marginal gap, or reliable gap. Labels are sorted by mean rating within each panel, with larger gaps lower in the panel.

above the dimension mean ($\Delta = -0.02$ and $\Delta = -0.01$ respectively). These three labels describe errors that coders could identify in the text, but that native speaker raters either did not notice or considered meaningful enough to penalize.

Culture errors Culture errors show the widest range of detection across any supercategory. Traditional activities and games are the most detectable error label in the entire table: $\Delta = 0.95$, $\bar{r} = 3.16$ ($n = 19$). Over-localization is second: $\Delta = 0.87$, $\bar{r} = 3.24$ ($n = 29$). Name form, the single most common culture label in the taxonomy has $\Delta = 0.71$, $\bar{r} = 3.40$ ($n = 135$). Food and drink items ($\Delta = 0.71$, $n = 25$), unrealistic or implausible scenarios ($\Delta = 0.60$, $n = 64$), and material culture objects ($\Delta = 0.53$, $n = 29$) are all reliably detected. When raters can identify a specific, concrete referent, a wrong name, an unfamiliar food, or an implausible scenario, it shows up in their ratings. Three culture labels are nearly invisible. Currency and denominations ($\Delta = 0.09$, $\bar{r} = 4.02$, $n = 56$) co-occur with above-average CA ratings despite being the most common domain-specific culture error in the taxonomy. Etiquette and respectful norms ($\Delta = 0.05$, $n = 33$) and humor ($\Delta = 0.11$, $n = 24$) are similarly undetected. The currency finding is particularly notable given that “units or currency did not match culture” was among the top checklist items flagged by India-region raters. The two instruments diverge: the structured checklist prompts raters to flag currency issues, but when coders identify currency errors in open-ended responses, those errors do not correspond to lower CA ratings. This suggests that currency mismatches are salient when named explicitly but do not strongly disrupt the holistic impression of cultural fit.

Math errors Math errors show the most extreme detection values in the dataset, in both directions. Numeric value errors have $\Delta = 1.36$, $\bar{r} = 2.82$ ($n = 22$): evaluations where a coder identified an incorrect number receive a mean RP rating nearly 1.4 points below the dimension average. This is the largest detection gap in the table. Conversion and rate

errors is second: $\Delta = 1.09$, $\bar{r} = 3.08$ ($n = 50$). Context completely changed ($\Delta = 0.79$, $n = 34$) and math problem confusing ($\Delta = 0.59$, $n = 129$) are also strongly detected. When the math is concretely wrong, the numbers are off, the conversion is incorrect, the problem no longer makes sense, raters seem to notice. Symbolic and notation errors sit at the opposite extreme: $\Delta = -0.26$, $\bar{r} = 4.44$ ($n = 16$, interpret with caution). Evaluations where coders identified symbolic or notation errors received ratings above the dimension mean. Internal contradiction in scenario ($\Delta = 0.04$, $n = 15$, interpret with caution) is also essentially invisible. Both labels have small n and should be interpreted cautiously, but the directional finding is consistent with the figure: structural and notational math errors that do not disrupt the numerical answer go undetected.

Resolving the Italian reversal The detection gap analysis provides a direct account of the Italian reversal. Italian translations accumulate errors predominantly in the grammar supercategory, word choice ($\Delta = 0.59$), fluency ($\Delta = 0.38$), word order ($\Delta = 0.37$), but the single grammar label that dominates the Italian taxonomy profile, word choice and lexical mistranslation, is a relatively visible error on the LQ dimension. It, however, does not directly affect CA. Italian raters are rating cultural appropriateness, not linguistic quality, when they give CA scores. What is invisible in Italian is not the grammar profile, but the cultural dimension of the register: register, formality, and tone ($\Delta = 0.04$), an error that captures whether the translation sounds culturally natural in its level of formality and communicative style is effectively undetected. Italian translations feel off in register; raters sense this holistically but cannot articulate it as a specific cultural category on a checklist, and the detection gap confirms it does not depress ratings even when coders identify it in the text. Italian also has the lowest total error count in the taxonomy (205 labels), yet the lowest CA ratings among HR languages. Fewer errors are being coded, but

the errors that are present are of the kind that are hardest to detect. The combination of a grammar-dominant error profile, low culture-error frequency, and near-invisible register failures explains why Italian raters sense cultural inadequacy without being able to name it.

No single instrument is sufficient The three measurement instruments used in this study, holistic Likert ratings, structured checklist flags, and open-ended taxonomy coding, each capture a distinct and partially non-overlapping signal. Holistic ratings detect diffuse quality failures but cannot locate them in specific error categories. Structured checklists identify concrete, nameable mismatches but are insensitive to abstract failures of register, tone, and cultural texture. Open-ended taxonomy coding identifies specific error types, but, as the detection gap analysis shows, the presence of a coded error does not guarantee that raters penalize it. Triangulation across all three is necessary for a complete characterization of LLM cultural translation quality. The problems that perform worst across all measures simultaneously, Q500 (venue comparison), Q131 (Santa's elves), and Q557 (Subway lunch), are those where a concrete, nameable failure is also holistically felt, and where the open-ended coding confirms the specific error type. These cases provide evidence of systematic LLM failure in culturally situated educational translation (see appendix for convergence profile).

Limitations

More participants, problems, and languages would sharpen the picture, but each axis carries a cost. We sized the study to the budget that let us run it carefully. The seven languages themselves were chosen partly by resource tier and partly by in-lab translation expertise, since we needed team members to help us interpret open-ended responses, which biases coverage toward South Asian and Romance languages and leaves African, East Asian, and indigenous American languages, where culturally situated scarcity may take different forms, untested. With seven languages across three regions, regional effects and language-specific artifacts are not fully separable at this sample size. HPLT word counts are a proxy for training exposure, but the best available one: open models rarely report per-language token counts, and closed models report none. Finally, we cannot verify whether participants consulted translation aids mid-survey; the bilingual screen establishes baseline competence, not unassisted evaluation. We cannot inspect the evaluated models' training data, so culturally situated resource scarcity remains an interpretive framing rather than a demonstrated mechanism. Finally, the model snapshots themselves (Claude Opus 4, GPT-4.1, and Gemini 2.5 Pro) were captured in mid-2025, and a year in this field is a long time. Whether the newer models will be better or worse at culturally situated translation is something to be tested.

Conclusion

Our results show that region predicts translation quality better than language resource tier. Urdu, classified as mid-resource by token count, scores below low-resource Sicilian on all three evaluation dimensions. Punjabi, low-resource by

the same measure, scores above mid-resource Urdu on every dimension. Italian, one of the highest-resource languages in our study, receives lower cultural appropriateness ratings than every Indian language we evaluated. These reversals signal that what determines translation quality in our data is not solely how much text a model has seen in a language globally, but region better explains this. We hypothesize that quality depends on how well the text represents the specific community, script, domain, and register, we call this culturally situated resource scarcity. A language can be data-rich at the global level and culturally thin in the narrow situation that matters, a Pakistani classroom, a Sicilian neighborhood, an Italian elementary school mathematics problem. This finding is also consistent with our prior work, which shows that cultural entity adaptation patterns also cluster by geographic region (Suchdev and Lovato 2026). Resource tier is a useful first approximation of LLM capability across languages, but it is an insufficient unit of analysis for evaluating culturally situated educational translation. No single instrument was sufficient to characterize translation failures: Likert ratings, checklists, and open-ended responses each made different errors visible.

These findings have implications for practitioners using LLMs to adapt educational materials. The ease of prompting a model to translate and culturally adapt a word problem creates an asymmetry, the output of the model may look finished on the surface level, but the errors it creates may be invisible to anyone who does not already know the culture well enough to catch them. A teacher reviewing an LLM-adapted problem in Punjab may miss a currency misattribution; in our data, currency mismatches did not lower holistic ratings even when coders identified them. A teacher in Italy will sense that something is off in register and tone but may struggle to articulate what. Deploying LLM-adapted materials without structured review instruments may cause harm to students in the classroom. Our error taxonomy and checklist together provide a starting point for structured review, not as a replacement for expert human judgment, but as a scaffold that makes specific failure categories visible. Future work will explore why these models fail in the ways we have identified and explore steering mechanisms.

Future work will include an NLP comparison, providing a third evaluation layer alongside human ratings and corpus analysis. It will also explore this work in the classroom and collaborate with teachers, testing whether the detection gap we identified in a survey setting persists when the evaluator has a direct stake in the outcome. What this paper establishes is the main problem, namely, that the failures that matter most for educational settings, like regional misattribution, cultural compression, register failures that feel wrong are not easy to classify. Closing that gap requires better data, better evaluation instruments, a mechanistic understanding, and a clearer account of what culturally situated competence actually means for this task.

Acknowledgments

The author would like to thank the reviewers for their feedback on this project. We would also like to thank the study participants for their time. We would like to thank Alice

Patania, Protiva Sen, Aviral Chawlwa, and Robin Sharma for their assistance with interpreting and translating text into target languages. A special thanks to Robin Sharma for assistance in labeling. We also thank Julia Witte Zimmerman, Tabia Tanzin Prama, and Calla Beauregard for their guidance on the literature review. We also thank Kate Mays for her guidance on qualitative methods. Our team acknowledges support from the Alfred P. Sloan Foundation (Grant #G-2024-22498) and the National Science Foundation (Grant #2242829).

Ethical Considerations Statement

This study was approved by the University of Vermont Institutional Review Board (IRB protocol STUDY00003740). All participants provided informed consent prior to completing the survey. Participants were recruited through the Qualtrics survey panel and compensated at standard panel rates for their region. No deception was involved. Survey responses were collected without personally identifying information beyond demographic covariates (age, education, region of residence, native language). Data are stored on university-managed servers in accordance with institutional data governance policies.

The bilingual comprehension screen used to verify participant eligibility was designed to exclude non-bilingual respondents and bot responses; it was not designed to rank or evaluate participants' language proficiency beyond the minimum threshold required for the task.

Our study involves evaluating LLM outputs in languages and cultural contexts that the research team does not fully represent. We recruited native speakers as evaluators precisely because their judgment is more reliable than our own in those contexts, and we treat their ratings and open-ended responses as the primary evidence in this paper rather than researcher-imposed assessments.

Researcher Positionality Statement

The authors are researchers based at University of Vermont working at the intersection of natural language processing, computational social science, and education. One author is from Pakistan and speaks Sindhi and Urdu as native languages, providing direct cultural and linguistic familiarity with the Pakistan languages in our study.

Neither author is a native speaker of all seven languages studied. The selection of target languages was deliberately shaped by language expertise available within our lab: we chose languages spoken by lab members and collaborators so that we could draw on their knowledge during interpretation, qualitative analysis, and review of translated materials. This was a principled constraint rather than an incidental one, we did not want to interpret findings in languages where we had no internal check on our readings. It also means our language sample reflects the communities represented in our lab, which is itself a positionality worth naming.

Our interpretations of qualitative findings across all seven languages are grounded in participant-reported ratings and consultation with lab members rather than researcher-imposed linguistic judgment alone. Our institution and disci-

plines, as researchers at a North American university working in English-dominant NLP, may nonetheless shape which findings we foreground and how we frame them. The concept of "culturally situated resource scarcity" emerged from patterns in the data, but the vocabulary we use to describe it reflects our training in NLP and computer science.

Adverse Impact Statement

This paper examines the errors in LLM-generated culturally adapted math word problems. We identify conditions under which LLM outputs are likely to underserve specific student populations, and argue that current evaluation instruments are insufficient to capture the most consequential failures. We intend these findings to support more careful deployment of LLMs in educational settings.

The error taxonomy and checklist we developed could be misappropriated as a compliance checklist. This would be a threshold that, once met, certifies an LLM-adapted problem as culturally adequate. Our results do not support this use case. We show that complying with a structured checklist does not guarantee cultural appropriateness. The failures most consequential for educational equity are those that structured instruments miss. Human interpretation and care should be paid toward all personalized learning in collaboration with large-scale analysis like this one. Any deployment framework that treats our instruments as sufficient review rather than as one layer of a broader quality process would misread our findings.

References

- Besacier, L.; Barnard, E.; Karpov, A.; and Schultz, T. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech communication*, 56: 85–100.
- Bird, S. 2022. Local Languages, Third Spaces, and other High-Resource Scenarios. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7817–7829. Dublin, Ireland: Association for Computational Linguistics.
- Callison-Burch, C.; Osborne, M.; and Koehn, P. 2006. Re-evaluating the Role of Bleu in Machine Translation Research. In McCarthy, D.; and Wintner, S., eds., *11th Conference of the European Chapter of the Association for Computational Linguistics*, 249–256. Trento, Italy: Association for Computational Linguistics.
- Chandler, P.; and Sweller, J. 1991. Cognitive Load Theory and the Format of Instruction. *Cognition and Instruction*, 8(4): 293–332.
- Chiu, Y. Y.; Jiang, L.; Lin, B. Y.; Park, C. Y.; Li, S. S.; Ravi, S.; Bhatia, M.; Antoniak, M.; Tsvetkov, Y.; Shwartz, V.; and Choi, Y. 2025. CulturalBench: A Robust, Diverse and Challenging Benchmark for Measuring LMs' Cultural Knowledge Through Human-AI Red-Teaming. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 25663–25701. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.

- Chyung, S. Y.; Hutchinson, D.; and Shamsy, J. A. 2020. Evidence-based survey design: Ceiling effects associated with response scales. *Performance Improvement*, 59(6): 6–13.
- Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; Hesse, C.; and Schulman, J. 2021. Training Verifiers to Solve Math Word Problems. *Computing Research Repository*, arXiv:2110.14168.
- Cummins, D. D.; Kintsch, W.; Reusser, K.; and Weimer, R. 1988. The Role of Understanding in Solving Word Problems. *Cognitive Psychology*, 20(4): 405–438.
- de Gibert, O.; Nail, G.; Arefyev, N.; Bañón, M.; van der Linde, J.; Ji, S.; Zaragoza-Bernabeu, J.; Aulamo, M.; Ramírez-Sánchez, G.; Kutuzov, A.; Pyysalo, S.; Oepen, S.; and Tiedemann, J. 2024. A New Massive Multilingual Dataset for High-Performance Language Technologies. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 1116–1128. Torino, Italia: ELRA and ICCL.
- Duong, L.; Cohn, T.; Verspoor, K.; Bird, S.; and Cook, P. 2014. What Can We Get From 1000 Tokens? A Case Study of Multilingual POS Tagging For Resource-Poor Languages. In Moschitti, A.; Pang, B.; and Daelemans, W., eds., *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 886–897. Doha, Qatar: Association for Computational Linguistics.
- ELRA. 2026. Language Resources and Evaluation. <https://link.springer.com/journal/10579>. Accessed 2026-05-11.
- Faisal, F.; Ahia, O.; Srivastava, A.; Ahuja, K.; Chiang, D.; Tsvetkov, Y.; and Anastasopoulos, A. 2024. DIALECT-BENCH: An NLP Benchmark for Dialects, Varieties, and Closely-Related Languages. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14412–14454. Bangkok, Thailand: Association for Computational Linguistics.
- Freitag, M.; Foster, G.; Grangier, D.; Ratnakar, V.; Tan, Q.; and Macherey, W. 2021. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9: 1460–1474.
- Good, C.; Rattan, A.; and Dweck, C. S. 2012. Why Do Women Opt Out? Sense of Belonging and Women’s Representation in Mathematics. *Journal of Personality and Social Psychology*, 102(4): 700–717.
- Gyllstad, H.; and Kupisch, T. 2025. Creating vocabulary tests for minoritized languages with typologically close contact languages: Challenges and a road map. *International Journal of Bilingualism*, 13670069251382050.
- Hämäläinen, M. 2021. Endangered languages are not low-resourced! *arXiv preprint arXiv:2103.09567*.
- Havaladar, S.; Cho, Y. M.; Rai, S.; and Ungar, L. 2025a. Culturally-Aware Conversations: A Framework & Benchmark for LLMs. In Blodgett, S. L.; Curry, A. C.; Dev, S.; Li, S.; Madaio, M.; Wang, J.; Wu, S. T.; Xiao, Z.; and Yang, D., eds., *Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP)*, 220–229. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-353-1.
- Havaladar, S.; Stein, A.; Wong, E.; and Ungar, L. 2025b. Towards Style Alignment in Cross-Cultural Translation. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 32213–32230. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.
- He, Y.; Benhaim, A.; Patra, B.; Vaddamanu, P.; Ahuja, S.; Chopra, P.; Chaudhary, V.; Zhao, H.; and Song, X. 2025. Scaling Laws for Multilingual Language Models. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 4257–4273. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.
- Hershcovich, D.; Frank, S.; Lent, H.; de Lhoneux, M.; Abdou, M.; Brandl, S.; Bugliarello, E.; Cabello Piqueras, L.; Chalkidis, I.; Cui, R.; Fierro, C.; Margatina, K.; Rust, P.; and Søgaard, A. 2022. Challenges and Strategies in Cross-Cultural NLP. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6997–7013. Dublin, Ireland: Association for Computational Linguistics.
- Joshi, P.; Santy, S.; Budhiraja, A.; Bali, K.; and Choudhury, M. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6282–6293. Online: Association for Computational Linguistics.
- Karim, A.; Karim, A.; Lohana, B.; Keon, M.; Singh, J.; and Sattar, A. 2025. Lost in Cultural Translation: Do LLMs Struggle with Math Across Cultural Contexts? *Computing Research Repository*, arXiv:2503.18018.
- Kintsch, W.; and Greeno, J. G. 1985. Understanding and Solving Word Arithmetic Problems. *Psychological Review*, 92(1): 109.
- Kocmi, T.; Zouhar, V.; Avramidis, E.; Grundkiewicz, R.; Karpinska, M.; Popović, M.; Sachan, M.; and Shmatova, M. 2024. Error Span Annotation: A Balanced Approach for Human Evaluation of Machine Translation. In Haddow, B.; Kocmi, T.; Koehn, P.; and Monz, C., eds., *Proceedings of the Ninth Conference on Machine Translation*, 1440–1453. Miami, Florida, USA: Association for Computational Linguistics.
- Koncel-Kedziorski, R.; Konstas, I.; Zettlemoyer, L.; and Hajishirzi, H. 2016. A theme-rewriting approach for generating algebra word problems. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, 1617–1628.
- Krauwert, S. 2003. The basic language resource kit

- (BLARK) as the first milestone for the language resources roadmap. In *Proceedings of SPECOM*, 8, 15.
- Kreutzer, J.; Caswell, I.; Wang, L.; Wahab, A.; van Esch, D.; Ulzii-Orshikh, N.; Tapo, A.; Subramani, N.; Sokolov, A.; Sikasote, C.; Setyawan, M.; Sarin, S.; Samb, S.; Sagot, B.; Rivera, C.; Rios, A.; Papadimitriou, I.; Osei, S.; Suarez, P. O.; Orife, I.; Ogueji, K.; Rubungo, A. N.; Nguyen, T. Q.; Müller, M.; Müller, A.; Muhammad, S. H.; Muhammad, N.; Mnyakeni, A.; Mirzakhlov, J.; Matangira, T.; Leong, C.; Lawson, N.; Kudugunta, S.; Jernite, Y.; Jenny, M.; Firat, O.; Dossou, B. F. P.; Dlamini, S.; de Silva, N.; Çabuk Ballı, S.; Biderman, S.; Battisti, A.; Baruwa, A.; Bapna, A.; Baljekar, P.; Azime, I. A.; Awokoya, A.; Ataman, D.; Ahia, O.; Ahia, O.; Agrawal, S.; and Adeyemi, M. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10: 50–72.
- Longpre, S.; Kudugunta, S.; Muennighoff, N.; Hsu, I.-H.; Caswell, I.; Pentland, A.; Arik, S.; Lee, C.-Y.; and Ebrahimi, S. 2026. ATLAS: Adaptive Transfer Scaling Laws for Multilingual Pretraining, Finetuning, and Decoding the Curse of Multilinguality. [arXiv:2510.22037](https://arxiv.org/abs/2510.22037).
- Lovato, J.; and Suchdev, P. 2026. Evaluating Large Language Model Translations of Elementary Math Word Problems Across Cultures and Languages. Preregistration.
- McDonald, N.; Schoenebeck, S.; and Forte, A. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on human-computer interaction*, 3(CSCW): 1–23.
- Messick, S. 1994. The interplay of evidence and consequences in the validation of performance assessments. *Educational researcher*, 23(2): 13–23.
- Messick, S. 1995. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, 50(9): 741.
- Moreno García, L. D.; and Mangiron, C. 2024. Exploring the Potential of GPT-4 as an Interactive Transcreation Assistant in Game Localisation: A Case Study on the Translation of Pokémon Names. *Perspectives*, 1–18.
- Mousi, B.; Durrani, N.; Ahmad, F.; Hasan, M. A.; Hasanain, M.; Kabbani, T.; Dalvi, F.; Chowdhury, S. A.; and Alam, F. 2025. AraDiCE: Benchmarks for Dialectal and Cultural Capabilities in LLMs. In Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B. D.; and Schockaert, S., eds., *Proceedings of the 31st International Conference on Computational Linguistics*, 4186–4218. Abu Dhabi, UAE: Association for Computational Linguistics.
- Myung, J.; Lee, N.; Zhou, Y.; Jin, J.; Putri, R. A.; Antypas, D.; Borkakoty, H.; Kim, E.; Perez-Almendros, C.; Ayele, A. A.; et al. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37: 78104–78146.
- Nakov, P.; and Ng, H. T. 2009. Improved Statistical Machine Translation for Resource-Poor Languages Using Related Resource-Rich Languages. In Koehn, P.; and Mihalcea, R., eds., *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 1358–1367. Singapore: Association for Computational Linguistics.
- Naous, T.; Ryan, M. J.; Ritter, A.; and Xu, W. 2024. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 16366–16393. Bangkok, Thailand: Association for Computational Linguistics.
- Nigatu, H. H.; Tonja, A. L.; Rosman, B.; Solorio, T.; and Choudhury, M. 2024. The Zeno's Paradox of 'Low-Resource' Languages. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 17753–17774. Miami, Florida, USA: Association for Computational Linguistics.
- OECD. 2026. *OECD Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education*. Paris: OECD Publishing.
- Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Isabelle, P.; Charniak, E.; and Lin, D., eds., *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.
- Probst, K.; Brown, R.; Carbonell, J.; Lavie, A.; Levin, L.; and Peterson, E. 2001. Design and implementation of controlled elicitation for machine translation of low-density languages. In Krauwer, S., ed., *Workshop on MT2010: Towards a Road Map for MT*. Santiago de Compostela, Spain.
- Rei, R.; Stewart, C.; Farinha, A. C.; and Lavie, A. 2020. COMET: A Neural Framework for MT Evaluation. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2685–2702. Online: Association for Computational Linguistics.
- Resnik, P. 1999. Mining the Web for Bilingual Text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 527–534. College Park, Maryland, USA: Association for Computational Linguistics.
- Resnik, P.; and Smith, N. A. 2003. The Web as a Parallel Corpus. *American Journal of Computational Linguistics*, 29(3): 349–380.
- Singh, N.; Bernal, G.; Savchenko, D.; and Glassman, E. L. 2023. Where to Hide a Stolen Elephant: Leaps in Creative Writing with Multimodal Machine Intelligence. *ACM Transactions on Computer-Human Interaction*, 30(5): 1–57.
- Suchdev, P.; and Lovato, J. 2026. Who Brought Easter Eggs to Eid? Auditing Cultural Translation of Math Word Problems Across Diverse Languages and Regions. *arXiv preprint arXiv:2606.11009*.
- Sweller, J. 1988. Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science*, 12(2): 257–285.
- Terrell, K.; and DeBay, D. 2020. Stories & Technology: Gateways to Mathematics for All. *International Journal for Technology in Mathematics Education*, 27(1).

- Walkington, C. 2025. The implications of generative artificial intelligence for mathematics education. *School Science and Mathematics*.
- Walton, G. M.; and Cohen, G. L. 2007. A Question of Belonging: Race, Social Fit, and Achievement. *Journal of Personality and Social Psychology*, 92(1): 82–96.
- Yarowsky, D.; Ngai, G.; and Wicentowski, R. 2001. Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora. In *Proceedings of the First International Conference on Human Language Technology Research*.
- Zhang, C.; Liao, Z.; and Feng, Y. 2025. Cross-Lingual Transfer of Cultural Knowledge: An Asymmetric Phenomenon. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 147–157. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-252-7.
- Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.