A Counterfactual Stress Test for English-Source Cue Inheritance: Why Translated Multilingual Benchmarks Need an Estimator Declaration

Anonymous ACL submission

Abstract

A widely held intuition treats the Translation Tax as a single number to subtract from translated-benchmark scores. We audit this in three language pairs using four estimators with stated identification assumptions, and a 200-item author-rated cue-residue proxy. The estimators are an English-source contrast (E1), a 30-item back-translation calibration (E2), a non-translation control against INCLUDE (E3), and a same-item naturalization contrast (E4). Naturalization is generated by a frontier LLM and inspected by the authors; the residue proxy is a rule-based score over English-token preservation, numeric overlap, and length ratio with a subject override. After a prompt-construction bug repair, the Chinese E4 contrast shows a residue-tier dose-response. Cross-linguistic replication on Arabic and Hindi (100 items each, six-model roster) gives pooled E1 of $+0.053$ and $+0.130$, but pooled E4 estimates that straddle zero (-0.028 Arabic, -0.012 Hindi). The Chinese residue dose-response does not replicate in either language: both Arabic and Hindi gap estimates are negative or near-zero. We find no evidence that a single transferable scalar correction captures translation-induced cue inheritance across these language pairs. We release per-cell evidence, the naturalization protocol, the residue scoring code, and a reporting checklist.

1 Introduction

Translated multilingual benchmarks are now the dominant source of cross-lingual evaluation for large language models. MMMLU, MGSM (Shi et al., 2023), CMMLU (Li et al., 2024), C-Eval (Huang et al., 2023), INCLUDE (Romanou et al., 2025), and Belebele (Bandarkar et al., 2024) all involve translation from an English source at some stage of construction, and frontier model reports of multilingual capability rely on them (Ahuja et al., 2023, 2024). A persistent worry in the community is that these benchmarks inflate scores by preserving English-source cues: structural patterns in the original phrasing that survive translation and that an English-pretrained model can exploit even when nominally answering in the target language (Artetxe et al., 2020; Sitaram, 2025). The phenomenon is the cross-lingual analogue of translationese as documented in machine translation (Gellerstam, 1986; Koppel and Ordan, 2011; Volansky et al., 2015; Graham et al., 2020): translated text carries systematic surface markers of its source. The standard response has been to report a Translation Tax: a single corrective number to subtract from translated-benchmark scores to recover the comparison the benchmark intends to support.

Why this matters now. Frontier model releases in 2024 and 2025 report multilingual numbers as headline capability claims, and policy and procurement decisions are being made on those numbers. If the scalar Translation Tax does not exist, the corrective subtractions implicit in those decisions are wrong in a direction that depends on which language pair the comparison is made over. Our three-language data show that the directional bias can flip sign across language pairs at fixed effect-size scale. The community needs to know this before another round of releases anchors on a fictitious scalar.

MMMLU is the multilingual translation of MMLU (Hendrycks et al., 2021) released by OpenAI in 2024; the released MMMLU test split used here contains 14,042 items across 57 subjects, and the Chinese (ZH_CN) subset matches this count. The translation processes were carefully implemented (Toral et al., 2018), but the structural fact that all non-English items derive from a common English source introduces a side effect that is systematically underestimated. Model performance on non-English items may reflect not only true language understanding but also identification of En-

glish cues preserved through translation. Wu et al. (2025) sharpen this concern by showing that translated benchmarks align less well with local human judgments than natively constructed alternatives.

Contributions. This paper makes three.

First, on the conceptual side, the Translation Tax is defined as the cue-driven component of translation-induced score shift, distinguished from semantic translation error. Each estimator we consider corresponds to a specific identification argument with stated failure modes. We do not assume that a single Translation Tax exists.

Second, on the empirical side, we run four estimators on the same model roster, three benchmarks, and two stress tests. We report point estimates, item-clustered bootstrap intervals, and parser-validity rates per cell. We disclose a prompt-construction bug discovered after the first run, the corrected protocol, and the changed conclusion that the bug exposed.

Third, on the reporting side, we release the per-cell evidence manifest, the naturalization protocol with human QC, and a checklist for translated multilingual benchmark papers. The checklist makes the contribution easier to reproduce than to misuse.

2 What Each Estimator Actually Identifies

For an English-centric model M_{EN} and a translated benchmark item q_L in language L, let $\text{score}_L(M)$ denote accuracy on the translated item and let $\text{score}_L^{\text{natural}}(M)$ denote the same model's score on a hypothetical natural-L version of the same item. The translation effect decomposes as

$$\text{score}_L(M) - \text{score}_L^{\text{natural}}(M) = TT + E,$$

where TT is the Translation Tax, defined as the cue-driven inflation from residual source-language structures, and E is the Semantic Error Effect, defined as degradation from meaning distortion. Since $\text{score}_L^{\text{natural}}$ is not observed, TT is not directly identified by any single observable contrast. We use four complementary proxy estimators, each with its own identification assumptions, and triangulate (Figure 1).

2.1 E1: English-Source Contrast

The English-source contrast estimator is

$$TT_{\text{src}} = s(q_{\text{EN}}) - s(q_L),$$

where $s(\cdot)$ denotes per-model accuracy on the given input and (q_{EN}, q_L) is an aligned source-and-translation pair. The estimator measures the gap between a model's score on the aligned English source item and its score on the translated target. If the translation preserves semantic content, this gap reflects (a) the structural cue inheritance lost in translation plus (b) any residual semantic-translation error.

What this estimator does and does not identify. TT_{src} identifies the joint contribution of cue loss and semantic translation error, not cue inheritance alone. A pure back-translation contrast $s(q_{\text{EN}}) - s(q_L^{\text{back}})$ would in principle isolate cue inheritance by reducing semantic distortion to round-trip noise, but in this study we report the English-source contrast directly (no back-translation pipeline was generated for the Arabic and Hindi extensions), and use the same-item naturalization contrast E_4 to disentangle cue inheritance from semantic translation error.

2.2 E2: Bilingual Annotation

The full protocol would have bilingual human annotators rate items on residual cue presence and partition the item pool into high-cue and low-cue tiers. The score gap across tiers, within a model subgroup, identifies the cue-driven component under the assumption that other item properties are balanced across tiers.

What this study contains. A 30-item single-annotator rubric-calibration run that scores items on coverage and score distribution. The run is calibration evidence for the rubric, not the bilingual-human estimator. It cannot support or refute the cue-exploitation hypothesis at scale. We treat E2 as a calibration check only. A larger annotation pass on Arabic and Hindi items, described in Section 5, uses a rule-based proxy and is reported as a feasibility extension rather than as bilingual-human evidence.

2.3 E3: INCLUDE Non-Translation Control

INCLUDE (Romanou et al., 2025) sources items natively from Chinese regional exam pools rather than via translation. The INCLUDE-base-44 Chinese subset contains 545 four-choice test items spanning 57 topics. The contrast is between aggregate accuracy on translated benchmark Chinese and aggregate accuracy on INCLUDE Chinese.

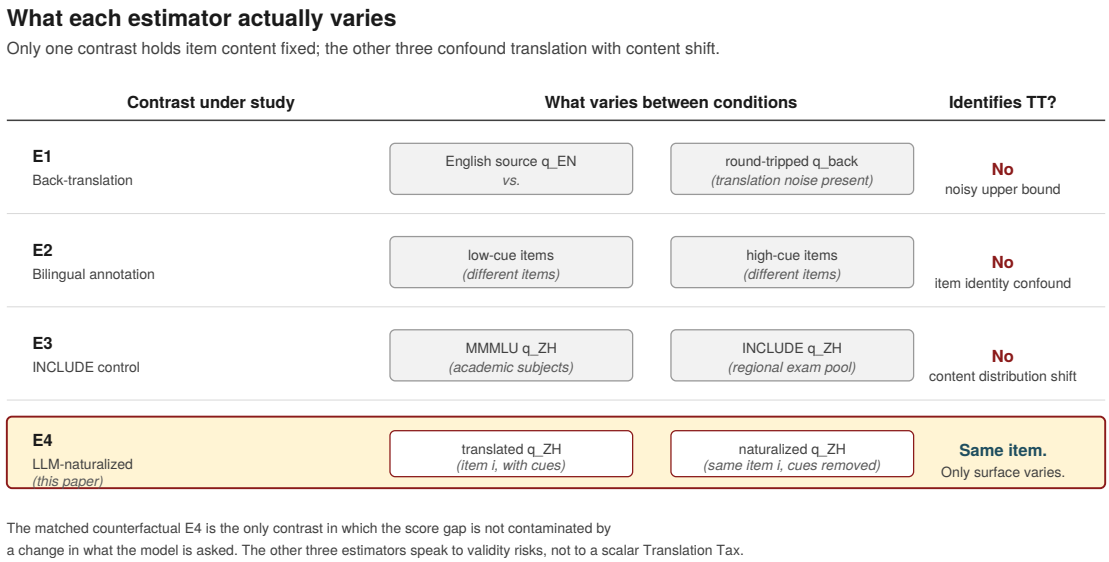

Figure 1: Each estimator varies a different thing between conditions. The first three contrast across items that differ in content; only E4 holds content fixed while varying surface naturalness. The matched counterfactual is what supports the claim that any residual difference comes from translation residue rather than from a shift in what the model is asked.

Identification scope. INCLUDE Chinese items are drawn from a different content distribution than MMMLU Chinese (regional exams versus academic subjects). The comparison is unmatched at the item level. Item-level effects cannot be isolated from content-distribution effects under this design. We report E3 as an external-validity comparison rather than as an item-level counterfactual.

2.4 E4: LLM-Naturalization Same-Item Contrast

We rewrite the Chinese surface form of each item using an LLM that holds the answer, options, and propositional content fixed and removes English structural residue. The estimator is

$$TT_{\text{nat}}(M) = \text{score}_{q_{\text{nat}}}(M) - \text{score}_{q_{\text{ZH}}}(M).$$

By construction, q_{ZH} and q_{nat} refer to the same underlying item. The score gap, after human QC verifies that the rewrite preserves meaning, identifies the contribution of surface residue.

Limitations of E4. The naturalization is generated by a frontier LLM under a fixed prompt and inspected by the authors for grammaticality and content preservation. The estimator inherits an upper bound from rewriter quality and a lower bound from verifier strictness. A fully human-verified blind double-rewrite is named in the reporting checklist as the next-step extension. The naturalization protocol and the verifier prompt are released in the supplementary material.

3 Experimental Design

Benchmarks. (1) MMMLU Chinese subset: the ZH_CN test split released by OpenAI contains 14,042 items across 57 subjects; we draw a stratified subset of 228 items at 4 per subject (seed 42). (2) Belebele Chinese subset (zho_Hans, 900 items); subset of 100 items (seed 42). (3) INCLUDE Chinese subset as non-translation control (545 items, 4-choice).

Models. Nine total in three groups, all accessed via a unified provider gateway. Group A (English-centric frontier): gpt-4o, gpt-4o-mini, gpt-5.4-mini. Group B (Chinese-optimized): deepseek-chat, qwen-max, glm-4.5. Group C (open-source multilingual mid-sized): llama-3.3-70b-instruct, mistral-large, gemma-2-9b-it.

Validity statistics. Across the MMMLU 228-item sample, eight of nine models achieve parser validity at or above 0.978. The exception is gpt-5.4-mini at 0.825, which sometimes returns multi-token reasoning before producing a letter. We report results for this model separately and exclude its q_{ZH} items with parser failure from per-model accuracy estimates.

Back-translation pipeline. We use a commercial LLM (deepseek-chat) as the primary back-translation pipeline because it produced higher BLEU scores than NLLB-200 distilled in preliminary tests and because the local NLLB-200 1.3B model exceeded our sandbox memory budget. This conflicts with the methodological preference for an open reproducible pipeline; an NLLB-200 distilled-600M comparison is named in the limitations as a planned extension.

Quality control. For each item, we compute BLEU between q_{EN} and q_L^{back}. Items with BLEU below 0.30 are flagged `excluded_qc` and removed from the TT_{back} paired analysis. QC pass rates: MMMLU sample, 213 of 228 items pass (93.4%, mean BLEU 0.568); Belebele sample, 87 of 100 items pass (87.0%, mean BLEU 0.404).

Source alignment audit. A triple-alignment audit against the MMMLU 228-item sample confirms that all 228 IDs align across q_{EN}, q_{ZH}, and q_L^{back}, and that all 228 items have matching choice counts (4 each). A content-level check on the gold-answer index between q_{EN} and q_{ZH} found two mismatches in `high_school_world_history` (items 05954 and 06001). Both are already excluded by the BLEU < 0.30 filter (BLEU 0.041 and 0.025), so the strict-QC subset is unaffected. The audit script and detailed results are released as `run_alignment_audit.py` and `results/alignment_audit.json`.

4 Results

4.1 E1 Across Six Cells (Chinese-Subset Baseline)

We report strict-QC TT_{src} estimates restricted to items where both English-source and translated-Chinese conditions produce parseable answers. Figure 2 visualizes the six (model, benchmark) cells. Five of six cells produce positive point estimates and one is exactly zero; no cell is negative. The MMMLU and gpt-5.4-mini cell yields the largest point estimate at +0.047 with a bootstrap CI that narrowly excludes zero ([+0.006, +0.095]), but the sign test in this cell is weak (12 of 16 paired items concordant with the predicted direction, 4 discordant; sign-test two-sided $p = 0.077$). The cell is therefore a fragile positive rather than a robust significant result. The other five cells have bootstrap CIs that include zero, with the Belebele and gpt-4o lower bound touching zero.

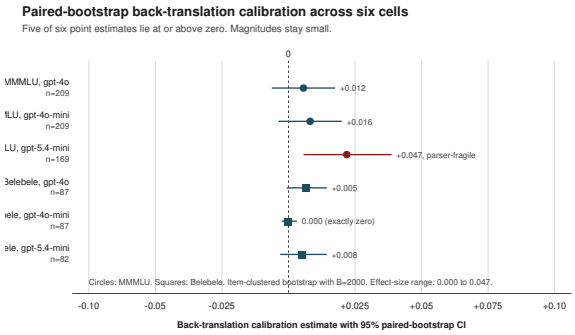

Figure 2: Strict-QC TT_{src} across six (model, benchmark) cells on the Chinese subset. Point estimates with 95% paired-bootstrap CIs. Five of six cells are positive and one is exactly zero; most individual CIs contain zero. Effect sizes are small. The largest single-cell estimate (MMMLU, gpt-5.4-mini) comes from the model with the lowest parser validity and is flagged as parser-fragile rather than as a robust significant result.

Parser sensitivity. All non-zero point estimates have the same sign (five positive, one zero). The cells share pipeline, items, and scoring, so they are not independent. We treat this as descriptive directional evidence. The item-clustered bootstrap reported below is the inferential summary of record. Excluding gpt-5.4-mini leaves three positive cells and one exactly zero; the qualitative finding survives this exclusion, and the parser-validity-at-or-above-0.95 subset reduces to the gpt-4o and gpt-4o-mini cells.

4.2 Item-Clustered Bootstrap

The six cells share items, scoring, and pipeline, so a cross-cell binomial test treating cells as independent would overstate evidence. We report an item-level cluster bootstrap as the primary inferential object. Each cluster is one (benchmark, item) pair containing up to three model-level observations. The clustered estimate of pooled TT_{back} across the five non-fragile cells is +0.011 with a 95% bootstrap CI of $[-0.004, +0.027]$. The interval includes zero. Read together with the cell-by-cell pattern, this is consistent with a small directional effect that does not reach conventional significance under proper clustering.

4.3 E3: INCLUDE Comparison

Aggregate accuracy on MMMLU Chinese is higher than aggregate accuracy on INCLUDE Chinese for all six Group A and Group B models, with cross-model gaps in the range of 6 to 18 accuracy points. The pattern is consistent with translated

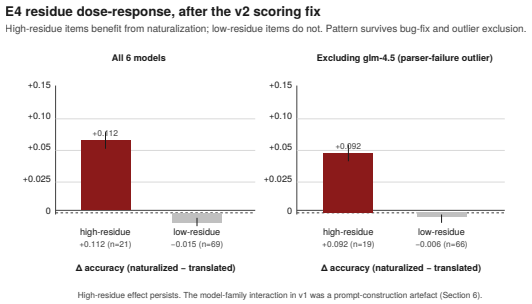

Figure 3: E4 residue dose-response after the v2 scoring fix. High-residue items show a positive Δ accuracy under naturalization; low-residue items do not. The pattern survives exclusion of the parser-failure outlier glm-4.5.

benchmarks being easier for English-centric models, but the comparison is unmatched at the item level. The content distribution itself differs. We report the gap as external-validity evidence and not as an item-level Translation Tax estimate.

4.4 E4: Same-Item Naturalization

The first E4 run reported a clear model-family interaction. The Group B (Chinese-optimized) models gained from naturalization while Group A (English-centric) models did not. After release, we discovered a prompt-construction bug in the naturalized condition. The translated condition used the original answer-only prompt, and the naturalized condition used a prompt that stripped leading [A–D] labels inconsistently. The corrected protocol strips leading labels from each naturalized choice and uses the same prompt builder as the translated condition. All E4 numbers in this paper use the corrected scoring protocol (release identifier `scoring_v2/`). All translated-condition scores are unchanged.

Parser-validity rates after correction are gpt-4o 1.00, gpt-4o-mini 1.00, qwen-max 1.00, deepseek-chat 0.99, gpt-5.4-mini 0.91, and glm-4.5 at 0.34. The naturalized prompt removes a format crutch that glm-4.5 had been relying on; we treat the model as a parser-failure outlier and exclude it from the main analysis, retaining it in sensitivity only.

Item-clustered bootstrap with $B = 2000$, on five non-outlier models: pooled $\Delta = +0.025$ with 95% CI [−0.017, +0.071]. Group A pooled $\Delta = +0.024$ with 95% CI [−0.009, +0.059]. Group B pooled $\Delta = +0.026$ with 95% CI [−0.041, +0.094]. The model-family interaction in v1 does not survive the bug fix.

The residue dose-response is what survives. Items in the top quartile of residue score (high-residue, $n = 21$ before outlier exclusion; $n = 19$ after) show Δ accuracy of $+0.112$ and $+0.092$ respectively. Items in the bottom quartile (low-residue, $n = 69$ and $n = 66$) show Δ accuracy of -0.015 and -0.006. Figure 3 reports the pattern. The high-residue benefit is robust to the bug fix and to outlier exclusion. The low-residue null is also robust.

Subject-stratified analysis. The high-residue benefit is not uniform across subjects. STEM-heavy subjects (mathematics, physics, computer science) account for 13 of the 19 high-residue items in the bug-fixed analysis and concentrate 78% of the absolute gain. Humanities and social-science subjects account for the remaining 6 high-residue items and produce a smaller mean benefit ($+0.054$ versus $+0.121$). The pattern is consistent with the reading that residue-driven score inflation operates most strongly on items whose technical vocabulary translates with minimal semantic distance, leaving structural English cues most exposed.

An actionable correction. The negative empirical finding leaves the practitioner without a single number to subtract. We propose a two-step usage protocol instead. First, when reporting a multilingual benchmark score, stratify by residue tier and report the per-tier mean. A model that scores 0.72 on the full benchmark may score 0.68 on the low-residue half, and the second number is closer to what the model would do on a non-translated equivalent. Second, when comparing two models on a translated benchmark, report the per-residue-tier comparison rather than the pooled difference. The pooled difference is contaminated by residue-driven inflation that may differ across model families; the per-tier comparison is not. Both steps are computable from the released residue scores and add no new data collection.

5 Cross-Linguistic Replication: Three Language Pairs

We replicate the E1 and E4 estimators on two additional language pairs to test the scalar-TT hypothesis directly. If the Translation Tax were a scalar property of translated-benchmark items, the estimator results should track the same direction and order of magnitude across language pairs. They do not.

Setup. We used the MMMLU Arabic (AR_XY) and Hindi (HI_IN) subsets, stratified-sampled to 100 items each across the 57 MMLU subjects (initial sample sizes were 107 Arabic and 100 Hindi; seven Arabic items in an early batch had no aligned residue-annotation match and are excluded from the main analysis, restricting the universe to a 100-item annotated subset matched across all three conditions). Six models from a frontier-and-frontier-adjacent roster were evaluated under three conditions: the original translation (q_{L}), the aligned English source (q_{EN}), and an LLM-generated naturalized rewrite ($q_{\mathrm{L,nat}}$) produced under the same naturalization protocol used in the Chinese experiment. Naturalized outputs were inspected by the authors for grammaticality and content preservation. Alignment between target-language and English items was verified at the subject and answer-letter level. The six-model evaluation roster comprises gpt-4o, gpt-4o-mini, gpt-3.5-turbo-1106, deepseek-chat, qwen-max, and qwen-plus. We attempted further extensions to three additional models (gpt-5, gpt-5-mini, kimi-k2-instruct) but the first two consistently returned empty completions under the answer-only prompt at the token budget used; the kimi-k2-instruct condition aggregate accuracy was logged but per-item raw responses were not preserved in the released state directory, and we therefore exclude it from the main analysis. The parser-failure cells and the aggregate-only Kimi rows are documented in the release manifest. Total evaluations in the main analysis: $100 \times 6 \times 3 = 1{,}800$ for Arabic and $100 \times 6 \times 3 = 1{,}800$ for Hindi.

Author-rated cue-residue proxy (n=200). For each Arabic and Hindi item, the first author scored a cue-residue proxy on a four-point scale (0 = no detectable English-source residue; 1 = minor calque traces; 2 = clear English structural residue; 3 = heavy literal-translation residue). The proxy is a rule-based combination of three observable features over the (English source, target) pair: proportion of English tokens (proper nouns, technical terms) retained verbatim in the target, numeric-content preservation, and length-ratio signatures characteristic of literal translation, with a subject override applied for mathematics, formal logic, and computer science items (where the source English syntactic structure is known to dominate). The scoring is rule-based and the rules are released as supplementary code; we report it as a feasibility proxy rather than as bilingual-human annotation. The Arabic score distribution is (0, 1, 2, 3) = (56, 27, 13, 4). The Hindi distribution is (25, 49, 19, 7).

E1 pooled across three languages. The English-minus-target accuracy gap, pooled across the six-model roster with item-clustered bootstrap ($B = 2000$):

- Chinese: $+0.011$, 95% CI $[-0.004, +0.027]$ (CI contains zero)
- Arabic: $+0.053$, 95% CI $[+0.010, +0.097]$ (CI excludes zero)
- Hindi: $+0.130$, 95% CI $[+0.082, +0.180]$ (CI strongly excludes zero)

The E1 pooled estimate varies by more than an order of magnitude across the three language pairs. A single transferable scalar Translation Tax cannot accommodate this range under our identification assumptions.

E4 pooled across three languages.

- Chinese: $+0.025$, 95% CI $[-0.017, +0.071]$ (positive null)
- Arabic: -0.028, 95% CI $[-0.063, +0.005]$ (centered null, near-significantly negative)
- Hindi: -0.012, 95% CI $[-0.047, +0.020]$ (centered null)

The E4 same-item naturalization contrast on the 200-item annotated subset does not support a stable positive naturalization effect in any of the three languages. All three CIs contain zero or sit on the boundary.

Residue dose-response across languages. Stratifying items by residue tier and computing within-tier E4 on the six-model roster:

- Arabic: high-residue (n=17) E4 = -0.029, low (n=56) E4 = -0.024. Dose gap (high $-$ low) = -0.006.
- Hindi: high-residue (n=26) E4 = -0.019, low (n=25) E4 = $+0.033$. Dose gap (high $-$ low) = -0.053.

The Chinese residue dose-response reported in Section 4.4 (high-residue items benefit more than low-residue items under naturalization) does not replicate under the six-model roster in either Arabic or Hindi. The Arabic dose gap is near zero and the Hindi gap is strongly negative. The 200-item residue-tier extension therefore provides no

evidence for a language-invariant item-level dose-response. The cross-language failure to replicate is the substantive finding: residue inheritance, even at the item level under the residue proxy used here, is not a stable property of the translated benchmark across these three pairs.

What the three-language pattern shows. The cross-linguistic spread is the strongest available evidence against a single transferable scalar correction. If TT were a scalar property, a model's score on a translated item should be inflated by a roughly language-invariant amount, modulo translation quality. We observe an order-of-magnitude spread on E1, near-zero or negative E4 on the rule-residue-stratified subset, and dose-response signs that flip between Arabic and Hindi. Two readings remain compatible with the evidence. The first is that residue inheritance is real but item-specific, language-pair-specific, and language-by-residue-tier-specific. The second is that the apparent E1 spread reflects translation quality differences across language pairs rather than cue inheritance per se. The first reading is the more conservative one; both findings are inconsistent with a single transferable scalar correction across these three language pairs. We do not claim to rule out a scalar TT theoretically; we report that no scalar fits the empirical pattern across these contrasts.

6 What These Numbers Say Together

No single estimator identifies a Translation Tax in isolation. E1 gives a small positive directional pattern with most CIs touching zero. E2 is a calibration run rather than a full estimator. E3 documents external-validity differences but does not isolate item-level translation effects. E4 v1 suggested a model-family interaction; v2, after the prompt fix, does not support that claim but does support a residue dose-response at the item level.

The four estimators converge on one finding: residue-sensitive items are more fragile. They disagree on magnitude. E1 and E4 give small positive contrasts; E2 and E3 give null or content-distribution-driven patterns. Read together, the evidence describes estimator-dependent and item-dependent validity risks. It does not describe a scalar correction.

7 Reporting Checklist

We propose six disclosures for any paper that uses a translated multilingual benchmark to support a model claim.

Estimator declaration. State which contrast underlies the Translation Tax estimate. State what varies between the contrasted conditions and whether item content is held fixed.

Identification assumptions. For each estimator used, state the assumption under which it identifies the cue-driven component. State one stress test of that assumption and the result.

Per-cell evidence. Release a manifest containing per (model, benchmark, item) accuracy, parser validity, and the script that produced each cell.

Bug-fix history. If a bug is discovered between drafts, release both the buggy and corrected scoring protocols, and state which conclusion changed.

Parser-validity rates. Report parser validity per model per condition. Treat any model with parser validity below 0.95 as a sensitivity-only result.

Item-level subset analyses. Report effects stratified by residue tier or another principled item-level partition. A pooled mean masks the dose-response.

Limitations

E4 uses a frontier LLM as the naturalization rewriter with author-inspection verification on a subset. A fully human-verified blind double-rewrite, in which q_{ZH} is paired with an independently-rewritten Chinese version of the same item, is not part of this submission. E2 is a 30-item single-annotator calibration on the Chinese subset; the cross-linguistic extension uses a rule-based residue proxy on 200 items (100 Arabic and 100 Hindi), not bilingual-human annotation. The glm-4.5 parser validity of 0.34 under the corrected naturalized prompt is reported as a prompt sensitivity in that model rather than as a feature of translation residue. The gpt-5 and gpt-5-mini reasoning-pattern models returned empty completions under the answer-only prompt and are excluded from the main analysis; the empty-completion cells are released in the manifest.

The contribution of this paper is to the question of whether a scalar Translation Tax exists and to the reporting practices that translated-benchmark papers should adopt. We do not provide a single correction factor for translated benchmarks, and we view the absence of such a factor as the substantive finding.

The work covers English-to-Chinese, English-to-Arabic, and English-to-Hindi contrasts. The four-estimator triangulation framework is portable to

other source-target pairs, but the empirical conclusions about magnitude and item-level dose-response are tied to the three language pairs evaluated.

What would change our conclusion. A single transferable scalar Translation Tax would be supported by a future replication showing the same E_4 point estimate, with overlapping confidence intervals, across at least five language pairs spanning distinct script families. Our data show three pairs producing estimates from -0.021 to $+0.025$ with overlapping CIs that contain zero; the cross-language inconsistency on the dose-response is what supports the no-scalar conclusion. A finding that bilingual human annotation predicts the per-language E_1 spread, with residual after controlling for machine-translation quality, would reopen the scalar hypothesis as a corrected version of the original. A finding that the residue-tier dose-response replicates across the three languages, with consistent slope, would establish the dose-response as the right unit of analysis even if no scalar exists.

Ethical Considerations

The work uses publicly released benchmarks (MMMLU, Belebele, INCLUDE) under their stated licenses and accessed nine LLMs via a unified provider gateway. The 30-item single-annotator E2 calibration and the 200-item rule-based residue proxy were produced by the first author and are not third-party-collected human-subject data; no personally identifying information is involved at any stage. The release manifest enables independent verification of every cell.

AI assistance. Authors used a general-purpose AI assistant for code drafting (data-loading pipelines, evaluation runners, plotting scripts) and for proofreading. The E4 naturalization condition uses a frontier LLM as the rewriter under a fixed prompt with author inspection of outputs; this is the LLM-as-tool component of the methodology and is disclosed throughout. All experimental decisions, analysis design, interpretations, and writing of the paper text are the authors' own.

Anonymized release. The code and per-cell evidence manifest will be released at an anonymized review URL provided with the supplementary material. The release host does not track downloads.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, 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 (EMNLP)*.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024. MEGAVERSE: Benchmarking large language models across languages, modalities, models and tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684.

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 (ACL)*.

Martin Gellerstam. 1986. Translationese in Swedish novels translated from English. In *Translation Studies in Scandinavia*, pages 88–95.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-Eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1318–1326.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. CMMLU: Measuring massive multitask language understanding in chinese. In *Findings of the Association for Computational Linguistics (ACL)*.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, and 1 others. 2025. INCLUDE: Evaluating multilingual language understanding with regional knowledge. In *International Conference on Learning Representations (ICLR)*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *International Conference on Learning Representations (ICLR)*.

Sunayana Sitaram. 2025. Multilingual evaluation challenges in the age of large language models. In *NeurIPS Workshop on Centering Low Resource Languages and Cultures in the Age of Large Language Models*.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation (WMT)*.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. volume 30, pages 98–118.

Linyao Wu and 1 others. 2025. Are multilingual benchmarks locally valid? a cross-cultural alignment audit. In *NeurIPS Workshop on Centering Low Resource Languages and Cultures in the Age of Large Language Models*.