

Same Problem, Different Answer: Accuracy Hides Cross-Lingual Inconsistency in Multilingual Inference

Ahsan Habib Shihab¹, Assel Ospan¹, Aman Mussa¹, Madina Mansurova²

¹Department of Artificial Intelligence and Big Data, ²Faculty of Mechanics and Mathematics

Al-Farabi Kazakh National University, Almaty, Kazakhstan

ahsan.h.shihab@gmail.com, assel.ospan@kaznu.kz

mussa.aman@kaznu.kz, madina.mansurova@kaznu.kz

Abstract

Multilingual natural language understanding is almost always reported as *per-language accuracy*: a model is judged reliable in a language if it scores well there. This lens overstates reliability, because it aggregates over items and hides *item-level cross-lingual inconsistency*: the same reasoning problem receiving different answers in different languages. On the parallel XNLI benchmark (identical premise–hypothesis pairs in 15 languages) we evaluate two off-the-shelf zero-shot multilingual NLI models from different architecture families. Despite per-language accuracies of 80.8 % and 71.3 %, the models return a unanimous answer across the 15 languages for only 38.9 % and 21.9 % of test problems; the prediction flips across languages for a majority of items (61.1 %, 78.1 %), and agreement beyond chance is only moderate ($\kappa = 0.70, 0.55$). These flips are typically *silent errors*: on most split items the cross-lingual majority is correct (80.2 %, 76.0 %), so a reader in the dissenting language receives a wrong answer the model gets right elsewhere, and they fall disproportionately on lower-resource languages (flip involvement vs. a resource proxy: $r = -0.67, -0.69$). Contrary to what the negation-is-hard literature would predict, this instability is *not* concentrated on explicit pragmatic cues: across both models, items containing negation or quantifiers flip slightly *less* than cue-free items, not more. Item-level cross-lingual consistency, not aggregate accuracy, is the reliability metric that matters when the same content must be understood identically across languages. Because the instability is pervasive rather than localized, it cannot be mitigated by flagging “hard” inputs.

1 Introduction

In high-stakes multilingual settings (a clinical instruction read in Swahili and English, a legal notice issued in Urdu and French, a public-health advisory pushed to a population that speaks fifteen

languages), the same propositional content must be understood the same way regardless of the language it arrives in. A system that is reliable in one language but draws a different conclusion from the translated version of the same case is a source of miscommunication risk, even if it is “accurate” everywhere.

Yet the standard way we measure multilingual language understanding is *per-language accuracy*: a model is evaluated separately in each language and declared reliable in a language if it scores well there (Conneau et al., 2018; Ahuja et al., 2023; Bandarkar et al., 2024). Accuracy aggregates over items, and this aggregation hides a distinct failure mode. Two languages can achieve *identical* accuracy while disagreeing on *which* items they get right; conversely, a model can answer a given problem correctly in one language and incorrectly in another without any drop in either language’s headline score. When the quantity of interest is whether the same input yields the same answer across languages, accuracy is the wrong measurement.

We make this concrete on Natural Language Inference (NLI), a canonical test of language understanding, using XNLI (Conneau et al., 2018): the exact same 5,010 premise–hypothesis problems, professionally translated into 15 languages and sharing one gold label per problem. This parallelism lets us ask an item-level question that per-language accuracy cannot: for a single reasoning problem, does the model return the *same* label when the problem is posed in Arabic, German, Swahili, Thai, and eleven other languages, and if not, where do the disagreements fall?

We run two off-the-shelf zero-shot multilingual NLI models from different architecture families: a small distilled model (MiniLM-L6) and a base model (mDeBERTa-v3-base). Cross-lingual answer *flips* are pervasive and invisible to accuracy reporting, and every qualitative result below replicates across both models. Despite per-language

accuracies of 80.8 % and 71.3 %, the models return a unanimous answer across the 15 languages for only 38.9 % and 21.9 % of problems; the prediction flips across languages for a majority of items (61.1 %, 78.1 %), and agreement beyond chance is only moderate ($\kappa = 0.70, 0.55$). On most flipped problems the cross-lingual majority is correct (80.2 %, 76.0 %), so a reader in the dissenting language receives an answer the model gets right elsewhere. The instability falls on lower-resource languages: flip involvement rises as pretraining resources fall ($r = -0.67, -0.69$), with English the most stable and Urdu and Swahili the least.

We also tested, and rejected, a natural hypothesis. Because NLI models are known to mishandle negation, we expected flips to concentrate on negation, modality, and quantifier cues. They do not: in both models, items whose source contains negation or a quantifier flip slightly less often than cue-free items (negation risk ratio 0.88 and 0.91; quantifier 0.92 and 0.93), and modality shows no effect. The instability is not localised to identifiable pragmatic pivots; it is pervasive.

Our contribution is one claim, replicated across models, and the evidence for it:

On parallel multilingual NLI, two multilingual models from different families change their prediction across languages for a majority of individual reasoning problems—silent errors that per-language accuracy conceals and that fall on lower-resource languages—and, contrary to the negation-is-hard intuition, this instability is pervasive rather than concentrated on explicit pragmatic cues.

We (i) formalise item-level cross-lingual consistency and its decomposition against accuracy; (ii) show the effect replicates across two architecture families, and localise where it falls (lower-resource languages) and where it does not (it is not cue-concentrated), using bootstrap and χ^2 tests; and (iii) provide a limited robustness check with three frontier LLMs (GPT-5.6, Claude Sonnet 5, and Gemini 3.1 Pro) on a fixed 300-item subset. The public repository contains the complete reproducibility artifacts needed to reproduce the reported results. Our aim is not a new model but a change of measurement.

2 Related Work

Multilingual understanding, measured by accuracy. Cross-lingual benchmarks such as XNLI (Conneau et al., 2018), and more recent parallel suites like BELEBELE (Bandarkar et al., 2024), together with large-scale evaluations of generative models (Ahuja et al., 2023), are almost universally summarised as per-language accuracy and its gap to English. This framing has driven progress in cross-lingual transfer (Conneau et al., 2020; He et al., 2023), but by construction it reports a marginal over items: it cannot see whether two languages that score alike do so on the same problems. We keep the same models and data but change the estimand from a per-language marginal to an item-level joint over languages.

Consistency of language models. Consistency (invariance of a model’s output under meaning-preserving input changes) has been studied mostly *within* a language. Elazar et al. (2021) show that English PLMs give inconsistent answers to paraphrased factual queries. Extending this *across* languages, Qi et al. (2023) introduce RankC and show that multilingual models encode factual knowledge inconsistently across languages, largely independently of accuracy. These works target *factual knowledge* (cloze/entity prediction). We instead target *reasoning*: whether the inferential relation a model draws between a premise and a hypothesis survives translation of that same premise and hypothesis. This is a different capability (inference, not recall) and, we argue, a more direct probe of understanding.

Negation, modality, and pragmatics in NLI. A recurring finding is that NLI models mishandle negation: Hossain et al. (2020) show negation is under-represented in NLI benchmarks and that models fail on negation-critical pairs, and Jeretič et al. (2020) show mixed competence on implicature and presupposition. Most recently and most closely related, Vrabcová et al. (2025) build a multilingual NLI negation dataset and report that both accuracy and negation *robustness* are language-dependent. Our work is complementary and distinct in three ways. First, our object is cross-lingual *agreement* on identical items, not per-language accuracy or robustness. Second, we do not restrict attention to negation: we test negation against modality and quantification within one consistent design, and find, contrary to that literature’s implication,

that cross-lingual flips do not concentrate on these cues (if anything the reverse). Third, our central methodological point, that accuracy hides item-level instability, is a statement about the joint distribution of answers over languages, not about any single language’s accuracy.

Why this matters for high-stakes multilingual deployment. In domains such as health, law, and governance, meaning frequently turns on a single negation or deontic modal (“must not” vs. “should”), and the same content is routinely delivered across many languages (Anastasopoulos et al., 2020). This includes lower-resource languages such as Kazakh, where machine translation of legal and medical text remains challenging (Rakhimova et al., 2025) and general-purpose language models underserve speakers (Bazarbek et al., 2025), motivating dedicated data resources (Mansurova et al., 2026) and model adaptation (Mussa and Mansurova, 2025). In such settings the operative reliability question is not “how accurate is the model in language ℓ ?” but “will the model reach the same conclusion for this case whether it is expressed in ℓ_1 or ℓ_2 ?” Item-level cross-lingual consistency answers that question, and we show it does so less reassuringly than accuracy alone suggests.

3 Method

3.1 Data: parallel NLI as a consistency probe

We use XNLI (Conneau et al., 2018), whose `all_languages` configuration stores the genuinely parallel structure we need: each item is one premise–hypothesis problem carried in all $L=15$ languages (ar, bg, de, el, en, es, fr, hi, ru, sw, th, tr, ur, vi, zh) with a *single* gold label $y_i \in \{\text{entailment, neutral, contradiction}\}$. The premises and hypotheses originate in English (from MultiNLI (Williams et al., 2018)) and were professionally translated, so item i denotes the *same* reasoning problem in every language. We report on the held-out test split ($5,010 \text{ items} \times 15 \text{ languages}$). We do not evaluate on the XNLI validation split: the public checkpoints we use were fine-tuned on it, so it is effectively training data for them (empirically we observed near-ceiling “accuracy” of > 0.99 there), and it cannot serve as a held-out test. We also avoid the per-language configurations, which are not row-aligned across languages. Replication is therefore across *models* on the same held-out test set, not across splits.

3.2 Models

We evaluate two off-the-shelf zero-shot multilingual NLI classifiers from different architecture families, so that any effect is not an artifact of a single model or family: **MiniLM-L6**, a small multilingual MiniLMv2 distilled from XLM-R (Wang et al., 2021; Conneau et al., 2020), and **mDeBERTa-v3-base** (He et al., 2023), a base model that is competitive with much larger models on XNLI. Both are public checkpoints fine-tuned on English MNLI and XNLI and used exactly as released. We read each model’s label ordering from its configuration and map outputs to the XNLI label space; we run full-precision inference on CPU with fixed seeds and length-bucketed batching (which does not change predictions), at a maximum sequence length of 256 tokens, which covers 100% of XNLI pairs in every language (longest observed: 216 tokens) so that no language is disproportionately truncated. Predictions and class probabilities for every (item, language, model) are saved verbatim. We report accuracy only as a reference point under this single fixed protocol and make no claim to reproduce any previously published score; our argument concerns the relationship between accuracy and consistency *within* this protocol, which is applied identically to every language.

3.3 Metrics

Let $\hat{y}_i^{(\ell)}$ be the model’s predicted label for item i in language ℓ , and $\hat{\mathbf{y}}_i = (\hat{y}_i^{(1)}, \dots, \hat{y}_i^{(L)})$ the vector of the L answers to the *same* problem.

Accuracy. Per-language accuracy $a_\ell = \frac{1}{N} \sum_i \mathbb{1}[\hat{y}_i^{(\ell)} = y_i]$, summarised by its mean over languages. This is our reference point, not our contribution.

Item-level cross-lingual consistency. For each item we ask whether the L answers agree:

- **Flip rate** (non-unanimity): the fraction of items for which $\hat{\mathbf{y}}_i$ is not constant, i.e. at least one language disagrees. Its complement is **unanimity**.
- **Mean pairwise agreement:** averaged over items, the probability that two randomly chosen languages give the same answer, $\frac{1}{L(L-1)} (\sum_c n_{ic}^2 - L)$ with n_{ic} the number of languages predicting class c on item i .

- **Fleiss’ κ** (Fleiss, 1971), treating the L languages as raters of each item, to quantify agreement *beyond chance*.

Accuracy does not explain instability. To show flips are a distinct phenomenon, we (i) decompose items into *unanimous-correct*, *unanimous-wrong*, and *split*; (ii) among split items, measure how often the majority label is nonetheless correct (these are silent, language-specific errors); and (iii) correlate each language’s **flip involvement** (how often it disagrees with the per-item majority) with its per-language accuracy and with a pretraining-resource proxy. Because our models provide class probabilities, we also report the confidence of the minority (flipping) predictions, distinguishing *confident* disagreement from low-confidence coin-flips.

Pragmatic-cue localisation. We tag each item on its *English source* (the pre-translation original, so the tag is defined once and applies to the aligned item in every language) for the presence of **negation**, **modality**, and **quantifier** cues using transparent surface lexica (Appendix A). For each cue we compare the flip rate on cue-bearing vs. cue-free items, reporting the **risk ratio**, a 2×2 χ^2 test (Yates-corrected), and a 95% bootstrap CI over items for the difference. Surface lexica trade recall for transparency and reproducibility; we treat them as a lower bound on the true pragmatic signal.

3.4 Frontier-LLM robustness study

All primary conclusions rest solely on the open-weight NLI models above. As a bounded robustness check we additionally evaluate three frontier LLMs (GPT-5.6, Claude Sonnet 5, and Gemini 3.1 Pro; exact identifiers in Appendix B) on a fixed random sample of 300 test items (seed 42), preserving the full 15-language parallel structure. Every model receives an *identical* English instruction and is constrained to output exactly one of the three labels; the only thing that varies is the language of the premise and hypothesis. We score each instance from its first successful recorded response under matched decoding settings, parse the label automatically, verify validity, and compute the *same* metrics as above. We treat this study strictly as corroboration and do not let it alter the primary claims unless it is consistent with them.

Model	Acc.	Agree	Unan.	κ	Flip
MiniLM-L6	71.3	69.9	21.9	0.55	78.1
mDeBERTa-base	80.8	79.9	38.9	0.70	61.1

Table 1: Accuracy vs. item-level cross-lingual consistency on XNLI test (5,010 items $\times$ 15 languages). Acc = mean per-language accuracy; Agree = mean pairwise cross-lingual agreement; Unan = unanimity; κ = Fleiss’ kappa; Flip = non-unanimity rate. High accuracy coexists with low agreement.

4 Results

4.1 Accuracy is high and clustered

Table 1 (column 1) confirms the setup our argument depends on: the models are accurate. mDeBERTa-v3-base reaches 80.8 % mean per-language accuracy, with the smaller MiniLM-L6 at 71.3 %. Per-language accuracy is tightly clustered (Table 5): for mDeBERTa it ranges from 88.3 % (en) down to only 74.0 % (ur). On the standard lens, these systems would appear reliable.

4.2 The same problem gets different answers across languages

The item-level picture is very different (Table 1, columns 2–5; Figure 1). For mDeBERTa, the 15 languages are unanimous on only 38.9 % of test problems, so the prediction flips across languages for 61.1 % of items. Mean pairwise cross-lingual agreement is 79.9 %, and agreement beyond chance is weak: Fleiss κ = 0.70. The pattern holds for both models: the small MiniLM-L6 flips on 78.1 % of items and the stronger mDeBERTa-base on 61.1 %, so higher accuracy does not yield cross-lingual consistency. Aggregate accuracy in the 70–80% range coexists with per-item cross-lingual answers that are unstable for a large fraction of problems.

4.3 Flips are silent errors, not noise

Decomposing items into *unanimous-correct*, *unanimous-wrong*, and *split* shows the flips are consequential. Unanimous-wrong items are rare (1.0 % for mDeBERTa, 0.8 % for MiniLM); the model is seldom uniformly confused. The dominant category is *split* items (61.1 %, 78.1 %), and on most of them the cross-lingual majority is correct (80.2 %, 76.0 %), so a reader in the dissenting language receives a wrong answer for a problem the model answers correctly in most languages. This is the failure mode per-language accuracy cannot see. The two models differ in one respect we report

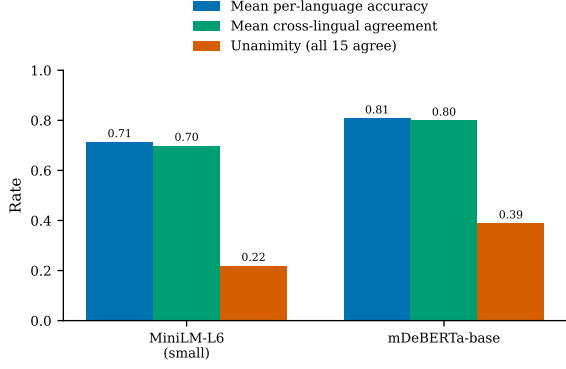


Figure 1: Accuracy hides instability. For each model, mean per-language accuracy (left) is far higher than cross-lingual agreement (middle) and unanimity (right): high headline accuracy coexists with frequent item-level disagreement across languages.

rather than average: for the stronger mDeBERTa, these dissenting predictions are usually confident (68.6 % are made with probability > 0.9), whereas for the smaller MiniLM only 25.0 % are, its flips sitting closer to the decision boundary. Confident silent errors are the sharper concern for the more accurate model. Table 2 makes this concrete: mDeBERTa returns the gold label in 14 languages and switches, at > 0.99 confidence, to contradiction in exactly one, an error no monolingual reader of that language could detect.

4.4 The instability is pervasive, not concentrated on pragmatic cues

Because NLI models are known to mishandle negation (Hossain et al., 2020; Vrabcová et al., 2025), the natural hypothesis is that cross-lingual flips concentrate on negation, modality, and quantifier cues. The data refute it. Figure 2 and Table 3 break the flip rate down by whether the item’s English source carries each cue. In both models, cue-bearing items flip no more (and for negation and quantifiers slightly less) than cue-free items: negation-bearing problems flip at risk ratio 0.88 (mDeBERTa) and 0.91 (MiniLM), quantifier items at 0.92 and 0.93 (all < 1 , Yates $\chi^2 > 12$, $df=1$), and modality shows no reliable effect (1.03, 1.00). Explicit negation and quantification, if anything, provide cross-lingual anchors: a strong, distinctive inference (e.g. negation signalling contradiction) that survives translation. The instability, by contrast, lives in the unmarked, semantically under-determined inferences that have no such anchor. This has a practical implication: because the insta-

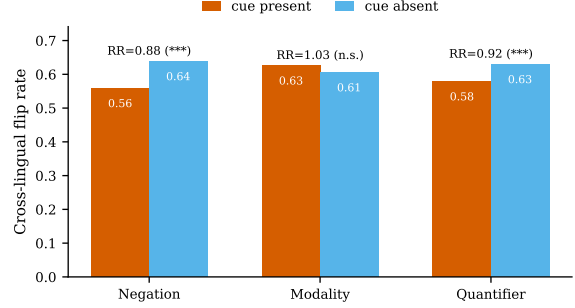


Figure 2: Cross-lingual flip rate on cue-bearing vs. cue-free items (mDeBERTa-v3-base). Contrary to the negation-is-hard intuition, items with negation or quantifier cues flip less than cue-free items; risk ratios (RR) and significance ($*p < .05$, $** .01$, $*** .001$) annotated. The pattern replicates in MiniLM (Table 3).

bility is pervasive rather than localised to identifiable hard inputs, it cannot be mitigated by flagging negated or quantified sentences for review.

4.5 The instability falls on lower-resource languages

Flip involvement (how often a language dissents from the per-item majority) rises as pretraining resources fall (Figure 3), correlating $r = -0.67$ (mDeBERTa) and $r = -0.69$ (MiniLM) with an ordinal resource proxy: English is the most stable language in both models and Urdu and Swahili the least. We do not overclaim here. A language’s dissent rate tracks its per-language accuracy very closely ($r = -0.95, -0.98$): the less accurate a language, the more it disagrees with the majority, so which languages carry the instability is largely a restatement of the familiar cross-lingual accuracy gap. The contribution is not this ranking but its consequence at the item level: the same reasoning problem receives systematically different answers depending on the language, and that inconsistency falls disproportionately on lower-resource languages, where additional human oversight may be especially important.

4.6 Robustness

Replication across models. Every result above is reported for both the distilled MiniLM-L6 and the base mDeBERTa-v3-base, two different architecture families at two accuracy levels (71.3 % and 80.8 %). The headline instability (Table 1), the silent-error decomposition, the resource asymmetry, and the non-concentration on pragmatic cues (Table 3) all hold in both; the only material differ-

Premise (English)	Hypothesis (English)	Gold	Cue	Dissent
But as she got older, um, she never acknowledged that she was wrong but she did change...	She never admitted that she was wrong.	ent	neg	ru:con
Please understand that Federal Regulations prohibit FAA personnel, XXXX Airlines, and ...	The FAA can't share some info about what they're doing.	ent	neg/qua	ur:con
Cala Mondrage is practically undeveloped (by the standards of coastal Mallorca), and m...	Cala Mondrage is not developed.	ent	neg/mod	el:con

Table 2: Confident single-language cross-lingual flips (mDeBERTa). Labels abbreviated ent/neu/con; “Dissent” is the one language that flips, with its (wrong) prediction. All dissents are made at >0.99 confidence.

Phenomenon		MiniLM-L6	mDeBERTa-base
Negation	cue present	73.4	55.9
	cue absent	80.6	63.9
	risk ratio	0.91	0.88
Modality	cue present	78.0	62.6
	cue absent	78.1	60.6
	risk ratio	1.00	1.03
Quantifier	cue present	74.5	57.9
	cue absent	80.2	63.0
	risk ratio	0.93	0.92

Table 3: Cross-lingual flip rate (%) on cue-bearing vs. cue-free items, and their risk ratio (RR), by source-side pragmatic cue and model. In both models, negation and quantifier cues yield $RR < 1$ (flip less); no cue increases flipping.

ence is the higher confidence of mDeBERTa’s flips (§4, “silent errors”). The effect is therefore not an artifact of a single checkpoint or family.

Frontier LLMs. On the fixed 300-item subset, all three frontier LLMs reproduce the central effect (Table 4): despite high per-language accuracy, they return a unanimous answer across the 15 languages for well under half of the problems and flip on a majority or near-majority, with only moderate chance-corrected agreement (accuracy / flip / unanimity / κ : GPT-5.6 80.1%/58.5%/41.5%/0.70; Claude Sonnet 5 85.2%/54.3%/45.7%/0.75; Gemini 3.1 Pro 85.1%/52.4%/47.6%/0.75). The open-weight models on the same 300 items behave comparably (mDeBERTa 59.0% flips), so the accuracy-hidden item-level instability is not an artifact of the smaller open-weight checkpoints; it persists for frontier reasoning LLMs. The 300-item subset is, however, underpowered for the source-cue analysis: no cue (negation, modality, quantifier) reaches significance for any model here (all Yates $\chi^2 < 3.4 < 3.84$), including the open-weight models whose full-set negation effect is highly significant (§4.4). The subset therefore neither confirms

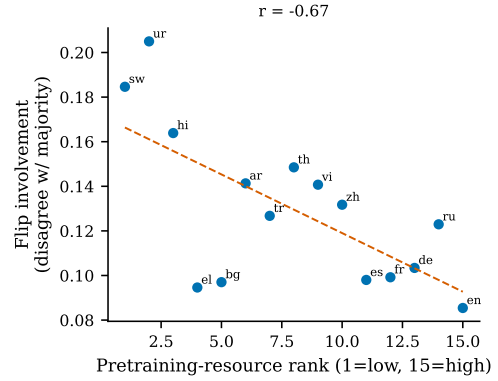


Figure 3: Per-language flip involvement vs. a pretraining-resource rank (mDeBERTa-v3-base). Lower-resource languages dissent from the cross-lingual majority more often.

nor refutes the (non-)concentration pattern for the LLMs; we treat the LLM study strictly as corroboration of the headline instability and base all cue-level and primary claims on the full open-weight test set.

5 Discussion

What the gap means. Per-language accuracy answers “how often is the model right in language ℓ ?”. It is silent on the question that matters when one population is served in many languages, or when a single case may be handled in different languages: “will the model reach the *same* conclusion regardless of the language of expression?” Our results show these questions have very different answers. Both models look uniformly reliable by the first measure while, by the second, changing their answer on a majority of individual problems, silent errors that land on lower-resource languages. This is the sense in which accuracy “hides” instability: item-level cross-lingual consistency is not recoverable from per-language accuracy, and it is the safety-relevant quantity we rarely report.

Model	Acc.	Agree	Unan.	κ	Flip	Neg. RR
MiniLM-L6	73.3	71.1	25.3	0.57	74.7	0.86
mDeBERTa-base	81.5	80.9	41.0	0.71	59.0	0.92
GPT-5.6	80.1	81.0	41.5	0.70	58.5	1.12
Claude Sonnet 5	85.2	83.6	45.7	0.75	54.3	0.98
Gemini 3.1 Pro	85.1	83.5	47.6	0.75	52.4	1.10

Table 4: Frontier-LLM robustness on the fixed 300-item subset (seed 42), compared to the open-weight models on the identical items. Acc = mean per-language accuracy; Agree = pairwise agreement; Unan = unanimity; Flip = cross-lingual non-unanimity; Neg. RR = negation flip risk ratio. The accuracy/consistency pattern persists for all frontier models; the cue risk ratios are not statistically significant on this 300-item subset (all $\chi^2 < 3.84$; see text), which is underpowered for cue analysis.

Why the instability is not where we expected.

We began with the standard intuition that negation, modality, and quantifiers (the cues that flip truth conditions with a single morpheme) would drive cross-lingual disagreement. The opposite is true: in both models, negation- and quantifier-bearing items are slightly more stable across languages. A plausible reading is that such cues are strong, distinctive inferential signals (negation typically forcing *contradiction*, universal quantifiers forcing *entailment* or *contradiction*) that translate robustly and anchor agreement, whereas the items that flip are the unmarked, semantically underdetermined inferences with no such anchor. This reframes the negation-is-hard literature (Hossain et al., 2020; Vrabcová et al., 2025): negation may be hard for *accuracy* within a language, but it is not the locus of cross-lingual *inconsistency*. The cross-lingual reliability problem is therefore diffuse, which is harder to guard against than a localized one.

A modest recommendation. Item-level cross-lingual consistency is cheap to report whenever a parallel evaluation set exists: it needs only the per-item predictions already produced during accuracy evaluation. We recommend that multilingual evaluations on parallel data report unanimity, pairwise agreement, and Fleiss κ alongside accuracy. Because the instability is pervasive rather than cue-localised, targeted safeguards (e.g. flagging negated inputs) will not suffice. Cross-lingual disagreement itself offers a free, unsupervised abstention signal: when a system’s answer would change with the language of the input, a human should be in the loop.

Implications for Human–Machine Understanding. Whether meaning is “represented, inferred, and negotiated” consistently across human and machine environments is, for multilingual models, an

empirical question. Our study gives one concrete, measurable answer: for current multilingual models, inference is not invariant to the language that carries it. Understanding that does not survive a change of language is a fragile basis for the high-stakes, cross-lingual communication these systems increasingly mediate.

Limitations

We measure cross-lingual *consistency*, not *correctness*: a model can be consistently wrong (unanimous-wrong items), and unanimity is necessary but not sufficient for reliability. Our headline claim is specifically that accuracy hides *instability*, not that consistency alone certifies a model.

The flip rate (non-unanimity across all 15 languages) mechanically increases with the number of languages compared, so its absolute magnitude should not be read as a per-pair disagreement probability; we therefore also report *pairwise* agreement and Fleiss κ , which are comparable across rater-set sizes, and our qualitative conclusions rest on those. Relatedly, *which* languages dissent tracks their per-language accuracy very closely (§4); the novel object is the item-level joint distribution of answers, not the per-language ranking.

Our pragmatic cues are detected with transparent surface lexica on the English source. This favours precision and reproducibility over recall: it misses morphological negation, implicit modality, and cross-linguistic realisations that differ from English, and it tags the *source*, not each translation. The reported cue effects are therefore best read as lower bounds on the true pragmatic signal.

XNLI is built from MultiNLI genres, not from actual clinical, legal, or governance text. We use it because it is the largest professionally-translated parallel NLI resource and because negation/modality/quantifier phenomena occur natu-

rally in it; the “high-stakes” framing motivates *why* item-level consistency matters, and we do not claim results transfer numerically to any specific safety-critical domain. Relatedly, some cross-lingual flips may originate in translation choices within XNLI rather than in the model; because the same translations are what a deployed system would receive, we regard this as part of the phenomenon, but it limits attribution to the model alone.

The open-weight models are zero-shot NLI classifiers fine-tuned on English MNLI and translated XNLI, which may bias them toward English-like realisations; we study 15 languages and two checkpoints, and larger or instruction-tuned models may behave differently (our frontier-LLM study is one probe of this, not a resolution). Finally, the LLM robustness study is deliberately small (300 items) and scores the first successful recorded response for each instance under the matched decoding settings; because GPT-5.6 supports only its default temperature, we cannot obtain greedy decoding for all models, so we cache all raw outputs to make scoring reproducible rather than claiming bit-level determinism of the API calls.

Ethics Statement

This work is an evaluation study on public, professionally-translated benchmark data (XNLI) and released model checkpoints; it involves no human subjects, no personal data, and no new data collection. Our motivating framing concerns high-stakes domains (health, law, governance), but we run no experiments on real clinical, legal, or governmental text and make no claims about any specific deployment. Readers should not treat our numbers as safety certifications for any product.

The central finding has a fairness dimension: item-level cross-lingual instability falls disproportionately on lower-resource languages, so a system judged reliable by aggregate accuracy may nonetheless deliver less consistent understanding to already-underserved language communities. We surface this so that multilingual evaluations do not mistake aggregate accuracy for equitable reliability. We also caution that “consistency” is not a virtue in itself (a model can be uniformly and confidently wrong) and should be reported together with accuracy and, ideally, human review on safety-critical inputs.

The frontier-LLM robustness study uses paid commercial APIs; we kept it to a fixed 300-item

subset to bound cost and environmental impact, cache all responses to avoid redundant calls, and report the exact model identifiers and settings used (Appendix B). All open-weight experiments run on CPU. The public repository contains the complete reproducibility artifacts needed to reproduce the reported results.

Acknowledgments

This research was conducted within the framework of the project “Developing a Large Language Model (LLM) to support the Kazakh language and technological progress” and was funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. BR24993001). AI-assisted tools were used during manuscript preparation for language editing and code assistance. All scientific decisions, analyses, and final text were reviewed and approved by the authors.

References

- Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4232–4267. Association for Computational Linguistics.
- Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Francisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Association for Computational Linguistics.
- Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabza. 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)*, pages 749–775. Association for Computational Linguistics.
- Zhaniya Bazarbek, Nazgul Toyganbaeva, Madina Mansurova, Talshyn Sarsembayeva, and Meruyert Sakypbekova. 2025. Developing a dataset for creating a large language model (LLM) for the Kazakh

- language. *Academic Scientific Journal of Computer Science*, 355(3):78–92. In Kazakh.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451. Association for Computational Linguistics.
- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.
- Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics (TACL)*, 9:1012–1031.
- Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations (ICLR)*.
- Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020. An analysis of natural language inference benchmarks through the lens of negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9106–9118. Association for Computational Linguistics.
- Paloma Jeretič, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPressive? Learning IMPLicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8690–8705. Association for Computational Linguistics.
- Madina Mansurova, Assel Ospan, Fakhridin Nuraliev, Rustam Khamdamov, and Talshyn Sarsembayeva. 2026. Building a deep search crawler for the Kazakh language: A reproducible web-scale pipeline. *Scientific Journal of Astana IT University*, 25:22–36.
- Aman Mussa and Madina Mansurova. 2025. Parameter-efficient fine-tuning of LLaMA 2 for the Kazakh language: Advancing low-resource language models. In *Proceedings of Tenth International Congress on Information and Communication Technology (ICICT)*, volume 1443 of *Lecture Notes in Networks and Systems*, pages 511–520. Springer Nature Singapore.
- Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 10650–10666. Association for Computational Linguistics.
- Diana Rakhimova, Aliya Zhiger, Madina Mansurova, Valentin Malykh, and XMagzhan Kairanbay. 2025. Improving post-editing of Kazakh translations with fine-tuned large language models: Dataset and evaluation. *International Journal of Innovative Research and Scientific Studies*, 8(8):220–233.
- Tereza Vrabcová, Marek Kadlčík, Petr Sojka, Michal Štefánik, and Michal Spiegel. 2025. Towards the roots of the negation problem: A multilingual NLI dataset and model scaling analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 25537–25551. Association for Computational Linguistics.
- Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021. MiniLMv2: Multi-head self-attention relation distillation for compressing pre-trained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2140–2151. Association for Computational Linguistics.
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1112–1122. Association for Computational Linguistics.

A Pragmatic-cue lexica

Cues are matched case-insensitively on the tokenised English source (premise + hypothesis). A tag fires if any listed marker is present.

- **Negation:** not, n’t (any token), no, never, none, nobody, nothing, nowhere, neither, nor, cannot, without, hardly, barely, scarcely.
- **Modality:** must, should, shall, may, might, can, could, would, ought, plus multiword have/has/had to, need(s/ed) to, supposed to, allowed to, required to.

- **Quantifier:** all, every, each, some, any, most, many, much, few, several, both, either, none, no, everyone/-thing/-body, someone/-thing, any-/no- pronouns.

On the test split the cue counts are: negation 1,747, modality 1,271, quantifier 1,878, any 3,258 (of 5,010 items). Categories may overlap (e.g. *none*); we also report an “any-cue” union.

B Frontier-LLM study details

Exact model identifiers (frozen at experiment time): GPT-5.6 = openai/gpt-5.6; Claude Sonnet 5 = anthropic/claude-sonnet-5; Gemini 3.1 Pro = google/gemini-3.1-pro-preview.

Access: Claude Sonnet 5 and Gemini 3.1 Pro were run through the OpenRouter API. For GPT-5.6, an initial subset of calls was completed directly through the OpenAI API; after that account’s quota was exhausted, the evaluation was switched to OpenRouter (openai/gpt-5.6) and resumed from cache, with identical prompt, temperature, minimal reasoning effort, and parsing. **Prompt:** a fixed English system instruction defining the NLI task and requiring a single word from {entailment, neutral, contradiction}, followed by the premise and hypothesis in the target language (template in the public repository). **Decoding:** temperature = 1 (the only value GPT-5.6 permits) with minimal/low reasoning effort where exposed, and a 256-token completion budget so the one-word answer is never truncated. We score the first successful recorded response for each instance under the matched decoding settings. **Parsing:** outputs are lowercased and matched to the three labels (exact, then prefix, then alphabetic-substring); unparseable outputs are recorded as invalid. Invalid rates were GPT-5.6 0.02%, Claude 0.00%, Gemini 3.1 Pro 0.09%. The sole GPT-5.6 invalid response (1 of 4,500, 0.02%; subset item 3545 in Swahili) was a reproducible empty completion (GPT-5.6 answered the same item correctly in all 14 other languages, every other model answered it, and an identical-protocol retry again returned empty), so we treat it as an isolated, input-specific generation anomaly, exclude that one item for GPT-5.6 by complete-case analysis, and note that it has no effect on any reported metric or conclusion. For 1,658 of the 4,500 GPT-5.6 instances the raw responses contain two successful answers (68 with different parsed labels); scoring deterministically uses the first successful response

in file order. **Subset:** 300 items sampled from the test split with seed 42, preserving all 15 languages (300×15 instances per model). All raw responses are committed to enable reproducible scoring.

C Reproducibility

All open-weight experiments use transformers and datasets with full-precision CPU inference, fixed seeds, and length-bucketed batching (which does not change predictions). We load XNLI’s all_languages parquet directly to guarantee cross-language row alignment. Every reported result value in the paper is emitted as a $\LaTeX$ macro by an analysis script that reads only the committed prediction files and the XNLI test data (gold labels, and the English source text used for cue tagging), so figures, tables, and in-text numbers cannot drift from the results. The public repository (<https://github.com/ahsan-h-shihab/same-problem-different-answer>; archived on Zenodo, DOI 10.5281/zenodo.22747570) contains the complete reproducibility artifacts needed to reproduce the reported results: run drivers, per-item predictions (.npz), raw LLM responses, analysis JSON, pinned dataset and model revisions, and a SHA256 provenance log covering the sampled subset, raw model predictions/API responses, parsed predictions, and final metrics, so every reported result can be tied to immutable experimental outputs.

D Full per-language results

Table 5 reports per-language accuracy for both models and, for the stronger model (mDeBERTa-v3-base), per-language flip involvement.

Lang	MiniLM-L6	mDeBERTa-base	Flip inv.
ar	68.7	80.2	14.1
bg	74.2	82.9	9.7
de	71.9	82.5	10.3
el	72.3	82.6	9.5
en	78.9	88.3	8.5
es	74.8	84.5	9.8
fr	74.2	83.4	9.9
hi	69.1	77.1	16.4
ru	71.4	81.3	12.3
sw	64.2	74.8	18.5
th	69.9	79.3	14.9
tr	69.6	80.7	12.7
ur	66.4	74.0	20.5
vi	72.3	79.6	14.1
zh	72.1	81.1	13.2

Table 5: Per-language accuracy (%) for both models and per-language flip involvement (%; mDeBERTa-v3-base) on XNLI test.