Biases in Translation: Assessing Opinion Distortion in Machine Translated Texts

Nazanin Shafiabadi, François Yvon
Sorbonne Université, CNRS, ISIR
75005, Paris, France
lastname@isir.upmc.fr

Abstract

Current machine translation (MT) evaluation practices largely assume that high lexical and semantic fidelity implies preservation of meaning. We question this assumption by introducing a framework for detecting and quantifying *translation-induced distortion*—the systematic alteration of a text's subjective properties during translation. Focusing on stance as a socially consequential property, we formalize stance preservation as an invariance problem and adapt two classical statistical tests, McNemar's test and the two-proportion Z-test, to diagnose systematic opinion shifts between source texts and their translations. Unlike standard MT metrics such as BLEU or COMET, which prioritize surface similarity and adequacy, our approach explicitly targets preservation of subjective meaning. In controlled experiments with synthetically distorted translations, we demonstrate that the proposed tests are sensitive to graded levels of stance manipulation. We apply our framework to evaluate twelve multilingual models and find that none reliably preserve stance across all tested language directions. Our findings reveal a critical gap in current MT evaluation practices and highlight the need for explicit evaluation of subjective meaning preservation in socially and politically sensitive contexts.

Keywords: Machine Translation, Stance Detection, Evaluation Metrics

1. Introduction

Machine translation (MT) is widely used to bridge language barriers in political discourse, news media, and social platforms. For instance, civic initiatives such as the *Conference on the Future of Europe* (CoFE) have enabled multilingual political deliberation by translating citizen proposals across EU languages. Newsrooms are likewise increasingly deploying MT to enable rapid multilingual coverage, as seen in *TVTechnology*'s live AI translation for real-time subtitling and cross-language news distribution (Dawson, 2025). Implicit in such deployments is the assumption that translation preserves not only literal meaning but also stance, opinion, or communicative framing. Yet this assumption is rarely tested, and evidence shows that even subtle translation artifacts can distort the speaker's intended position. Figure 1 illustrates an example where a seemingly minor lexical change introduced by translation—"government control" becoming "contrôle gouvernemental (*control by the government*)"—leads to a change in perceived opinion.

We propose a framework for detecting and quantifying distortion introduced by MT, measuring the extent to which MT systems preserve—or alter—the subjective meaning expressed in the source text. We focus on *stance*, defined in prior work as the expressed position (FAVOR, AGAINST, or NEUTRAL) toward a given target or proposition (Mohammad et al., 2016a), due to its central role in politically sensitive domains and in shaping public discourse. However, the framework is extensible to other subjective dimensions such as sentiment, framing, or toxicity.

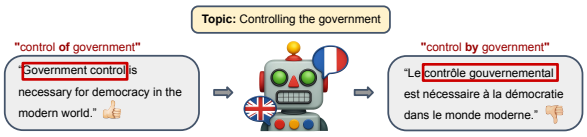


Figure 1: An example of stance distortion through translation. The English sentence expresses a favorable opinion about "controlling the government" in a democratic society (framing citizens as agents), but the French translation subtly alters the framing to "control by government" (state as the agent), which conveys a markedly different point of view.

Ensuring reliable preservation of such subjective dimensions is also critical in research settings that rely on automatically translated datasets. Previous studies have demonstrated that original texts and their machine-translated counterparts exhibit distinct characteristics, and found this divergence to negatively impact model performance (Artetxe et al., 2020). More recent work has advanced MT-based *translate-test* pipelines for cross-lingual classification (Artetxe et al., 2023; Bell et al., 2025), but their focus lies in improving classification accuracy *after* translation rather than examining the translation step itself. Our framework complements this line of research by shifting the focus to the fidelity of the translation step, explicitly measuring whether stance is preserved or not.

As with other dimensions of subjective texts, stance preservation can be framed as a problem of *invariance*: a translation should not alter the stance conveyed by the source (Bianchi et al., 2022). To diagnose violations, we adapt classical statistical

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8596–8614
11–16 May 2026.

tests to detect when translations introduce systematic stance shifts. Our objectives are: (a) to determine and quantify whether a given translation system introduces systematic stance distortion, and (b) to build on this quantification to compare multiple translation systems in terms of their relative stance-preserving fidelity. To this end, we design a statistical testing framework that compares stance predictions on translations with those on source or native-language texts. Our experiments incorporate controlled synthetic stance perturbation into MT outputs, allowing us to evaluate the framework's sensitivity to varying distortion levels.

To support these experiments, we introduce two resources. First, *BiMultiSD-XLT*, a harmonized multilingual stance detection corpus compiled from six existing datasets and their automatic translations, standardized for binary stance classification, which we use to train both encoder- and decoder-based stance classifiers for quantifying stance shifts in translations. Second, a curated set of 100 high-quality French stance-reversed examples from the test split of *X-Stance* (Vamvas and Sennrich, 2020), created by native speakers to invert stance while preserving content, style, and maximum surface similarity.

We make the following contributions:

1. Casting stance preservation as an invariance problem and adapting hypothesis tests to detect stance shifts in translation;
2. Designing a manually curated calibration set that enables controlled stance perturbation, allowing us to probe the sensitivity of stance classifiers to systematic shifts;
3. Implementing and releasing a full evaluation framework and demonstrating that our method detects stance-level distortions invisible to standard MT metrics.
4. Providing the first quantitative assessment of stance preservation in translation.

All code and data are publicly available to facilitate replication and future research.[1]

2. Related Work

2.1. Limits of MT Quality Metrics

Reference-based metrics like BLEU (Papineni et al., 2002), COMET (Rei et al., 2020), and BERTScore (Zhang et al., 2020) rely on lexical or semantic similarity to human references, making them sensitive to reference quality (Freitag et al., 2023). More recent reference-free approaches—such as CometKiwi (Rei et al., 2022), ReFreeEval (Wu et al., 2023), and BLASER 2.0-QE (Dale and Costa-jussà, 2024)—leverage large language models (LLMs) to assess fluency and adequacy without references, but still overlook deeper communicative traits like register, style, or stance.

Consequently, subtle distortions in subjective meaning often go undetected. For instance, sentiment can shift under translation (Mohammad et al., 2016b)—sometimes even inverting polarity in otherwise fluent outputs (Lohar et al., 2017)—and automatic translation tools frequently mistranslate emotions in multilingual tweets, leading to sentiment inversion or loss despite standard metrics reporting high scores (Saadany et al., 2023). We empirically demonstrate that this limitation of standard metrics also extends to *stance*. Building on efforts by toolkits like MT-Lens (García Gilabert et al., 2025) that broadened coverage to bias and toxicity, we formulate stance preservation as a measurable evaluation criterion and demonstrate that widely used metrics like BLEU or CometKiwi are equally blind to stance distortion.

2.2. Invariance and Language-Invariant Properties

Bianchi et al. (2022) introduce the notion of *language-invariant properties*, arguing that certain semantic or pragmatic attributes of discourse should remain stable across transformations such as translation, which they evaluate by computing distributional divergence in classification outcomes before and after transformation. Our work operates within this invariance perspective but specializes it to the evaluation of MT in politically and socially consequential settings, extending their population-level analysis by also measuring paired equality of proportions. We focus on stance as a high-impact subjective property and operationalize its preservation under translation using statistical hypothesis tests. In addition to measuring both aggregate distributional shifts and instance-level distortions, we validate the sensitivity of the proposed statistical tests to graded levels of opinion manipulation. By systematically applying this methodology across multiple multilingual MT systems, we provide an empirically grounded assessment of translation-induced distortion in contemporary MT models.

2.3. Automatically Translated Benchmarks

A growing number of multilingual benchmarks rely on automatic translation to expand English-centric resources into other lower-resource languages (Gureja et al., 2025; Asai et al., 2024; Chen et al., 2024; Hu et al., 2020). This approach makes multilingual evaluation more scalable, but rests on

[1] https://github.com/NazaninShafiabadi/mt-invariance

the assumption that MT outputs are sufficiently accurate despite evidence to the contrary (Park et al., 2024). Moreover, translated texts, even when fluent, often systematically diverge from native texts, a phenomenon known as *translationese* (Wang et al., 2023). Work in translation studies has documented characteristic features of translationese—such as simplification, explicitation, and normalization—(Baker et al., 1993), with differences in lexical distributions, syntax, and discourse patterns confirmed by large-scale analyses (Volansky et al., 2015). Benchmarks created through MT therefore risk embedding systematic distortions, undermining their reliability as proxies for naturally occurring data. This underscores the need for evaluation frameworks that can detect such distortions—especially in subjective properties like stance or framing—that remain largely untested by current metrics.

3. Methodology

Our methodology is designed to systematically evaluate whether MT systems preserve the stance expressed in opinionated text. The core idea is to compare stance predictions for translated texts against those for either (a) their source-language counterparts, or (b) native texts from the same distribution that share the same target language and stance. As we develop below, if translation faithfully preserves stance, the predicted stance distributions in the two conditions should not significantly differ.

3.1. Invariance in Stance Translation

Let f be an MT system that maps a source-language text s in L_1 to a target-language output $t = f(s)$ in L_2. Since modern MT systems are often non-deterministic, we model f as a probabilistic function assigning scores $f(s, t)$ to candidate translations t in L_2. Our goal is to quantify whether f systematically distorts stance.

Our evaluation pipeline relies on two key resources: a set of reference texts with gold stance annotations, and a stance detection model operating in one or both languages. The specific requirements vary depending on the statistical test applied; see §3.3 for a detailed breakdown. These resources support a three-stage evaluation of f:

1. **Translation**: using f to translate stance-labeled texts from L_1 to L_2;
2. **Stance detection**: predicting stance for translations and source/native texts in L_1 or L_2 ;
3. **Statistical testing**: applying hypothesis tests to detect significant differences in prediction distributions, signaling systematic distortion.

The formal testing procedure is described in §3.2.

3.2. Statistical Testing

We begin with a set of source-language texts $S_y = \{s_1, \ldots, s_n\}$ in L_1, all annotated with a *known* stance y (either FAVOR or AGAINST). These texts are translated into the target language L_2 using the MT system f, producing a set of translations $T_{\tilde{y}} = \{t_1, \ldots, t_n\} = f(S_y)$, each conveying an *unknown* stance $\tilde{y}_i$. If f perfectly preserves stance, then each translated text t_i should reflect the same stance as its source, i.e., $\forall i, \tilde{y}_i = y_i$.

We then apply a stance detection model g to each translated text $t_i \in T_y$, to get a predicted stance $\hat{y}_i = g(t_i)$. If g were perfectly accurate—i.e., $\forall i, \hat{y}_i = \tilde{y}_i$—then identifying systematic shifts in f would reduce to counting how often $\hat{y}_i$ matches the original stance y_i. However, since we cannot assume perfect accuracy, we use statistical hypothesis tests to determine whether prediction errors are more frequent in translations than in native texts. We apply two hypothesis tests with distinct assumptions and interpretations: McNemar's test and the two-proportion Z-test.

3.2.1. McNemar's Test

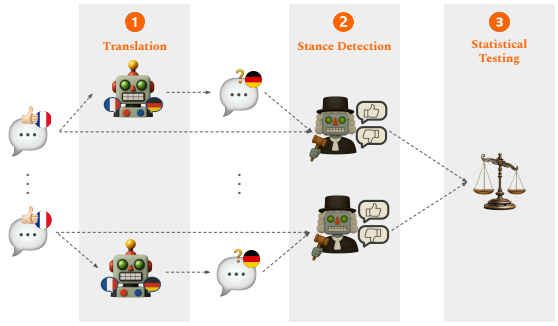
McNemar's test evaluates whether predictions for the same items change systematically across two conditions. In our setup, each source text s_i and its translation t_i are classified by the same binary stance detector g, yielding a 2×2 contingency table:

		$g(t_i)$	
		FAVOR	**AGAINST**
$g(s_i)$	**FAVOR**	a	b
	AGAINST	c	d

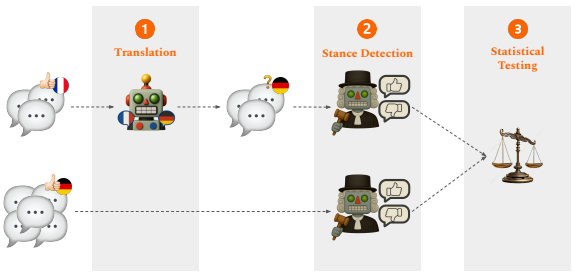
Here, a counts examples classified as FAVOR both before and after translation, d those classified as AGAINST in both cases, while b and c are the discordant pairs where translation flips a prediction from FAVOR to AGAINST or vice versa. McNemar's test targets the off-diagonal cells b and c, as only these reflect changes due to translation.

Under the null hypothesis (H_0), McNemar's test evaluates whether flips occur equally in both directions. It rejects H_0 only if stance flips are asymmetric, indicating *directional bias*. A key limitation is that flips in opposite directions cancel out. For instance, if every FAVOR flips to AGAINST and vice versa, then $b = c$ and the test fails to detect any shift, despite complete stance distortion.

To avoid this cancellation effect, we stratify the data by gold stance label and apply the test separately to FAVOR- and AGAINST-annotated subsets. This ensures that even if translation flips stance in both directions equally, changes within each group



(a) **McNemar's test**: Compares stance predictions before and after MT (here, French → German) for texts sharing the same gold stance label (here, FAVOR). The test assesses whether translation systematically flips individual stance predictions.



(b) **Two-proportion Z-test**: Compares stance predictions for machine-translated texts (here, French → German) and native target-language texts (here, German), both with the same gold stance label (here, FAVOR). The test evaluates whether translation shifts overall stance distributions.

Figure 2: Pipelines for detecting stance distortion using **McNemar's test** (prediction flips) and the **Two-proportion Z-test** (distribution shifts).

are still captured. Aggregating results across both subsets then reveals consistent stance shifts regardless of direction.

For moderate or large discordant totals ($b + c \geq 25$), we use the standard McNemar statistic:

$$\chi^2 = \frac{(b-c)^2}{b+c}, \tag{1}$$

For smaller samples, a *one-tailed* exact binomial test is used to specifically detect cases where translation *degrades* stance fidelity. We model the number of flips in the wrong direction (b) as a binomial variable and compute the p-value by summing the probabilities of outcomes at least as extreme as the observed b. In all cases, a significance level of $\alpha = 0.05$ is used. The evaluation pipeline for McNemar's test is illustrated in Figure 2a.

3.2.2. Two-Proportion Z-Test

We complement McNemar's pairwise analysis with a two-proportion Z-test that evaluates stance fidelity by comparing the distribution of correctly predicted stances in a set of translated texts to a separate set of native texts in the same language with the same gold stance label. Let p_t and p_o denote the proportions of predicted stances matching the gold stance y in the translated and native sets, respectively, with sample sizes n_t and n_o. These quantities reflect how well the stance detector g aligns with the gold label under each condition.

We then apply a *one-tailed* two-proportion Z-test (Equation (2)) with a 5% significance level to evaluate whether p_o is significantly *greater than* p_t. A significant difference would suggest that translations generated by f do not reliably preserve the intended stance y. An illustration of this setup is provided in Figure 2b.

$$Z = \frac{p_o - p_t}{\sqrt{P(1-P)\left(\frac{1}{n_o} + \frac{1}{n_t}\right)}}, \qquad P = \frac{p_o n_o + p_t n_t}{n_o + n_t} \tag{2}$$

The test is applied to each stance label (FAVOR and AGAINST) separately, and the results are aggregated into a final diagnostic score summarizing the stance-preserving fidelity of the translation system f. For instance, for a multilingual system, one could count the number of language directions for which a systematic shift is not detected.

3.3. Discussion and Extensions

As outlined in §3.1, the proposed evaluation framework relies on two key resources: stance-annotated reference texts and a stance detection model. However, the specific requirements vary depending on the statistical test applied.

3.3.1. McNemar's Test

This test relies on paired source-translation inputs with gold stance annotations in L_1, and a stance detection model g that operates in both L_1 and L_2 with comparable accuracy. This latter condition is crucial as it ensures that any observed differences in stance predictions can be attributed to translation effects rather than model asymmetry. In practice, however, multilingual models are rarely calibrated across languages, likely due to uneven training resources or the inherent difficulty of the languages (see, e.g., the performance disparities across languages for each variant of g in Table 3).

This requirement can be relaxed through *round-trip translation* (RTT), wherein L_2 translations of L_1 texts are translated back into L_1 using the same

Requirement	McNemar	Z-Test
Gold labels in L_1	✔	✔
Gold labels in L_2	✘	✔
Stance detector in L_1	✔	✘
Stance detector in L_2	✔	✔

Table 1: Requirements for McNemar's test and the Two-proportion Z-test.

MT system (Somers, 2005). This enables stance detection to be performed entirely *in the source language*, removing the need for bilingual model parity. However, RTT introduces an additional translation step in the opposite direction, which may compound noise and obscure the source of distortion.[2]

3.3.2. Two-Proportion Z-Test

Compared to McNemar's test, this setup removes the need for a bilingual stance detector, but introduces a dependency on the availability of annotated data in both L_1 and L_2. This dependency can also be mitigated via RTT, which eliminates the need for gold-annotated target-language texts. In this variant, translations are mapped back into L_1, and stance predictions for RTTs are compared to those for native L_1 texts with the same gold label. However, the same limitations of RTT discussed above apply. Moreover, the Z-test assumes that g performs reliably in the evaluation language. While this assumption is more tractable than requiring balanced performance across languages, it remains sensitive to the quality of the stance detector, which may vary depending on language resources: as can be observed in Table 3, different variants of g yield varying levels of accuracy, which can influence the sensitivity of the test. A summary of the requirements for each test is presented in Table 1.

3.4. A Calibration Set

To validate our framework's ability to detect stance distortion, we construct a controlled synthetic translation setup designed to simulate stance-altering behavior. To the best of our knowledge, no established tools currently exist for reliably quantifying stance distortion in real-world MT outputs, making a synthetic setup essential for calibration.

Our approach involves a *two-step process*, conceptually mirroring the prompting-based framework described in Appendix §A.1, to generate a dataset of stance-distorted translations. First, native French speakers manually created 100 high-confidence, monolingual rewrites of original French texts from the *X-Stance* dataset. These rewrites were designed to invert the original stance (50 FAVOR, 50 AGAINST) while preserving content, style, and surface similarity as much as possible.[3]

Second, this curated set of stance-flipped texts was translated into German using `EuroLLM-9B-Instruct` (Martins et al., 2025),[4] creating our final pool of stance-distorted translations. We then simulated varying degrees of distortion by combining increasing proportions of these inverted translations (from 0% to 50% in 10% increments) with standard, non-distorted translations from the same model, yielding six sets per stance label. These sets allow us to quantify our framework's sensitivity to progressively increasing stance shifts.

4. Implementation

4.1. Multilingual Stance Dataset

As outlined in §3.3, evaluating stance distortion in machine translation using the two-proportion Z-test requires a set of reference texts in both L_1 and L_2 with annotated stance labels. To meet this requirement, we use the test split of *X-Stance* (Vamvas and Sennrich, 2020), a dataset containing $67k$ candidate-authored stance-bearing comments in German, French, and Italian on political issues, collected from the Swiss voting advice platform *smartvote.ch*.[5] To minimize the risk of truncation–particularly the loss of stance-relevant content during model inference–we retain only comments of at most three sentences. These filtered samples serve as the foundation for our experiments in §5.

4.2. Stance Detection Model

The second required resource for evaluating stance distortion is a stance detector (see §3.1). We experiment with different architectures and capacities to assess whether the choice of stance detector influences our final estimates. Specifically, we fine-tune the base version of `XLM-RoBERTa` (Conneau et al., 2020), a transformer encoder pretrained on 100+ languages, and `LLaMA-3.2-3B` (Grattafiori et al., 2024), a multilingual auto-regressive transformer,

[2] This assumes that translated texts are statistically indiscernible from original texts, which might not be entirely true due to translationese effects (cf. §2.3).

[3] Attempts were made to automate this step using an instruction-tuned language model, `LLaMA-3.2-3B-Instruct` (Grattafiori et al., 2024) (see Appendix §A.3 for examples). However, this approach frequently failed to produce valid counterfactual rewrites, particularly for ambiguous or inconsistent comments, for which even human annotators found stance reversal to be infeasible. Therefore, manual curation was necessary to ensure the quality of the calibration set.

[4] Translation quality was validated by ensuring prediction error rates on translations and RTTs fell within the 95% confidence interval of those on original texts.

[5] https://www.smartvote.ch/fr/all

for binary classification. Full hyperparameter configurations are provided in Appendix A.2.

4.2.1. Training Data.

To build a robust, generalist multilingual classifier, we combine six publicly available stance detection datasets, covering multiple domains and languages. Table 2 summarizes the processed datasets, including class distributions and language coverage. A representative example from each dataset is presented in Table 8 in Appendix A.4 to illustrate the diversity of text-target pairs used for fine-tuning. The resulting corpus, ***BiMultiSD***, contains over $80k$ training examples in English, German, and French.

While all source datasets in *BiMultiSD* consist of naturally produced texts, the resulting stance detector is applied, at evaluation time, to both native and machine-translated inputs. This introduces a domain mismatch between training and inference, as translated texts are known to exhibit systematic distributional differences from original texts (cf. §2.3), and results in systematically poorer predictions for real than to translated texts (Artetxe et al., 2023). To mitigate this mismatch, we construct an extended training corpus, ***BiMultiSD-XLT***, which augments the original data with automatically translated instances. The resulting dataset nearly doubles the amount of training data and exposes the classifier to a balanced mix of native and translated texts across all languages, following a *translate-train* paradigm. Full construction details are provided in Appendix A.4. We use *BiMultiSD-XLT* to train a second LLaMA-based stance detector designed to better handle translated inputs at inference time.

4.2.2. Assessment.

Table 3 reports accuracies on two test sets[6]—*CoFE* and *X-Stance*—for four variants of the stance detector g {okayedthat differ in architecture and/or training data. These include `XLM-RoBERTa` fine-tuned on *X-Stance* (**XLM-R+X-Stance**), `XLM-RoBERTa` and `LLaMA-3.2-3B` fine-tuned on *BiMultiSD* (**XLM-R+BiMultiSD**, **LLaMA+BiMultiSD**), and `LLaMA-3.2-3B` fine-tuned on the translation-augmented *BiMultiSD-XLT* (**LLaMA+BiMultiSD-XLT**). A majority-class baseline (**MFC**) is included for reference.

Fine-tuning on the aggregated datasets improves accuracy in all settings. The LLaMA-based classifiers achieve the best results, with only a minor drop on natural data when translations are added to the training. All models are included in the calibration study, and the top-performing detector is used for downstream stance-preservation experiments.

4.2.3. Qualitative Error Analysis.

To better understand the limitations of the stance detection model, we conducted a manual review of instances where the `LLaMA+BiMultiSD` predictions diverged from the gold annotations on native texts. This inspection revealed that a significant portion of these "errors" stems from inherent ambiguity or noise within the ground truth data rather than model failure. The model frequently flags contradictory gold labels—for instance, correctly predicting AGAINST for a sarcastic, rhetorical comment opposing gun control despite an erroneous FAVOR annotation (see Appendix A.5). Consequently, the classifier's effective accuracy on natural texts may be higher than the reported metrics, as it is occasionally penalized for disagreeing with noisy human annotations.

5. Measuring Stance Distortion in MT

5.1. Evaluation with Synthetic Bias

To validate our methodology, we conduct a controlled evaluation using translations with synthetically induced stance shifts (§3.4), and assess the sensitivity of McNemar's test and the two-proportion Z-test to increasing levels of distortion. Figure 3 shows p-values for both stance labels in translation and RTT settings. Figure 3b compares Z-scores (converted to p-values) across classifier variants, while Figure 3a focuses on `LLaMA+BiMultiSD`, showing McNemar p-values alongside CometKiwi[7] scores for translations and BLEU[8] scores for RTTs. Table 4 provides full results for both tests and all classifiers.

Both tests exhibit increasing sensitivity as distortion levels rise, confirming that the framework can successfully flag stance shifts of varying magnitude. Comparing test results on direct translations versus RTTs reveals that while the tests identify significant shifts in both conditions, their detection thresholds often diverge. In some cases, RTT leads the tests to flag significance at lower distortion levels (e.g., `XLM-R+X-Stance`, FAVOR), while in others significance emerges only at much higher levels (e.g., `XLM-R+X-Stance`, AGAINST). This pattern suggests that RTT introduces confounding effects of its own–sometimes amplifying distortions, sometimes masking them. We therefore regard RTT as a practical fallback for easing the constraints discussed in §3.3 when necessary, but direct translation continues to offer a more reliable basis for diagnosing stance distortion in MT outputs.

A key factor in interpreting these results is the quality of the stance detector itself. More accurate

[6] In both cases, we use the preprocessed test subsets distributed by the respective data providers.

[7] `wmt22-cometkiwi-da`

[8] Using SacreBLEU (Post, 2018).

Datasets	Domain	# Samples			Class dist. (train)		Languages
		Train	Dev	Test	Favor	Against	(train)
ArgMin (Stab et al., 2018)	Web	8,676	965	1,068	44.5%	55.5%	en
CoFE (Barrière et al., 2022)	Debates	2,674	433	449	77.1%	22.9%	en, de, fr
FNC_ARC (Hanselowski et al., 2018)	News	6,416	713	380	68.7%	31.3%	en
PERSPECTRUM (Chen et al., 2019)	debates	9,537	1,060	1,178	51.8%	48.2%	en
VAST (Allaway and McKeown, 2020)	debates	9,989	1,345	383	48.8%	51.2%	en
X-Stance (Vamvas and Sennrich, 2020)	Debates	42,958	3,926	16,427	50.2%	49.8%	de, fr
BiMultiSD	**Mixed**	**80,250**	**8,442**	**19,885**	**52%**	**48%**	**en, de, fr**
BiMultiSD-XLT	**Mixed**	**160,500**	**16,036**	**37,286**	**51.4%**	**48.6%**	**en, de, fr**

Table 2: Pre-processed dataset statistics: domains, splits, class distributions, and language coverage. The merged corpus includes 45% English, 41% German, and 14.1% French instances.

Test	Lang.	Baseline (MFC)	XLM-R +X-Stance	XLM-R +BiMultiSD	LLaMA +BiMultiSD	LLaMA +BiMultiSD-XLT
CoFE	de	81.03%	77.59% (-3.44)	86.21% (+8.62)	96.55% (+10.34)	94.83% (-1.72)
	en	76.13%	81.53% (+5.40)	88.29% (+6.76)	93.69% (+5.40)	92.79% (-0.90)
	fr	73.33%	77.14% (+3.81)	79.05% (+1.91)	90.48% (+11.43)	92.38% (+1.90)
	it	90.7%	88.37% (-2.33)	88.37% (+0.00)	93.02% (+4.65)	88.37% (-4.65)
X-Stance	de	50.88%	72.24% (+21.36)	75.48% (+3.24)	84.16% (+8.68)	84.13% (-0.03)
	fr	54.4%	76.84% (+22.44)	77.21% (+0.37)	85.66% (+8.45)	85.86% (+0.20)
	it	53.91%	73.58% (+19.67)	76.45% (+2.87)	84.42% (+7.97)	82.77% (-1.65)

Table 3: Accuracy of classifier variants on *CoFE* and *X-Stance* test sets (original data). Parentheses indicate gains over the preceding setup (in green for gains, red for losses), and shaded rows correspond to cross-lingual evaluation.

classifiers yield more stable and interpretable outcomes because their predictions on native texts are already reliable. In contrast, weaker detectors introduce noise that can be mistaken for translation-induced shifts. With a stronger model such as `LLaMA+BiMultiSD`, any systematic discrepancies between native and translated predictions are more likely to reflect genuine MT distortions rather than classification errors. This pattern is evident in our results: the detection thresholds obtained with `LLaMA+BiMultiSD` fall in a reasonable range—neither hypersensitive to noise nor overly desensitized—indicating that the tests respond coherently to meaningful distortions (at around 10%). In contrast, weaker classifiers exhibit far greater variability, sometimes registering significant differences at 0% distortion—false positives likely due to classifier noise rather than translation bias—and sometimes as late as 46%.

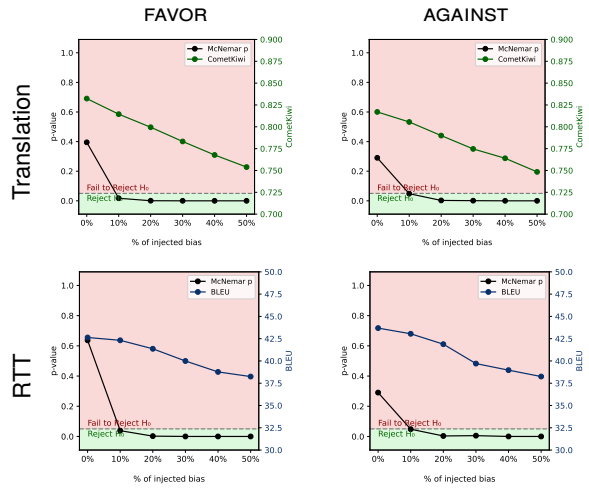
The results also highlight a limitation of standard MT metrics. Figure 3a shows that although BLEU and CometKiwi scores decline, they do so on a much narrower scale and offer no indication of the noise level at which stance errors begin to emerge. In contrast, our detector pinpoints the onset of distortion with precision—even when shifts are subtle.

5.2. Can current models be trusted?

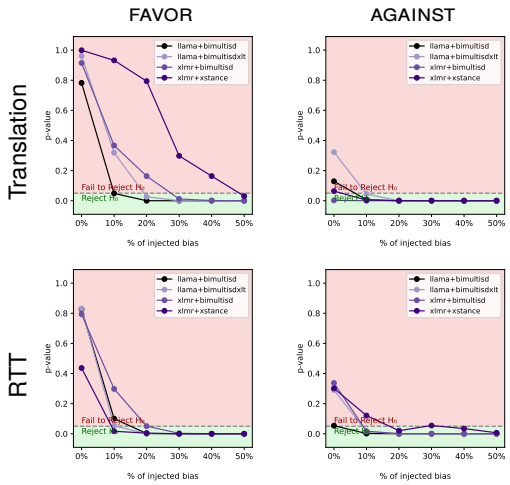
We next apply our statistical testing framework to evaluate whether existing multilingual LLMs can serve as reliable MT systems in stance-sensitive contexts. We evaluate twelve models spanning five major MT families, summarized in Table 5. When available, we include multiple parameter sizes within each family to examine the effect of model size on stance preservation. All models are evaluated on the *X-Stance* dataset, which contains user comments in German, French, and Italian.

We adapt our testing procedure to the languages supported by the stance detector and to the specific assumptions of each statistical test. For McNemar's test, we use *round-trip translations* to avoid applying the stance classifier to languages with uneven coverage (see §3.3). Concretely, we translate comments from German and French—the detector's supported languages—into four pivot languages (German, French, English, Italian) and back, yielding six language pairs. Since we evaluate the FAVOR and AGAINST labels separately, this produces 12 McNemar's tests in total.

For the two-proportion Z-test, which requires labeled data in both source and target languages, we evaluate *direct translations* only between the languages available in *X-Stance*: German, French, and Italian. As Italian falls outside the detector's training domain, it is excluded as a target language,



(a) McNemar p-values from `LLaMA+BiMultiSD` predictions, plotted alongside CometKiwi scores for direct translations and BLEU scores for RTTs.



(b) p-values of the two-proportion Z-test for all classifier variants.

Figure 3: Calibration results using McNemar's test (left) and two-proportion Z-test (right) for FAVOR and AGAINST labels in translation and RTT settings.

Model	Test	Translation		RTT	
		FAVOR	AGAINST	FAVOR	AGAINST
XLM-R+X-Stance	McNemar	29% (p=0.044)	0% (p=0.048)	6% (p=0.04)	29% (p=0.047)
	Z-Test	46% (p=0.049)	1% (p=0.041)	12% (p=0.03)	25% (p=0.034)
XLM-R+BiMultiSD	McNemar	0% (p=0.046)	0% (p=0.013)	5% (p=0.042)	8% (p=0.045)
	Z-Test	30% (p=0.034)	0% (p=0.003)	23% (p=0.031)	9% (p=0.049)
LLaMA+BiMultiSD	McNemar	9% (p=0.041)	9% (p=0.044)	11% (p=0.038)	9% (p=0.046)
	Z-Test	12% (p=0.048)	3% (p=0.047)	13% (p=0.034)	2% (p=0.029)
LLaMA+BiMultiSD-XLT	McNemar	4% (p=0.046)	24% (p=0.045)	7% (p=0.046)	13% (p=0.041)
	Z-Test	18% (p=0.045)	7% (p=0.045)	13% (p=0.029)	8% (p=0.029)

Table 4: Minimum distortion levels at which each classifier detects significant stance shifts for FAVOR and AGAINST labels across Translation and RTT settings. Each model is evaluated using both McNemar's Test and the two-proportion Z-Test. Results are averaged over 10 randomized runs using distinct seeds (1–10) to shuffle the distorted translations before injection, with significance defined as $p < 0.05$.

resulting in a total of 8 Z-tests.

To quantify stance preservation, we compute a *stance consistency score* for each model, defined as the number of language–stance pairs (out of all tested combinations[9]) for which we *do not detect* statistically significant stance shifts. Scores are computed at two significance levels: $\alpha = 0.05$ (stricter) and $\alpha = 0.01$ (more lenient). We report the aggregated results for $\alpha = 0.01$ in Figure 4.[10]

[9]Appendix §A.7 details p-values for all language–stance combinations under direct MT and RTT across both statistical tests. This breakdown helps readers identify the precise conditions under which each model struggled and observe how RTT reshaped the statistical outcomes.

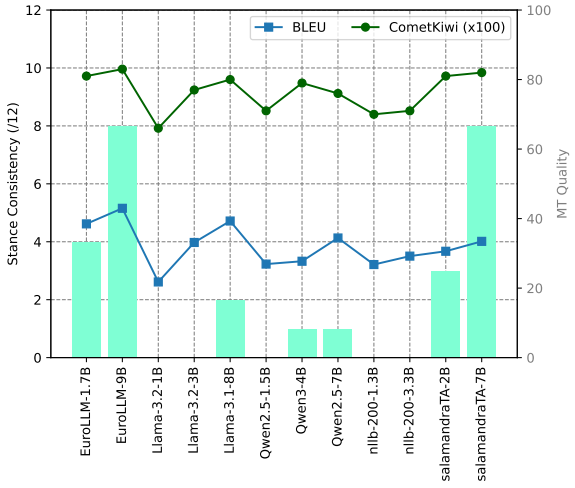
[10]All reported results use predictions from `LLaMA+BiMultiSD`; corresponding analyses with `LLaMA+BiMultiSD-XLT` are included in Appendix §A.6, where scores are generally higher but the overall tendencies and conclusions remain unchanged.

The results show that **none of the twelve systems preserve stance reliably across all conditions**: all models introduce statistically significant stance shifts in at least one translation direction or polarity. Smaller instruction-tuned models are particularly unstable, displaying significant shifts even at high BLEU and CometKiwi scores—confirming that conventional MT metrics fail to capture stance-level distortions. Larger models exhibit more stable behavior, but increased capacity alone does not guarantee higher stance fidelity.

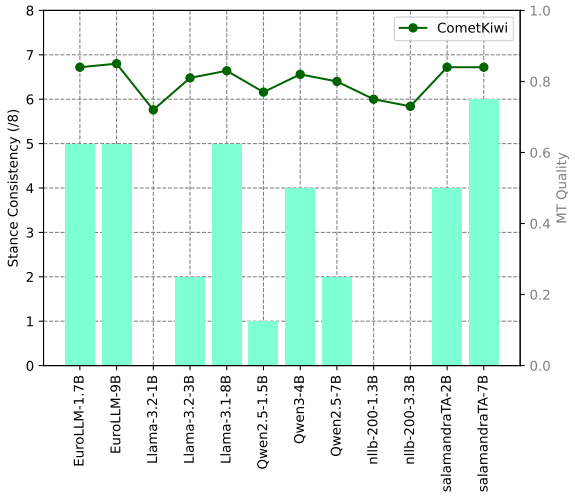
Across families, `EuroLLM-9B-Instruct` and `salamandraTA-7b-instruct` emerge as the least distortion-prone systems, yielding the highest stance-consistency scores under both tests. Nonetheless, even these models introduce statistically significant stance shifts in several translation directions. In contrast, both `NLLB` variants consistently produce significant alterations—even at the lowest threshold ($\alpha = 0.001$).

Family	Small	Medium	Large
EuroLLM	`EuroLLM-1.7B-Instruct`	–	`EuroLLM-9B-Instruct`
LLaMA	`LLaMA-3.2-1B-Instruct`	`LLaMA-3.2-3B-Instruct`	`LLaMA-3.1-8B-Instruct`
Qwen	`Qwen-2.5-1.5B-Instruct`	`Qwen-3-4B-Instruct-2507`	`Qwen-2.5-7B-Instruct`
NLLB	`nllb-200-distilled-1.3B`	`nllb-200-3.3B`	–
Salamandra	`salamandraTA-2b-instruct`	–	`salamandraTA-7b-instruct`

Table 5: Evaluated multilingual MT systems grouped by model family and size (small, medium, large).



(a) McNemar's test (RTT).



(b) Two-proportion Z-test (direct translation).

Figure 4: Stance consistency scores for twelve multilingual MT systems. Scores reflect the number of language–stance pairs without statistically significant stance shifts ($\alpha = 0.01$).

Overall, these results indicate that current multilingual MT systems cannot yet be trusted to faithfully preserve stance. Even when standard MT metrics such as BLEU and CometKiwi suggest high translation quality, our statistical tests uncover systematic stance shifts that these metrics overlook. This study constitutes the first large-scale quantitative analysis of stance preservation in MT, revealing a critical gap in current evaluation practices and highlighting the need for stance-aware assessment in socially or politically sensitive applications.

6. Conclusion

Our work presents the first quantitative framework for systematically evaluating stance preservation in machine translation. We have demonstrated that despite high scores on traditional metrics like BLEU and CometKiwi, current multilingual models consistently introduce statistically significant stance shifts. By adapting statistical tests—McNemar's test for paired comparisons and the two-proportion Z-test for group comparisons—we've provided a robust method for diagnosing this critical form of translation effect. Our findings, based on an evaluation of twelve popular MT systems, show that none can be fully trusted to preserve the original stance of a text, even when translating between high-resourced languages. This is a particularly serious issue in politically sensitive domains, where subtle opinion distortions can have significant real-world implications. This research highlights the inadequacy of existing evaluation paradigms and underscores the urgent need for new metrics that can assess the fidelity of subjective content in translations. As such, it can also be read as a warning against transferring subjective annotation via automatic translation.

7. Limitations

Our work presents a first attempt to design a framework for quantifying stance distortion in machine translation. It is subject to several limitations that warrant consideration. First, the calibration set includes 100 French examples, reflecting a deliberate emphasis on quality over scale. Native human rewrites were employed to ensure reliable stance reversal, as automated methods failed in this regard—resulting in a final set constrained in both size and language coverage. Second, our stance detection relies on relatively small models—`XLM-RoBERTa` (250M) and `LLaMA-3.2-3B`—whose capacity directly impacts diagnostic sensitivity. Larger classifiers may yield more stable results.

Third, the multilingual scope of our evaluation is limited by the availability of annotated data and

classifier coverage. While we tested twelve MT systems, our analysis focused on high-resource languages (German, French, Italian), with classifiers fine-tuned only on English, German, and French. Future work could extend this analysis to additional language families and incorporate human translations as a comparative benchmark.

Finally, our methodology depends on two hypothesis tests—McNemar's and the two-proportion Z-test—which determine whether statistically significant stance shifts occur but do not provide a continuous measure of distortion magnitude. Developing continuous metrics that quantify the *degree* of stance alteration would both enrich interpretability and enable the natural extension of the framework to multi-class or graded stance representations.

8. Ethical Considerations

All experiments were conducted on publicly available datasets, and synthetic distortions were introduced solely for calibration purposes. We acknowledge, however, the dual-use potential of our metric: while designed to diagnose and mitigate unintended stance shifts, it could in principle be repurposed to tune models that deliberately distort stance while remaining undetectable to conventional metrics. This risk underscores the importance of transparency in evaluation practices and the need for robust safeguards when deploying MT systems in sensitive domains.

9. Acknowledgements

We thank the reviewers for their thoughtful and constructive feedback. This research was supported by BPI-France through the project *AI for Democracy – Democratic Commons*, one of the seven laureates of the "Digital Commons for Generative AI" call for projects funded under the France 2030 investment plan. We are grateful to all members of the Democratic Commons program at Sorbonne Université, Sciences Po, and make.org, in particular Paul Lerner and Léo Labat, for their discussions and feedback on earlier versions of this work, and we acknowledge the contribution of Hal Daumé III during his visit to our group in spring 2025. This work was performed using HPC resources from GENCI–IDRIS (Grant 2024-AD011015928).

10. Bibliographical References

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

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, Online. Association for Computational Linguistics.

Mona Baker, Gill Francis, and Elena Tognini-Bonelli. 1993. *'Corpus Linguistics and Translation Studies: Implications and Applications'*. John Benjamins Publishing Company, Netherlands.

Samuel J. Bell, Eduardo Sánchez, David Dale, Pontus Stenetorp, Mikel Artetxe, and Marta R. Costa-jussà. 2025. Translate, then Detect: Leveraging Machine Translation for Cross-Lingual Toxicity Classification.

Federico Bianchi, Debora Nozza, and Dirk Hovy. 2022. Language invariant properties in natural language processing. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 84–92, Dublin, Ireland. Association for Computational Linguistics.

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*, pages 8440–8451, Online. Association for Computational Linguistics.

David Dale and Marta R. Costa-jussà. 2024. BLASER 2.0: a metric for evaluation and quality estimation of massively multilingual speech and text translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16075–16085, Miami, Florida, USA. Association for Computational Linguistics.

Fred Dawson. 2025. Broadcasters push AI to new levels. Tvtech online, visited on March 05th, 2026.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Javier García Gilabert, Carlos Escolano, Audrey Mash, Xixian Liao, and Maite Melero. 2025. MT-LENS: An all-in-one toolkit for better machine translation evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 51–60, Albuquerque, New Mexico. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Hol-

land, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The Llama 3 Herd of Models.

Pintu Lohar, Haithem Afli, and Andy Way. 2017. Maintaining Sentiment Polarity in Translation of User-Generated Content. *The Prague Bulletin of Mathematical Linguistics*, 108:73 – 84.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. Eurollm-9b: Technical report.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016a. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Saif M. Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016b. How Translation Alters Sentiment. *J. Artif. Intell. Res.*, 55:95–130.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

ChaeHun Park, Koanho Lee, Hyesu Lim, Jaeseok Kim, Junmo Park, Yu-Jung Heo, Du-Seong Chang, and Jaegul Choo. 2024. Translation deserves better: Analyzing translation artifacts in cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5193–5221, Bangkok, Thailand. Association for Computational Linguistics.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Hadeel Saadany, Constantin Orasan, Rocio Caro Quintana, Felix Do Carmo, and Leonardo Zilio. 2023. Analysing mistranslation of emotions in multilingual tweets by online MT tools. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 275–284, Tampere, Finland. European Association for Machine Translation.

Harold Somers. 2005. Round-trip translation: What is it good for? In *Proceedings of the Australasian Language Technology Workshop 2005*, pages 127–133, Sydney, Australia.

V. Volansky, N. Ordan, and S. Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

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

Hanming Wu, Wenjuan Han, Hui Di, Yufeng Chen, and Jinan Xu. 2023. A holistic approach to reference-free evaluation of machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 623–636, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT.

11. Language Resource References

Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931, Online. Association for Computational Linguistics.

Akari Asai, Sneha Kudugunta, Xinyan Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2024. BUFFET: Benchmarking large language models for few-shot cross-lingual transfer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1771–1800, Mexico City, Mexico. Association for Computational Linguistics.

Valentin Barrière, Guillaume Jacquet, and Leo Hémamou. 2022. CoFE: A new dataset of intra-multilingual multi-target stance classification from an online European participatory democracy platform. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 418–422, Online only. Association for Computational Linguistics.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2024. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7001–7016, Miami, Florida, USA. Association for Computational Linguistics.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing

things from a different angle:discovering diverse perspectives about claims. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 542–557, Minneapolis, Minnesota. Association for Computational Linguistics.

Srishti Gureja, Lester James Validad Miranda, Shayekh Bin Islam, Rishabh Maheshwary, Drishti Sharma, Gusti Triandi Winata, Nathan Lambert, Sebastian Ruder, Sara Hooker, and Marzieh Fadaee. 2025. M-RewardBench: Evaluating reward models in multilingual settings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 43–58, Vienna, Austria. Association for Computational Linguistics.

Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1859–1874, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium. Association for Computational Linguistics.

Jannis Vamvas and Rico Sennrich. 2020. X-Stance: A multilingual multi-target dataset for stance detection. In *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*, Zurich, Switzerland.

A. Appendix

A.1. Two-Step Prompting Strategy for Controlled Stance Manipulation

To generate controlled stance-preserving and stance-inverting translations, we employ a two-step prompting strategy. An instruction-tuned model first transforms a source-language statement into its semantic opposite. The resulting text is then translated into the target language using either the same model or a different model specialized for MT. This decomposition allows independent control over stance manipulation and translation.

All prompts follow a chat-style format compatible with instruction-tuned models and consist of a system instruction and a user query. Outputs are explicitly constrained to contain only the transformed or translated text, with no additional commentary.

Step 1: Stance Reversal The transformation stage applies the following system instruction:

Stance Reversal Instruction (Step 1)

```
Transform the user's statement into
its semantic opposite.  The target
is only given for your reference.
Guidelines:
 1. Maintain minimal changes to the
    wording while ensuring the
    stance toward the target is
    fully reversed.
 2. Avoid simple negation.
 3. Respond with ONLY the
    transformed statement-no
    preambles, repetition of the
    input, or explanations.
```

The corresponding user query template is:

```
Target: {T}
Statement: {S}
```

Step 2: Translation The translation stage applies the following system instruction:

Translation Instruction (Step 2)

```
You are an impartial translator
dedicated to conveying the user's
input with complete accuracy while
maintaining the original tone and
stance.  Your output must consist
only of the translation itself.
```

The corresponding user query template is:

```
Translate the following text from
{L_1} into {L_2}.
{L_1}: {S}
{L_2}:
```

Hyperparameter	Value
Learning rate	2×10^{-5}
Optimizer	AdamW, weight decay 0.01
Scheduler	Cosine with 10% warmup
Batch size	16
Max sequence length	512 tokens
Epochs	10
Gradient clipping	5.0
Early stopping patience	3 epochs
Loss function	Cross-entropy (binary)
Mixed precision	bf16 (fallback to fp16)
Evaluation metric for checkpoint selection	Validation loss on *BiMultiSD*

Table 6: Hyperparameters for fine-tuning `xlm-roberta-base` and `Llama-3.2-3B`.

This modular design enables controlled generation under two experimental conditions: (i) stance-preserving translation (translation step only) and (ii) stance-inverting translation (transformation followed by translation). Separating stance manipulation from translation allows (1) the use of specialized models or human annotators for either step, and (2) reuse of the same stance-reversed source statements across multiple target languages, ensuring consistency in cross-lingual comparisons.

A.2. Hyperparameters and Training Setup

Table 6 lists the hyperparameters used for fine-tuning `xlm-roberta-base` and `Llama-3.2-3B`. Training was conducted using the HuggingFace Transformers library with PyTorch as the backend.

A.3. Examples of Oppositional Rewrite Failures and Successes

Table 7 presents representative examples of a successful and unsuccessful oppositional rewrite generated by `LLaMA-3.2-3B-Instruct` for the calibration set described in 3.4. All transformations were manually reviewed and, if needed, revised by native French-speaking annotators. The top example illustrates a case in which the original comment was too vague, indirect, or internally inconsistent to permit a reliable stance inversion, even for human annotators; such items were excluded from the final set. In contrast, the bottom example shows a successful rewrite that met the criteria for stance inversion while preserving close surface-level similarity, and was therefore retained.

A.4. Dataset Details

Table 8 provides representative examples from the datasets used to construct *BiMultiSD*, showing the topic, comment, stance label (FAVOR/AGAINST), and comment language for each.

To reduce the domain mismatch between native and translated inputs at inference time, we construct an extended training corpus, *BiMultiSD-XLT*, by augmenting *BiMultiSD* with automatically translated instances. The goal of this augmentation is not to create new gold-standard annotations, but to expose the stance classifier to translated text during training.

The construction of *BiMultiSD-XLT* proceeds as follows. First, all samples in each language are translated into the other two languages present in the dataset using `EuroLLM-9B-Instruct`. This yields multiple translated variants for each original instance. Second, the stance of each translated text is predicted using a LLaMA-based stance classifier trained on the original *BiMultiSD*, which was the best-performing variant in preliminary experiments. Third, translated instances are filtered to retain only those that (i) exceed a COMETKiwi quality threshold of 0.7, and (ii) receive the same predicted stance label as the original gold annotation. Finally, for each language, we select the top-n scoring translated instances, where n corresponds to the number of original samples in that language, to maintain approximate balance between natural and translated data.

The resulting *BiMultiSD-XLT* nearly doubles the size of the original corpus and contains comparable proportions of native and translated texts in English, German, and French. In a small number of cases, the filtering criteria remove too many translated candidates, resulting in slightly fewer translated samples than natural ones for a given language. This augmented dataset is used to train a second stance classifier that has been exposed to both natural and translated inputs during training.

A.5. Qualitative Error Analysis Examples

Here we provide specific examples from our manual review of the `LLaMA+BiMultiSD` classifier's predictions on native texts. Table 9 highlights instances where the model's prediction diverges from the gold label. In these cases, the model successfully identifies the underlying pragmatic meaning—such as sarcasm or disentangling mixed sentiments within a single statement—revealing that the mismatch is driven by noisy labels in the original dataset rather than classification errors.

Target	Une initiative demande que les subsides aux assurances maladie soient conçus de manière à ce que personne n'ait à consacrer plus de dix pour cent de son revenu disponible au paiement des primes. Soutenez-vous cette proposition?
Comment	Une réforme structurelle des coûts de la santé doit néanmoins être proposée notamment avec la mise en place d'une caisse unique. Notre système de santé doit aussi s'adapter très rapidement aux besoins des seniors avec l'arrivée du 'Papy boom'.
Label	FAVOR
LLaMA-3 Transformation	Une réforme structurelle des coûts de la santé doit néanmoins ne pas être proposée notamment avec la mise en place d'une caisse unique. Notre système de santé doit aussi s'adapter très lentement aux besoins des seniors avec l'arrivée du 'Papy boom'.
Annotator suggestion	Non mais une réforme structurelle des coûts de la santé doit néanmoins être proposée notamment avec la mise en place d'une caisse unique. Notre système de santé doit aussi s'adapter très rapidement aux besoins des seniors avec l'arrivée du 'Papy boom'.
Status	REMOVED
Target	Êtes-vous favorable à la légalisation du statut des sans-papiers par un octroi collectif et unique de permis de séjour?
Comment	Cela créerait une inégalité de traitement par rapport aux migrants qui arrivent en Suisse légalement.
Label	AGAINST
LLaMA-3 Transformation	Cela créerait une égalité de traitement par rapport aux migrants qui arrivent en Suisse légalement.
Annotator suggestion	-
Status	KEPT

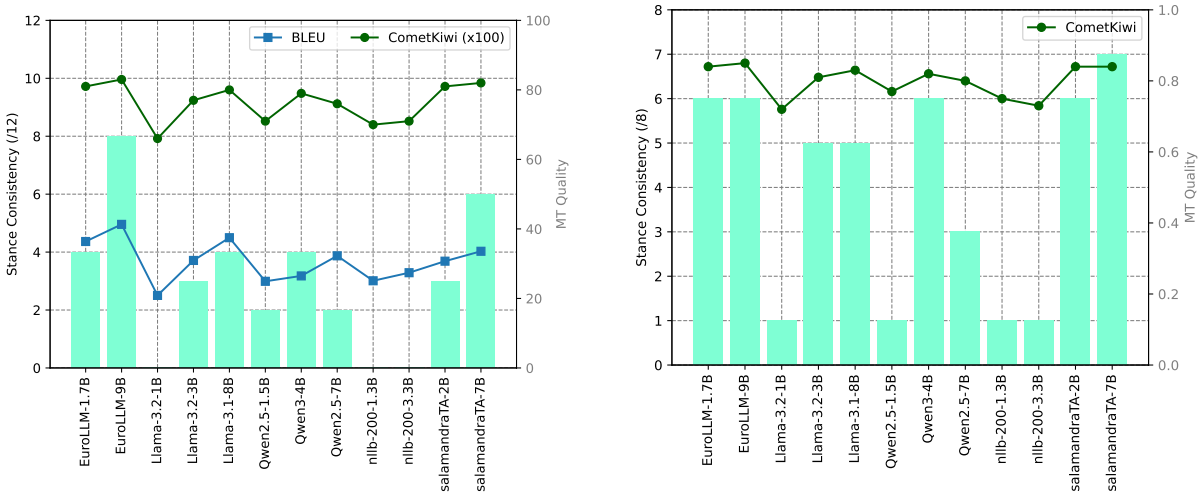
Table 7: Examples of two oppositional rewrites: a rejected case in which the original comment does not directly address the question (top), and a retained case that successfully inverts the stance while preserving high surface-level similarity (bottom).

Dataset	Topic	Comment	Label	Lang.
ArgMin	marijuana legalization	At best , that patchwork of laws , differing from one locality to another , will be yet another unintended and predictable problem arising from legalization as envisioned under this act	AGAINST	en
CoFE	We need to have migrant camps in syria, tunisia, gambia, guiana and all other countries that migrants come from, so we can provide help to them there, this will cost less to our governement, will protect more people, and no jihadist will enter here	Agree. If the asylum seekers don't have to cross the Mediterranean to apply for it less of them will die trying and less human trafficking will take place. And whoever is rejected won't be in the country and eat up resources that could be used to process others.	FAVOR	en
FNC_ARC	Sport leagues should enjoy nonprofit status	America's sport? Of course they should pay taxes, like any profit making corporate entity. Another aberration of our tax code that our legislators provide for the wealthy, privileged, connected and very unworthy, few. Ouch! I got a concussion thinking about it.	AGAINST	en
PERSPECTRUM	School Uniforms Should Be Mandatory	School uniforms may deter crime and increase student safety.	FAVOR	en
VAST	a tax break	No to tax breaks for private education - it takes public money away from public schools, which need. Furthermore, this will further erode separation of church and state. Finally, the is a conservative scam to take public monies for private benefit.	AGAINST	en
X-Stance	La Confédération devrait-elle se retirer de la promotion de la culture?	La culture est quelque chose de très personnel et il est risqué de faire la promotion de celle-ci sans fâcher personne. Un soutient peut toutefois être maintenu pour la préservation du patrimoine.	FAVOR	fr

Table 8: Example from each dataset used for fine-tuning the stance detection models.

Topic	Comment	Gold	Pred.
Au cours des dernières années, les règles d'acquisition et de possession d'armes se sont renforcées. Êtes-vous favorables à cette évolution? *(There has been an increasing tightening of rules on the acquisition and possession of weapons in recent years. Do you welcome this development?)*	Est-ce que l'Etat, les chasseurs et les délinquants doivent-ils avoir le monopole de la violence ? *(Should the state, hunters, and criminals have a monopoly on violence?)*	FAVOR	AGAINST
Befürworten Sie eine strengere Kontrolle der Lohngleichheit von Frauen und Männern? *(Are you in favour of stricter monitoring of pay equity for women and men?)*	Lohngleichheit finde ich super wichtig. Kontrolle finde ich weniger cool. *(I think equal pay is very important. Control is less cool.)*	FAVOR	AGAINST

Table 9: Examples of stance detection mismatches driven by noisy ground-truth labels. The classifier successfully navigates underlying pragmatics (e.g., sarcasm or separating general sentiment from policy stance) that contradict the self-reported labels.



(a) McNemar's test (RTT).

(b) Two-proportion Z-test (direct translation).

Figure 5: Stance consistency scores for twelve multilingual MT systems based on predictions from `LLaMA+BiMultiSD-XLT`. Scores reflect the number of language–stance pairs *without statistically significant stance shifts* ($\alpha = 0.01$).

A.6. MT-system evaluation using `LLaMA+BiMultiSD-XLT` predictions

To verify that our findings are not specific to a single stance detector, we replicate the full MT-system evaluation using predictions from `LLaMA+BiMultiSD-XLT`, a translationese-protected variant of the classifier used in the main text. Figures 5a and 5b report the corresponding McNemar and two-proportion Z-test results under the same experimental setup described in §5.2.

Overall, the XLT-based evaluation yields **higher stance-consistency scores** across most MT systems, reflecting the model's reduced sensitivity to translationese artifacts. However, the **qualitative tendencies remain unchanged**. No MT system preserves stance consistently across all translation directions or polarities, and the relative ranking of model families is stable: `EuroLLM-9B` and `SalamandraTA-7B` remain the least distortion-prone, while both `NLLB` variants continue to yield substantial stance shifts across nearly all tested conditions.

These results confirm that **our main findings are robust to the choice of stance detector**. While `LLaMA+BiMultiSD-XLT` produces slightly more optimistic scores, the overarching pattern persists: current multilingual MT systems introduce systematic stance distortions that are not captured by standard MT metrics, underscoring the need for stance-aware evaluation in sensitive applications.

A.7. Comprehensive p-Value Tables for All Language–Polarity Pairs

Tables 10 and 11 detail test outcomes for all tested combinations across direct machine translation (Direct MT) and round-trip translation (RTT).

Model Capacity and Family Trends Across both statistical tests, smaller models (`Llama3.2` 1B/3B and `NLLB` 1.3B/3.3B) consistently fail to preserve stance, scoring 0 in nearly all conditions. This indi-

cates that these models introduce statistically significant stance shifts regardless of the language direction or original polarity. Conversely, increased size correlates with improved stance fidelity, though it does not guarantee immunity to distortion: `EuroLLM-9B` and `SalamandraTA-7B` emerge as the most robust models in the evaluated group. However, their sub-optimal Direct MT scores confirm that even state-of-the-art multilingual LLMs remain vulnerable to systematic opinion shifts.

The Confounding Effects of RTT These tables highlight how RTT reshapes the statistical outcomes. While overall stance consistency scores remain broadly stable between the direct translation and RTT settings, our evaluations indicate that RTT introduces distinct confounding effects: it appears to compound translation noise for mid-tier models, while frequently inflating consistency scores for top-performing ones. For example, the `Qwen` models (specifically `Qwen3-4B-Instruct-2507` and `Qwen2.5-7B-Instruct`) experience a noticeable drop in stance consistency when transitioning from Direct MT to RTT across both hypothesis tests. Conversely, `SalamandraTA-7B` preserves stance in 3/8 language-polarity scenarios under direct translation, but 5/8 under RTT (Table 10). This asymmetry may stem either from discrepancies in the stance classifier's performance across languages in the direct translation setting, or from the back-translation process inadvertently "correcting" the stance in the RTT setting. Consequently, while RTT serves as a necessary fallback for mitigating classifier limitations, these results empirically demonstrate that it provides an occasionally misleading diagnostic signal compared to direct translation.

Directional and Polarity Asymmetries Finally, the itemized p-values reveal that translation-induced stance shifts are highly asymmetrical: models frequently distort FAVOR and AGAINST labels at unequal rates within a given language direction. This asymmetry validates the methodological decision to stratify evaluations by gold label, as aggregating these predictions would otherwise allow bidirectional flips to cancel each other out.

Model	Src	Direct MT							RTT						
		de_{tgt}		fr_{tgt}		en_{tgt}		Score (/8)	de_{pvt}		fr_{pvt}		en_{pvt}		Score (/8)
		+	-	+	-	+	-		+	-	+	-	+	-	
EuroLLM-1.7B	de	—	—	0.0	0.002	0.0	0.135	2	—	—	0.0	0.0	0.0	0.024	2
	fr	0.001	0.0	—	—	0.05	0.008		0.0	0.001	—	—	0.028	0.002	
EuroLLM-9B	de	—	—	0.0	0.022	0.004	0.104	4	—	—	0.001	0.041	0.019	0.06	4
	fr	0.025	0.0	—	—	0.149	0.004		0.004	0.007	—	—	0.009	0.384	
Llama3.2-1B	de	—	—	0.0	0.0	0.0	0.0	0	—	—	0.0	0.0	0.0	0.0	0
	fr	0.0	0.0	—	—	0.0	0.0		0.0	0.0	—	—	0.0	0.0	
Llama3.2-3B	de	—	—	0.0	0.0	0.0	0.0	0	—	—	0.0	0.0	0.0	0.0	0
	fr	0.0	0.0	—	—	0.0	0.0		0.0	0.0	—	—	0.0	0.0	
Llama3.1-8B	de	—	—	0.0	0.0	0.0	0.023	2	—	—	0.0	0.0	0.0	0.0	1
	fr	0.0	0.0	—	—	0.001	0.09		0.0	0.003	—	—	0.0	0.042	
Qwen2.5-1.5B	de	—	—	0.0	0.0	0.0	0.0	1	—	—	0.0	0.0	0.0	0.0	0
	fr	0.0	0.0	—	—	0.0	0.054		0.0	0.0	—	—	0.0	0.01	
Qwen3-4B	de	—	—	0.0	0.0	0.0	0.047	3	—	—	0.0	0.0	0.0	0.0	1
	fr	0.001	0.0	—	—	0.033	0.549		0.0	0.003	—	—	0.0	0.26	
Qwen2.5-7B	de	—	—	0.0	0.0	0.0	0.117	3	—	—	0.0	0.0	0.0	0.0	1
	fr	0.0	0.0	—	—	0.054	0.012		0.0	0.0	—	—	0.0	0.029	
NLLB-1.3B	de	—	—	0.0	0.0	0.0	0.0	0	—	—	0.0	0.0	0.0	0.0	0
	fr	0.0	0.0	—	—	0.0	0.0		0.0	0.0	—	—	0.0	0.0	
NLLB-3.3B	de	—	—	0.0	0.0	0.0	0.0	0	—	—	0.0	0.0	0.0	0.0	0
	fr	0.0	0.0	—	—	0.0	0.0		0.0	0.0	—	—	0.0	0.0	
SalamandraTA-2B	de	—	—	0.0	0.001	0.0	0.195	2	—	—	0.0	0.001	0.002	0.0	1
	fr	0.003	0.0	—	—	0.003	0.047		0.001	0.0	—	—	0.001	0.12	
SalamandraTA-7B	de	—	—	0.0	0.0	0.0	0.324	3	—	—	0.024	0.001	0.092	0.001	5
	fr	0.005	0.024	—	—	0.072	0.001		0.0	0.349	—	—	0.033	0.14	

Table 10: McNemar's test p-values for all models across Direct MT and RTT settings, evaluated at the $\alpha = 0.01$ significance level (Red : significant shifts; Green : non-significant shifts). The **Src** column specifies the source language, while the **+** and **-** columns correspond to the FAVOR and AGAINST stance labels, respectively. Under **Direct MT**, the columns denote the **target** language; here, the classifier is applied cross-lingually to evaluate both the source text and its translation. Under **RTT**, the columns denote the intermediate **pivot** language; because the text is back-translated, the classifier is applied monolingually to the source text and its round-trip translation. The final **Score** is the number of test instances with *insignificant* stance shifts for each evaluated model.

Model	Src	Direct MT					RTT				
		de_{tgt}		fr_{tgt}		Score (/4)	de_{pvt}		fr_{pvt}		Score (/4)
		+	-	+	-		+	-	+	-	
EuroLLM-1.7B	de	—	—	0.0	0.001	2	—	—	0.0	0.003	2
	fr	0.181	0.125	—	—		0.016	0.023	—	—	
EuroLLM-9B	de	—	—	0.0	0.003	2	—	—	0.03	0.107	4
	fr	0.504	0.095	—	—		0.072	0.071	—	—	
Llama3.2-1B	de	—	—	0.0	0.0	0	—	—	0.0	0.0	0
	fr	0.0	0.0	—	—		0.0	0.0	—	—	
Llama3.2-3B	de	—	—	0.0	0.0	1	—	—	0.0	0.0	0
	fr	0.006	0.012	—	—		0.0	0.002	—	—	
Llama3.1-8B	de	—	—	0.0	0.0	2	—	—	0.0	0.003	1
	fr	0.115	0.106	—	—		0.008	0.034	—	—	
Qwen2.5-1.5B	de	—	—	0.0	0.0	1	—	—	0.0	0.0	0
	fr	0.0	0.024	—	—		0.0	0.001	—	—	
Qwen3-4B	de	—	—	0.0	0.0	2	—	—	0.0	0.0	1
	fr	0.198	0.09	—	—		0.006	0.026	—	—	
Qwen2.5-7B	de	—	—	0.0	0.0	2	—	—	0.0	0.0	0
	fr	0.024	0.087	—	—		0.0	0.005	—	—	
NLLB-1.3B	de	—	—	0.0	0.0	0	—	—	0.0	0.0	0
	fr	0.0	0.0	—	—		0.0	0.0	—	—	
NLLB-3.3B	de	—	—	0.0	0.0	0	—	—	0.0	0.0	0
	fr	0.0	0.0	—	—		0.0	0.0	—	—	
SalamandraTA-2B	de	—	—	0.0	0.001	2	—	—	0.005	0.001	1
	fr	0.284	0.029	—	—		0.031	0.009	—	—	
SalamandraTA-7B	de	—	—	0.0	0.0	2	—	—	0.093	0.027	4
	fr	0.319	0.476	—	—		0.013	0.288	—	—	

Table 11: Two-proportion Z-test p-values evaluated at the $\alpha = 0.01$ significance level. The table structure and notation are identical to those detailed in Table 10.