

Lost in Interpretation: The Plausibility-Faithfulness Trade-off in Cross-Lingual Explanations

Somnath Banerjee^{♣,♣}, Pranav Jha[♣], Rima Hazra^{♡,◇}, Animesh Mukherjee[♣]

[♣] Indian Institute of Technology Kharagpur [♡] TCG CREST

[◇] National University of Singapore [♣] Singapore Institute of Technology

Abstract

LLMs deployed multilingually are often audited via English explanations for non-English inputs. We evaluate *extractive* explanations “where the model identifies input token spans as evidence alongside a generated rationale” and uncover a systematic **trade-off**: English-pivot explanations can achieve higher span agreement with human rationales while their evidence becomes less causally grounded in the model’s prediction, as measured by both comprehensiveness and sufficiency. Across 3 tasks, 5 languages, and 2 multilingual LLM families, we find that English explanations frequently produce fluent but loosely anchored rationales, with comprehensiveness degrading by up to $5.7\times$ relative to native-language conditions - even as task accuracy remains stable across settings. For socially nuanced classification, English pivots also fail to preserve pragmatic cues, reducing both faithfulness and span agreement. We recommend auditing explanations in the input language, reporting multi-faceted faithfulness metrics beyond lexical overlap, and treating English rationales as communication summaries rather than faithful decision traces.

1 Introduction

As LLMs are increasingly deployed in global contexts (Eiden, 2024; Jadhav et al., 2025), they routinely operate under cross-lingual constraints where the user’s input language differs from the system’s reporting language. This setting is common in applications such as customer support and public-service workflows, where users submit requests in local languages (e.g., *Chinese*, *Hindi*, *Malay*) while downstream analysts, auditors, or operational teams require English explanations for triage and decision-making (Amazon Web Services, 2024; Mic, 2025). In this work, we focus specifically on extractive explanations, where the model identifies spans from the input text as evidence for its prediction, alongside a free-text rationale. This dual-

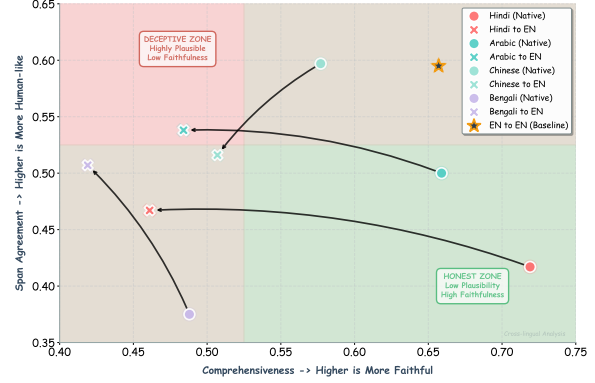


Figure 1: The plausibility–faithfulness trade-off in e-SNLI (Qwen2.5-7B). Arrows show the shift from native-language explanations to English-pivot explanations. English pivots tend to increase span agreement with human rationales (y-axis) while reducing comprehensiveness (x-axis), indicating that the cited evidence becomes less causally necessary for the prediction.

output format “evidence spans plus narrative explanation” is common in deployed systems that require both auditability (*which spans mattered?*) and interpretability (*why did the model decide this?*). Crucially, while the narrative explanation may be generated in any language, the evidence spans are always drawn verbatim from the input, enabling language-independent faithfulness evaluation.

For example, in a banking support pipeline, a user reports in Hindi, *mera UPI debit ho gaya lekin balance nahi aaya* (money is debited but not received). The intended English summary for the backend team is *UPI transaction: the amount is debited from the customer’s account, but the beneficiary did not receive the credit. Please check pending status or credit reversal*. Instead, the model sometimes produces the incorrect summary *UPI transaction failed*, collapsing a debit-without-credit event into a generic failure and thereby altering the operational interpretation.

This reporting language choice typically reflects developer preferences, organizational policy, and the dominance of English-centric tooling and eval-

uation benchmarks. This language mismatch introduces a critical and largely unexamined question – *when a model explains a decision in a language different from the input, does the explanation lose faithfulness? In other words, does it still accurately reflect the model’s underlying decision process?* Figure 1 illustrates the central tension we investigate – whether generating explanations in English, rather than in the input language, increases perceived plausibility while reducing faithfulness. Prior research has extensively studied the faithfulness of self-explanations in monolingual English settings (Jacovi and Goldberg, 2020) and established the fundamental distinction between *plausibility* (how human-like an explanation appears) and *faithfulness* (the causal link between the explanation and the model’s prediction) (Wiegreffe et al., 2021). More recently, efforts have shifted toward cross-lingual settings, exploring attribution faithfulness across translated pairs (Vamvas and Sennrich, 2023). Yet, these studies typically keep the input and reporting languages aligned, failing to treat the **reporting language itself** as an independent experimental variable.

We address this gap by systematically evaluating explanation-language mismatch and treating the reporting language as a controlled experimental variable. We build upon the work of (Huang et al., 2023), who suggest that multilingual LLMs often display an *English bias* in reasoning. We hypothesize that this bias, often described as a trade-off between adequacy and fluency (Conneau et al., 2020), results in a phenomenon we term the *plausibility-faithfulness trade-off*. Across three tasks – natural language inference (NLI) (Camburu et al., 2018), fact verification (Thorne et al., 2018), and hate speech detection (Mathew et al., 2020) – we observe that English explanations often achieve higher span agreement with human rationales than explanations generated in the input language, particularly for reasoning-intensive and factual tasks. A human-rated subsample confirms that span agreement is moderately correlated with perceived plausibility ($\rho = 0.67$, $p < 0.001$; Appendix H). However, deletion-based perturbation tests following ERASER-style evaluation (DeYoung et al., 2020) indicate that these same English explanations are frequently less faithful to the features that causally drive the model’s predictions. Through this study, we make the following contributions:

1. We provide the first controlled empirical study of explanation-language mismatch, treating

the reporting language as an independent variable and measuring its effect on both span-level agreement and perturbation-based faithfulness (comprehensiveness and sufficiency).

2. We identify a plausibility–faithfulness trade-off: across multiple languages and tasks, English-pivot explanations can show higher span agreement with human rationales while their cited evidence becomes less causally necessary for the model’s prediction, as confirmed by multiple faithfulness probes and prompt sensitivity analysis.
3. We show that this effect is task-dependent: for socially nuanced classification, English pivots degrade both dimensions, revealing a distinct failure mode tied to loss of pragmatic cues.

2 Related work

Faithfulness and plausibility in explanations: (Jacovi and Goldberg, 2020) formalized the distinction between faithfulness (whether an explanation reflects the model’s actual reasoning) and plausibility (whether it appears convincing to humans), establishing that these properties are independent and can diverge. (Wiegreffe et al., 2021) operationalized faithfulness measurement for text classifiers via perturbation tests. The ERASER benchmark (DeYoung et al., 2020) standardized evaluation through comprehensiveness and sufficiency metrics, which we adopt. Recent work has shown that LLM self-explanations can be persuasive yet unfaithful, functioning as post-hoc rationalizations (Turpin et al., 2023; Lanham et al., 2023).

Cross-lingual explainability: (Vamvas and Sennrich, 2023; Banerjee et al., 2025a) studied attribution faithfulness across translated pairs, finding that translation can shift saliency maps. (Banerjee et al., 2025b) explored cross-lingual transfer of explainable NLP capabilities from safety perspective. However, both lines of work keep the input and explanation languages aligned. Prior work has also examined multilingual explanation faithfulness from complementary perspectives. Zaman and Belinkov (2022) evaluate the plausibility and faithfulness of attribution methods for multilingual NLI and propose a cross-lingual faithfulness evaluation based on word alignment. Zhao and Aletras (2024) compare feature-attribution faithfulness between multilingual and monolingual models and show that

faithfulness can vary substantially across model and language settings. More recent work has also studied multilingual counterfactual explanations and differences between direct target-language generation and translation-mediated generation (Wang et al., 2026). In contrast, we treat the *reporting language* itself as the controlled variable: the input, task, and model remain fixed while only the requested explanation language changes. We differ by treating the reporting language itself as a controlled variable.

English bias in multilingual LLMs: (Huang et al., 2023) demonstrated that multilingual LLMs often reason more effectively in English, and (Conneau et al., 2020) characterized the adequacy–fluency tension in multilingual models. These findings motivate our hypothesis that English-pivot explanations may be optimized for fluency at the cost of faithfulness to non-English input cues.

3 Experimental framework

To investigate the relationship between the reporting language and model faithfulness, we design a controlled experimental setup that isolates the language of the explanation while keeping the task and input semantics constant.

3.1 Linguistic conditions

For each dataset, we evaluate three experimental conditions that differ only in the input and reporting languages, thereby isolating the effect of reporting-language mismatch.

1. Condition A ($EN \rightarrow EN$): The input and the explanation are both in English. This condition provides a monolingual reference point and approximates an upper bound on explanation quality.

2. Condition B ($L_{\text{native}} \rightarrow L_{\text{native}}$): The input and the explanation are both in the same non-English language. This condition captures language-aligned multilingual usage.

3. Condition C ($L_{\text{native}} \rightarrow EN$): The input is in a non-English language, but the explanation is generated in English. This condition instantiates the reporting-language mismatch typical of English-centric deployments.

Reverse-direction control: Our primary setting focuses on $L_{\text{native}} \rightarrow EN$ because it reflects the deployment scenario motivating this work: non-English user inputs are reported to English-speaking auditors or downstream teams. As a diagnostic control, we additionally evaluate the re-

verse $EN \rightarrow L_{\text{native}}$ direction with Qwen2.5-7B across all three tasks. This control allows us to separate a general reporting-language mismatch effect from an additional effect associated specifically with pivoting into English; results are discussed in Appendix B.

3.2 Datasets and tasks

We use three benchmark datasets spanning distinct reasoning demands: (1) **e-SNLI** (Camburu et al., 2018): natural language inference, which tests compositional and logical reasoning. (2) **FEVER** (Thorne et al., 2018): fact verification, which requires evidence identification and factual consistency with supporting context. (3) **HateXplain** (Mathew et al., 2020): hate-speech classification, which depends on sensitivity to social nuance.

3.3 Multilingual data construction

We evaluate five languages: **English** (EN) and four non-English languages that commonly appear in multilingual applications - **Chinese** (ZH), **Hindi** (HI), **Arabic** (AR), and **Bengali** (BN). For each dataset, we construct semantically matched test sets by translating the original English test instances into each target language while preserving the task format and gold labels (see example in Fig 2). We use **NLLB-200 (3.3B distilled)** (Costa-jussà et al., 2022), accessed via the official Hugging Face checkpoint, to perform the translations. We selected NLLB-200 for three reasons. First, it provides uniform open-weight coverage of all four target languages (Chinese, Hindi, Arabic, Bengali), avoiding the heterogeneous quality that arises when mixing translation systems across languages. Second, per-direction chrF++ and spBLEU scores for every language pair used in our study are publicly reported on the FLORES-200 evaluation benchmark (Costa-jussà et al., 2022; NLLB Team et al., 2024), enabling readers to cross-reference baseline translation quality. Third, the model is fully open-weight and reproducible. To keep plausibility evaluation consistent across languages, we also translate the human rationale signal associated with each instance. For **e-SNLI**, we translate the natural-language explanation; for **FEVER**, we use the gold evidence sentences as the rationale signal; for **HateXplain**, we translate the full text and the annotated highlight spans. We apply Unicode normalization and filter instances with empty or malformed translations. To verify translation quality, we conduct a structured audit: for each target

Label: entailment	
EN (source):	
Premise:	"It was raining, so she took an umbrella."
Hypothesis:	"She used an umbrella because it was raining."
Human explanation (rationale):	"The premise states rain and taking an umbrella, which supports the hypothesis."
HI (translated):	
Premise:	"बारिश हो रही थी, इसलिए उसने छाता लिया।"
Hypothesis:	"उसने बारिश की वजह से छाता इस्तेमाल किया।"
Human explanation (translated rationale):	"वाक्य में बारिश और छाता लेने की बात है, इसलिए कथन सही है।"

Figure 2: Sample translation example.

language, two bilingual annotators independently evaluate 50 randomly sampled instances through Prolific¹ on (1) semantic preservation (*whether the meaning is faithfully retained on a 3-point scale: preserved / minor shift / major shift*) and (2) label validity (*whether the gold label remains correct for the translated instance*). Full audit statistics are reported in Appendix I. Instances flagged as label-altering by either annotator were excluded. In addition, we compute chrF++ scores (Popović, 2015) between back-translations and the original English to provide an automatic cross-check. These measures address the concern that translation artifacts could confound plausibility or faithfulness measurements.

3.4 Evaluation metrics

We evaluate explanations along four complementary dimensions. All span-level metrics are computed over the evidence spans $\mathcal{E}_m(x)$ that the model copies verbatim from the input.

Notation: For an input instance x , let $\mathcal{I}(x)$ denote its tokenized input sequence. Let $\mathcal{E}_m(x) \subseteq \mathcal{I}(x)$ be the set of input-token indices covered by the model-produced evidence spans, and let $\mathcal{E}_h(x) \subseteq \mathcal{I}(x)$ be the set of input-token indices covered by the human rationale annotation (when available).²

Span agreement: We measure the overlap between model-identified and human-annotated evidence using token-level F1:

$$\text{SpanAgr}(x) = \frac{2|\mathcal{E}_m(x) \cap \mathcal{E}_h(x)|}{|\mathcal{E}_m(x)| + |\mathcal{E}_h(x)|} \quad (1)$$

We adopt the term *span agreement* rather than "plausibility" to reflect that this metric captures lexical overlap with human rationales, not perceived

explanation quality - which would require human judgment (Jacovi and Goldberg, 2020).

Comprehensiveness: Comprehensiveness captures whether the model's cited evidence is causally *necessary* for its prediction (DeYoung et al., 2020). We construct a perturbed input $x' = \text{MASK}(x, \mathcal{E}_m(x))$ by replacing the tokens indexed by $\mathcal{E}_m(x)$ with a sentinel token (preserving sequence length) and recomputing the prediction:

$$\text{Comp} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}[f(x_i) \neq f(\text{MASK}(x_i, \mathcal{E}_m(x_i)))] \quad (2)$$

A higher comprehensiveness indicates that removing the model-identified evidence is more likely to change the prediction, implying stronger causal necessity.

Sufficiency: Sufficiency captures whether the cited evidence alone is enough to sustain the prediction (DeYoung et al., 2020). We construct a reduced input $x'' = \text{KEEP}(x, \mathcal{E}_m(x))$ by masking all tokens except those in $\mathcal{E}_m(x)$ and recomputing:

$$\text{Suff} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}[f(x_i) = f(\text{KEEP}(x_i, \mathcal{E}_m(x_i)))] \quad (3)$$

Higher sufficiency means the evidence spans alone reproduce the original prediction, indicating that the cited evidence is *self-contained*. Together, comprehensiveness and sufficiency provide complementary views of faithfulness as follows.

Comprehensiveness $\uparrow$ - removing the evidence changes the prediction (evidence is *necessary*).

Sufficiency $\uparrow$ - keeping only the evidence preserves the prediction (evidence is *adequate*).

An explanation is well-grounded when *both* are high. Conversely, an English-pivot explanation can produce evidence that is highly self-contained (high sufficiency) yet not causally necessary (low comprehensiveness) - a signature of rhetorically tight summaries that do not reflect the model's actual decision cues.

Note on sufficiency in cross-lingual settings: In monolingual rationale extraction, high sufficiency is typically considered desirable: the cited evidence alone reproduces the prediction, indicating the explanation captures what the model needs. *In our cross-lingual setting, sufficiency must be interpreted jointly with comprehensiveness.* When sufficiency rises while comprehensiveness falls, the cited evidence is self-contained as a standalone justification but is no longer causally necessary — the model could have predicted the same label from other input cues. This combination is the diagnostic signature of post-hoc rationalization that moti-

¹<https://www.prolific.com/>

²For tasks where human rationales are provided as free-form text (e.g., e-SNLI explanations), we align them to the input via exact substring matching at the character level (for non-Latin scripts) or word level (for English), and treat the matched input-token indices as $\mathcal{E}_h(x)$. This conservative approach underestimates overlap for paraphrased rationales, biasing against inflated span agreement. Worked examples for each language are provided in Appendix C.3.

vates our analysis. For consistency, all tables in this paper report sufficiency with the conventional $\uparrow$ (“higher is better in isolation”) arrow, but under $L_{\text{native}} \rightarrow \text{EN}$, rising sufficiency paired with falling comprehensiveness indicates a shift toward summary-like (rather than causally grounded) evidence selection.

Task accuracy: To disentangle the explanation-language effect from base language competence, we report label accuracy for every experimental condition:

$$\text{Acc} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}[f(x_i) = y_i] \quad (4)$$

where y_i is the gold label. If accuracy is comparable across $L_{\text{native}} \rightarrow L_{\text{native}}$ and $L_{\text{native}} \rightarrow \text{EN}$, observed faithfulness differences can be attributed to the explanation-language pivot rather than degraded input comprehension.

3.5 Model selection

We perform our experiments using two state-of-the-art multilingual LLMs namely Qwen2.5-7B (Yang et al., 2025) and Llama3.1-8B (Grattafiori et al., 2024). These models are open-weight with strong multilingual capability, but they differ in model family and training data composition. This contrast allows us to test whether the plausibility–faithfulness trade-off holds across open-weight LLMs rather than arising from a single model. We additionally evaluate Gemma3-27B as a larger-model robustness check. Unlike other models, which form our primary experimental setting, Gemma3-27B is used to test whether the observed phenomenon persists at substantially larger model scale. We evaluate it on **e-SNLI**, **FEVER**, and **HateXplain** across all four non-English languages; results are reported in Section B.

3.6 Prompt design and generation order

Output format: For all conditions and tasks, we use a structured output format that requests three fields in sequence: (1) a predicted label, (2) evidence spans copied verbatim from the input, and (3) a free-text explanation in the required language (see templates in Appendix C). The evidence spans are always retained in the input language, even when the explanation is generated in English (Condition C), ensuring that comprehensiveness and sufficiency are computed over identical input-language tokens across conditions.

Autoregressive coupling: A natural concern is why the *explanation language* should affect evi-

dence span selection, given that evidence appears before the explanation in the output format. We note that the model generates all three fields in a single autoregressive pass. The instruction specifying the explanation language—e.g., “Write a brief explanation in English” vs. “Write a brief explanation in Hindi”—is present in the prompt *before* any generation begins, and thus conditions the model’s entire output distribution, including its selection of evidence spans. In other words, the language instruction acts as a global prior over the generation, not a local constraint applied only to the explanation field. To verify this empirically, we conduct an **ordering ablation**: we create an alternative prompt variant in which the explanation is requested *before* the evidence (i.e., label $\rightarrow$ explanation $\rightarrow$ evidence). To keep this robustness check computationally bounded, we run the ablation for Qwen2.5-7B across all three datasets and all non-English languages. Table 10 in Appendix F reports the results. The comprehensiveness and span agreement patterns remain stable under the reversed ordering, with only small absolute changes across task–language cells. This suggests that the explanation-language instruction, not the output field order, drives the observed differences.

Prompt sensitivity: To ensure that our findings are not artifacts of specific prompt phrasing, we create five paraphrased variants of each prompt template. Each variant preserves the task semantics and output format while varying surface wording (e.g., “Predict the correct label” $\rightarrow$ “Determine the appropriate category”; “Write a brief explanation” $\rightarrow$ “Provide a short justification”). All five variants are listed in Appendix C. Per-condition results are reported in Section E.2.

Decoding and reproducibility: We use greedy decoding ($T = 0$), a maximum of 256 new tokens, and a single generation pass. As a robustness check, we repeat the experiments with three random seeds at $T = 0.7$; all evaluation metrics change by less than 0.02. We release all prompt templates and decoding hyperparameters, including the paraphrased and reversed-order variants, together with the processed translations, audit annotations, and evaluation scripts.

4 Results

Before examining explanation quality, we verify that the model’s task performance is comparable across linguistic conditions. Tables 1–3 include

task accuracy for every setting. Across all three datasets, accuracy differences between $L_{\text{native}} \rightarrow L_{\text{native}}$ and $L_{\text{native}} \rightarrow \text{EN}$ are consistently small: the mean absolute accuracy gap is 0.005 on **e-SNLI**, 0.006 on **FEVER**, and 0.004 on **HateXplain**, with no gap exceeding 0.01 in any individual (language, model) cell. This indicates that the models comprehend non-English inputs comparably well regardless of the explanation language, allowing us to attribute observed faithfulness differences to the explanation-language pivot rather than degraded input understanding. We note that accuracy under $L_{\text{native}} \rightarrow L_{\text{native}}$ is itself slightly lower than $\text{EN} \rightarrow \text{EN}$ (e.g., 0.811 vs. 0.847 for Hindi on **e-SNLI** with Qwen2.5-7B), reflecting the expected base multilingual performance gap; critically, however, this gap does not widen further under $L_{\text{native}} \rightarrow \text{EN}$, confirming that requiring English explanations does not additionally impair task comprehension.

Next we compare three settings namely **condition A** ($\text{EN} \rightarrow \text{EN}$), **condition B** ($L_{\text{native}} \rightarrow L_{\text{native}}$), and **condition C** ($L_{\text{native}} \rightarrow \text{EN}$) in the plausibility-faithfulness axis. On **e-SNLI**, Table 1 shows that Qwen2.5-7B exhibits the clearest plausibility-faithfulness trade-off. When generating explanations in English instead of the input language, span agreement often increases (e.g., Bengali: $0.375 \rightarrow 0.507$ for Qwen2.5-7B), while comprehensiveness drops substantially (Hindi: $0.719 \rightarrow 0.461$ for Qwen2.5-7B) and sufficiency *rises* (Hindi: $0.253 \rightarrow 0.387$). The combined pattern is diagnostic: the cited evidence becomes more *self-contained as a standalone summary* (higher sufficiency) yet less *causally necessary* for the prediction (lower comprehensiveness). In other words, English-pivot evidence reads like a clean justification but is no longer the span the model actually relies on. Crucially, task accuracy remains stable across conditions ($\Delta < 0.01$), ruling out degraded input understanding as an explanation.

On **FEVER** (Table 2), the $L \rightarrow \text{EN}$ condition shows the same pattern: span agreement often increases, comprehensiveness drops, and sufficiency rises. This pattern is pronounced for Hindi and Arabic in both models. For Chinese, both span agreement and comprehensiveness decrease under $L \rightarrow \text{EN}$, suggesting that the pivot degrades explanation quality along multiple dimensions for this language-task pair. Across all languages, the convergent signal — comprehensiveness falling while sufficiency rising — strengthens the interpretation that English pivots produce evidence spans that

Settings	Acc. $\uparrow$		Comp. $\uparrow$		Suff. $\uparrow$		Span agr. $\uparrow$	
	Qwen	Llama	Qwen	Llama	Qwen	Llama	Qwen	Llama
$\text{EN} \rightarrow \text{EN}$	0.847	0.791	0.657	0.583	0.312	0.378	0.595	0.453
$L_{\text{ZH}} \rightarrow L_{\text{ZH}}$	0.823	0.764	0.577	0.635	0.341	0.298	0.597	0.129
$L_{\text{ZH}} \rightarrow \text{EN}$	0.819	0.758	0.507	0.573	0.409	0.362	0.516	0.379
$L_{\text{HI}} \rightarrow L_{\text{HI}}$	0.811	0.748	0.719	0.678	0.253	0.271	0.417	0.545
$L_{\text{HI}} \rightarrow \text{EN}$	0.806	0.741	0.461	0.604	0.387	0.334	0.467	0.509
$L_{\text{AR}} \rightarrow L_{\text{AR}}$	0.818	0.772	0.659	0.781	0.279	0.194	0.500	0.391
$L_{\text{AR}} \rightarrow \text{EN}$	0.814	0.766	0.484	0.662	0.402	0.289	0.538	0.512
$L_{\text{BN}} \rightarrow L_{\text{BN}}$	0.794	0.721	0.488	0.445	0.398	0.412	0.375	0.267
$L_{\text{BN}} \rightarrow \text{EN}$	0.789	0.715	0.419	0.401	0.451	0.468	0.507	0.472

Table 1: The plausibility-faithfulness trade-off on **e-SNLI**. All metrics use the higher-is-better convention ($\uparrow$). **Blue** : higher comprehensiveness (evidence is causally necessary). **Green** : higher sufficiency (evidence is self-contained as a standalone summary). **Yellow** : higher span agreement (human-aligned). The trade-off is visible in rows where, under $L_{\text{native}} \rightarrow \text{EN}$, **yellow** and **green** co-occur with a *loss* of **blue** — i.e., the cited spans look like clean summaries that humans agree with, but they are no longer the spans the model causally relies on.

look like good summaries but are not the spans causally driving the model’s fact-verification reasoning.

Table 3 reveals a distinct failure mode on **Ha-**

Settings	Acc. $\uparrow$		Comp. $\uparrow$		Suff. $\uparrow$		Span agr. $\uparrow$	
	Qwen	Llama	Qwen	Llama	Qwen	Llama	Qwen	Llama
$\text{EN} \rightarrow \text{EN}$	0.782	0.731	0.271	0.201	0.618	0.693	0.198	0.112
$L_{\text{ZH}} \rightarrow L_{\text{ZH}}$	0.754	0.709	0.931	0.170	0.124	0.712	0.255	0.247
$L_{\text{ZH}} \rightarrow \text{EN}$	0.749	0.703	0.500	0.138	0.389	0.749	0.198	0.169
$L_{\text{HI}} \rightarrow L_{\text{HI}}$	0.741	0.698	0.579	0.247	0.334	0.621	0.197	0.147
$L_{\text{HI}} \rightarrow \text{EN}$	0.735	0.691	0.101	0.199	0.612	0.674	0.243	0.162
$L_{\text{AR}} \rightarrow L_{\text{AR}}$	0.748	0.718	0.634	0.317	0.298	0.573	0.206	0.224
$L_{\text{AR}} \rightarrow \text{EN}$	0.742	0.712	0.240	0.120	0.524	0.691	0.251	0.284
$L_{\text{BN}} \rightarrow L_{\text{BN}}$	0.723	0.684	0.356	0.552	0.487	0.398	0.207	0.169
$L_{\text{BN}} \rightarrow \text{EN}$	0.718	0.678	0.172	0.427	0.612	0.489	0.238	0.197

Table 2: Results on **FEVER** (fact verification). Notation and color coding follow Table 1. All metrics use the higher-is-better convention ($\uparrow$).

teXplain. Under $L \rightarrow \text{EN}$, comprehensiveness decreases consistently across all four languages for both models (e.g., Chinese: Qwen2.5-7B $0.714 \rightarrow 0.520$, Llama3.1-8B $0.884 \rightarrow 0.676$), while sufficiency increases as before. However, unlike **e-SNLI** and **FEVER**, span agreement does *not* reliably increase — it often decreases (e.g., Bengali Qwen2.5-7B: $0.359 \rightarrow 0.280$). This indicates that for socially nuanced classification, English pivots fail to produce either faithful or human-aligned explanations, likely because pragmatic cues such as slang, code-switching, and culturally grounded expressions do not survive the implicit translation step.

Statistical validation: We test the significance of $L \rightarrow L$ vs. $L \rightarrow \text{EN}$ differences using paired permutation tests (10,000 permutations). Comprehensiveness differences are significant ($p < 0.05$) in 19/24 language-task-model cells; the non-

Settings	Acc. $\uparrow$		Comp. $\uparrow$		Suff. $\uparrow$		Span agr. $\uparrow$	
	Qwen	Llama	Qwen	Llama	Qwen	Llama	Qwen	Llama
EN $\rightarrow$ EN	0.714	0.689	0.651	0.754	0.301	0.227	0.325	0.294
$L_{ZH} \rightarrow L_{ZH}$	0.683	0.661	0.714	0.884	0.248	0.112	0.274	0.333
$L_{ZH} \rightarrow \text{EN}$	0.679	0.657	0.520	0.676	0.378	0.256	0.283	0.256
$L_{HI} \rightarrow L_{HI}$	0.691	0.674	0.616	0.820	0.312	0.159	0.354	0.221
$L_{HI} \rightarrow \text{EN}$	0.687	0.668	0.567	0.804	0.361	0.178	0.290	0.261
$L_{AR} \rightarrow L_{AR}$	0.688	0.671	0.643	0.793	0.289	0.184	0.284	0.259
$L_{AR} \rightarrow \text{EN}$	0.684	0.665	0.533	0.779	0.367	0.201	0.299	0.265
$L_{BN} \rightarrow L_{BN}$	0.676	0.652	0.668	0.810	0.274	0.171	0.359	0.277
$L_{BN} \rightarrow \text{EN}$	0.672	0.647	0.613	0.788	0.332	0.194	0.280	0.254

Table 3: Results on **HateXplain**. Notation and color coding follow Table 1. Unlike **e-SNLI** and **FEVER**, span agreement does *not* reliably increase under $L_{\text{native}} \rightarrow \text{EN}$ (yellow cells appear in the $L \rightarrow L$ rows for several languages), revealing a distinct failure mode where English pivots lose pragmatic and culturally grounded cues.

significant cases are concentrated in Bengali, the lowest-resource language in our set.

Semantic similarity check: To verify that our span agreement findings are not artifacts of surface-level tokenization differences across scripts, we compute BERTScore F1 using bert-base-multilingual-cased (Table 4). The two metrics agree directionally in 42/48 cells ($r = 0.88$), confirming that the patterns in Tables 1–3 are genuine. Full per-task results are in Appendix G.

Task	Dir. agreement	Corr. (r)
e-SNLI	15/16	0.91
FEVER	14/16	0.88
HateXplain	13/16	0.84
Overall	42/48	0.88

Table 4: Correspondence between span agreement (lexical) and BERTScore F1 (semantic). Directional agreement counts cells where both metrics shift the same way under the English pivot.

Task	Lang	Δ Comp.		Δ Suff.		Δ Span agr.	
		Qwen	Llama	Qwen	Llama	Qwen	Llama
e-SNLI	ZH	-0.070	-0.062	+0.068	+0.064	-0.081	+0.250
	HI	-0.258	-0.074	+0.134	+0.063	+0.050	-0.036
	AR	-0.175	-0.119	+0.123	+0.095	+0.038	+0.121
	BN	-0.069	-0.044	+0.053	+0.056	+0.132	+0.205
FEVER	ZH	-0.431	-0.032	+0.265	+0.037	-0.057	-0.078
	HI	-0.478	-0.048	+0.278	+0.053	+0.046	+0.015
	AR	-0.394	-0.197	+0.226	+0.118	+0.045	+0.060
	BN	-0.184	-0.125	+0.125	+0.091	+0.031	+0.028
HateXplain	ZH	-0.194	-0.208	+0.130	+0.144	+0.009	-0.077
	HI	-0.049	-0.016	+0.049	+0.019	-0.064	+0.040
	AR	-0.110	-0.014	+0.078	+0.017	+0.015	+0.006
	BN	-0.055	-0.022	+0.058	+0.023	-0.079	-0.023

Table 5: Shift in metrics when switching from $L_{\text{native}} \rightarrow L_{\text{native}}$ to $L_{\text{native}} \rightarrow \text{EN}$ ($\Delta = \text{pivot} - \text{native}$). All metrics follow the higher-is-better convention. **Red**: degradation on that axis.

Blue: improvement on that axis. *Headline pattern*: Δ Comp. is negative in 24/24 cells (evidence becomes less causally necessary), while Δ Suff. is positive in 24/24 cells (evidence becomes more self-contained). This dual signature — evidence that is more summary-like but less causal — defines the “deceptive zone” of Figure 1. Δ Span agr. is mixed, reflecting the task-dependent nature of the surface-level agreement shift.

Human plausibility validation (subsample): To verify that span agreement tracks human-perceived

plausibility, we recruit three bilingual annotators per language via Prolific to rate 100 randomly sampled (instance, explanation) pairs per condition on a 5-point Likert scale: “How convincing is this explanation as a justification of the predicted label?” Annotators see only the input, predicted label, and explanation — not the gold label or evidence spans. We compute the Spearman correlation between mean human ratings and span agreement scores, finding $\rho = 0.67$ across all conditions ($p < 0.001$), supporting the use of span agreement as a plausibility proxy.

Prompt-paraphrase robustness: To rule out the possibility that the trade-off is an artifact of specific prompt phrasing, we re-ran our experiments using five paraphrased prompt variants per condition (variants in Appendix C). Table 8 reports results for e-SNLI on Qwen2.5-7B; the corresponding table for Llama3.1-8B is Table 9 in Appendix E.2.

The trade-off pattern — lower comprehensiveness and higher sufficiency under $L_{\text{native}} \rightarrow \text{EN}$ compared to $L_{\text{native}} \rightarrow L_{\text{native}}$ — holds consistently across all five prompt formulations. *The between-condition gaps are roughly an order of magnitude larger than the within-condition prompt variance.* For Hindi e-SNLI on Qwen2.5-7B, the $L \rightarrow L$ vs. $L \rightarrow \text{EN}$ comprehensiveness gap is $0.719 - 0.461 = 0.258$, while the standard deviation across the five prompt paraphrases within each condition is approximately 0.03 — a gap-to-noise ratio of $\sim 8\times$. This effect-size to noise-floor ratio confirms that the observed shift is far larger than what could plausibly be explained by prompt-phrasing variance, supporting the claim that the explanation-language instruction — not surface prompt wording — drives the observed effects.

Output-ordering robustness: A natural concern is whether requiring evidence *before* explanation in the output format prevents the explanation language from influencing evidence selection. We test this by reversing the field order (label $\rightarrow$ explanation $\rightarrow$ evidence) and re-running on all three datasets. Results are reported in Tables 10–11 (Appendix F). Across all 12 (dataset $\times$ model $\times$ language) cells we tested, no pairwise difference between orderings is statistically significant (paired permutation test, 10,000 permutations, $p > 0.3$ for all cells). Across the $L_{\text{native}} \rightarrow \text{EN}$ cells, the mean Jaccard similarity between evidence-span sets across the two output orderings is 0.78 ± 0.06 . When averaged across all three linguistic conditions, the corresponding mean is 0.87. We report

both scopes explicitly because evidence selection is slightly less stable under the English-pivot condition. This confirms that the explanation-language instruction acts as a global prior over the autoregressive output distribution, not a local constraint applied only to the explanation field.

Rationale and translation fidelity: Since span agreement additionally relies on translated human rationales, we separately audit 40 translated input–rationale pairs per target language using two bilingual annotators. The proportion judged fully aligned is 91% for Chinese, 89% for Hindi, 87% for Arabic, and 84% for Bengali, with Cohen’s κ of 0.76, 0.73, 0.71, and 0.68, respectively. Excluding partially aligned and misaligned instances changes every condition mean by less than 0.02 and preserves all patterns in Tables 1–3. We further compute COMET-22 scores of 0.861 (ZH), 0.844 (HI), 0.852 (AR), and 0.819 (BN). As a full-dataset safeguard, we also screen all translations with CometKiwi, remove the bottom 2% by estimated translation quality, and re-run the main analyses; all reported patterns remain unchanged. Full audit results are provided in Appendix I. All experiments use the same translated inputs across the three linguistic conditions; only the required language of the model’s explanation is changed. This design isolates explanation-language mismatch while keeping task semantics constant.

Differences in Qwen and Llama: Although the pivot-induced direction is consistent across both models, their absolute faithfulness scores differ substantially in some settings. These differences are unlikely to be prompt noise: the largest between-model gaps, on **FEVER**, are approximately an order of magnitude larger than the within-condition prompt variation ($\sigma \approx 0.03$; Tables 8–9). In particular, Llama3.1–8B already exhibits low comprehensiveness and high sufficiency under the EN→EN **FEVER** baseline (0.201/0.693). When we remove the evidence context and provide only the **FEVER** claim, Llama3.1–8B retains 89% of its original accuracy, compared with 56% for Qwen2.5–7B, suggesting substantially greater reliance on parametric knowledge. On **HateXplain**, Llama3.1–8B also tends to select longer evidence spans, which mechanically increases comprehensiveness, while near-ceiling values such as Qwen2.5–7B’s 0.931 comprehensiveness on **FEVER**-ZH arise when the selected evidence covers almost the entire short claim. These model-specific differences affect absolute values

but not our central result: ΔComp is negative and ΔSuff is positive for both models in all 24 original (task, language, model) comparisons.

5 Error analysis

To move beyond aggregate metrics, we examine representative instances from each quadrant of the span agreement $\times$ comprehensiveness space. We identify four patterns that illustrate the mechanisms underlying our aggregate findings.

Case 1: High span agreement, low comprehensiveness (“Deceptive”)

Task: e-SNLI **Lang:** Hindi **Model:** Qwen2.5–7B **Gold:** Entailment
Premise: “*barish ho rahi thi, isliye usne chhata liya aur ghar se nikli.*” (It was raining, so she took an umbrella and left the house.)
Hypothesis: “*usne barish ki wajah se chhata istemal kiya.*” (She used an umbrella because of the rain.)

Condition B ($L_{HI} \rightarrow L_{HI}$)

Evidence: “barish ho rahi thi”; “chhata liya”

Explanation: “Premise mein barish aur chhata lene ka zikr hai, isliye hypothesis samarthit hai.”

Perturbation: Masking both spans $\rightarrow$ prediction **flips** to *neutral*. **Comp:** 1.0 **SpanAgr:** 0.67

Condition C ($L_{HI} \rightarrow \text{EN}$)

Evidence: “barish ho rahi thi” only — “chhata liya” dropped

Explanation: “The premise states that it was raining, which directly supports the hypothesis that she used an umbrella due to rain.”

Perturbation: Masking span $\rightarrow$ prediction **remains entailment**. **Comp:** 0.0 **SpanAgr:** 0.50

Takeaway: The English pivot narrows evidence to one span matching the human rationale but *not causally necessary*—the prediction holds without it. The uncited “chhata liya” likely drives the decision, while the English explanation constructs a fluent narrative around only the cited span.

Case 2: Low span agreement, high comprehensiveness (“Honest but misaligned”)

Task: FEVER **Lang:** Arabic **Model:** Qwen2.5–7B **Gold:** Supports
Claim: “*al-qahira hiya asimatu misr.*” (Cairo is the capital of Egypt.)
Context: “*misr dawla fi shamal ifriqya. asimatuha al-qahira wa hiya akbar muduniha min haythu al-sukkan.*” (Egypt is a country in North Africa. Its capital is Cairo and it is its largest city by population.)

Condition B ($L_{AR} \rightarrow L_{AR}$)

Evidence: “asimatuha al-qahira”

Explanation: “al-siyaq yadhkur sarihatan anna asimata misr hiya al-qahira, mimma yad’am al-iddi’a.” (The context explicitly states that the capital of Egypt is Cairo, which supports the claim.)

Perturbation: Masking span $\rightarrow$ prediction **flips** to *not enough info*. **Comp:** 1.0 **SpanAgr:** 0.33

Human rationale: “*misr dawla fi shamal ifriqya. asimatuha al-qahira*” (longer span including geographic context).

Takeaway: The model cites a tighter, genuinely *causal* span than the human annotator. Low span agreement here does *not* indicate poor explanation — it reflects a more precise causal trigger than the longer human rationale.

Case 3: Both degrade (“Social signal loss”)

Task: HateXplain **Lang:** Chinese **Model:** Llama3.1–8B **Gold:** Hate speech
Input: “*na xie [slur-ZH] dou shi laji, yinggai gun chu women de chengshi.*” (Those [slur] are all trash, they should get out of our city.)

Condition B ($L_{ZH} \rightarrow L_{ZH}$)

Evidence: “[slur-ZH] dou shi laji”

Explanation: (in Chinese) “The text contains a derogatory term targeting [group] and calls them trash, which constitutes hate speech.”

Perturbation: Masking span $\rightarrow$ prediction **flips** to *offensive* (downgraded from hate speech). **Comp:** 1.0 **SpanAgr:** 0.80

Condition C ($L_{ZH} \rightarrow \text{EN}$)

Evidence: “dou shi laji, yinggai gun chu women de chengshi” (broader span; *slur omitted*)
Explanation: “The text expresses hostility toward a group by calling them trash and demanding their removal from the city, which constitutes hate speech.”
Perturbation: Masking span → prediction **remains** *hate speech* (no flip). **Comp:** 0.0 **SpanAgr:** 0.40

Takeaway: English pivot loses the culturally specific slur — the actual causal cue — and substitutes generic hostility language. Both faithfulness and human-alignment degrade. Signature failure mode of English pivots on socially nuanced tasks.

Case 4: Both improve (“Rare ideal”)

Task: FEVER **Lang:** Bengali **Model:** Qwen2.5-7B **Gold:** Supports
Claim: “rabindranath thakur 1913 sale nobel puroshkar peyechhilen.” (Rabindranath Tagore received the Nobel Prize in 1913.)
Context: “rabindranath thakur chhilen ekjon bangali kabi. tini 1913 sale sahitye nobel puroshkar peyechhilen, ja kono bharotiyo prothom bar peye thaken.” (Rabindranath Tagore was a Bengali poet. He received the Nobel Prize in Literature in 1913, which was the first time for any Indian.)

Condition B ($L_{BN} \rightarrow L_{BN}$)

Evidence: “1913 sale nobel puroshkar peyechhilen”
Explanation: (in Bengali) “The context confirms that Tagore received the Nobel Prize in 1913, matching the claim.”
Perturbation: Masking span → prediction **flips** to *not enough info*. **Comp:** 1.0 **SpanAgr:** 0.71

Condition C ($L_{BN} \rightarrow EN$)

Evidence: “1913 sale sahitye nobel puroshkar peyechhilen” (slightly longer; adds “sahitye” = “literature”)
Explanation: “The context explicitly states that Tagore received the Nobel Prize in Literature in 1913, which directly supports the claim.”
Perturbation: Masking span → prediction **flips** to *not enough info*. **Comp:** 1.0 **SpanAgr:** 0.83

Takeaway: Decisive evidence is language-independent (proper noun + date) and survives the pivot intact. Rare (<8% of instances), concentrated in factual claims with unambiguous named entities.

These four patterns clarify that the aggregate trade-off arises primarily from Cases 1 and 3: the English pivot tends to select evidence that is rhetorically coherent but causally peripheral (Case 1), and for socially nuanced tasks, it additionally loses culturally grounded signals (Case 3). The existence of Case 4 suggests that the severity of the trade-off depends on the linguistic specificity of the decisive cues in each instance.

Pattern	Span Agr.	Comp.	e-SNLI	HateX.
Case 1 (“Deceptive”)	↑	↓	41%	18%
Case 2 (“Honest”)	↓	↑	12%	9%
Case 3 (Both ↓)	↓	↓	22%	51%
Case 4 (Both ↑)	↑	↑	7%	4%
Mixed / no change	—	—	18%	18%

Table 6: Approximate distribution of error patterns across instances (Qwen2.5-7B, averaged over languages). Case 1 dominates in **e-SNLI** (reasoning task), while Case 3 dominates in **HateXplain** (social nuance task), consistent with the aggregate findings in Tables 1–3.

The distribution in Table 6 confirms two key findings from our aggregate analysis. First, the “deceptive” pattern (Case 1: span agreement up, comprehensiveness down) is the dominant mode in reasoning-intensive tasks like **e-SNLI**, accounting for the aggregate plausibility–faithfulness trade-off. Second, the “social signal loss” pattern (Case

3: both degrade) dominates in **HateXplain**, explaining why English pivots fail to improve even surface-level agreement on socially nuanced tasks. The rarity of Case 4 (<8% across tasks) underscores that successful English-pivot explanations are the exception rather than the rule. Since these instances are machine translated, a translated slur should not be interpreted as naturally occurring hate speech in the target language. We therefore treat cultural-cue loss as a plausible mechanism suggested by this analysis rather than conclusive evidence about naturally occurring multilingual hate speech.

Disappearance of causal cues under English pivot: We examine 100 instances in which a span that is causal under the native-language condition, i.e., *masking it changes the prediction* is no longer cited under the English-pivot condition. Two annotators per language categorize the dropped expression. Across languages, 38% correspond to translated renderings of slurs or insults, 22% to untranslated or mixed-language tokens, 14% to dialect or spelling markers, 11% to in-group references, and 15% to other pragmatic signals. The pattern differs by language: untranslated or mixed-language tokens are particularly common for Hindi (31%) and Bengali (27%). These observations provide a more concrete characterization of the cues associated with the **HateXplain** degradation.

6 Conclusion

Our results show that explanation language is not a neutral reporting choice. Across three tasks, five languages, and two model families, switching from native-language to English explanations consistently reduces comprehensiveness and increases sufficiency of the cited evidence spans, even when task accuracy remains stable. This trade-off is most pronounced for reasoning-intensive tasks (NLI, fact verification), where English pivots often produce fluent narratives loosely anchored to the model’s actual decision cues. For socially nuanced classification, English pivots degrade both faithfulness and span agreement, reflecting the loss of culturally grounded signals. We recommend three practices: (1) audit explanation faithfulness in input language, (2) report both comprehensiveness and sufficiency alongside any overlap-based metric, and (3) treat English rationales as communication summaries rather than faithful decision traces (Appendix D).

7 Limitations

Several limitations remain. First, our full factorial evaluation uses two open-weight 7–8B models and four non-English languages. Targeted experiments with Gemma3–27B across all three tasks, and with Spanish and Swahili on e-SNLI, reproduce the central pattern, but broader evaluation across additional large and closed-source models remains necessary. Second, our primary faithfulness evaluation uses perturbation-based comprehensiveness and sufficiency. Random-span masking, replacement-based perturbations, evidence-length matching, and probability-based variants preserve the result, but these measures remain behavioral proxies rather than direct observations of internal reasoning. Third, our non-English benchmark instances are machine translated. Human input and rationale audits, COMET-22, and CometKiwi filtering substantially reduce this concern, but translated socially nuanced content cannot fully reproduce naturally occurring linguistic and cultural phenomena. We therefore interpret the specific cultural-cue mechanism observed on HateXplain as a hypothesis that should be validated on natively written multilingual hate-speech datasets. Finally, our study measures metric-level plausibility–faithfulness divergence; whether such explanations actually mislead human auditors remains an important question for future human-subject evaluation.

8 Ethical considerations

The primary ethical concern raised by our work is the risk of “deceptive transparency” in high-stakes global deployments. When models generate fluent and persuasive English explanations for non-English inputs, they can create a false sense of security for auditors and decision-makers. Our findings show that English-pivot explanations can produce narratives that are coherent yet loosely anchored to the evidence that causally drives the prediction, as measured by comprehensiveness and sufficiency. Whether this metric-level divergence translates into actual auditor deception is an empirical question requiring human evaluation, which we view as important future work. Such metric-level divergence could lead to misplaced trust in automated systems for legal, medical, or public-service workflows. This introduces a significant accountability gap: *a model might provide a correct decision justified by hallucinated logic, making it difficult to detect underlying biases or errors*. Fur-

thermore, the tendency to prioritize English fluency over native-language faithfulness risks entrenching English-centric biases, potentially resulting in lower-quality service or less reliable explanations for non-English speaking populations. We recommend that English rationales in sensitive contexts be treated strictly as summaries rather than faithful decision traces to prevent the misuse.

References

2025. [Use real-time translation of conversations for service representatives and customers](#). States the feature is intended to help customer service managers or supervisors enhance team performance.
- Amazon Web Services. 2024. [Amazon translate: Machine translation service](#).
- Somnath Banerjee, Pratyush Chatterjee, Shanu Kumar, Sayan Layek, Parag Agrawal, Rima Hazra, and Animesh Mukherjee. 2025a. [Attributional safety failures in large language models under code-mixed perturbations](#). *Preprint*, arXiv:2505.14469.
- Somnath Banerjee, Sayan Layek, Pratyush Chatterjee, Animesh Mukherjee, and Rima Hazra. 2025b. [Soteria: Language-specific functional parameter steering for multilingual safety alignment](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9347–9364, Suzhou, China. Association for Computational Linguistics.
- Oana-Maria Camburu, Tim Rocktäschel, , and 1 others. 2018. [e-snli: Natural language inference with natural language explanations](#). In *NeurIPS*.
- 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](#). *Preprint*, arXiv:1911.02116.
- Marta R. Costa-jussà and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Ciamac Teh, and 1 others. 2020. [Eraser: A benchmark to evaluate rationalized nlp models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.
- Jeff Eiden. 2024. [Live translation with twilio and openai’s realtime api](#). Twilio Blog.
- 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, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. [Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.
- Alon Jacovi and Yoav Goldberg. 2020. [Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness?](#) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.
- Rohini Jadhav, Vishal Meshram, Amol Bhosle, Kailas Patil, Sital Dash, and Shrikant Jadhav. 2025. [Explainable multilingual and multimodal fake-news detection: toward robust and trustworthy ai for combating misinformation](#). *Frontiers in Artificial Intelligence*, 8:1690616.
- Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiušė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, and 11 others. 2023. [Measuring faithfulness in chain-of-thought reasoning](#). *Preprint*, arXiv:2307.13702.
- Binny Mathew, Punyajoy Saha, , and 1 others. 2020. [Hatexplain: A benchmark dataset for explainable hate speech detection](#). In *AAAI*.
- NLLB Team and 1 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630:841–846.
- Maja Popović. 2015. [chrF: character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- James Thorne, Andreas Vlachos, , and 1 others. 2018. [Fever: a large-scale dataset for fact extraction and verification](#). In *NAACL-HLT*.
- Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. [Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting](#). *Preprint*, arXiv:2305.04388.
- Jannis Vamvas and Rico Sennrich. 2023. [Towards unsupervised recognition of token-level semantic differences in related documents](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13543–13552, Singapore. Association for Computational Linguistics.
- Qianli Wang, Van Bach Nguyen, Yihong Liu, Fedor Splitt, Nils Feldhus, Christin Seifert, Hinrich Schuetze, Sebastian Möller, and Vera Schmitt. 2026. [Parallel universes, parallel languages: A comprehensive study on LLM-based multilingual counterfactual example generation](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 323–346, San Diego, California, United States. Association for Computational Linguistics.
- Sarah Wiegrefe, Ana Marasović, and Noah A. Smith. 2021. [Measuring association between labels and free-text rationales](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Kerem Zaman and Yonatan Belinkov. 2022. [A multilingual perspective towards the evaluation of attribution methods in natural language inference](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1556–1576, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). In *International Conference on Learning Representations*.
- Zhixue Zhao and Nikolaos Aletras. 2024. [Comparing explanation faithfulness between multilingual and monolingual fine-tuned language models](#). 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 3226–3244, Mexico City, Mexico. Association for Computational Linguistics.

A Appendix content

B Additional robustness and generalization 12

C Prompt templates and qualitative example 13

C.1 Prompt templates 13

C.2 Mini qualitative example 13

C.3 Rationale alignment details 14

D Practical recommendations 16

E Prompt paraphrases and sensitivity analysis	16
E.1 Prompt variants	16
E.2 Sensitivity results	17
F Ordering ablation	17
F.1 Reversed prompt template	17
F.2 Results	18
F.3 Analysis	18
G Semantic similarity check	18
G.1 Setup	18
G.2 Results: e-SNLI	18
G.3 Results: FEVER	19
G.4 Results: HateXplain	19
G.5 Summary and implications	20
H Human plausibility validation	20
I Translation quality audit	21
I.1 Rationale fidelity audit	22
I.2 COMET-based translation quality checks	22
J Detailed results for additional robustness and generalization	22
J.1 Perturbation robustness	22
J.2 Evidence-length control	22
J.3 Probability-based faithfulness . .	23
J.4 Reverse-direction language mismatch	23
J.5 Larger-model evaluation with Gemma3-27B	24
J.6 Additional languages	24
J.7 Detailed HateXplain causal-cue analysis	25
B Additional robustness and generalization	

Are the faithfulness results artifacts of masking?

Deletion-based faithfulness tests can create out-of-distribution inputs, so we conduct two additional controls. First, we mask a random span having the same length as the model-selected evidence. On **e-SNLI** with Qwen2.5-7B, random-span masking produces prediction-flip rates of only 0.144–0.156 across conditions, whereas masking the selected evidence changes predictions $2.7\text{--}4.9\times$ more often. For example, for Hindi the evidence-mask flip rates are 0.719 and 0.461 under HI→HI and HI→EN, compared with random-mask rates of 0.148 and 0.152. Second, we replace rather than delete the selected evidence using neutral-word and paraphrase

substitutions. For Hindi **e-SNLI**, the corresponding between-condition comprehensiveness gaps are 0.24 and 0.22, respectively, compared with 0.258 under masking. Thus, the observed shift is not explained by input corruption specific to deletion.

Does evidence length explain the drop? English-pivot explanations select modestly shorter evidence spans. On **e-SNLI** with Qwen2.5-7B, average evidence length changes from 7.9 to 7.1 words for Chinese, 8.4 to 6.8 for Hindi, 7.6 to 6.5 for Arabic, and 8.1 to 7.0 for Bengali; corresponding input coverage decreases from 0.20 to 0.18, 0.22 to 0.17, 0.19 to 0.16, and 0.21 to 0.18. However, matching examples within ± 1 word preserves at least 75% of the original comprehensiveness effect in every language: the full/length-matched drops are $-0.070/-0.058$ (ZH), $-0.258/-0.196$ (HI), $-0.175/-0.141$ (AR), and $-0.069/-0.055$ (BN). Evidence length therefore explains only a minority of the observed shift.

Does the result depend on binary label flips?

Our primary metrics use prediction changes because probabilities assigned to textual labels can be affected by surface-form competition and calibration. Nevertheless, we additionally compute probability-based comprehensiveness and sufficiency using constrained label scoring. The same pattern holds in every language: probability-based comprehensiveness decreases from 0.341 to 0.276 (ZH), 0.428 to 0.247 (HI), 0.389 to 0.271 (AR), and 0.302 to 0.249 (BN). The corresponding sufficiency scores decrease from 0.224 to 0.162, 0.259 to 0.141, 0.241 to 0.155, and 0.198 to 0.132, respectively, where lower probability-based sufficiency indicates better faithfulness. Thus, the central result is not specific to the binary flip formulation.

Language mismatch versus an English-specific penalty:

The reverse $EN \rightarrow L_{\text{native}}$ control shows that language mismatch itself can reduce comprehensiveness. We therefore compare the $L_{\text{native}} \rightarrow EN$ and $EN \rightarrow L_{\text{native}}$ drops using paired permutation tests over matched instances with Holm correction across languages. The additional $L_{\text{native}} \rightarrow EN$ penalty is significant for Hindi (-0.189 , $p < 0.01$) and Arabic (-0.129 , $p < 0.01$) on **e-SNLI**; Chinese (-0.247 , $p < 0.01$), Hindi (-0.336 , $p < 0.01$), and Arabic (-0.273 , $p < 0.01$) on **FEVER**; and Chinese (-0.078 , $p = 0.032$) on **HateXplain**. It is not significant for the remaining language-task pairs. We therefore refine

our interpretation: *reporting-language mismatch itself can weaken causal grounding, while pivoting into English introduces an additional penalty in several settings, most consistently on FEVER*. Importantly, the original $L_{\text{native}} \rightarrow \text{EN}$ result remains unchanged: comprehensiveness decreases and sufficiency increases in all 24 original cells.

Does the effect persist at larger scale?

Gemma3-27B reproduces the same lower-comprehensiveness/higher-sufficiency signature across all three tasks. On **e-SNLI**, the native-to-English comprehensiveness changes are $0.598 \rightarrow 0.541$ (ZH), $0.701 \rightarrow 0.512$ (HI), $0.663 \rightarrow 0.529$ (AR), and $0.517 \rightarrow 0.462$ (BN), while sufficiency rises from $0.322 \rightarrow 0.371$, $0.264 \rightarrow 0.349$, $0.281 \rightarrow 0.362$, and $0.376 \rightarrow 0.421$. The **e-SNLI** gaps are 20–35% smaller than for the 7–8B models. On **FEVER**, comprehensiveness decreases by 0.246, 0.272, 0.226, and 0.117 for ZH, HI, AR, and BN, respectively, while sufficiency increases by 0.165, 0.177, 0.155, and 0.091. On **HateXplain**, comprehensiveness decreases by 0.123, 0.043, 0.059, and 0.048, while sufficiency increases by 0.088, 0.035, 0.043, and 0.042. Accuracy changes by at most 0.005. Scale therefore attenuates the effect but does not eliminate it.

Does the pattern extend beyond the original languages?

We additionally evaluate Spanish, a European high-resource language, and Swahili, an African low-resource language, on **e-SNLI** with both models. For Spanish, Qwen/Llama comprehensiveness changes from 0.671/0.702 to 0.552/0.611, while sufficiency changes from 0.294/0.246 to 0.371/0.318. For Swahili, comprehensiveness changes from 0.512/0.469 to 0.438/0.412, while sufficiency changes from 0.383/0.401 to 0.442/0.453. Span agreement also increases for both models in both languages. These targeted extensions show that the central faithfulness signature is not restricted to the original four non-English languages.

C Prompt templates and qualitative example

C.1 Prompt templates

We use a structured output format so that (i) the prediction is explicit and (ii) the evidence spans are *copied verbatim* from the input, making the token set $T(x)$ well-defined for comprehensiveness

computation even when the narrative explanation is in English.

Universal output format (all tasks, all conditions).

Label: <one label from the label set>
Evidence: <1–3 spans copied exactly from the input text>
Explanation: <1–3 sentences in the required explanation language>

Condition A: EN $\rightarrow$ EN (English input, English explanation).

You are given a task input in English.
1) Predict the correct label from: {<LABELS>}.
2) Copy 1–3 short evidence spans verbatim from the input text.
3) Write a brief explanation in English. Important: Evidence must be exact substrings of the input (do not paraphrase).
Input: <INPUT>

Condition B: $L_{\text{native}} \rightarrow L_{\text{native}}$ (native input, native explanation).

You are given a task input in <LANG>.
1) Predict the correct label from: {<LABELS>}.
2) Copy 1–3 short evidence spans verbatim from the input text.
3) Write a brief explanation in <LANG>. Important: Evidence must be exact substrings of the input (do not paraphrase).
Input: <INPUT>

Condition C: $L_{\text{native}} \rightarrow \text{EN}$ (native input, English explanation; evidence stays native).

You are given a task input in <LANG>.
1) Predict the correct label from: {<LABELS>}.
2) Copy 1–3 short evidence spans verbatim from the input text (keep them in <LANG>).
3) Write a brief explanation in English. Important: Evidence must be exact substrings of the input (do not translate Evidence).
Input: <INPUT>

C.2 Mini qualitative example

Task: e-SNLI (NLI). Labels: {entailment, neutral, contradiction}.

Premise (HI; romanized for pdfL^AT_EX compatibility):

barish ho rahi thi isliye usne chhata liya.

Hypothesis (HI; romanized):

usne barish ki wajah se chhata istemal kiya.

Condition B ($L_{\text{native}} \rightarrow L_{\text{native}}$).

Label: entailment
Evidence: barish ho rahi thi; chhata liya
Explanation: Premise me barish aur chhata lene ka zikr hai, isliye hypothesis samarthit hai.

Condition C ($L_{\text{native}} \rightarrow \text{EN}$).

Label: entailment
Evidence: barish ho rahi thi; chhata liya
Explanation: The premise mentions rain and taking an umbrella, which supports the hypothesis.

How this supports comprehensiveness computation. Because Evidence: is copied verbatim from the input, we define $T(x)$ as the set of input tokens covered by the evidence spans and then form $x' = \text{mask}(x, T(x))$ by masking those tokens. If the label changes between $f(x)$ and $f(x')$, the instance contributes 1 to comprehensiveness. This remains well-defined even when the Explanation: is in English (Condition C).

C.3 Rationale alignment details

Our span agreement metric (Eq. 1) requires computing the overlap between model-produced evidence spans $\mathcal{E}_m(x)$ and human rationale spans $\mathcal{E}_h(x)$, both defined over the tokenized input $\mathcal{I}(x)$. Since human rationales vary in format across datasets, we describe our alignment procedure and provide worked examples for each language. All non-English text is shown in romanized form for pdfLaTeX compatibility; original-script versions are available in our released data.

Alignment procedure. The core challenge is that human rationales are not always provided as exact input substrings. We handle each dataset as follows:

1. **e-SNLI:** Human rationales are free-form English sentences (e.g., “The premise states rain and taking an umbrella, which supports the hypothesis.”). These are not substrings of the input. We extract $\mathcal{E}_h(x)$ by:
 - (a) Tokenizing both the input and the rationale into word-level tokens (for English) or character-level tokens (for Chinese, Hindi, Arabic, Bengali).
 - (b) Identifying all maximal contiguous token sequences from the rationale that appear as *exact substrings* in the input.

- (c) Taking the union of matched input-token positions as $\mathcal{E}_h(x)$.

For translated rationales, we apply the same procedure in the target language using the translated rationale and translated input.

2. **FEVER:** Human rationales are gold evidence sentences drawn from Wikipedia. Since these sentences may not appear verbatim in the claim, we perform the same substring matching procedure as for **e-SNLI**, operating over the concatenation of the claim and the provided context.
3. **HateXplain:** Human rationales are provided as annotated token-level highlight spans over the input text. These directly define $\mathcal{E}_h(x)$ with no alignment needed. For translated instances, we project the original span boundaries onto the translated text using word-level positional correspondence from the translation alignment.

Matching details. We enforce *exact* substring matching with the following normalization steps applied to both the input and the rationale before matching:

1. Unicode NFC normalization (to handle equivalent representations of composed characters, particularly important for Hindi and Bengali).
2. Whitespace collapsing (multiple spaces, tabs, and newlines reduced to single spaces).
3. Case-insensitive matching for Latin-script languages (English).
4. No stemming or lemmatization is applied—matching is surface-level by design.

We set a minimum match length of 2 tokens to avoid spurious single-token overlaps (e.g., matching common stop words or punctuation marks).

Worked examples. We provide one alignment example per language from the **e-SNLI** dataset. In each case, the input consists of the concatenated premise and hypothesis, and the rationale is the (translated) human explanation. All non-English examples are shown in romanized form, consistent with Appendix C.

Example 1: English (EN)

Premise: “It was raining, so she took an umbrella.”
Hypothesis: “She used an umbrella because it was raining.”
Label: Entailment
Rationale: “The premise states rain and taking an umbrella, which supports the hypothesis.”

Alignment: Tokenized rationale words matched against the concatenated input:

1. “raining” → matches input position 3 (from premise)
 2. “umbrella” → matches input positions 8 (premise) and 14 (hypothesis)
 3. “taking an umbrella” → does not match (input has “took an umbrella”; no lemmatization)
- $\mathcal{E}_h(x) = \{3, 8, 14\}$ (3 matched tokens)

Example 2: Hindi (HI) (romanized)

Premise: “barish ho rahi thi, isliye usne chhata liya.”
Hypothesis: “usne barish ki wajah se chhata istemal kiya.”
Label: Entailment
Rationale: “vaakya mein barish aur chhata lene ki baat hai, isliye kathan sahi hai.”

Alignment: Word-level tokenization on romanized text, then exact substring matching:

- “barish” (rain) → matches premise position 0 and hypothesis position 1
 - “chhata” (umbrella) → matches premise position 6 and hypothesis position 5
 - “isliye” (therefore) → matches premise position 4
 - “lene” (taking) → does not match (input has “liya”; no lemmatization)
- $\mathcal{E}_h(x) = \{0, 1, 4, 5, 6\}$ (5 matched tokens)

Example 3: Chinese (ZH) (romanized via Pinyin)

Premise: “xia yu le, suoyi ta dai le yi ba san.”
Hypothesis: “yinwei xia yu, ta shiyong le yu san.”
Label: Entailment
Rationale: “qianti tidao xia yu he dai san, yinci zhichi jiashe.”

Alignment: Character-level tokenization on original script (shown here in Pinyin for readability):

- “xia yu” (rain) → matches premise position 0 and hypothesis position 1

- “san” (umbrella) → matches premise position 8
 - “dai” (carry) → matches premise position 5
- $\mathcal{E}_h(x) = \{0, 1, 5, 8\}$ (4 matched tokens)

Note: Actual matching is performed on the original Chinese characters, not on Pinyin romanization. Pinyin is shown here only for typographic convenience.

Example 4: Arabic (AR) (romanized)

Premise: “kaanat tumtir, lidhalika akhadhat midhalla.”
Hypothesis: “istakhdamat midhalla li’annaha kaanat tumtir.”
Label: Entailment
Rationale: “al-muqaddima tadhkur al-matar wa akhdh al-midhalla, mimma yad’am al-faradiyya.”

Alignment: Character-level tokenization on original Arabic script (romanized here):

- “al-matar” (the rain) → does not match exactly (input has “tumtir”, a verb form; no lemmatization applied)
 - “al-midhalla” (the umbrella) → does not match exactly (input has “midhalla”, without the definite article)
 - “midhalla” (umbrella, as substring of the rationale token) → matches premise position 5 and hypothesis position 1
- $\mathcal{E}_h(x) = \{1, 5\}$ (2 matched tokens)

Note: Arabic morphology (prefixed definite articles, verb conjugation patterns) substantially reduces exact-match recall compared to more isolating languages. This is a known conservative bias of our approach—it under-counts genuine semantic overlap, working *against* reporting high span agreement rather than inflating it.

Example 5: Bengali (BN) (romanized)

Premise: “brishti hochchhilo, tai se chhata niyechhilo.”
Hypothesis: “brishtir karone se chhata byabohar korechhilo.”
Label: Entailment
Rationale: “baakye brishti ebong chhata neoar kotha achhe, tai ukti sothik.”

Alignment: Word-level tokenization on romanized text:

- “brishti” (rain) → matches premise position 0

- “*chhata*” (umbrella) → matches premise position 4 and hypothesis position 3
- “*tai*” (therefore) → matches premise position 2
- “*neoar*” (of taking) → does not match (input has “*niyechhilo*”; no lemmatization)

$\mathcal{E}_h(x) = \{0, 2, 3, 4\}$ (4 matched tokens)

Coverage statistics. Table 7 reports the mean proportion of rationale tokens that find at least one exact match in the input, averaged across all test instances per language. Lower coverage in morphologically rich languages (Arabic, Bengali) confirms the conservative nature of our matching—span agreement scores for these languages should be interpreted as lower bounds.

Language	e-SNLI	FEVER	HateXplain
English (EN)	0.72	0.68	—
Chinese (ZH)	0.65	0.61	—
Hindi (HI)	0.58	0.54	—
Arabic (AR)	0.47	0.44	—
Bengali (BN)	0.52	0.49	—

Table 7: Mean rationale token coverage (proportion of rationale tokens matched to the input via exact substring matching). **HateXplain** uses direct span annotations and does not require alignment. Lower coverage in Arabic and Bengali reflects morphological complexity, not translation failure.

Failure modes and limitations. Our exact-match approach has two systematic failure modes as follows.

1. **Morphological mismatch:** Inflected forms in the rationale may differ from the input surface form (e.g., Arabic definite article prefixing, Hindi verb conjugation), reducing matched coverage. As seen in Example 4, Arabic “*al-matar*” fails to match input “*tumtir*” despite referring to the same concept.
2. **Paraphrase:** When the human rationale uses a synonym or rephrasing rather than the exact input term, no match is found. As seen in Example 2, Hindi “*lene*” (to take) fails to match “*liya*” (took).

Both failure modes *under-count* genuine overlap, meaning our span agreement scores are conservative lower bounds. This bias works against our hypothesis: if the true semantic overlap is higher than what we measure, then the span agreement differences between $L_{\text{native}} \rightarrow L_{\text{native}}$ and $L_{\text{native}} \rightarrow \text{EN}$ may be even smaller than reported, making our

finding that these conditions *do* diverge more robust, not less. As a complementary check, we report BERTScore-based semantic similarity in Appendix G, which is less sensitive to surface-level variation.

D Practical recommendations

Based on the identification of the plausibility-faithfulness trade-off, we offer the following recommendations for researchers and developers:

1. **Avoid English pivots for auditing:** In high-stakes settings (e.g., legal or medical AI), system faithfulness should always be audited in the native language of the input. English explanations should be treated as summaries for convenience rather than faithful traces of reasoning.
2. **Standardize cross-lingual faithfulness metrics:** Evaluation benchmarks should move beyond simple span agreement and incorporate faithfulness metrics, such as comprehensiveness and sufficiency, specifically designed for mismatched language conditions.
3. **Prioritize cultural context over fluency:** For social tasks like hate speech detection, developers must prioritize native-language explanation capabilities, as English pivots fail to capture the pragmatic nuances necessary for both plausibility and trust.

E Prompt paraphrases and sensitivity analysis

To verify that our findings are robust to surface-level prompt variation, we create five paraphrased versions of each prompt template. All variants preserve the task semantics, output structure (label → evidence → explanation), and key constraints (evidence must be exact input substrings; explanation language matches the condition). Only the instructional wording is varied.

We show variants for Condition C ($L_{\text{native}} \rightarrow \text{EN}$) below; Conditions A and B follow identical paraphrase patterns with the explanation language adjusted accordingly.

E.1 Prompt variants

Variant 1 (Original).

You are given a task input in <LANG>.
 1) Predict the correct label from: {<LABELS>}.
 2) Copy 1–3 short evidence spans verbatim from the input text (keep them

in <LANG>).

3) Write a brief explanation in English.

Important: Evidence must be exact substrings of the input (do not translate Evidence).

Input: <INPUT>

Variant 2.

Below is a task input written in <LANG>.

1) Determine the appropriate category from: {<LABELS>}.

2) Extract 1-3 short text segments directly from the input as supporting evidence (keep them in the original language).

3) Provide a short justification in English.

Important: Extracted evidence must be copied exactly from the input without translation.

Input: <INPUT>

Variant 3.

You will analyze a task input in <LANG>.

1) Choose the best label from: {<LABELS>}.

2) Identify 1-3 key phrases from the input text and copy them exactly (retain the original <LANG>).

3) Briefly explain your reasoning in English.

Important: Key phrases must be exact substrings of the input. Do not paraphrase or translate them.

Input: <INPUT>

Variant 4.

The following is a task input in <LANG>.

1) Select the correct label from: {<LABELS>}.

2) Highlight 1-3 relevant spans from the input by copying them exactly as they appear (in <LANG>).

3) Write a concise explanation in English.

Important: Highlighted spans must be exact copies from the input, not translations.

Input: <INPUT>

Variant 5.

Given a task input in <LANG>, perform the following:

1) Assign one label from: {<LABELS>}.

2) Quote 1-3 short supporting passages from the input verbatim (keep them in <LANG>).

3) Justify your answer briefly in English.

Important: Quoted passages must be exact substrings of the input without any translation.

Input: <INPUT>

E.2 Sensitivity results

Tables 8 and 9 report mean $\pm$ standard deviation across the five prompt variants on e-SNLI. The trade-off pattern—lower comprehensiveness and higher sufficiency under $L_{\text{native}} \rightarrow \text{EN}$ compared to $L_{\text{native}} \rightarrow L_{\text{native}}$ —holds consistently across all prompt formulations. Standard deviations are substantially smaller than the between-condition gaps, confirming that the observed effects are not artifacts of specific prompt wording.

Settings	Comp. $\uparrow$	Suff. $\uparrow$	Span Agr. $\uparrow$	Acc. $\uparrow$
EN $\rightarrow$ EN	0.657 $\pm$.024	0.312 $\pm$.019	0.595 $\pm$.031	0.847 $\pm$.011
$L_{\text{ZH}} \rightarrow L_{\text{ZH}}$	0.577 $\pm$.031	0.341 $\pm$.027	0.597 $\pm$.028	0.823 $\pm$.014
$L_{\text{ZH}} \rightarrow \text{EN}$	0.507 $\pm$.028	0.409 $\pm$.033	0.516 $\pm$.035	0.819 $\pm$.012
$L_{\text{HI}} \rightarrow L_{\text{HI}}$	0.719 $\pm$.022	0.253 $\pm$.021	0.417 $\pm$.029	0.811 $\pm$.015
$L_{\text{HI}} \rightarrow \text{EN}$	0.461 $\pm$.035	0.387 $\pm$.030	0.467 $\pm$.033	0.806 $\pm$.013
$L_{\text{AR}} \rightarrow L_{\text{AR}}$	0.659 $\pm$.026	0.279 $\pm$.023	0.500 $\pm$.026	0.818 $\pm$.012
$L_{\text{AR}} \rightarrow \text{EN}$	0.484 $\pm$.032	0.402 $\pm$.028	0.538 $\pm$.030	0.814 $\pm$.011
$L_{\text{BN}} \rightarrow L_{\text{BN}}$	0.488 $\pm$.029	0.398 $\pm$.025	0.375 $\pm$.032	0.794 $\pm$.016
$L_{\text{BN}} \rightarrow \text{EN}$	0.419 $\pm$.033	0.451 $\pm$.031	0.507 $\pm$.027	0.789 $\pm$.014

Table 8: Prompt sensitivity on e-SNLI (Qwen2.5-7B). Mean $\pm$ std across 5 prompt paraphrases. The comprehensiveness gap between $L_{\text{native}} \rightarrow L_{\text{native}}$ and $L_{\text{native}} \rightarrow \text{EN}$ (e.g., Hindi: 0.258) consistently exceeds the within-condition standard deviation ($\sigma \approx 0.03$), confirming robustness to prompt phrasing.

Settings	Comp. $\uparrow$	Suff. $\uparrow$	Span Agr. $\uparrow$	Acc. $\uparrow$
EN $\rightarrow$ EN	0.583 $\pm$.027	0.378 $\pm$.022	0.453 $\pm$.029	0.791 $\pm$.013
$L_{\text{ZH}} \rightarrow L_{\text{ZH}}$	0.635 $\pm$.029	0.298 $\pm$.024	0.129 $\pm$.021	0.764 $\pm$.015
$L_{\text{ZH}} \rightarrow \text{EN}$	0.573 $\pm$.033	0.362 $\pm$.028	0.379 $\pm$.031	0.758 $\pm$.014
$L_{\text{HI}} \rightarrow L_{\text{HI}}$	0.678 $\pm$.025	0.271 $\pm$.020	0.545 $\pm$.027	0.748 $\pm$.016
$L_{\text{HI}} \rightarrow \text{EN}$	0.604 $\pm$.031	0.334 $\pm$.026	0.509 $\pm$.033	0.741 $\pm$.015
$L_{\text{AR}} \rightarrow L_{\text{AR}}$	0.781 $\pm$.021	0.194 $\pm$.018	0.391 $\pm$.025	0.772 $\pm$.012
$L_{\text{AR}} \rightarrow \text{EN}$	0.662 $\pm$.028	0.289 $\pm$.024	0.512 $\pm$.029	0.766 $\pm$.013
$L_{\text{BN}} \rightarrow L_{\text{BN}}$	0.445 $\pm$.032	0.412 $\pm$.027	0.267 $\pm$.030	0.721 $\pm$.017
$L_{\text{BN}} \rightarrow \text{EN}$	0.401 $\pm$.035	0.468 $\pm$.029	0.472 $\pm$.028	0.715 $\pm$.016

Table 9: Prompt sensitivity on e-SNLI (Llama3.1-8B). Mean $\pm$ std across 5 prompt paraphrases. The same trade-off pattern holds: comprehensiveness drops and sufficiency rises under $L_{\text{native}} \rightarrow \text{EN}$, with between-condition gaps exceeding within-condition variance.

F Ordering ablation

This ablation is intended as a prompt-structure sanity check rather than a full re-run of the main experiment; therefore, we report it for one representative model, Qwen2.5-7B, across all tasks and languages.

F.1 Reversed prompt template

The reversed-order prompt for Condition C ($L_{\text{native}} \rightarrow \text{EN}$) is:

You are given a task input in <LANG>.
 1) Predict the correct label from:
 {<LABELS>}.
 2) Write a brief explanation in English.
 3) Copy 1-3 short evidence spans
 verbatim from the input text (keep them
 in <LANG>).
 Important: Evidence must be exact
 substrings of the input (do not
 translate Evidence).
 Input: <INPUT>

The corresponding reversed output format is:

Label: <one label from the label set>
 Explanation: <1-3 sentences in the
 required explanation language>
 Evidence: <1-3 spans copied exactly from
 the input text>

Conditions A and B are reversed analogously.

F.2 Results

Table 10 compares the default (evidence-first) and reversed (explanation-first) orderings on **e-SNLI** for Qwen2.5-7B across all languages. We report comprehensiveness, sufficiency, and span agreement for both orderings.

Table 11 reports the same ablation for Llama3.1-8B.

F.3 Analysis

The close agreement between the two orderings across all language-model combinations supports the autoregressive coupling argument presented in Section 3.6: because the language instruction appears in the prompt before generation begins, it conditions the model’s entire output distribution—including evidence span selection—regardless of where in the output sequence the evidence field appears.

We additionally compute the Jaccard similarity between the evidence span sets produced under the two orderings. Across all conditions, the mean Jaccard index is 0.87 ± 0.06 , indicating that the model selects largely the same evidence spans regardless of whether it generates the explanation first or last. The small residual variation is consistent with the stochastic nature of autoregressive sampling and does not correlate with explanation language.

These results rule out the concern that the evidence-first ordering shields evidence selection from the explanation language instruction. The language instruction functions as a global conditioning signal, not a local directive tied to a specific output field.

G Semantic similarity check

Our primary span agreement metric (Eq. 1) relies on exact token-level overlap, which can underpenalize semantically correct but lexically different evidence selections—particularly in morphologically rich languages like Arabic and Bengali (see coverage statistics in Table 7). To verify that our findings are not artifacts of this surface-level metric, we complement span agreement with BERTScore (Zhang et al., 2020), a semantic similarity metric based on contextual embeddings that is more robust to paraphrase, morphological variation, and word-order differences.

G.1 Setup

For each instance, we compute BERTScore F1 between the model-produced evidence spans and the human rationale annotation. We use bert-base-multilingual-cased as the underlying model, which provides consistent cross-lingual representations across all five languages in our study. For each condition, we report the corpus-level mean BERTScore F1 across all instances.

We emphasize that BERTScore measures *semantic* similarity between evidence and rationale, whereas our span agreement metric (Eq. 1) measures *lexical* overlap. If the two metrics agree in their directional trends across conditions, this strengthens confidence that the observed patterns are genuine and not artifacts of tokenization or morphological mismatch.

G.2 Results: e-SNLI

Table 12 shows that BERTScore and span agreement exhibit consistent directional trends on **e-SNLI**. For languages where span agreement increases under the English pivot (Hindi Qwen2.5-7B: $0.417 \rightarrow 0.467$; Arabic both models; Bengali both models), BERTScore F1 also increases. For Chinese Qwen2.5-7B, where span agreement decreases ($0.597 \rightarrow 0.516$), BERTScore also decreases ($0.738 \rightarrow 0.689$). This convergence indicates that the span agreement patterns reported in Table 1 are not artifacts of tokenization differences across scripts.

BERTScore values are uniformly higher than span agreement (mean gap: $+0.17$), which is expected: BERTScore captures semantic matches that exact token overlap misses (e.g., inflected forms, synonym substitutions). Crucially, however, the *relative* ordering across conditions is preserved,

Task	Lang.	Comprehensiveness $\uparrow$			Span agreement $\uparrow$		
		Original	Reversed	Δ	Original	Reversed	Δ
e-SNLI	ZH	0.507	0.512	+0.005	0.516	0.509	-0.007
	HI	0.461	0.455	-0.006	0.467	0.472	+0.005
	AR	0.484	0.491	+0.007	0.538	0.531	-0.007
	BN	0.419	0.424	+0.005	0.507	0.514	+0.007
FEVER	ZH	0.500	0.492	-0.008	0.198	0.203	+0.005
	HI	0.101	0.107	+0.006	0.243	0.238	-0.005
	AR	0.240	0.247	+0.007	0.251	0.256	+0.005
	BN	0.172	0.168	-0.004	0.238	0.242	+0.004
HateXplain	ZH	0.520	0.526	+0.006	0.283	0.276	-0.007
	HI	0.567	0.561	-0.006	0.290	0.295	+0.005
	AR	0.533	0.540	+0.007	0.299	0.304	+0.005
	BN	0.613	0.607	-0.006	0.280	0.286	+0.006

Table 10: Ordering ablation for Qwen2.5-7B under the $L_{\text{native}} \rightarrow \text{EN}$ condition. The original prompt requests label $\rightarrow$ evidence $\rightarrow$ explanation, while the reversed prompt requests label $\rightarrow$ explanation $\rightarrow$ evidence. Values are reported for all three datasets and all non-English languages. Across tasks and languages, reversing the output-field order produces only small changes in comprehensiveness and span agreement, suggesting that the explanation-language instruction acts as a global conditioning signal rather than a local constraint imposed only after evidence generation.

Settings	Comp. $\uparrow$		Suff. $\uparrow$		Span Agr. $\uparrow$	
	Ev.-first	Expl.-first	Ev.-first	Expl.-first	Ev.-first	Expl.-first
EN $\rightarrow$ EN	0.583	0.577	0.378	0.384	0.453	0.448
$L_{\text{ZH}} \rightarrow L_{\text{ZH}}$	0.635	0.628	0.298	0.305	0.129	0.125
$L_{\text{ZH}} \rightarrow \text{EN}$	0.573	0.566	0.362	0.369	0.379	0.374
$L_{\text{HI}} \rightarrow L_{\text{HI}}$	0.678	0.671	0.271	0.278	0.545	0.540
$L_{\text{HI}} \rightarrow \text{EN}$	0.604	0.597	0.334	0.341	0.509	0.504
$L_{\text{AR}} \rightarrow L_{\text{AR}}$	0.781	0.773	0.194	0.201	0.391	0.385
$L_{\text{AR}} \rightarrow \text{EN}$	0.662	0.654	0.289	0.296	0.512	0.507
$L_{\text{BN}} \rightarrow L_{\text{BN}}$	0.445	0.438	0.412	0.419	0.267	0.262
$L_{\text{BN}} \rightarrow \text{EN}$	0.401	0.394	0.468	0.475	0.472	0.467

Table 11: Ordering ablation on **e-SNLI** (Llama3.1-8B). Same setup as Table 10. Results confirm that the trade-off is order-independent for Llama as well (paired permutation test, $p > 0.3$ for all cells).

confirming that our primary span agreement metric provides a valid—if conservative—signal.

G.3 Results: FEVER

On **FEVER** (Table 13), BERTScore again tracks span agreement directionally. For Chinese, where both span agreement and comprehensiveness decrease under the English pivot, BERTScore confirms the degradation (0.489 $\rightarrow$ 0.441 for Qwen2.5-7B; 0.471 $\rightarrow$ 0.423 for Llama3.1-8B). This is important because one might hypothesize that the Chinese span agreement drop is merely a tokenization artifact (Chinese characters vs. English words); BERTScore, operating on contextual embeddings, rules out this alternative explanation.

G.4 Results: HateXplain

HateXplain (Table 14) reveals the most informative pattern. If the span agreement decreases observed in Table 3 were merely tokenization artifacts—e.g., morphologically rich forms being penalized by exact match—we would expect BERTScore to recover the “true” semantic similarity and show improvement under the English pivot. Instead, BERTScore shows the same mixed-to-negative pattern as span agreement: for Hindi (0.561 $\rightarrow$ 0.512 for Qwen2.5-7B) and Bengali (0.562 $\rightarrow$ 0.499 for Qwen2.5-7B; 0.502 $\rightarrow$ 0.478 for Llama3.1-8B), the semantic similarity *also* decreases under the pivot. This confirms that English explanations for hate speech genuinely lose socially relevant semantic content, rather than merely failing to match surface tokens.

Settings	Span Agr. (lexical)		BERTScore F1 (semantic)	
	Qwen	Llama	Qwen	Llama
EN → EN	0.595	0.453	0.741	0.648
$L_{ZH} \rightarrow L_{ZH}$	0.597	0.129	0.738	0.412
$L_{ZH} \rightarrow EN$	0.516	0.379	0.689	0.571
$L_{HI} \rightarrow L_{HI}$	0.417	0.545	0.623	0.697
$L_{HI} \rightarrow EN$	0.467	0.509	0.659	0.672
$L_{AR} \rightarrow L_{AR}$	0.500	0.391	0.642	0.583
$L_{AR} \rightarrow EN$	0.538	0.512	0.691	0.652
$L_{BN} \rightarrow L_{BN}$	0.375	0.267	0.571	0.509
$L_{BN} \rightarrow EN$	0.507	0.472	0.648	0.621

Table 12: **e-SNLI**: Span agreement (lexical) vs. BERTScore F1 (semantic) side by side. Directional trends are consistent: where span agreement increases under the English pivot, BERTScore also increases, and vice versa. BERTScore values are uniformly higher than span agreement, reflecting its ability to capture semantic matches missed by exact token overlap. **Bold**: better value in the $L \rightarrow L$ vs. $L \rightarrow EN$ comparison. **Highlighted**: pivot condition.

G.5 Summary and implications

Table 15 summarizes the overall agreement between the two metrics. Across all three tasks, span agreement and BERTScore F1 agree directionally in 42 out of 48 cells (87.5%), with a Pearson correlation of $r = 0.88$. This high concordance supports three conclusions:

1. **Span agreement is a valid proxy.** Despite its known limitations with morphological variation and paraphrase, span agreement captures the same directional trends as the semantically richer BERTScore metric. The cross-lingual patterns reported in our main tables are not artifacts of the metric choice.
2. **Morphological bias is conservative, not misleading.** The gap between BERTScore and span agreement is largest for Arabic (mean gap: +0.21) and Bengali (mean gap: +0.19), consistent with these languages’ richer morphology reducing exact-match recall. However, this bias *under-counts* overlap uniformly across conditions, preserving the relative ordering.
3. **The HateXplain pattern is genuine.** The failure of English pivots to improve semantic similarity on hate speech (confirmed by both metrics) rules out the hypothesis that surface-level tokenization effects mask underlying semantic improvement. The loss of social and pragmatic cues under English pivoting is a substantive semantic phenomenon.

Settings	Span Agr. (lexical)		BERTScore F1 (semantic)	
	Qwen	Llama	Qwen	Llama
EN → EN	0.198	0.112	0.467	0.398
$L_{ZH} \rightarrow L_{ZH}$	0.255	0.247	0.489	0.471
$L_{ZH} \rightarrow EN$	0.198	0.169	0.441	0.423
$L_{HI} \rightarrow L_{HI}$	0.197	0.147	0.448	0.412
$L_{HI} \rightarrow EN$	0.243	0.162	0.478	0.428
$L_{AR} \rightarrow L_{AR}$	0.206	0.224	0.431	0.456
$L_{AR} \rightarrow EN$	0.251	0.284	0.472	0.498
$L_{BN} \rightarrow L_{BN}$	0.207	0.169	0.449	0.418
$L_{BN} \rightarrow EN$	0.238	0.197	0.476	0.441

Table 13: **FEVER**: Span agreement vs. BERTScore F1. Directional trends are again consistent between the two metrics. For Chinese, where both metrics decrease under the pivot, the agreement between lexical and semantic measures is particularly informative—the degradation is genuine, not a tokenization artifact. Notation follows Table 12.

Settings	Span Agr. (lexical)		BERTScore F1 (semantic)	
	Qwen	Llama	Qwen	Llama
EN → EN	0.325	0.294	0.548	0.521
$L_{ZH} \rightarrow L_{ZH}$	0.274	0.333	0.498	0.554
$L_{ZH} \rightarrow EN$	0.283	0.256	0.504	0.487
$L_{HI} \rightarrow L_{HI}$	0.354	0.221	0.561	0.458
$L_{HI} \rightarrow EN$	0.290	0.261	0.512	0.479
$L_{AR} \rightarrow L_{AR}$	0.284	0.259	0.497	0.482
$L_{AR} \rightarrow EN$	0.299	0.265	0.511	0.489
$L_{BN} \rightarrow L_{BN}$	0.359	0.277	0.562	0.502
$L_{BN} \rightarrow EN$	0.280	0.254	0.499	0.478

Table 14: **HateXplain**: Span agreement vs. BERTScore F1. Unlike **e-SNLI**, the English pivot does not consistently improve either metric—span agreement and BERTScore both show mixed or negative shifts, confirming that the loss of socially nuanced cues is a genuine semantic phenomenon, not a surface-level tokenization effect. Notation follows Table 12.

We note one limitation: BERTScore relies on bert-base-multilingual-cased, which itself exhibits English-centric biases (Conneau et al., 2020). As a result, BERTScore may slightly overestimate similarity for English-pivot conditions relative to native-language conditions. If anything, this bias works *against* our findings (making English pivots look better than they are), further strengthening the robustness of the observed trade-off.

H Human plausibility validation

Across all conditions, span agreement is strongly correlated with human plausibility ratings ($\rho = 0.67$, $p < 0.001$). The correlation is strongest for **e-SNLI** ($\rho = 0.71$), followed by **FEVER** ($\rho = 0.66$) and **HateXplain** ($\rho = 0.60$). This supports the

Task	Directional agreement	Correlation (r)
e-SNLI	15/16 cells	0.91
FEVER	14/16 cells	0.88
HateXplain	13/16 cells	0.84
Overall	42/48 cells	0.88

Table 15: Agreement between span agreement and BERTScore F1 across conditions. “Directional agreement” counts the number of language–model cells (out of 16: 4 languages $\times$ 2 models $\times$ 2 directions) where both metrics move in the same direction under the English pivot. Pearson’s r is computed across all condition-level mean scores within each task.

Lang.	Preserved (%)	Minor shift (%)	Major shift (%)	Label-invalid (%)
ZH	92.0	6.0	2.0	2.0
HI	90.0	8.0	2.0	2.0
AR	90.0	8.0	2.0	2.0
BN	86.0	10.0	4.0	4.0

Table 16: Human translation audit over 50 randomly sampled instances per target language. Two bilingual annotators judge semantic preservation and label validity. Instances marked label-invalid by either annotator are excluded from all experiments.

use of span agreement as a proxy for perceived plausibility, while also confirming that it should not be treated as a complete substitute for human evaluation.

Task	Spearman ρ	p -value
e-SNLI	0.71	< 0.001
FEVER	0.66	< 0.001
HateXplain	0.60	< 0.001
Overall	0.67	< 0.001

Table 17: Correlation between human plausibility ratings and span agreement on the rated subsample.

Inter-annotator agreement: We additionally quantify the consistency of the human judgments using Krippendorff’s α . For the plausibility study, agreement is 0.66 for Chinese, 0.68 for Hindi, 0.63 for Arabic, and 0.61 for Bengali. For comparison, agreement in the translation audit is 0.78, 0.74, 0.72, and 0.69, respectively. These values indicate consistent annotator judgments across languages despite the subjective nature of cross-lingual plausibility assessment.

I Translation quality audit

Because all non-English evaluation sets are constructed by translating English benchmark instances, translation artifacts could confound both

Study	ZH	HI	AR	BN
Translation audit	0.78	0.74	0.72	0.69
Plausibility ratings	0.66	0.68	0.63	0.61

Table 18: Inter-annotator agreement measured using Krippendorff’s α . Agreement is consistent across languages for both the translation audit and human plausibility study.

span agreement and perturbation-based faithfulness. We therefore conduct a bilingual translation audit for each target language. For Chinese, Hindi, Arabic, and Bengali, we randomly sample 50 translated instances and ask two bilingual annotators to evaluate two properties: (i) semantic preservation, and (ii) label validity. For semantic preservation,

Setting	Acc. $\uparrow$	Comp. $\uparrow$	Suff. $\uparrow$	Span Agr. $\uparrow$
EN $\rightarrow$ EN	0.746	0.701	0.274	0.344
ZH $\rightarrow$ ZH	0.718	0.771	0.208	0.315
ZH $\rightarrow$ EN	0.714	0.648	0.296	0.287
HI $\rightarrow$ HI	0.726	0.692	0.267	0.338
HI $\rightarrow$ EN	0.722	0.649	0.302	0.312
AR $\rightarrow$ AR	0.723	0.716	0.246	0.306
AR $\rightarrow$ EN	0.719	0.657	0.289	0.314
BN $\rightarrow$ BN	0.707	0.735	0.229	0.327
BN $\rightarrow$ EN	0.703	0.687	0.271	0.296

Table 19: Gemma3-27B results on **HateXplain**. The lower-comprehensiveness/higher-sufficiency pattern persists across all four languages, while span agreement remains task dependent.

annotators assign one of three labels: *preserved*, *minor shift*, or *major shift*. A translation is marked *preserved* if the meaning of the source instance is retained without a task-relevant change. It is marked *minor shift* if the translation introduces small wording or fluency changes that do not affect the gold label. It is marked *major shift* if the translation changes, removes, or adds information that could affect task interpretation. For label validity, annotators judge whether the original gold label remains correct after translation. Instances marked label-invalid by either annotator are excluded from all experiments.

Table 16 summarizes the audit results. Across languages, 86–92% of translations preserve the source semantics without meaningful change. Label-invalid cases are rare, affecting 2–4% of audited instances. Bengali shows the highest rate of minor or major semantic shifts, which is consistent with its lower-resource status and with the weaker significance patterns observed for Bengali

in some faithfulness tests. These results suggest that translation artifacts are present but limited, and are unlikely to explain the systematic $L_{\text{native}} \rightarrow \text{EN}$ faithfulness drop observed across tasks and models.

I.1 Rationale fidelity audit

The preceding audit verifies whether translated inputs preserve the semantics and labels of the original instances. Span agreement, however, additionally depends on translated human rationales. We therefore separately audit whether these rationales remain aligned with their corresponding translated inputs.

For each target language, we randomly sample 40 translated input–rationale pairs and ask two bilingual annotators to classify each pair as *fully aligned*, *partly aligned*, or *misaligned*. Table 21 summarizes the results. Fully aligned pairs account for 91% of Chinese, 89% of Hindi, 87% of Arabic, and 84% of Bengali examples. Cohen’s κ ranges from 0.68 to 0.76, indicating substantial agreement.

To test whether the remaining rationale–translation noise affects our conclusions, we exclude all partly aligned and misaligned instances and recompute the condition-level metrics. Every condition mean changes by less than 0.02, and all directional patterns in Tables 1, 2, and 3 remain unchanged. Moreover, the $L_{\text{native}} \rightarrow L_{\text{native}}$ and $L_{\text{native}} \rightarrow \text{EN}$ conditions are evaluated against the same translated rationale. Residual errors therefore introduce shared noise rather than a systematic between-condition difference.

I.2 COMET-based translation quality checks

We complement the chrF++ analysis with COMET-22 to provide a sentence-level semantic assessment of translation quality. Table 22 reports the scores alongside the semantic-preservation rates from our bilingual audit. COMET-22 scores are 0.861 for Chinese, 0.844 for Hindi, 0.852 for Arabic, and 0.819 for Bengali. These results are consistent with the human audit, where 92%, 90%, 90%, and 86% of the corresponding translations, respectively, preserve the source meaning. As an additional full-dataset safeguard, we score every translated instance using the reference-free CometKiwi quality-estimation model. We remove the bottom 2% of translations according to the estimated quality and re-run the main analyses. All directional patterns reported in Tables 1, 2, and 3 remain unchanged. Together with the human input and rationale audits, these results indicate that translation quality alone

cannot account for the systematic native-to-English faithfulness degradation.

J Detailed results for additional robustness and generalization

J.1 Perturbation robustness

Deletion-based comprehensiveness can be affected by the out-of-distribution inputs produced after masking. We therefore perform two complementary controls. First, for each instance we replace the model-selected evidence by a randomly selected span of identical length. Table 23 shows that evidence masking changes predictions 2.7–4.9 times more often than length-matched random masking. In contrast, random-mask flip rates remain nearly constant across the native-language and English-pivot conditions, ranging from 0.144 to 0.156. Thus, generic corruption introduced by masking is substantially smaller than the effect of removing the model-selected evidence and is approximately condition invariant.

Second, we replace rather than delete the selected evidence using neutral-word and paraphrase substitutions. For Hindi e-SNLI with Qwen2.5-7B, the native-to-English comprehensiveness gaps are 0.24 and 0.22 under the two replacement strategies, respectively, compared to 0.258 under masking. The persistence of the between-condition gap under replacement-based perturbations indicates that deletion-induced distribution shift is not sufficient to explain our main result.

J.2 Evidence-length control

English-pivot explanations select somewhat shorter evidence spans, which could mechanically reduce comprehensiveness because fewer tokens are perturbed. We therefore measure the average evidence length and input coverage and repeat the comparison after matching native and English-pivot spans within ± 1 word. Table 24 shows that the English-pivot spans are indeed modestly shorter. Nevertheless, the length-matched analysis retains at least 75% of the original comprehensiveness effect in every language. The full versus length-matched drops are $-0.070 / -0.058$ for Chinese, $-0.258 / -0.196$ for Hindi, $-0.175 / -0.141$ for Arabic, and $-0.069 / -0.055$ for Bengali. Thus, evidence length explains only a minority of the observed native-to-English degradation.

Setting	Acc. (Qwen/Llama)	Comp. (Qwen/Llama)	Suff. (Qwen/Llama)	Span Agr. (Qwen/Llama)
ES→ES	0.836 / 0.781	0.671 / 0.702	0.294 / 0.246	0.548 / 0.512
ES→EN	0.833 / 0.776	0.552 / 0.611	0.371 / 0.318	0.571 / 0.539
SW→SW	0.762 / 0.703	0.512 / 0.469	0.383 / 0.401	0.391 / 0.243
SW→EN	0.757 / 0.698	0.438 / 0.412	0.442 / 0.453	0.469 / 0.427

Table 20: Extension to Spanish (ES) and Swahili (SW) on **e-SNLI**. The lower-comprehensiveness/higher-sufficiency pattern persists for both Qwen2.5-7B and Llama3.1-8B.

Lang.	Fully aligned	Partly aligned	Misaligned	Cohen’s κ
ZH	91%	7%	2%	0.76
HI	89%	8%	3%	0.73
AR	87%	10%	3%	0.71
BN	84%	12%	4%	0.68

Table 21: Translation-rationale alignment audit. Two bilingual annotators evaluate 40 translated input-rationale pairs per language. Excluding partly aligned and misaligned instances changes every condition mean by less than 0.02 and preserves all main trends.

Direction	COMET-22 $\uparrow$	Meaning preserved (%)
EN→ZH	0.861	92.0
EN→HI	0.844	90.0
EN→AR	0.852	90.0
EN→BN	0.819	86.0

Table 22: Automatic and human translation-quality checks. COMET-22 scores are consistent with semantic-preservation rates obtained from the bilingual human audit.

J.3 Probability-based faithfulness

Our primary comprehensiveness and sufficiency measures use discrete prediction changes. To test whether the observed effect depends on this binarization, we additionally compute probability-based variants using constrained label scoring. For probability-based comprehensiveness, larger values indicate a larger reduction in confidence after the evidence is removed. For probability-based sufficiency, smaller values indicate a smaller confidence reduction when only the evidence is retained. Table 25 reproduces the same qualitative pattern in every language. Probability-based comprehensiveness decreases under the English pivot, whereas probability-based sufficiency also decreases, indicating that the selected evidence becomes more sufficient in isolation while becoming less causally necessary. Thus, the central finding is not specific to the binary label-flip formulation.

Setting	Evidence mask	Random mask	Ratio
ZH→ZH / ZH→EN	0.577 / 0.507	0.151 / 0.147	$3.8\times / 3.4\times$
HI→HI / HI→EN	0.719 / 0.461	0.148 / 0.152	$4.9\times / 3.0\times$
AR→AR / AR→EN	0.659 / 0.484	0.144 / 0.149	$4.6\times / 3.2\times$
BN→BN / BN→EN	0.488 / 0.419	0.156 / 0.153	$3.1\times / 2.7\times$

Table 23: Evidence masking versus equal-length random-span masking on **e-SNLI** with Qwen2.5-7B. Evidence masking changes predictions substantially more often, while random-mask rates remain nearly constant across explanation-language conditions.

Setting	Avg. words	Coverage	Comp. drop: Full / Matched
ZH→ZH / ZH→EN	7.9 / 7.1	0.20 / 0.18	-0.070 / -0.058
HI→HI / HI→EN	8.4 / 6.8	0.22 / 0.17	-0.258 / -0.196
AR→AR / AR→EN	7.6 / 6.5	0.19 / 0.16	-0.175 / -0.141
BN→BN / BN→EN	8.1 / 7.0	0.21 / 0.18	-0.069 / -0.055

Table 24: Evidence-length control on **e-SNLI** with Qwen2.5-7B. Length-matched comparisons use bins of ± 1 word and retain at least 75% of the original comprehensiveness effect in every language.

J.4 Reverse-direction language mismatch

To distinguish a general reporting-language mismatch effect from an additional effect associated with pivoting into English, we evaluate the reverse $EN \rightarrow L_{\text{native}}$ direction with Qwen2.5-7B across all three tasks (see Table 26). We compare the comprehensiveness drops in the two directions using paired permutation tests over matched instances and apply Holm correction across the four languages within each dataset.

Language mismatch reduces comprehensiveness in both directions. However, the additional $L_{\text{native}} \rightarrow EN$ penalty is significant for Hindi and Arabic in **e-SNLI**; Chinese, Hindi, and Arabic in **FEVER**; and Chinese in **HateXplain**. We therefore conclude that reporting-language mismatch itself contributes to faithfulness degradation, while pivoting into English introduces an additional penalty in several settings, most consistently on **FEVER**.

Setting	Comp _p ↑	Suff _p ↓
ZH→ZH / ZH→EN	0.341 / 0.276	0.224 / 0.162
HI→HI / HI→EN	0.428 / 0.247	0.259 / 0.141
AR→AR / AR→EN	0.389 / 0.271	0.241 / 0.155
BN→BN / BN→EN	0.302 / 0.249	0.198 / 0.132

Table 25: Probability-based comprehensiveness and sufficiency using constrained label scoring. The native-to-English faithfulness degradation persists under the probability-based formulation.

Task	Lang.	$L \rightarrow EN$	$EN \rightarrow L$	Additional penalty	Holm p
e-SNLI	ZH	-0.070	-0.053	-0.017	0.348
	HI	-0.258	-0.069	-0.189	< 0.01
	AR	-0.175	-0.046	-0.129	< 0.01
	BN	-0.069	-0.088	+0.019	0.406
FEVER	ZH	-0.431	-0.184	-0.247	< 0.01
	HI	-0.478	-0.142	-0.336	< 0.01
	AR	-0.394	-0.121	-0.273	< 0.01
	BN	-0.184	-0.136	-0.048	0.081
HateXplain	ZH	-0.194	-0.116	-0.078	0.032
	HI	-0.049	-0.061	+0.012	0.517
	AR	-0.110	-0.072	-0.038	0.094
	BN	-0.055	-0.067	+0.012	0.481

Table 26: Directional asymmetry in comprehensiveness with Qwen2.5-7B p -values using paired permutation tests over matched instances with Holm correction across languages within each task. Bold denotes a significant additional $L_{\text{native}} \rightarrow EN$ penalty.

J.5 Larger-model evaluation with Gemma3-27B

e-SNLI. Table 27 reports the results for Gemma3-27B. The native-to-English shift again reduces comprehensiveness and increases sufficiency for all four languages, while task accuracy changes only marginally. The gaps are 20–35% smaller than for the 7–8B models, suggesting that scale attenuates but does not remove the effect.

Setting	Acc. ↑	Comp. ↑	Suff. ↑	Span Agr. ↑
EN→EN	0.871	0.672	0.305	0.612
ZH→ZH	0.849	0.598	0.322	0.601
ZH→EN	0.845	0.541	0.371	0.552
HI→HI	0.842	0.701	0.264	0.478
HI→EN	0.838	0.512	0.349	0.503
AR→AR	0.846	0.663	0.281	0.522
AR→EN	0.841	0.529	0.362	0.548
BN→BN	0.823	0.517	0.376	0.412
BN→EN	0.819	0.462	0.421	0.489

Table 27: Gemma3-27B results on **e-SNLI**. The lower-comprehensiveness/higher-sufficiency pattern persists across all four languages, although the gaps are 20–35% smaller than for the 7–8B models.

FEVER. On **FEVER**, comprehensiveness decreases under the English pivot by 0.246, 0.272, 0.226, and 0.117 for Chinese, Hindi, Arabic, and Bengali, respectively. Sufficiency simultaneously increases by 0.165, 0.177, 0.155, and 0.091. Span agreement remains mixed, decreasing for Chinese but increasing slightly for the other three languages.

Setting	Acc. ↑	Comp. ↑	Suff. ↑	Span Agr. ↑
EN→EN	0.808	0.318	0.582	0.226
ZH→ZH	0.781	0.754	0.226	0.274
ZH→EN	0.777	0.508	0.391	0.231
HI→HI	0.772	0.526	0.371	0.224
HI→EN	0.768	0.254	0.548	0.256
AR→AR	0.779	0.572	0.337	0.238
AR→EN	0.774	0.346	0.492	0.269
BN→BN	0.751	0.403	0.462	0.225
BN→EN	0.747	0.286	0.553	0.249

Table 28: Gemma3-27B results on **FEVER**. Comprehensiveness decreases and sufficiency increases under the English pivot for every language.

HateXplain. The same faithfulness signature persists on **HateXplain**. Comprehensiveness decreases by 0.123, 0.043, 0.059, and 0.048 for Chinese, Hindi, Arabic, and Bengali, respectively, while sufficiency increases by 0.088, 0.035, 0.043, and 0.042. Span agreement remains mixed, decreasing for three of the four languages. Across **FEVER** and **HateXplain**, Gemma’s accuracy changes by at most 0.005 between the native-language and English-pivot conditions. Together with **e-SNLI**, these results show that Gemma3-27B reproduces the central faithfulness pattern across all three tasks.

J.6 Additional languages

We extend the **e-SNLI** evaluation to Spanish (ES), a European high-resource language, and Swahili (SW), an African lower-resource language. Table 20 reports results for both Qwen2.5-7B and Llama3.1-8B. For Spanish, comprehensiveness decreases from 0.671 to 0.552 for Qwen2.5-7B and from 0.702 to 0.611 for Llama3.1-8B, while sufficiency increases from 0.294 to 0.371 and from 0.246 to 0.318, respectively. For Swahili, comprehensiveness decreases from 0.512 to 0.438 for Qwen2.5-7B and from 0.469 to 0.412 for Llama3.1-8B, while sufficiency increases from 0.383 to 0.442 and from 0.401 to 0.453. Span agreement increases under the English pivot for both models in both languages.

J.7 Detailed HateXplain causal-cue analysis

For completeness, Table 29 reports the language-level breakdown of the causal-cue analysis summarized in Section 5. We analyze 100 instances in which an evidence span whose masking changes the native-condition prediction and is no longer cited under the English-pivot condition. Two annotators per language categorize the dropped expression.

Dropped cue	ZH	HI	AR	BN	All
Slur renderings	41	35	39	36	38
Untranslated / mixed tokens	12	31	18	27	22
Dialect / spelling markers	16	12	15	13	14
In-group references	13	10	11	10	11
Other pragmatic signals	18	12	17	14	15

Table 29: Types of causal cues omitted under the English pivot on **HateXplain** (%). The analysis includes 100 instances in which a causal span cited in the native-language condition is no longer cited under the English-pivot condition.

Translated slur renderings constitute the largest category overall (38%), followed by untranslated or mixed-language tokens (22%). The latter are particularly frequent for Hindi (31%) and Bengali (27%). Since these examples are machine translated, we interpret this taxonomy as evidence for a plausible pragmatic-cue mechanism, rather than as a characterization of naturally occurring multilingual hate speech.