Lost in Interpretation: The Plausibility-Faithfulness Paradox in Cross-Lingual Explanations

Anonymous ACL submission

Abstract

LLMs are often audited with English explanations for non-English inputs, yet these pivot rationales may not reflect how decisions are made. We uncover a *Plausibility–Faithfulness Paradox*: English pivots can sound more human while becoming less evidence-grounded. Across 3 diverse tasks, 5 different languages, and 2 multilingual LLM families, english explanations often turn into fluent post-hoc stories, masking brittle cues and degrading faithfulness by up to 5.7×. In safety-sensitive classification, pivots can also wash out social signals and reduce plausibility. We therefore recommend auditing explanations in the original language and reporting faithfulness alongside plausibility, using english rationales only as a secondary communication layer.

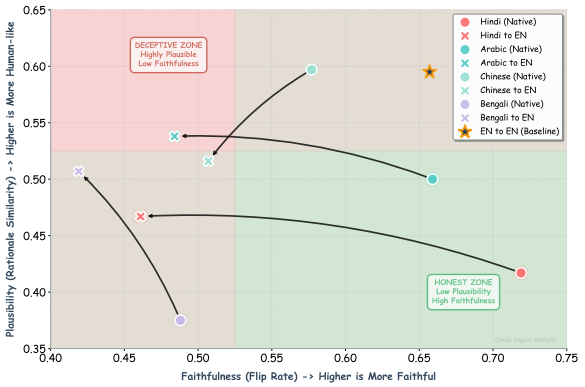

Figure 1: **The Plausibility-Faithfulness Paradox** in SNLI (Qwen2.5-7B). Arrows show the shift from native-aligned explanations to English-pivot explanations. While explaining in English makes the model appear more "human-like" (higher plausibility), it simultaneously makes the explanation less representative of the model's actual logic (lower faithfulness). The figure divides the space into four quadrants to aid visualization.

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 (AWS; Mic, 2025).

For example, in a banking support pipeline, a user reports *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 does 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 evaluation 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 paradox 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) and the transferability of explainable NLP capabilities (Desai and Durrett, 2021). 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 paradox*. 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 receive higher plausibility scores than explanations generated in the input language, particularly for reasoning-intensive and factual tasks. 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 systematic empirical analysis of the impact of **explanation-language mismatch** on LLM faithfulness and plausibility.
2. We identify a `plausibility-faithfulness paradox`, demonstrating that English-pivot explanations can be deceptively persuasive while being logically decoupled from the model's decisions.

2 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.

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

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

2.2 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.

2.3 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). 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. We additionally spot-check a random sample in each language to verify semantic preservation and label integrity.

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.

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.**

2.4 Evaluation metrics

We evaluate explanations along two complementary dimensions shown below.

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).[1]

Plausibility. We measure plausibility as token-level overlap between model-identified and human-annotated evidence:

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

Faithfulness (Flip rate). Faithfulness captures whether the model's stated evidence is *causally necessary* for its prediction. We construct a perturbed input $x' = \text{mask}(x, \mathcal{E}_m(x))$ by masking the tokens indexed by $\mathcal{E}_m(x)$ with a sentinel token (preserving sequence length), and then recompute the prediction. The *flip rate* is:

$$\text{FlipRate} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[f(x_i) \neq f\big(\text{mask}(x_i, \mathcal{E}_m(x_i))\big)\right]. \quad (2)$$

A higher flip rate indicates that removing the model-identified evidence is more likely to change the prediction, and is thus interpreted as higher faithfulness.

2.5 Model selection

We perform our experiments using two state-of-the-art multilingual LLMs namely **Qwen2.5-7B** (Qwen 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 paradox holds across open-weight LLMs rather than arising from a single model.

[1]For tasks where human rationales are provided as free-form text (e.g., e-SNLI explanations), we align them to the input via token matching and treat the matched input-token indices as $\mathcal{E}_h(x)$.

3 Results

We compare three settings namely **condition A** ($EN \rightarrow EN$), **condition B** ($L_{native} \rightarrow L_{native}$), and **condition C** ($L_{native} \rightarrow EN$).
On e-SNLI dataset, Table 1 shows that qwen exhibits the clearest plausibility–faithfulness paradox. When we generate explanations in English instead of the input language, plausibility increases, but faithfulness decreases, which indicates weaker alignment between the explanation and the factors that drive the model's prediction.

Settings	Faithfulness (flip rate) ↑		Plausibility ↑	
	Qwen	Llama	Qwen	Llama
$EN \rightarrow EN$	0.657	0.583	0.595	0.453
$L_{ZH} \rightarrow L_{ZH}$	0.577	0.635	0.597	0.129
$L_{ZH} \rightarrow EN$	**0.507**	**0.573**	0.516	**0.379**
$L_{HI} \rightarrow L_{HI}$	0.719	0.678	0.417	0.545
$L_{HI} \rightarrow EN$	**0.461**	**0.604**	**0.467**	0.509
$L_{AR} \rightarrow L_{AR}$	0.659	0.781	0.500	0.391
$L_{AR} \rightarrow EN$	**0.484**	**0.662**	**0.538**	**0.512**
$L_{BN} \rightarrow L_{BN}$	0.488	0.445	0.375	0.267
$L_{BN} \rightarrow EN$	**0.419**	**0.401**	**0.507**	**0.472**

Table 1: The trade-off between plausibility and faithfulness observed in the e-SNLI task.

In case of FEVER dataset, table 2 shows that the mismatch setting ($L_{native} \rightarrow EN$) often increases plausibility but reduces `faithfulness` on FEVER. This pattern holds in ($L_{HI} \rightarrow EN$), ($L_{AR} \rightarrow EN$), and ($L_{BN} \rightarrow EN$), where plausibility rises for both Qwen and Llama while faithfulness drops relative to ($L_{native} \rightarrow L_{native}$). For ZH, ($L_{ZH} \rightarrow EN$) reduces plausibility for both models, while faithfulness also decreases. We observe that generating English explanations for non-English inputs tends to increase perceived plausibility while weakening causal alignment with the evidence that drives the fact-verification decision.

In case of HateXplain dataset, table 3 shows that the setting $L_{\text{native}} \rightarrow EN$ consistently reduces `faithfulness` relative to the $L_{native} \rightarrow L_{native}$), which indicates that English explanations weaken alignment with the decision cues needed for socially nuanced classification. This degradation appears across all four languages, with Qwen dropping from 0.714 to 0.520 in Chinese and from 0.643 to 0.533 in Arabic, and with Llama dropping from

Settings	Faithfulness (flip rate) ↑		Plausibility ↑	
	Qwen	Llama	Qwen	Llama
$EN \rightarrow EN$	0.271	0.201	0.198	0.112
$L_{ZH} \rightarrow L_{ZH}$	1.000	0.170	0.255	0.247
$L_{ZH} \rightarrow EN$	**0.500**	**0.138**	0.198	0.169
$L_{HI} \rightarrow L_{HI}$	0.579	0.247	0.197	0.147
$L_{HI} \rightarrow EN$	**0.101**	**0.199**	**0.243**	**0.162**
$L_{ZH} \rightarrow L_{ZH}$	0.634	0.317	0.206	0.224
$L_{ZH} \rightarrow EN$	**0.240**	**0.120**	**0.251**	**0.284**
$L_{BN} \rightarrow L_{BN}$	0.356	0.552	0.207	0.169
$L_{BN} \rightarrow EN$	**0.172**	**0.427**	**0.238**	**0.197**

Table 2: Results on FEVER dataset concisely describes the performance breakdown for the fact verification task.

0.884 to 0.676 in Chinese and from 0.810 to 0.788 in Bengali. Plausibility does not show a uniform gain under ($L_{native} \rightarrow EN$) and often decreases, which suggests that generating English explanations does not reliably improve perceived quality for this task.

Settings	Faithfulness (flip rate) ↑		Plausibility ↑	
	Qwen	Llama	Qwen	Llama
$EN \rightarrow EN$	0.651	0.754	0.325	0.294
$L_{ZH} \rightarrow L_{ZH}$	0.714	0.884	0.274	0.333
$L_{ZH} \rightarrow EN$	**0.520**	**0.676**	**0.283**	**0.256**
$L_{HI} \rightarrow L_{HI}$	0.616	0.820	0.354	0.221
$L_{HI} \rightarrow EN$	**0.567**	**0.804**	**0.290**	**0.261**
$L_{AR} \rightarrow L_{AR}$	0.643	0.793	0.284	0.259
$L_{AR} \rightarrow EN$	**0.533**	**0.779**	**0.299**	**0.265**
$L_{BN} \rightarrow L_{BN}$	0.668	0.810	0.359	0.277
$L_{BN} \rightarrow EN$	**0.613**	**0.788**	**0.280**	**0.254**

Table 3: The result for the HateXplain dataset highlights the failure of English pivots to capture social nuances.

4 Discussion

Our results show that reporting language functions as a substantive experimental factor rather than a presentation choice. In e-SNLI and FEVER, $L_{native} \rightarrow EN$ often increases plausibility relative to $L_{native} \rightarrow L_{native}$, while it consistently reduces faithfulness as measured by flip rate. These task-dependent patterns indicate that English explanations for non-English inputs do not consistently reflect the evidence that drives the model prediction, and they can instead act as a narrative layer whose quality depends on the linguistic and pragmatic demands of the task.

Reporting language as an optimization target: We attribute the plausibility gains under ($L_{native} \rightarrow EN$) to the fact that instruction-tuned LLMs are strongly optimized to produce fluent, benchmark-style rationales in English. This objective improves the surface quality of English explanations, but it does not force the explanation to reference the specific cues present in the original (L_{native}) input, such as the decisive words or spans that trigger the prediction. As a result, the model can produce an English rationale that is rhetorically coherent yet only loosely anchored to those ($L_{native} \rightarrow EN$) cues, which increases plausibility while reducing `faithfulness`.

Cross-lingual evidence drift in reasoning and factual tasks: To produce an English explanation from a non-English input, the model implicitly performs translation, abstraction, and selection of salient cues. Each step can shift the rationale away from the features that the classifier actually uses, especially when multiple cues support the same label. In e-SNLI, this drift produces post hoc rationales that remain logically consistent with the predicted label and therefore score well on plausibility, while perturbation tests show that the cited cues are often not causally necessary. These trends align with our deletion-based tests, which show that removing tokens suggested by English explanations often does not flip the prediction, despite high perceived plausibility.

Why social nuance behaves differently?: HateXplain dataset reveals a different failure mode. Under $L_{native} \rightarrow EN$, `faithfulness` decreases across languages, but plausibility does not reliably increase and often decreases. This behavior is consistent with the task requirement to preserve social and language-specific signals such as slang, code-switching, and pragmatic cues that do not translate cleanly into English. When the model explains in English, it often loses or normalizes these signals, which degrades both the causal alignment and the perceived adequacy of the explanation.

5 Conclusion

Our results show that explanation language is not a neutral reporting choice: requiring English rationales for non-English inputs often increases fluency but can reduce causal alignment with the evidence driving the prediction, consistent with cross-lingual drift from implicit translation and abstraction. The effect is task-dependent, especially for socially nuanced classification, English explanations may wash out social- and language-specific cues, lowering adequacy. In practice, fluency is not transparency: evaluate and audit faithfulness in the input language, and treat English explanations mainly as communication summaries (optionally grounded by cited input-language evidence). Detailed practical recommendations are provided in Appendix B.

6 Limitations

While our study provides the first systematic analysis of reporting-language mismatch, several limitations remain. First, our investigation is restricted to two open-weight multilingual models and five languages. The extent of the effect may vary in larger closed-source models or in lower-resource languages where the model's English-centric optimization might be less pronounced. Second, our evaluation of faithfulness relies on deletion-based perturbation tests following the ERASER benchmark; while established, these metrics may not capture all nuances of the model's internal reasoning. Third, we observe that the Plausibility-Faithfulness Paradox is task-dependent. While it is stark in reasoning-heavy tasks like NLI and fact verification, the effect is less uniform in socially nuanced tasks like hate speech detection, where English pivots fail to capture culturally grounded cues. Finally, our methodology assumes that human rationales, even when translated, serve as a valid ground truth for plausibility, which may overlook subtle linguistic shifts introduced during the translation process.

7 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 these "post-hoc storytellers" produce narratives that mask fragile or decoupled decision triggers, which 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*. Furthermore, 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 ethically recommend that English rationales in sensitive contexts be treated strictly as summaries rather than faithful decision traces to prevent the misuse.

References

Machine translation service — amazon translate. Describes real-time translation for chat/helpdesk/ticketing enabling an English-speaking agent to communicate with customers across multiple languages.

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.

Oana-Maria Camburu, Tim Rocktäschel, , and 1 others. 2018. e-snli: Natural language inference with natural language explanations. In *NeurIPS*.

Alexis Conneau and 1 others. 2020. The curse of multilinguality: Adequacy vs. fluency. In *arXiv preprint arXiv:2004.04511*.

Shrey Desai and Greg Durrett. 2021. Cross-lingual transfer learning for explainable nlp. *ACL-IJCNLP*.

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 and 1 others. 2023. Are multilingual llms better reasoners in english? *arXiv preprint arXiv:2308.04031*.

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.

Binny Mathew, Punyajoy Saha, , and 1 others. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection. In *AAAI*.

Qwen, :, 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, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

James Thorne, Andreas Vlachos, , and 1 others. 2018. Fever: a large-scale dataset for fact extraction and verification. In *NAACL-HLT*.

Jannis Vamvas and Rico Sennrich. 2023. Towards cross-lingual faithfulness in explainable ai. *arXiv preprint arXiv:2305.15065*.

Sarah Wiegreffe, Ana Marasovic, and Noah A. Smith. 2021. Measuring faithfulness of text classifications. *ACL-IJCNLP*.

A Prompt Templates and Qualitative Example

A.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 Flip Rate 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 → 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: $\mathbf{L_{native}} \rightarrow \mathbf{L_{native}}$ (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: $\mathbf{L_{native}} \rightarrow$ 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>
```

A.2 Mini Qualitative Example (Illustrative)

Task: e-SNLI (NLI). Labels: {entailment, neutral, contradiction}.
Premise (HI; romanized for pdfLaTeX compatibility):

```
barish ho rahi thi isliye usne chhata liya.
```

Hypothesis (HI; romanized):

```
usne barish ki wajah se chhata istemal kiya.
```

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

```
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 ($\mathbf{L_{native}} \rightarrow$ 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 Flip Rate 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 $x' = \mathrm{mask}(x, T(x))$ by masking those tokens. If the label changes between $f(x)$ and $f(x')$, the instance contributes 1 to Flip Rate. This remains well-defined even when the `Explanation:` is in English (Condition C).

B Practical Recommendations

Based on the identification of the Plausibility-Faithfulness Paradox, 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 Plausibility and incorporate Faithfulness metrics, such as Flip Rates, 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.