Parallel Universes, Parallel Languages: A Comprehensive Study on LLM-based Multilingual Counterfactual Example Generation

Anonymous ACL submission

Abstract

Counterfactuals refer to minimally edited inputs that cause a model's prediction to change, serving as a promising approach to explaining the model's behavior. Large language models (LLMs) excel at generating English counterfactuals and demonstrate multilingual proficiency. However, their effectiveness in generating multilingual counterfactuals remains unclear. To this end, we conduct a comprehensive study on multilingual counterfactuals. We first conduct automatic evaluations on both directly generated counterfactuals in the target languages and those derived via English translation across six languages. Although translation-based counterfactuals offer higher validity than their directly generated counterparts, they demand substantially more modifications and still fall short of matching the quality of the original English counterfactuals. Second, we find the patterns of edits applied to high-resource European-language counterfactuals to be remarkably similar, suggesting that cross-lingual perturbations follow common strategic principles. Third, we reveal that multilingual counterfactual data augmentation (CDA) yields larger model performance improvements than cross-lingual CDA, especially for lower-resource languages. Yet, the imperfections of the generated counterfactuals limit gains in model performance and robustness. Finally, we identify and categorize four main types of errors that consistently appear in the generated counterfactuals across languages.[1]

1 Introduction

The importance of providing explanations in multiple languages and illuminating the behavior of multilingual models has been increasingly recognized (Cui et al., 2022; Zhao and Aletras, 2024; Resck et al., 2025; Dumas et al., 2025). Counterfactual examples, minimally edited inputs that lead to different model predictions than their original counterparts, shed light on a model's black-box behavior in a contrastive manner (Wu et al., 2021; Madaan et al., 2021; Zhao et al., 2024). However, despite significant advancements in counterfactual generation methods (Ross et al., 2021; Bhan et al., 2023b; Wang et al., 2025a) and the impressive multilingual capabilities of LLMs (Üstün et al., 2024; Gao et al., 2025), these approaches have been applied almost exclusively to English (McAleese and Keane, 2024; Nguyen et al., 2024b). Moreover, cross-lingual analyses have revealed systematic behavioral variations between English and non-English contexts (Lai et al., 2023; Poelman and de Lhoneux, 2025), suggesting that English-only counterfactuals are insufficient for capturing the full scope of model behaviors. Nevertheless, the effectiveness of LLMs in generating high-quality multilingual counterfactuals remains an open question.

To bridge this gap, we conduct a comprehensive study on multilingual counterfactuals generated by three LLMs of varying sizes across two multilingual datasets, covering six languages: *English*, *Arabic*, *German*, *Spanish*, *Hindi*, and *Swahili* (Figure 1). **First**, we assess the effectiveness of (1) counterfactuals generated directly in the target language (DG-CFs), and (2) translation-based counterfactuals obtained by translating English counterfactuals (TB-CFs). We observe that DG-CFs in high-resource European languages can frequently successfully change the model prediction, as identified by higher label flip rate (LFR). In particular, English counterfactuals generally surpass the LFR of those in other languages. In comparison, TB-CFs outperform DG-CFs in terms of LFR, although they require substantially more modifications. Moreover, TB-CFs show lower LFR compared to the original English counterfactuals from which they are translated. **Second**, we investigate the extent to which analogous modifications are applied in counterfactuals across different languages

[1] Code and evaluation results are available at: https://anonymous.4open.science/r/multicfe.

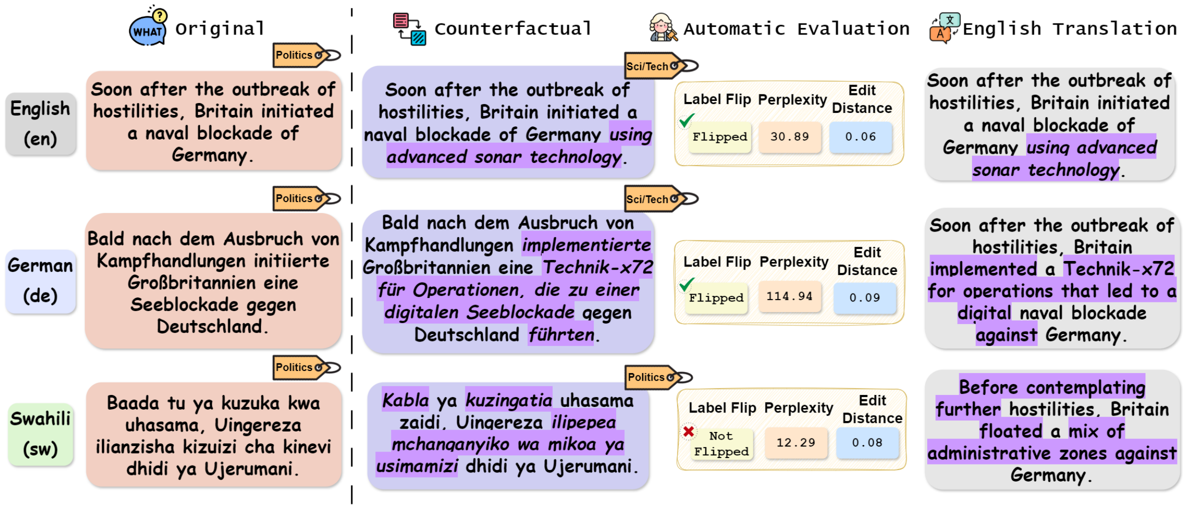


Figure 1: ❶ Parallel original inputs from the SIB200 dataset classified as **"Politics"** in *English* (en), *German* (de), and *Swahili* (sw), ❷ their corresponding counterfactuals aimed at changing the label towards "**Science/Technology**" (Sci/Tech), ❸ automatic evaluation results and ❹ English translations of the generated counterfactuals. Multilingual counterfactuals are evaluated using three automatic metrics (*label flip*↑, *perplexity*↓ and *edit distance*↓). In the multilingual counterfactuals and their English translations, words modified by LLMs are highlighted in purple.

to alter the semantics of the original input. Our analysis demonstrates that input modifications in English, German, and Spanish exhibit a high degree of similarity; specifically, similar words are edited across languages (cf. Figure 14). **Third**, we investigate the impact of cross-lingual and multilingual counterfactual data augmentation (CDA) on model performance and robustness (Liu et al., 2021). While there are mixed signals regarding performance and robustness gains, multilingual CDA generally achieves better model performance than cross-lingual CDA, particularly for low-resource languages. **Lastly**, we report four common error patterns observed in the generated counterfactuals: *copy-paste*, *negation*, *inconsistency* and *language confusion*.

2 Related Work

Counterfactual Example Generation. MICE produces contrastive edits that shift a model's output to a specified alternative prediction (Ross et al., 2021). Polyjuice leverages a fine-tuned GPT2 (Radford et al., 2019) to determine the type of transformation needed for generating counterfactual instances (Wu et al., 2021). Bhan et al. (2023a) propose a method to determine impactful input tokens with respect to generated counterfactual examples. CREST (Treviso et al., 2023) generates counterfactual examples by combining rationalization with span-level masked language modeling. Bhattacharjee et al. (2024b) uncover latent representations in the input and link them back to observable features to craft counterfactuals. FIZLE (Bhattacharjee et al., 2024a) uses LLMs as pseudo-oracles in a zero-shot setting, guided by important words generated by the same LLM, to create counterfactual examples. ZeroCF (Wang et al., 2025a) utilizes feature importance methods to pinpoint the important words that steer the generation of counterfactual examples. However, all of these methods have been evaluated exclusively on English datasets, leaving the ability of LLMs to generate multilingual counterfactuals underexplored.

Counterfactual Explanation Evaluation. The quality of counterfactuals can be assessed using various automatic evaluation metrics. **Label Flip Rate (LFR)** is positioned as the primary evaluation metric for assessing the effectiveness and validity of generated counterfactuals (Ross et al., 2021; Ge et al., 2021; Nguyen et al., 2024b). LFR is defined as the percentage of instances in which labels are successfully flipped, relative to the total number of generated counterfactuals. **Similarity** measures the extent of textual modification, typically quantified by edit distance, required to generate the counterfactual (Bhattacharjee et al., 2024a; Wang et al., 2025a). **Diversity** quantifies the average pairwise distance between multiple counterfactual examples for a given input (Wu et al., 2021; Chen et al., 2023). **Fluency** assesses the degree to which a counterfactual resembles human-written text (Robeer et al.,

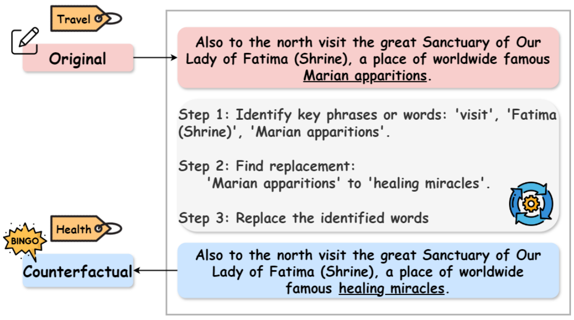


Figure 2: An overview of counterfactual generation process. Given an original instance x from SIB200 classified as "**Travel**", the corresponding counterfactual $\tilde{x}$ is classified as "**Health**". Edits to x are underlined.

2021; Madaan et al., 2021).

Multilingual Counterfactuals. Liu et al. (2021) propose using multilingual counterfactuals as additional training data for machine translation – an approach known as counterfactual data augmentation (CDA). The counterfactuals employed in CDA flip the ground-truth labels rather than the model predictions, and therefore differ from counterfactual explanations explored in this paper. Barriere and Cifuentes (2024) leverage counterfactuals to evaluate nationality bias across diverse languages. Saha Roy et al. (2025) use counterfactuals to measure answer attribution in a bilingual retrieval augmentation generation system. Nevertheless, none of existing work has investigated how well LLMs are capable of generating high-quality multilingual counterfactual explanations.

3 Experimental Setup

3.1 Counterfactual Generation

Our goal is to generate a counterfactual $\tilde{x}$ for the input x, changing the original model prediction y to the target label $\tilde{y}$. We adopt the counterfactual generation approach proposed by Nguyen et al. (2024b), which fulfills the following properties:

- Generated counterfactuals can be used for counterfactual data augmentation (§5.3).
- Human intervention or additional training of LLMs is not required, thereby ensuring computational feasibility.
- Generated counterfactuals rely solely on the evaluated LLM to avoid confounding factors, e.g., *extrinsic* important feature signals (Bhan et al., 2023b; Wang et al., 2025a; Nguyen et al., 2025).

We directly generate counterfactuals $\tilde{x}$ (DG-CFs, Table 1a) in target languages through a three-step process as shown in Figure 2[2]:

(1) Identify the important words in the original input that are most influential in flipping the model's prediction.
(2) Find suitable replacements for these identified words that are likely to lead to the target label.
(3) Substitute the original words with the selected replacements to construct the counterfactual.

Furthermore, we investigate the effectiveness of translation-based counterfactuals $\tilde{x}_{\text{en-}\ell}$ (TB-CFs, Table 1b), where $\ell \in \{\text{ar,de,es,hi,sw}\}$. Specifically, LLMs first follow the three-step process in Figure 2 to generate counterfactuals in English. We then apply the same LLM to translate these generated counterfactuals into the target languages (Figure 8). Translation quality is evaluated in Appendix D by automatic evaluation metrics (§D.1) and human annotators (§D.2).

3.2 Datasets

We focus on two widely studied classification tasks in the counterfactual generation literature: natural language inference and topic classification. Accordingly, we select two task-aligned multilingual datasets and evaluate the resulting multilingual counterfactual examples.[3]

XNLI (Conneau et al., 2018) is designed for cross-lingual natural language inference (NLI) tasks. It extends the English MultiNLI (Williams et al., 2018) corpus by translating into 14 additional languages. XNLI categorizes the relationship between a *premise* and a *hypothesis* into *entailment*, *contradiction*, or *neutral*.

SIB200 (Adelani et al., 2024) is a large-scale dataset for topic classification across 205 languages. SIB200 categorizes sentences into seven distinct topics: *science/technology*, *travel*, *politics*, *sports*, *health*, *entertainment*, and *geography*.

Language Selection We identify six overlapping languages between the XNLI and SIB200 datasets: *English*, *Arabic*, *German*, *Spanish*, *Hindi*, and *Swahili*. These languages are representative of their typological diversity, spanning a spectrum from widely spoken to low-resource languages and encompassing a variety of scripts.

[2] Prompt instructions are provided in Appendix A.
[3] Dataset examples and label distributions are presented in Appendix B.

3.3 Models

We select three state-of-the-art, open-source, instruction fine-tuned LLMs with increasing parameter sizes: `Qwen2.5-7B` (Qwen et al., 2024), `Gemma3-27B` (Team, 2025), `Llama3.3-70B` (Grattafiori et al., 2024).[4] These models offer multilingual support and have been trained on data that include multiple selected languages (§3.2, Appendix C.1.1). Furthermore, in our experiments, we aim to use counterfactuals to explain a multilingual `BERT` (Devlin et al., 2019), which is fine-tuned on the target dataset (§3.2).[5]

4 Evaluation Setup

4.1 Automatic Evaluation

We evaluate the generated multilingual counterfactuals using three automated metrics widely adopted in the literature (Ross et al., 2021; Bhan et al., 2023a; Nguyen et al., 2024a; Wang et al., 2025a):

Label Flip Rate (LFR) quantifies how often counterfactuals lead to changes in their original model predictions (Ge et al., 2021; Nguyen et al., 2024b; Bhattacharjee et al., 2024b). For a dataset containing N instances, LFR is calculated as:

$$LFR = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\big(\mathcal{M}(\tilde{x}_i) \neq \mathcal{M}(x_i)\big)$$

where $\mathbb{1}$ is the indicator function, which returns 1 if the condition is true and 0 otherwise. $\mathcal{M}$ denotes the model to be explained. x_i represents the original input and $\tilde{x}_i$ is the corresponding counterfactual.

Textual Similarity (TS) The counterfactual $\tilde{x}$ should closely resemble the original input x (Madaan et al., 2021), with smaller distances signifying higher similarity. We measure this similarity using the Levenshtein distance d on token level, a standard metric in existing research literature (Ross et al., 2021; Treviso et al., 2023; Wang et al., 2025a), which captures all edits between inputs:

$$TS = \frac{1}{N}\sum_{i=1}^{N}\frac{d(x_i, \tilde{x}_i)}{|x_i|}$$

Perplexity (PPL) is the exponential of the average negative log-likelihood computed over a sequence. It measures the naturalness of text distributions and indicates how fluently a model can predict the subsequent word based on preceding words (Fan et al., 2018). For a given sequence $\mathcal{S} = (t_1, t_2, \cdots, t_n)$, PPL is computed as follows:

$$PPL(\mathcal{S}) = \exp\left\{\frac{1}{n}\sum_{i=1}^{n}\log p_\theta(t_i|t_{<i})\right\}$$

While `GPT2` parameterized by θ is commonly used in the counterfactual literature to calculate PPL (Le et al., 2023; Nguyen et al., 2024a), it is trained on English data only and is therefore unsuitable for multilingual counterfactual evaluation. Consequently, we use `mGPT-1.3B` (Shliazhko et al., 2024), which excels at modeling text distributions and provides coverage across all target languages, to compute PPL in our experiments.[6]

4.2 Cross-lingual Edit Similarity

Following the concept of cross-lingual consistency (Qi et al., 2023), we investigate the extent to which cross-lingual modifications are consistently applied in counterfactuals across different languages to alter the semantics of the original input.[7] To this end, we employ a pretrained multilingual `SBERT`[8] to measure the sentence embedding similarity by (1) computing pairwise cosine similarity among directly generated counterfactuals $\tilde{x}_\ell$ across different target languages ℓ; (2) back-translating the directly generated counterfactuals $\tilde{x}_\ell$ from language ℓ into English $\tilde{x}_{\ell\text{-EN}}$ and quantifying the pairwise cosine similarity among these (back-translated) English counterfactuals.

4.3 Counterfactual Data Augmentation

To validate whether, and to what extent, counterfactual examples enhance model performance and robustness (Kaushik et al., 2020; Gardner et al., 2020; Dixit et al., 2022; Wang et al., 2025b), we conduct cross-lingual and multilingual CDA experiments using a pretrained multilingual `BERT`. The baseline for CDA is denoted as $\mathcal{M}_{base}$, which is fine-tuned on $D_{\text{base}_c} = \{(x_{i,\text{en}}, y_i) \mid i \in \mathcal{N}\}$ for cross-lingual CDA, and on $D_{\text{base}_m} =$

[4]Further details about the used models for counterfactual generation and inference time can be found in Appendix C.

[5]Information about the explained models and the training process is detailed in Appendix C.1.2.

[6]Average perplexity scores of data points in different target languages across each dataset are provided in Table 11.

[7]We instruct LLMs to edit each original input in multiple languages while keeping the target counterfactual label fixed.

[8]https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

Model	Language	XNLI LFR ↑	XNLI PPL ↓	XNLI TS ↓	SIB200 LFR ↑	SIB200 PPL ↓	SIB200 TS ↓
Qwen2.5-7B	en	45.42%	36.68	0.0518	**92.16%**	46.30	0.0334
	ar	44.10%	36.75	0.0586	89.22%	124.37	0.0845
	de	46.63%	32.85	0.0453	77.45%	34.42	0.0364
	es	**49.44%**	30.36	0.0480	72.55%	26.97	0.0368
	hi	39.92%	**8.12**	0.0498	89.71%	**4.84**	0.0484
	sw	38.31%	24.04	**0.0400**	84.80%	22.57	**0.0327**
Gemma3-27B	en	**43.37%**	38.26	0.0695	**87.75%**	53.66	0.0498
	ar	37.59%	36.32	0.0822	87.75%	81.96	0.1342
	de	38.19%	33.69	**0.0590**	79.41%	34.94	**0.0463**
	es	39.92%	31.18	0.0604	80.88%	30.73	0.0507
	hi	36.43%	**11.30**	0.0745	81.37%	**4.35**	0.0778
	sw	33.90%	23.30	0.0610	87.25%	16.70	0.0770
Llama3.3-70B	en	**50.88%**	39.47	0.0483	87.25%	52.84	0.0663
	ar	36.91%	37.85	0.0511	88.73%	77.32	0.2008
	de	42.25%	33.59	0.0378	78.43%	31.58	**0.0485**
	es	44.70%	31.20	0.0405	83.33%	29.41	0.0581
	hi	41.33%	**10.46**	0.0505	85.29%	**4.39**	0.0602
	sw	34.42%	22.67	**0.0361**	**91.18%**	14.43	0.0612

(a) Directly generated counterfactuals $\tilde{x}_\ell$.

Model	Language	XNLI LFR ↑	XNLI PPL ↓	XNLI TS ↓	SIB200 LFR ↑	SIB200 PPL ↓	SIB200 TS ↓
Qwen2.5-7B	en-ar	43.49%	110.76	0.2208	90.20%	45.11	0.1920
	en-de	44.54%	73.59	**0.1549**	**93.63%**	39.90	**0.1415**
	en-es	**45.98%**	52.24	0.1754	92.16%	28.26	0.1594
	en-hi	41.45%	**9.40**	0.2055	90.20%	**9.19**	0.1875
	en-sw	43.73%	57.39	0.1623	92.65%	46.35	0.1555
Gemma3-27B	en-ar	42.49%	48.27	0.2389	88.73%	27.09	0.1936
	en-de	42.49%	52.77	**0.1745**	**90.20%**	27.01	**0.1454**
	en-es	42.69%	50.43	0.1958	89.22%	24.31	0.1599
	en-hi	42.41%	**5.73**	0.2249	89.22%	**4.10**	0.1941
	en-sw	**43.01%**	28.28	0.1924	85.78%	13.66	0.1605
Llama3.3-70B	en-ar	45.14%	169.00	0.2331	86.27%	34.47	0.1932
	en-de	47.58%	60.86	**0.1613**	**86.27%**	31.00	**0.1418**
	en-es	**50.04%**	54.28	0.1813	73.53%	28.80	0.1588
	en-hi	44.66%	**5.51**	0.2145	83.82%	**4.02**	0.1921
	en-sw	44.78%	38.65	0.1784	85.29%	13.53	0.1593

(b) Translation-based counterfactuals $\tilde{x}_{\text{en-}\ell}$.

Table 1: Automatic evaluation results of counterfactuals based on label flip rate (LFR), perplexity (PPL), and edit distance (TS) on XNLI and SIB200 across *English* (en), *Arabic* (ar), *German* (de), *Spanish* (es), *Hindi* (hi), and *Swahili* (sw). **Bold**-faced languages indicate the best performance on a given metric, while underlined languages denote the worst.

$\{(x_{i,\ell}, y_i) \mid i \in \mathcal{N}, \ell \in \{\text{en,ar,de,es,hi,sw}\}\}$ for multilingual CDA, where $\mathcal{N}$ denotes the total number of data points. The counterfactually augmented models $\mathcal{M}_c$ and $\mathcal{M}_m$ are fine-tuned using D_{base_c} and D_{base_m}, respectively, along with their corresponding counterfactuals $\tilde{x}_\ell$ in the target languages ℓ, generated either directly (§5.3) or through translation (Appendix E.3) with different LLMs.

5 Results

5.1 Multilingual Counterfactual Quality

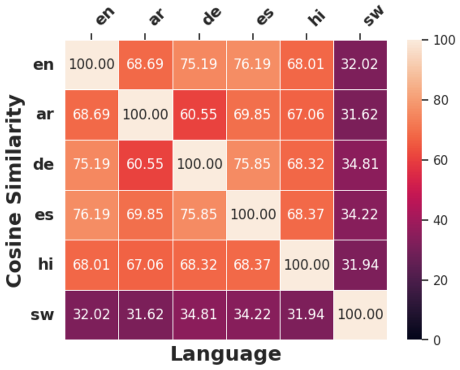
5.1.1 Directly Generated Counterfactuals

Table 1a displays that LFR is dramatically higher for all models on SIB200 than on XNLI, reflecting the greater inherent difficulty of the NLI task. **Counterfactuals in English tend to achieve the highest LFR on both XNLI and SIB200.** On XNLI, the gap between high- and low-resource languages widens with model scale, reaching up to 16.46%. In contrast, on SIB200, this gap narrows, where, for instance, counterfactuals in Swahili generated by `Llama3.3-70B` attain the highest LFR. **Nevertheless, higher-resource European languages (English, German, and Spanish) generally exhibit higher LFRs than lower-resource languages (Arabic, Hindi and Swahili).** Furthermore, counterfactuals in Hindi consistently achieve the best perplexity scores across all three models, indicating superior fluency, whereas counterfactuals in Arabic are generally less fluent. Meanwhile, counterfactuals in Arabic involve more extensive modifications to the original texts indicated by higher edited distances, whereas those in Swahili and German are generally less edited. Additionally, no single model produces counterfactuals that are optimal across every metrics and language. **Likewise, counterfactuals in none of the languages consistently excel across all evaluation metrics.** For example, English counterfactuals achieve higher LFR, but exhibit lower fluency and require more edits than those in other languages, underscoring that the idea of an "optimal" or "sub-optimal" language for counterfactual quality is inherently contextual and metric-dependent.

5.1.2 Translation-based Counterfactuals

Comparison with DG-CFs. Table 1b demonstrates that, in most cases[9], TB-CFs $\tilde{x}_{\text{en-}\ell}$ yield higher LFR than DG-CFs $\tilde{x}_\ell$ in the target language ℓ (Table 1a). The LFR improvement is most pronounced for German and least significant for Hindi, although the validity of counterfactuals in Hindi consistently benefits from the translation. Despite TB-CFs $\tilde{x}_{\text{en-}\ell}$ achieving higher LFR compared to DG-CFs $\tilde{x}_\ell$, overall, the LFR of $\tilde{x}_{\text{en-}\ell}$ is lower than that of the original English counterfactuals $\tilde{x}_{\text{en}}$. In addition, TB-CFs $\tilde{x}_{\text{en-}\ell}$ generally require significantly more edits (on average 335% more than DG-CFs) to the original input due to artifacts introduced by machine translation, and they tend to exhibit lower fluency (38% lower on average) owing to limitations in translation quality.

[9] In other cases, impairments in translation-based counterfactual quality may suffer from imperfect translations and limitations in LLMs' counterfactual generation capabilities, particularly pronounced on XNLI.



(a) XNLI

(b) SIB200

Figure 3: Cosine similarity scores of counterfactuals $\tilde{x}_\ell$ across different languages measured by SBERT.

Correlation between TB-CFs and Machine Translation. The degree of LFR improvement is weakly positively correlated with the machine translation quality, measured by automatic evaluation (Spearman's $\rho = 0.27$, Table 7) and by human evaluation (Spearman's $\rho = 0.07$, Table 8) (Appendix D). The weak observed correlations suggest that improvements are driven primarily by the quality of the English counterfactuals, with translation quality contributing only to a limited extent.

5.2 Cross-lingual Edit Similarity

Figure 3 and Figure 13 indicate that LLMs generally edit inputs for Swahili and Arabic counterfactuals in a substantially different manner than other languages, as evidenced by lower cosine similarity scores.[10] Notably, for European languages (English, German and Spanish), LLMs tend to apply similar modifications to the original input during counterfactual generation (Figure 14), likely because of structural and lexical similarities among these languages (Haspelmath, 2005; Holman et al., 2011) (Appendix E.2.3). Additionally, the edits applied across different languages when generating counterfactuals on SIB200 differ markedly from those on XNLI, as reflected in noticeable differences in cosine similarity scores between the two datasets. This disparity likely stem from SIB200's focus on topic classification. When a target label is specified, compared to XNLI, there might be more distinct ways to construct valid counterfactuals that elicit the required prediction change.

[10]Cosine similarity scores for original input and back-translated counterfactuals $\tilde{x}_{\ell\text{-en}}$ in English from XNLI and SIB200, are provided in Appendix E.2.

Model	Counter-factual	Lang-uage	Cross-lingual XNLI	Cross-lingual SIB200	Multilingual XNLI	Multilingual SIB200
$\mathcal{M}_{base}$	-	en	68.70	83.80	72.22	82.83
	-	ar	60.12	25.30	63.21	54.55
	-	de	63.33	88.90	67.60	87.88
	-	es	66.05	87.90	68.72	87.88
	-	hi	56.09	74.70	62.04	80.81
	-	sw	48.66	64.60	59.00	78.79
$\mathcal{M}_c/\mathcal{M}_m$	Qwen2.5-7B	en	69.86$_{+1.16}$	82.80$_{-1.00}$	73.45$_{+1.23}$	85.86$_{+3.03}$
		ar	58.10$_{-2.02}$	26.30$_{+1.00}$	64.89$_{+1.68}$	53.54$_{-1.01}$
		de	63.49$_{+0.16}$	84.80$_{-4.10}$	68.42$_{+0.82}$	84.85$_{-3.03}$
		es	65.43$_{-0.62}$	84.80$_{-3.10}$	69.94$_{+1.22}$	88.89$_{+1.01}$
		hi	55.33$_{-0.76}$	75.80$_{+1.10}$	62.32$_{+0.28}$	75.76$_{-5.05}$
		sw	48.92$_{+0.26}$	63.60$_{-1.00}$	57.74$_{-1.26}$	76.77$_{-2.02}$
	Gemma3-27B	en	71.66$_{+2.96}$	85.90$_{+2.10}$	74.61$_{+2.39}$	86.87$_{+4.04}$
		ar	56.01$_{-4.11}$	23.20$_{-2.10}$	65.11$_{+1.90}$	49.49$_{-5.06}$
		de	62.53$_{-0.80}$	87.90$_{-1.00}$	68.66$_{+1.06}$	86.87$_{-1.01}$
		es	64.35$_{-1.70}$	86.90$_{-1.00}$	70.98$_{+2.26}$	89.90$_{+2.02}$
		hi	52.38$_{-3.71}$	73.70$_{-1.00}$	61.10$_{-0.94}$	83.84$_{+3.03}$
		sw	46.81$_{-1.85}$	64.60$_{0.00}$	55.57$_{-3.43}$	70.71$_{-8.08}$
	Llama3.3-70B	en	70.86$_{+2.16}$	83.80$_{0.00}$	74.61$_{+2.39}$	83.84$_{+1.01}$
		ar	55.01$_{-5.11}$	25.30$_{0.00}$	64.77$_{+1.56}$	56.57$_{+2.02}$
		de	61.58$_{-1.75}$	83.80$_{-5.10}$	68.26$_{+0.66}$	87.88$_{0.00}$
		es	63.51$_{-2.54}$	84.80$_{-3.10}$	71.32$_{+2.60}$	88.89$_{+1.01}$
		hi	51.28$_{-4.81}$	73.70$_{-1.00}$	62.46$_{+0.42}$	79.80$_{-1.01}$
		sw	46.89$_{-1.77}$	59.60$_{-5.00}$	55.21$_{-3.79}$	73.74$_{-5.05}$

Table 2: Cross-lingual and multilingual CDA results (*accuracy* in %) for the base model $\mathcal{M}_{base}$ and the counterfactually augmented models $\mathcal{M}_c$ and $\mathcal{M}_m$.

5.3 Counterfactual Data Augmentation

Table 2 reflects that for the base model $\mathcal{M}_{base}$, multilingual CDA generally leads to a substantial improvement in performance compared to cross-lingual CDA across two datasets. This effect is especially compelling for Arabic, with average accuracy gains of 64.45%, while for English, the improvement is least observable due to its already satisfactory performance in the cross-lingual setting.

For XNLI, cross-lingual CDA enhances model performance only on English, while degrading performance on the other languages. In the context of multilingual CDA, overall, model performance improves across languages other than Swahili. For SIB200, in the cross-lingual setting, CDA generally has an adverse impact on model performance. Meanwhile, although the generated counterfactuals are more effective and valid (Table 1), in the multilingual setting, augmenting with these counterfactuals only yields reliable gains in English and Spanish, while it even consistently hampers per-

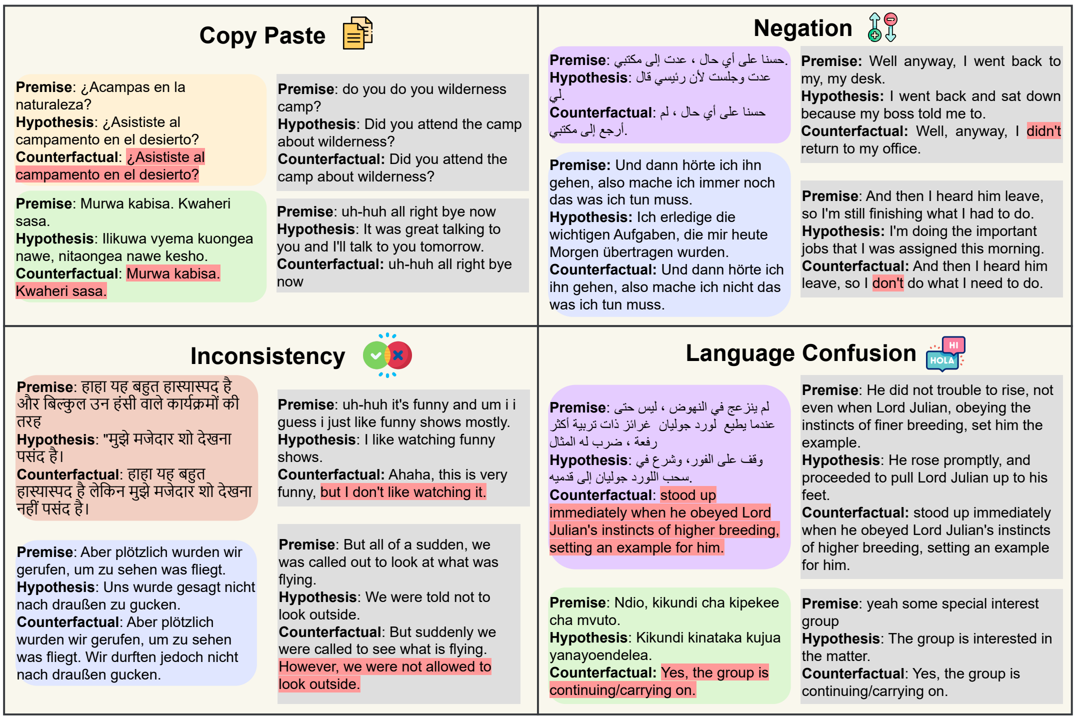


Figure 4: Categorization of error types in generating multilingual counterfactuals across five languages: *Arabic*, *German*, *Spanish*, *Hindi*, and *Swahili*. For each error type, we present two examples and their corresponding *English* translations. Error spans are marked with red highlights to indicate the exact locations of the issues.

formance for Swahili. This effect is remarkably pronounced when using smaller LLMs.

The limited performance improvement from augmenting with counterfactuals can be attributed to the imperfection of generated counterfactuals, which stems from both the limited multilingual capabilities of LLMs and suboptimal multilingual counterfactual generation method (cf. Figure 4). Furthermore, while counterfactuals for SIB200 often succeed in flipping model predictions, they frequently fail to flip the ground-truth labels due to insufficient revision – an essential requirements for CDA – resulting in noisy labels that can even deteriorate performance (Zhu et al., 2022; Song et al., 2023; Wang et al., 2025b).[11]

5.4 Error Analysis

Generating counterfactuals is not immune to errors, possibly due to the suboptimal instruction-following ability of LLMs and their difficulty in handling fine-grained semantic changes. Nguyen et al. (2024b) have identified three common categories of errors in English counterfactuals. We hypothesize that similar issues may arise in multilingual counterfactual generation. Building on this insight, we examine the directly generated counterfactuals $\tilde{x}_\ell$ in all target languages, analyzing them both manually and automatically, depending on the type of error. To facilitate our investigation, we translate the counterfactuals into English when necessary and compare them against their original texts. Based on this process, we identify four distinct error types, which we summarize below (see Figure 4 for examples in each error type).

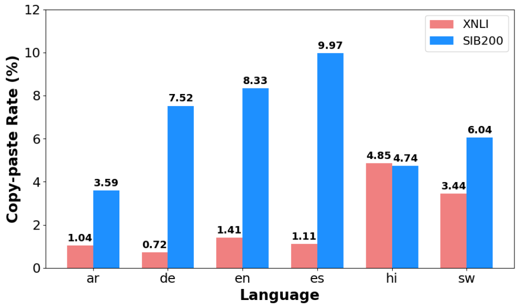
Copy-Paste. When LLMs are prompted to generate counterfactuals by altering the model-predicted label, they occasionally return the original input unchanged as the counterfactual. Figure 5a shows that the copy-paste rate is considerably higher on SIB200 (average: 6.7%) than on XNLI (average: 2.1%). However, the trend in two datasets is not consistent across languages. High-resource languages like English and Spanish in SIB200 present higher copy-paste rates. In contrast, lower-resource languages like Hindi and Swahili are most affected by the copy-paste issue. A closer inspection suggests that LLMs often struggle to sufficiently revise the input to align with the target label, resulting in incomplete or superficial edits.

[11]Further details on CDA, including training-data selection, model training, and additional results evaluated using human-annotated counterfactuals, are offered in Appendix E.3.

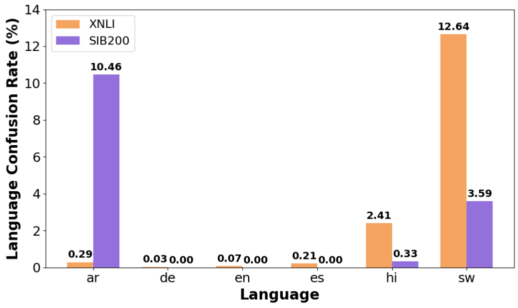
Negation. For counterfactual generation, LLMs often attempt to reverse the original meaning by introducing explicit negation while preserving most of the context. However, this strategy frequently fails to trigger the intended label change, resulting in semantically ambiguous or label-preserving outputs (Wang et al., 2025b). A likely reason is that LLMs may rely on shallow heuristics – negation being a common surface-level cue for meaning reversal learned during pretraining. Especially in languages with simple and explicit negation markers (e.g., English, German), LLMs tend to perform minimal edits (e.g., adding "not") rather than making deeper structural changes required for a true semantic shift.

Inconsistency. Counterfactuals may introduce statements that are logically contradictory or incoherent relative to the original input. This often results from the model appending or modifying content without fully reconciling the semantic implications of the added text with the existing context. In such cases, the counterfactual may contain mutually exclusive statements, e.g., simultaneously asserting that an event occurred and that it was prohibited (cf. Figure 4). These inconsistencies highlight the model's difficulty in preserving global meaning while introducing label-altering edits, particularly when attempting to retain much of the original phrasing.

Language Confusion. We further identify the language of directly generated counterfactuals $\tilde{x}$ and examine whether it aligns with the intended target language.[12] Figure 5b illustrates the language confusion rate (Marchisio et al., 2024) across different languages on XNLI and SIB200. Overall, counterfactuals in high-resource languages, i.e., German, English, and Spanish, can be generated consistently in the expected target language. In contrast, when relatively lower-resource languages, such as Arabic or Swahili, are specified as the target language, LLMs frequently misinterpret the prompts[13] or default to generate counterfactuals in the predominant language of English (Hwang et al., 2025).



(a) Copy-paste rate



(b) Language confusion rate

Figure 5: (a) Copy-paste rates and (b) language confusion rates for counterfactuals across different languages.

6 Conclusion

In this work, we first conducted automatic evaluations on directly generated counterfactuals in the target languages and translation-based counterfactuals generated by three LLMs across two datasets covering six languages. Our results show that directly generated counterfactuals in high-resource European languages tend to be more valid and effective. Translation-based counterfactuals yield higher LFR than directly generated ones but at the cost of substantially greater editing effort. Nonetheless, these translated variants still fall short of the original English counterfactuals from which they derive. Second, we revealed that the nature and pattern of edits in English, German, and Spanish counterfactuals are strikingly similar, indicating that cross-lingual perturbations follow common strategies. Third, we extended our study to CDA. Evaluations across languages show that multilingual CDA outperforms cross-lingual CDA, particularly for low-resource languages. However, given that the multilingual counterfactuals are imperfect, CDA does not reliably improve model performance or robustness. Lastly, we cataloged four principal error types that emerge in the generated counterfactuals. Of these, the tendency to copy and paste segments from the source text is by far the most pervasive issue across languages and models. Based on our findings, we want to emphasize the urgent need to develop more advanced approaches to improve multilingual counterfactual generation and further enhance the explainability of multilingual LLMs.

[12] https://github.com/zafercavdar/fasttext-langdetect

[13] More discussion about the selection of languages for prompts can be found in Appendix A.

Limitations

We use token-level Levenshtein distance to measure textual similarity (§5.1), which is widely adopted in prior studies (Ross et al., 2021; Treviso et al., 2023; Wang et al., 2025a); however, it might not fully capture similarity when non-Latin scripts are involved. This underscores the need for further investigation of this issue in a multilingual setting, e.g., through the development of new metrics for textual similarity.

We do not exhaustively explore all languages common to SIB200 and XNLI; instead, we select 6 languages spanning from high-resource to low-resource to ensure typological diversity and cover a variety of scripts (§3.2). Thus, expanding the evaluation to more languages and exploring more models with different architectures and sizes are considered as directions for future work.

In this work, following prior research on comprehensive studies of English counterfactuals (Nguyen et al., 2024b; Wang et al., 2024; McAleese and Keane, 2024), we focus exclusively on automatic evaluations of multilingual counterfactuals along three dimensions – validity, fluency and minimality (§5.1), rather than on subjective aspects such as usefulness, helpfulness, or coherence of counterfactuals (Domnich et al., 2025; Wang et al., 2025c), which can only be assessed through user study. As future work, we plan to conduct a user study to subjectively assess the quality of the multilingual counterfactuals.

Ethics Statement

The participants in the machine translation evaluation (Appendix D.2) were compensated at or above the minimum wage, in accordance with the standards of our host institutions' regions. The annotation took each annotator approximately an hour on average.

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Valentin Barriere and Sebastian Cifuentes. 2024. A study of nationality bias in names and perplexity using off-the-shelf affect-related tweet classifiers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 569–579, Miami, Florida, USA. Association for Computational Linguistics.

Milan Bhan, Jean-noel Vittaut, Nicolas Chesneau, and Marie-jeanne Lesot. 2023a. Enhancing textual counterfactual explanation intelligibility through counterfactual feature importance. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 221–231, Toronto, Canada. Association for Computational Linguistics.

Milan Bhan, Jean-Noël Vittaut, Nicolas Chesneau, and Marie-Jeanne Lesot. 2023b. Tigtec: Token importance guided text counterfactuals. In *Machine Learning and Knowledge Discovery in Databases: Research Track*, pages 496–512, Cham. Springer Nature Switzerland.

Amrita Bhattacharjee, Raha Moraffah, Joshua Garland, and Huan Liu. 2024a. Zero-shot LLM-guided Counterfactual Generation: A Case Study on NLP Model Evaluation . In *2024 IEEE International Conference on Big Data (BigData)*, pages 1243–1248, Los Alamitos, CA, USA. IEEE Computer Society.

Amrita Bhattacharjee, Raha Moraffah, Joshua Garland, and Huan Liu. 2024b. Towards llm-guided causal explainability for black-box text classifiers. In *AAAI 2024 Workshop on Responsible Language Models, Vancouver, BC, Canada.*

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Zeming Chen, Qiyue Gao, Antoine Bosselut, Ashish Sabharwal, and Kyle Richardson. 2023. DISCO: Distilling counterfactuals with large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5514–5528, Toronto, Canada. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of*

the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Yiming Cui, Wei-Nan Zhang, Wanxiang Che, Ting Liu, Zhigang Chen, and Shijin Wang. 2022. Multilingual multi-aspect explainability analyses on machine reading comprehension models. *iScience*, 25(5):104176.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tanay Dixit, Bhargavi Paranjape, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. CORE: A retrieve-then-edit framework for counterfactual data generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2964–2984, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Marharyta Domnich, Julius Välja, Rasmus Moorits Veski, Giacomo Magnifico, Kadi Tulver, Eduard Barbu, and Raul Vicente. 2025. Towards unifying evaluation of counterfactual explanations: Leveraging large language models for human-centric assessments. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(15):16308–16316.

Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2025. Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31822–31841, Vienna, Austria. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Changjiang Gao, Xu Huang, Wenhao Zhu, Shujian Huang, Lei Li, and Fei Yuan. 2025. Could thinking multilingually empower llm reasoning? *Preprint*, arXiv:2504.11833.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, and 7 others. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Yingqiang Ge, Shuchang Liu, Zelong Li, Shuyuan Xu, Shijie Geng, Yunqi Li, Juntao Tan, Fei Sun, and Yongfeng Zhang. 2021. Counterfactual evaluation for explainable ai. *Preprint*, arXiv:2109.01962.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Martin Haspelmath. 2005. *The world atlas of language structures*. Oxford University Press.

Eric W Holman, Cecil H Brown, Søren Wichmann, André Müller, Viveka Velupillai, Harald Hammarström, Sebastian Sauppe, Hagen Jung, Dik Bakker, Pamela Brown, and 1 others. 2011. Automated dating of the world's language families based on lexical similarity. *Current Anthropology*, 52(6):841–875.

Kaiyu Huang, Fengran Mo, Xinyu Zhang, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, Jinan Xu, Jian-Yun Nie, and Yang Liu. 2025. A survey on large language models with multilingualism: Recent advances and new frontiers. *Preprint*, arXiv:2405.10936.

Jaedong Hwang, Kumar Tanmay, Seok-Jin Lee, Ayush Agrawal, Hamid Palangi, Kumar Ayush, Ila Fiete, and Paul Pu Liang. 2025. Learn globally, speak locally: Bridging the gaps in multilingual reasoning. *Preprint*, arXiv:2507.05418.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Viet Dac Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Tiep Le, Vasudev Lal, and Phillip Howard. 2023. COCO-counterfactuals: Automatically constructed counterfactual examples for image-text pairs. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2025. Is translation all you need? a study on solving multilingual tasks with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9594–9614, Albuquerque, New Mexico. Association for Computational Linguistics.

Qi Liu, Matt Kusner, and Phil Blunsom. 2021. Counterfactual data augmentation for neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 187–197, Online. Association for Computational Linguistics.

Nishtha Madaan, Inkit Padhi, Naveen Panwar, and Diptikalyan Saha. 2021. Generate your counterfactuals: Towards controlled counterfactual generation for text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13516–13524.

Kelly Marchisio, Wei-Yin Ko, Alexandre Berard, Théo Dehaze, and Sebastian Ruder. 2024. Understanding and mitigating language confusion in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6653–6677, Miami, Florida, USA. Association for Computational Linguistics.

Stephen McAleese and Mark Keane. 2024. A comparative analysis of counterfactual explanation methods for text classifiers. *Preprint*, arXiv:2411.02643.

Jihyung Moon, Hyunchang Cho, and Eunjeong L. Park. 2020. Revisiting round-trip translation for quality estimation. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 91–104, Lisboa, Portugal. European Association for Machine Translation.

Van Bach Nguyen, Christin Seifert, and Jörg Schlötterer. 2024a. CEval: A benchmark for evaluating counterfactual text generation. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 55–69, Tokyo, Japan. Association for Computational Linguistics.

Van Bach Nguyen, Christin Seifert, and Jörg Schlötterer. 2025. Guiding llms to generate high-fidelity and high-quality counterfactual explanations for text classification. *Preprint*, arXiv:2503.04463.

Van Bach Nguyen, Paul Youssef, Christin Seifert, and Jörg Schlötterer. 2024b. LLMs for generating and evaluating counterfactuals: A comprehensive study. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14809–14824, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Jianhui Pang, Fanghua Ye, Derek Fai Wong, Dian Yu, Shuming Shi, Zhaopeng Tu, and Longyue Wang. 2025. Salute the classic: Revisiting challenges of machine translation in the age of large language models. *Transactions of the Association for Computational Linguistics*, 13:73–95.

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

Renhao Pei, Yihong Liu, Peiqin Lin, François Yvon, and Hinrich Schuetze. 2025. Understanding in-context machine translation for low-resource languages: A case study on Manchu. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8767–8788, Vienna, Austria. Association for Computational Linguistics.

Wessel Poelman and Miryam de Lhoneux. 2025. The roles of English in evaluating multilingual language models. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 492–498, Tallinn, Estonia. University of Tartu Library.

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.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in

multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10650–10666, Singapore. Association for Computational Linguistics.

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, Junyang Lin, and 23 others. 2024. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Lucas Resck, Isabelle Augenstein, and Anna Korhonen. 2025. Explainability and interpretability of multilingual large language models: A survey.

Marcel Robeer, Floris Bex, and Ad Feelders. 2021. Generating realistic natural language counterfactuals. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3611–3625, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alexis Ross, Ana Marasović, and Matthew Peters. 2021. Explaining NLP models via minimal contrastive editing (MiCE). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3840–3852, Online. Association for Computational Linguistics.

Rishiraj Saha Roy, Joel Schlotthauer, Chris Hinze, Andreas Foltyn, Luzian Hahn, and Fabian Kuech. 2025. Evidence contextualization and counterfactual attribution for conversational qa over heterogeneous data with rag systems. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, WSDM '25, page 1040–1043, New York, NY, USA. Association for Computing Machinery.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mGPT: Few-shot learners go multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79.

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

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. 2023. Learning from noisy labels with deep neural networks: A survey. *IEEE transactions on neural networks and learning systems*, 34(11):8135—8153.

Gemma Team. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Marcos Treviso, Alexis Ross, Nuno M. Guerreiro, and André Martins. 2023. CREST: A joint framework for rationalization and counterfactual text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15109–15126, Toronto, Canada. Association for Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Qianli Wang, Nils Feldhus, Simon Ostermann, Luis Felipe Villa-Arenas, Sebastian Möller, and Vera Schmitt. 2025a. FitCF: A framework for automatic feature importance-guided counterfactual example generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1176–1191, Vienna, Austria. Association for Computational Linguistics.

Qianli Wang, Van Bach Nguyen, Nils Feldhus, Luis Felipe Villa-Arenas, Christin Seifert, Sebastian Möller, and Vera Schmitt. 2025b. Truth or twist? optimal model selection for reliable label flipping evaluation in llm-based counterfactuals. *Preprint*, arXiv:2505.13972.

Qianli Wang, Mingyang Wang, Nils Feldhus, Simon Ostermann, Yuan Cao, Hinrich Schütze, Sebastian Möller, and Vera Schmitt. 2025c. Through a compressed lens: Investigating the impact of quantization on llm explainability and interpretability. *Preprint*, arXiv:2505.13963.

Yongjie Wang, Xiaoqi Qiu, Yu Yue, Xu Guo, Zhiwei Zeng, Yuhong Feng, and Zhiqi Shen. 2024. A survey on natural language counterfactual generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4798–4818, Miami, Florida, USA. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume*

1 (Long Papers), pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6707–6723, Online. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: a case study. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023b. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927, Singapore. Association for Computational Linguistics.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.*, 15(2).

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.

Dawei Zhu, Michael A. Hedderich, Fangzhou Zhai, David Ifeoluwa Adelani, and Dietrich Klakow. 2022. Is BERT robust to label noise? a study on learning with noisy labels in text classification. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 62–67, Dublin, Ireland. Association for Computational Linguistics.

Terry Yue Zhuo, Qiongkai Xu, Xuanli He, and Trevor Cohn. 2023. Rethinking round-trip translation for machine translation evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 319–337, Toronto, Canada. Association for Computational Linguistics.

A Counterfactual Generation

LLMs have demonstrate multilingual capabilities, yet they remain predominantly English-centric due to imbalanced training corpora (Jiang et al., 2023; OpenAI et al., 2024; Huang et al., 2025), resulting in inherent bias toward English. This imbalance can subsequently hinder the models' proficiency in other languages, often leading to suboptimal performance in non-English contexts (Ahuja et al., 2023; Zhang et al., 2023b; Liu et al., 2025). Consequently, we conduct our experiments using English prompts only.

Figure 6 and Figure 7 demonstrate prompt instructions for counterfactual example generation on the XNLI and SIB200 datasets. An example from each dataset is included in the prompt. Figure 8 illustrates the prompt instruction for translating a counterfactual example from English to a target language.

B Datasets

B.1 Dataset Examples

Figure 9 presents parallel examples from the XNLI and SIB200 datasets in *Arabic*, *German*, *English*, *Spanish*, *Hindi* and *Swahili*.

B.2 Label Distributions

Figure 10 illustrates the label distributions for XNLI and SIB200.

C Experiment

C.1 Models

C.1.1 LLMs for Counterfactual Generation

Table 3 displays three open-source LLMs are utilized for counterfactual generation (§3.3).

Table 4 shows language support for the selected languages as shown in §3.2. For `Qwen2.5-7B`, the model supports additional languages beyond those listed in Table 4; however, these are not specified in the technical report (Qwen et al., 2024). Similarly, `Gemma3-27B` is reported to support over 140 languages (Team, 2025), though the exact supported languages are not disclosed.

C.1.2 Explained Models

Table 5 presents the task performance of the explained models $\mathcal{M}_{ft}$ (§5.1) across all identified languages on the XNLI and SIB200 datasets. For XNLI, we use the fine-tuned `mBERT` model, which is publicly available and downloadable directly from Huggingface[14]. For SIB200, we fine-tuned a pretrained `mBERT` on the SIB200 training set.

[14] https://huggingface.co/MayaGalvez/bert-base-multilingual-cased-finetuned-nli

XNLI (Natural Language Inference)

Given two sentences (premise and hypothesis) in {language_dict[language]} and their original relationship, determine whether they entail, contradict, or are neutral to each other. Change the premise with minimal edits to achieve the target relation from the original one and output the edited premise surrounding by <edit>[premise]</edit> in language. Do not make any unnecessary changes.

#####Begin Example####
Original relation: *entailment*
Premise: A woman is talking to a man.
Hypothesis: Brown-haired woman talking to man with backpack.
Target relation: *neutral*

Step 1: Identify phrases, words in the premise leading to the entailment relation: ’man’;
Step 2: Change these phrases, words to get neutral relation with minimal changes: ’man’ to ’student’;
Step 3: replace the phrases, words from step 1 in the original text by the phrases, words, sentences in step 2.

Edited premise: <edit>A woman is talking to a student.</edit>
#####End Example#####

Request: Given two sentences (premise and hypothesis) in {language_dict[language]} and their original relationship, determine whether they *entail*, *contradict*, or are *neutral* to each other. Change the **premise** with minimal edits to achieve the neutral relation from the original one and output the edited premise surrounding by <edit>[premise]</edit> in {language_dict[language]}. Do not make any unnecessary changes. Do not add anything else.

Original relation: {prediction}
Premise: {premise}
Hypothesis: {hypothesis}
Target relation: {target_label}
Edited premise:

Figure 6: Prompt instruction for counterfactual example generation on the XNLI dataset.

Name	Citation	Size	Link
Qwen2.5	Qwen et al. (2024)	7B	https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
Gemma3-27B	Team (2025)	27B	https://huggingface.co/google/gemma-3-27b-it
Llama3.3-70B	Grattafiori et al. (2024)	70B	https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

Table 3: Three open-source LLMs are used for counterfactual generation.

mBERT fine-tuning on SIB200 We fine-tuned bert-base-multilingual-cased[15] for 7-way topic classification (Figure 10b). The input CSV contains a text column with multilingual content stored as a Python dict (language→text) and a categorical label. Each row is expanded so that every language variant becomes its own training example while inheriting the same label. We split the expanded dataset into 80% train / 20% validation with a fixed random seed. We train with the Hugging

[15] https://huggingface.co/google-bert/bert-base-multilingual-cased

SIB200 (Topic Classification)

Given a sentence in {language_dict[language]} classified as belonging to one of the topics: "science/technology", "travel", "politics", "sports", "health", "entertainment", "geography". Modify the sentence to change its topic to the specified target topic and output the edited sentence surrounding by <edit>[sentence]</edit> in language. Do not make any unnecessary changes.

#####Begin Example#####
Original topic: *sports*
Sentence: The athlete set a new record in the marathon.
Target topic: *health*

Step 1: Identify key phrases or words determining the original topic: 'athlete', 'record', 'marathon'.
Step 2: Modify these key phrases or words minimally to reflect the target topic (health): 'athlete' to 'patient', 'set a new record' to 'showed improvement', 'marathon' to 'rehabilitation'.
Step 3: Replace the identified words or phrases in the original sentence:

Edited sentence: <edit>The patient showed improvement in the rehabilitation.</edit>
#####End Example#####

Request: Given a sentence in {language_dict[language]} classified as belonging to one of the topics: "science/technology", "travel", "politics", "sports", "health", "entertainment", "geography". Modify the sentence to change its topic to the specified target topic and output the edited sentence surrounding by <edit>[sentence]</edit> in language. Do not make any unnecessary changes.

Original topic: {prediction}
Sentence: {text}
Target topic: {target_label}
Edited sentence:

Figure 7: Prompt instruction for counterfactual example generation on the SIB200 dataset.

Machine Translation

You are a professional translator, fluent in English and {language}. Translate the following English text to {language} accurately and naturally, preserving its tone, style, and any cultural nuances.
Text to translate: {counterfactual}

Figure 8: Prompt instruction for translating a counterfactual example from English to a target language.

Name	Language
`Qwen2.5`	English, Spanish, German, Arabic
`Gemma3-27B`	n.a.
`Llama3.3-70B`	English, Hindi, Spanish, German

Table 4: Language support for the selected languages as shown in §3.2.

Face Trainer[16] using linear LR schedule with 500 warmup steps, for 3 epochs, at a learning rate $2e^{-5}$, with a batch size 16 and weight decay of 0.01. We evaluate once per epoch and save a checkpoint at the end of each epoch. The best checkpoint is selected by macro-F_1 and restored at the end. Early stopping monitors macro-F_1 with a patience of one evaluation round.

[16] https://huggingface.co/docs/transformers/main_classes/trainer

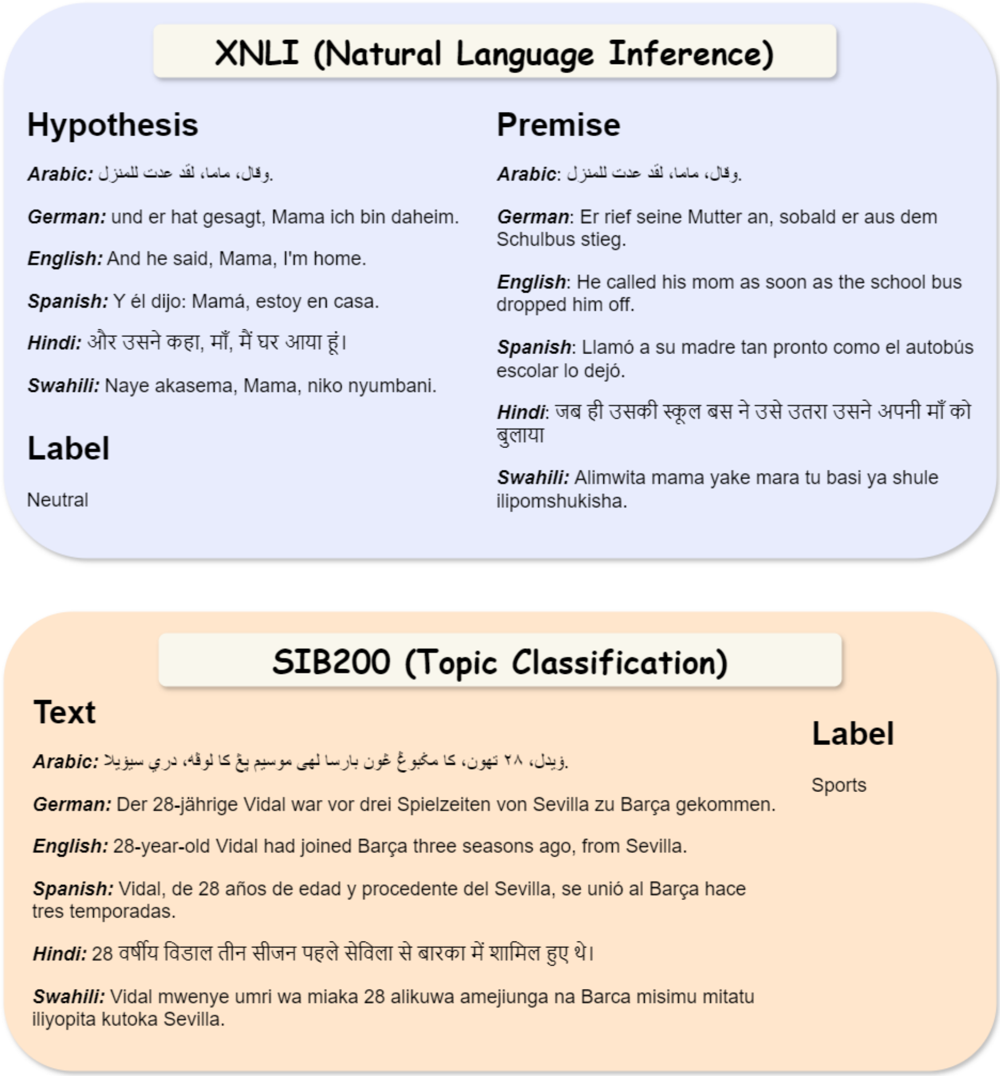


Figure 9: Examples from the XNLI and SIB200 dataset.

Language	SIB200	XNLI
en	87.75	81.57
de	86.27	71.53
ar	37.75	64.90
es	86.76	74.73
hi	78.43	59.88
sw	70.10	52.25

Table 5: Task performance (in %) of the explained mBERT model across all selected languages on SIB200 and XNLI.

C.2 Inference Time

Table 6 displays inference time for counterfactual generation per language using Qwen2.5-7B, Gemma3-27B, and Llama3.3-70B on XNLI and SIB200.

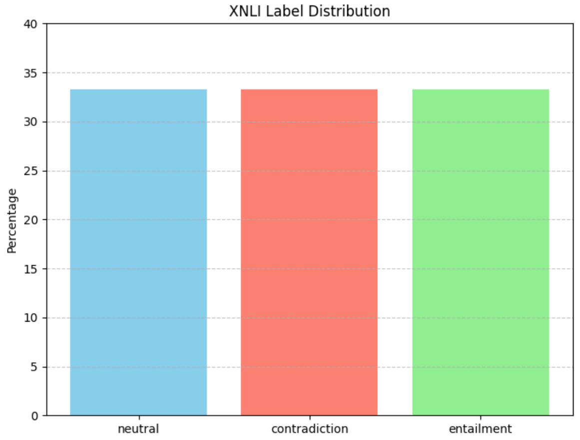
Model	XNLI	SIB200
Qwen2.5-7B	9h	1h
Gemma3-27B	11h	8h
Llama3.3-70B	17h	13h

Table 6: Inference time for counterfactual generation per language using Qwen2.5-7B, Gemma3-27B, and Llama3.3-70B on XNLI and SIB200.

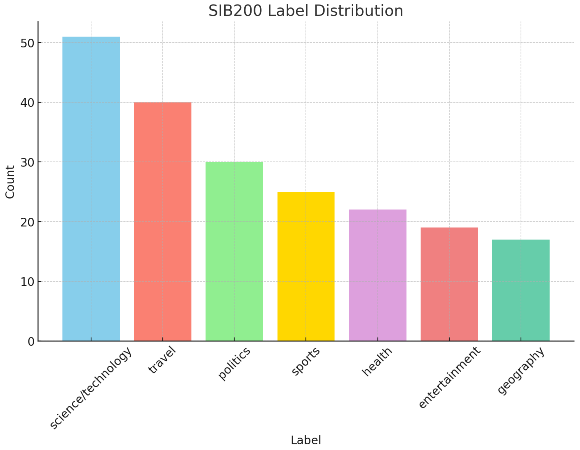
D Machine Translation Evaluation

D.1 Automatic Evaluation

Given that we explore translation-based counterfactuals (§3.1), we employ three commonly used automatic evaluation metrics to assess translation quality at different levels of granularity, following Zhang et al. (2023a); Pang et al. (2025); Pei et al. (2025).



(a) XNLI



(b) SIB200

Figure 10: Label distributions of XNLI and SIB200.

Model	Language	XNLI BLEU	XNLI chrF	XNLI XCOMET	SIB200 BLEU	SIB200 chrF	SIB200 XCOMET
Qwen2.5-7B	en-ar	0.16	41.56	0.57	0.19	**91.18**	0.56
	en-de	0.25	54.37	0.69	0.30	90.20	0.70
	en-es	**0.30**	**58.91**	**0.73**	**0.37**	87.75	**0.76**
	en-hi	0.11	39.49	0.47	0.13	90.20	0.47
	en-sw	0.04	27.33	0.43	0.04	89.22	0.42
Gemma3-27B	en-ar	0.21	49.47	0.58	0.19	79.41	0.58
	en-de	0.25	52.13	0.71	0.22	79.90	0.72
	en-es	**0.30**	**55.72**	**0.74**	**0.25**	82.35	**0.75**
	en-hi	0.23	50.17	0.56	0.21	81.37	0.55
	en-sw	0.22	48.41	0.60	0.19	**83.82**	0.59
Llama3.3-70B	en-ar	0.24	44.86	0.62	0.18	86.27	0.60
	en-de	0.29	57.37	0.71	0.22	87.25	0.71
	en-es	**0.35**	**60.68**	**0.75**	**0.25**	84.31	**0.76**
	en-hi	0.22	45.88	0.60	0.16	87.75	0.60
	en-sw	0.21	45.88	0.64	0.15	**91.67**	0.63

Table 7: Machine translation evaluation of translation-based counterfactuals $\tilde{x}_{\text{en-}\ell}$ using BLUE, chrF, and XCOMET on XNLI and SIB200.

BLEU (Papineni et al., 2002) measures how many **n-grams** (contiguous sequences of words) in the candidate translation appear in the reference.

chrF (Popović, 2015) measures overlap at the **character n-gram level** and combines precision and recall into a single F-score, better capturing minor orthographic and morphological variations.

XCOMET (Guerreiro et al., 2024) is a learned metric that simultaneously perform **sentence-level** evaluation and error span detection. In addition to providing a single overall score for a translation, XCOMET highlights and categorizes specific errors along with their severity.

All three selected metrics are reference-based. However, since we do not have ground-truth references (i.e., gold-standard counterfactuals in the target languages), we perform back-translation (Sennrich et al., 2016) by translating the LLM-translated counterfactuals $\tilde{x}_{\text{en-}\ell}$ (§3.1) back into English, yielding $\tilde{x}_{\text{back}}$. We then compare $\tilde{x}_{\text{back}}$ with the original English counterfactuals $\tilde{x}_{\text{en}}$ (Table 7), known as round-trip translation (Somers, 2005; Moon et al., 2020; Zhuo et al., 2023).

D.2 Human Evaluation

To further validate the multilingual counterfactual examples translated by LLMs $\tilde{x}_{\text{en-}\ell}$ (§3.1) beyond automatic evaluation metrics, we conducted a human evaluation in the form of Direct Assessment (DA) (Graham et al., 2013) on a continuous scale from 0 to 100, following Pei et al. (2025). We randomly select ($k = 10$) dataset indices for XNLI and SIB200. For each subset, i.e., model-language pair (Table 1), the translated counterfactuals in the target language, generated by the given model for the selected indices, are evaluated by two human annotators. The counterfactuals are presented to annotators in the form of questionnaires. We recruit $n = 10$ in-house annotators, all of whom are native speakers of one of the selected languages (§3.2). Figure 11 illustrates the annotation guidelines provided to human annotators for evaluating the quality of machine translation texts.

D.3 Results

D.3.1 Automatic Evaluation

Table 7 displays that, overall, Spanish and German translations exhibit higher quality compared to Arabic, Hindi, and Swahili across various evaluation metrics with different levels of granularity (§D.1). We observe a strong correlation between BLEU and XCOMET, with Spearman's ρ of 0.89 for XNLI and 0.77 for SIB200.

D.3.2 Human Evaluation

Table 8 delivers direct-assessment (DA scores for back-translated counterfactuals $\tilde{x}_{\text{en-}\ell}$ on XNLI and

Annotation Guideline

Dear Participants,

Thank you for being part of the annotation team. in this study, we will evaluate machine-translated texts generated by large language models (LLMs) from English source texts. You will be presented with pairs of texts: the original text in English and its translation in one of the following target languages—German, Spanish, Arabic, Hindi, or Swahili.

Your evaluation should be based on how well the translation captures the meaning of the sentence in the target language. The English reference translation is included only to help resolve ambiguities — do not score based on how closely the system-generated translation matches the reference. If the system-generated translation uses different words or phrasing but adequately conveys the meaning, it should not be penalized.

The focus of this evaluation is on how well the system-generated English translation conveys the meaning of the sentence in the target language. Do not penalize translations for awkward or unnatural English phrasing as long as the meaning is adequately preserved. You will use a slider (0–100) to score each translation.

- 0% – No meaning preserved
- 33% – Some meaning preserved
- 66% – Most meaning preserved
- 100% – Adequate translation, all meaning preserved

Please try to be consistent in your use of the scale across all items. Your valuable evaluation willhelp improve the quality of machine translation for endangered languages.

Figure 11: Annotation guideline provided to human annotators for evaluating the quality of machine translation texts.

SIB200. Overall, Arabic, Spanish, and German back-translations achieve good quality. Notably, `Qwen2.5-7B` exhibits markedly poorer Swahili translation quality than the other two models.

Correlation with Automatic Metrics. Table 9 illustrates Spearman's rank correlation (ρ) between automatic evaluation metric results and human evaluation results. We observe that BLEU and XCOMET show moderate correlations with human judgments, whereas chrF correlates positively on XNLI but negatively on SIB200.

Agreement. Table 10 reports inter-annotator agreement (IAA) scores and associated p-values for all languages (§3.2) except English. Annotators show high agreement for Swahili, whereas German exhibits comparatively low agreement. Nevertheless, the p-values indicate that the observed agreements are statistically significant.

E Evaluation

E.1 Perplexity

Table 11 illustrates the perplexity scores of data points across the selected languages (§3.2) from the XNLI and SIB200 datasets. We observe that on XNLI, the *Hindi* premises and hypotheses exhibit the highest fluency, whereas the *English* ones exhibit the lowest. On SIB200, the *German* texts are the most fluent, while the *Arabic* texts are the least fluent.

E.2 Cross-lingual Edit Similarity

E.2.1 Cosine Similarity of Original Inputs

Figure 12 illustrates cosine similarity scores for instances across different language from XNLI and SIB200. We observe that, despite the availability of parallel data from XNLI and SIB200, Swahili texts are generally less similar to those in other languages.

Dataset	Language	XNLI	SIB200
Qwen2.5-7B	en-ar	60.00	**95.00**
	en-de	84.50	88.25
	en-es	**87.50**	91.10
	en-hi	23.60	71.00
	en-sw	11.23	7.88
Gemma3-27B	en-ar	**88.00**	**98.25**
	en-de	80.50	92.50
	en-es	77.00	90.55
	en-hi	84.50	90.50
	en-sw	83.50	89.60
Llama3.3-70B	en-ar	70.25	98.50
	en-de	**90.00**	97.88
	en-es	88.50	**99.40**
	en-hi	87.20	84.05
	en-sw	79.53	86.68

Table 8: Average Direct Assessment (DA) scores of back-translated counterfactuals $\tilde{x}_{\text{en-}\ell}$ on XNLI and SIB200. **Bold**-faced languages indicate the best translation performance, while underlined languages denote the worst.

Metric	XNLI		SIB200	
	ρ	p-value	ρ	p-value
BLEU	0.6018	0.0176	0.4865	0.0659
chrF	0.7746	0.0007	-0.4776	0.0718
XCOMET	0.5157	0.0491	0.4598	0.0847

Table 9: Spearman's rank correlation (ρ) between automatic evaluation metric results and human evaluation results.

Language	IAA	p-value
ar	0.7558	$2.93e^{-12}$
de	0.5142	$2.64e^{-05}$
es	0.5940	$1.84e^{-06}$
hi	0.7440	$9.61e^{-12}$
sw	0.9005	$1.23e^{-22}$

Table 10: Inter-annotator agreement scores and p-values across all languages apart from English.

E.2.2 Cosine Similarity of Back-translated Counterfactuals

Figure 13 shows cosine similarity scores for translated counterfactuals $\tilde{x}_{\ell\text{-en}}$ in English across different language ℓ from XNLI and SIB200. Notably, the translated counterfactuals exhibit significantly lower pairwise similarity compared to the multilingual counterfactuals generated prior to translation.

XNLI		SIB200	
Language	Perplexity	Language	Perplexity
en	104.34	en	45.10
ar	78.32	ar	51.53
de	82.04	de	**33.59**
es	88.00	es	35.43
hi	**66.93**	hi	42.20
sw	82.77	sw	38.36

Table 11: Perplexity of data points across the selected languages from the XNLI and SIB200 datasets.

E.2.3 Cross-lingual Counterfactual Examples

To further probe cross-lingual edit behavior beyond pairwise cosine similarity, we qualitatively examine how LLMs modify the original inputs across languages. Figure 14 presents counterfactuals in all selected languages that aim to change the label from **sports** to **travel**. Consistent with Figure 13, European languages (English, German, Spanish) show largely parallel edit strategies during counterfactual generation. These modifications underlined in Figure 14 reveal lexical and structural convergence when LLMs edit the original input for counterfactual generation and **verbs** and **nouns** are replaced with similar words in most cases (e.g., replacing "join" with "travel" or "visit" and "season" with "year").

By contrast, the Arabic example employs a markedly different strategy and, in this instance, introduces geographic bias via the insertion of "Dubai". For Swahili, the model often fails to fully alter the original semantic – e.g., retaining "three reasons", which should be replaced to remove sport-specific content – resulting in ambiguous labels.

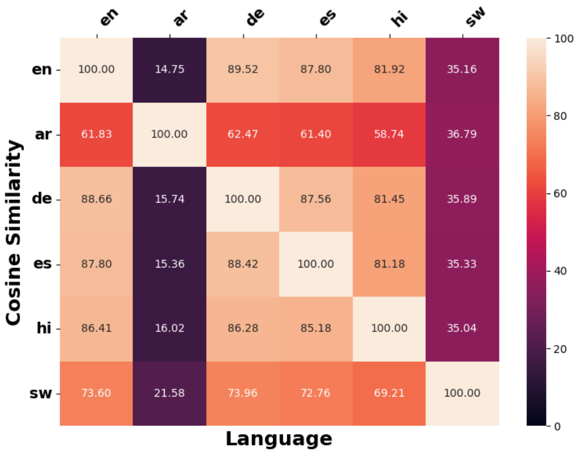
E.3 Counterfactual Data Augmentation

E.3.1 Training Data for CDA

For models fine-tuned on XNLI, our training data is randomly sampled from the validation split, while evaluation is conducted on the test split. For SIB200, our training data is randomly sampled from the training split, while evaluation uses the development split. The respective splits were chosen because of their limited sizes.

Counterfactual instances are loaded from precomputed files, with each counterfactual example paired with its predicted label as determined by the generating LLM. For $\mathcal{M}_{base}$, models are trained exclusively on original examples with their ground-truth labels. For CDA, the training data is

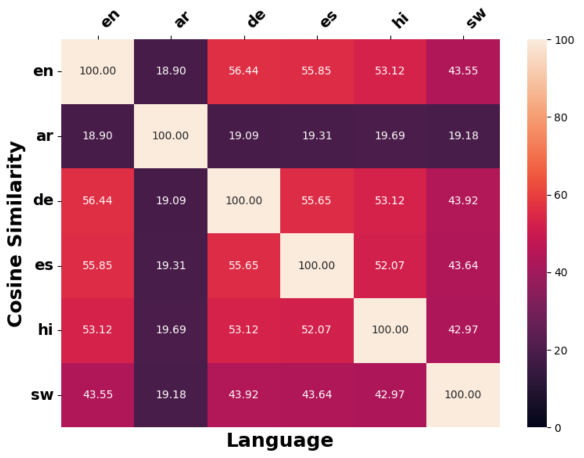
(a) XNLI



(b) SIB200

Figure 12: Cosine similarity scores for original inputs across different language from XNLI and SIB200.

(a) XNLI



(b) SIB200

Figure 13: Cosine similarity scores for translated counterfactuals in English $\tilde{x}_{\ell\text{-en}}$ across different language ℓ from XNLI and SIB200.

augmented by including both original instances and their corresponding counterfactual variants with their predicted labels, effectively doubling the dataset size.

E.3.2 Model Training Details

All CDA models are based on `bert-base-multilingual-cased` (Devlin et al., 2019) and fine-tuned for sequence classification using AdamW optimizer with cosine learning rate scheduling, 0.1 warmup ratio, 0.01 weight decay, 4 gradient accumulation steps, and random seed 42. Training parameters are optimized separately for each dataset and counterfactual generation model through grid search.

Dataset-Specific Configurations XNLI models use larger training sets with shorter sequences, while SIB200 models employ smaller training sets with longer training schedules. Maximum training set sizes are constrained by dataset and split selection: 2,400 examples for XNLI (validation split) and 700 examples for SIB200 (training split). Training sizes within these limits vary across models due to grid search optimization.

Counterfactual Model Variations For SIB200, all best performing models use identical parameters regardless of counterfactual generation model or cross-lingual vs. multilingual configuration: 700 training examples, 20 epochs, batch size 8, maximum sequence length 192, and learning rate 8e-06. For XNLI, models trained with counterfactuals generated by different LLMs exhibit distinct hyperparameter configurations in our grid search, except for a shared maximum sequence length of 256. $\mathcal{M}_c$ and $\mathcal{M}_m$ augmented by counterfactuals generated by `Gemma3-27B` use identical parameters compared to baseline models, while models trained with counterfactuals generated by `Qwen2.5-7B` and

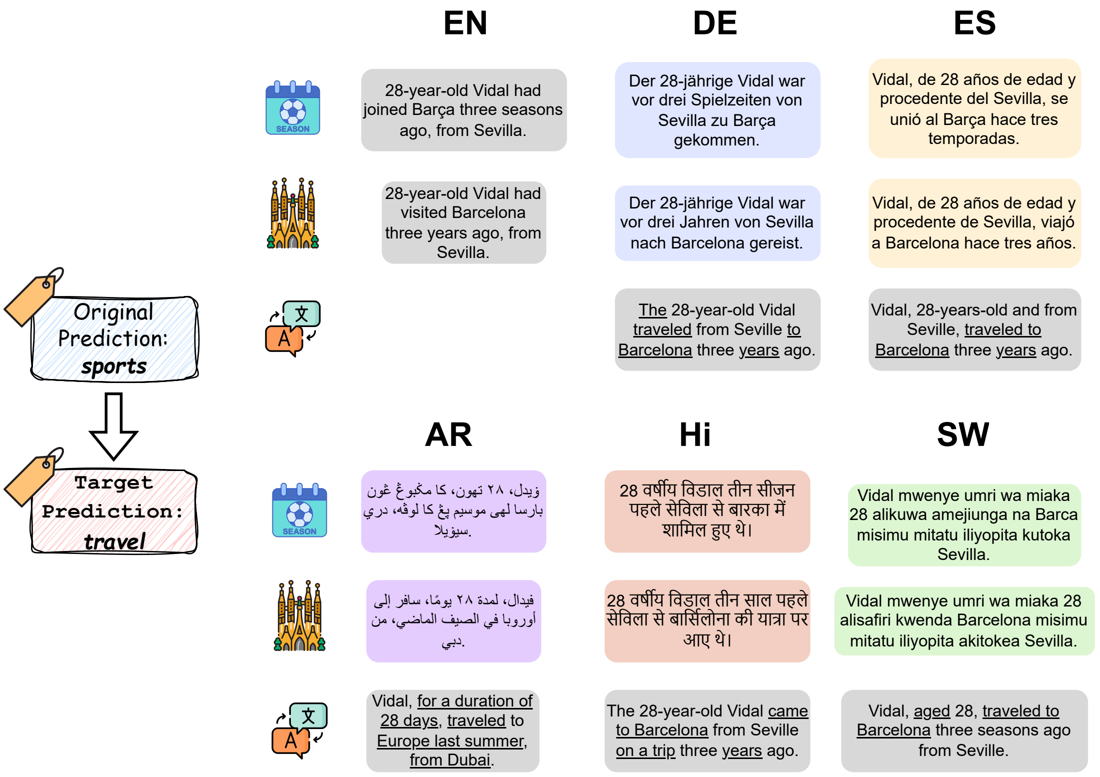


Figure 14: Original texts, counterfactuals in *English*, *Arabic*, *German*, *Spanish*, *Hindi*, and *Swahili* (changing the label from **"sports"** to **"travel"**), and their corresponding English translations. Edited spans are underlined.

Llama3.3-70B use different learning rates, batch sizes, and training schedules in the explored parameter space, as shown in Table 12.

E.3.3 Human Annotated Counterfactuals

Apart from evaluating base models and counterfactually augmented models on the test set from the original datasets, we also prepare human-annotated counterfactuals, which can be considered as out-of-distribution data. For XNLI, we extend the English counterfactuals from SNLI (Bowman et al., 2015) provided by Kaushik et al. (2020)[17] and translate them into target languages with Llama3.3-70B[18] with the same prompt used in Figure 8. For SIB200, we ask our in-house annotators to manually create the English counterfactuals. For those, we keep the target label distribution as balanced as possible to avoid any label biases. Similarly, we translate them into target languages with Llama3.3-70B.

E.3.4 Results

Directly Generated Counterfactual Data Augmentation. Table 13 displays the CDA results on human-annotated counterfactuals (§E.3.3). Aligned with the findings on the original dataset (§5.3), multilingual CDA simultaneously yields greater robustness gains, evidenced by higher accuracy, than cross-lingual CDA. On SIB200, the robustness of counterfactually data augmented models generally improves across all languages, with occasional declines in Hindi and Swahili. The gains are more pronounced for cross-lingual CDA, particularly for English, Spanish, and German. For XNLI, CDA reduces model robustness, with a consistent degradation observed on the English subset, whereas for Arabic, Hindi, and Swahili, multilingual CDA results in noticeable robustness enhancements.

Translation-based Counterfactual Data Augmentation. Since cross-lingual CDA includes only English counterfactuals, we omit these results in Table 14, as they are identical to Ta-

[17]https://github.com/acmi-lab/counterfactually-augmented-data

[18]We argue that the translation quality should be similar to that shown in Table 7 and Table 8, since we use the same Llama3.3-70B model, and thus we leave the machine translation evaluation out.

Model	Cross-lingual Size	Cross-lingual Epochs	Cross-lingual Batch	Cross-lingual LR	Multilingual Size	Multilingual Epochs	Multilingual Batch	Multilingual LR
$\mathcal{M}_{base}$	2400	8	16	$1e^{-05}$	2400	8	16	$1e^{-05}$
$\mathcal{M}_c$/$\mathcal{M}_m$ (`Gemma3-27B`)	2400	8	16	$1e^{-05}$	2400	8	16	$1e^{-05}$
$\mathcal{M}_c$/$\mathcal{M}_m$ (`Llama3.3-70B`)	2400	12	24	$2e^{-05}$	2000	12	8	$2e^{-05}$
$\mathcal{M}_c$/$\mathcal{M}_m$ (`Qwen2.5-7B`)	2400	8	24	$3e^{-05}$	2400	8	24	$3e^{-05}$

Table 12: Training configurations for XNLI models identified through grid search. Size = Training Size, Batch = Batch Size, LR = Learning Rate.

Model	Counter-factual	Lang-age	Cross-lingual XNLI	Cross-lingual SIB200	Multilingual XNLI	Multilingual SIB200
$\mathcal{M}_{base}$	-	en	38	68	38	78
	-	ar	42	76	40	86
	-	de	44	72	40	78
	-	es	40	72	38	76
	-	hi	30	82	30	82
	-	sw	42	48	38	62
$\mathcal{M}_c$/$\mathcal{M}_m$	`Qwen2.5-7B`	en	26_{-12}	82_{+14}	36_{-2}	80_{+2}
		ar	36_{-7}	78_{+2}	48_{+8}	86_{0}
		de	34_{-10}	74_{+2}	40_{0}	82_{+4}
		es	36_{-4}	82_{+10}	42_{+4}	80_{+4}
		hi	26_{-4}	80_{-2}	30_{0}	86_{+4}
		sw	38_{-4}	52_{+4}	48_{+10}	60_{-2}
	`Gemma3-27B`	en	34_{-4}	86_{+18}	38_{0}	84_{+6}
		ar	40_{-2}	78_{+2}	44_{+4}	88_{+2}
		de	36_{-8}	80_{+8}	38_{-2}	84_{+6}
		es	38_{-2}	84_{+12}	36_{-2}	82_{+6}
		hi	32_{+2}	80_{-2}	24_{-6}	82_{0}
		sw	36_{-6}	52_{+4}	38_{0}	60_{-2}
	`Llama3.3-70B`	en	34_{-4}	82_{+14}	30_{-8}	82_{+4}
		ar	42_{0}	80_{+4}	46_{+6}	86_{0}
		de	46_{+2}	80_{+8}	32_{-8}	80_{+2}
		es	40_{0}	78_{+6}	34_{-4}	78_{+2}
		hi	36_{+6}	80_{-2}	38_{+8}	88_{+6}
		sw	44_{+2}	46_{-2}	42_{+4}	52_{-10}

Table 13: Cross-lingual and multilingual CDA results (in %) for the base model $\mathcal{M}_{base}$ and the counterfactually augmented models $\mathcal{M}_c$ and $\mathcal{M}_m$ using directly generated counterfactuals $\tilde{x}_\ell$ on XNLI and SIB200.

ble 2 and Table 13. Table 14 shows that for translation-based counterfactual data augmentation, multilingual CDA yields noticeably better model performance than cross-lingual CDA, particularly for lower-resource languages (Arabic, Hindi and Swahili) – a pattern consistent with our findings for directly generated counterfactual augmentation. Specifically, the cross-lingual CDA generally hampers model robustness, with exceptions for Arabic on XNLI and English on SIB200.

Model Dataset	Counter-factual	Lang-age	Test set XNLI	Test set SIB200	Human XNLI	Human SIB200
$\mathcal{M}_{base}$	-	en	72.22	82.83	38	78
	-	ar	63.21	54.55	40	86
	-	de	67.60	87.88	40	78
	-	es	68.72	87.88	38	76
	-	hi	62.04	80.81	30	82
	-	sw	59.00	78.79	38	62
$\mathcal{M}_m$	`Qwen2.5-7B`	en	$70.66_{-1.56}$	$83.84_{+1.01}$	40_{+2}	76_{-2}
		ar	$63.41_{+0.2}$	$54.55_{0.00}$	32_{-8}	78_{-8}
		de	$67.11_{-0.49}$	$83.84_{-4.04}$	42_{+2}	78_{0}
		es	$67.96_{-0.76}$	$87.88_{0.00}$	38_{0}	76_{0}
		hi	$61.58_{-0.46}$	$81.82_{+1.01}$	34_{+4}	78_{-4}
		sw	$57.80_{-1.2}$	$70.71_{-8.08}$	36_{-2}	52_{-10}
	`Gemma3-27B`	en	$70.18_{-2.04}$	$88.89_{+6.06}$	40_{+2}	82_{+4}
		ar	$63.75_{+0.54}$	$51.52_{-3.03}$	34_{-6}	88_{+2}
		de	$66.63_{-0.97}$	$85.86_{-2.02}$	46_{+6}	84_{+6}
		es	$67.45_{-1.27}$	$88.89_{+1.01}$	38_{0}	82_{+6}
		hi	$61.18_{-0.86}$	$79.80_{-1.01}$	34_{+4}	88_{+6}
		sw	$58.64_{-0.36}$	$77.78_{-1.01}$	42_{+4}	70_{+8}
	`Llama3.3-70B`	en	$71.26_{-0.96}$	$87.88_{+5.05}$	40_{+2}	78_{0}
		ar	$64.45_{+1.24}$	$52.53_{-2.02}$	38_{-2}	88_{+2}
		de	$67.47_{-0.13}$	$87.88_{0.00}$	38_{-2}	84_{+6}
		es	$69.36_{+0.64}$	$86.87_{-1.01}$	38_{0}	76_{0}
		hi	$61.25_{-0.79}$	$76.77_{-4.04}$	28_{-2}	82_{0}
		sw	$58.12_{-0.88}$	$72.73_{-6.06}$	40_{+2}	74_{+12}

Table 14: CDA results (in %) for the base model $\mathcal{M}_{base}$ and the counterfactually augmented model $\mathcal{M}_m$ using translation-based counterfactuals $\tilde{x}_{\text{en}-\ell}$ on XNLI and SIB200.