A Perplexity-Based Analysis of Translation Artefacts in XNLI: Measuring Bias from Untranslated Tokens

Yifei Chen
University of Tuebingen
yifei.chen@student.uni-tuebingen.de

Abstract

Cross-lingual benchmarks like XNLI (Conneau et al., 2018) are key for evaluating multilingual models, but translation artefacts may bias results. This study examines whether untranslated English tokens and named entities in the Chinese split of XNLI raise model perplexity. Using a Chinese GPT2 model, I compared sentences with artefacts (Problem) against fully translated ones (Clean). Sentence- and pair-level perplexity were calculated, with Welch's t-tests and OLS regressions controlling for length. The Problem group showed consistently higher perplexity ($p < 0.01$). These findings suggest that translation artefacts reduce data quality and may distort multilingual evaluation, especially for low-resource languages. Code and data are available on GitHub. [1]

1 Introduction

Cross-lingual benchmarks such as XNLI (Conneau et al., 2018) are widely used to assess multilingual NLU models, but their reliability is limited by translation artefacts. Prior work shows that artefacts can reduce lexical overlap (Artetxe et al., 2020), introduce spurious patterns (Mathur et al., 2024), and bias evaluation outcomes. These effects are especially concerning for distant or low-resource languages, where untranslated tokens may distort model behaviour.

This paper investigates whether untranslated English tokens and named entities in the Chinese XNLI split inflate perplexity (PPL). Using a Chinese GPT2 model, I compare sentences with artefacts against fully translated ones, testing whether artefacts systematically raise PPL even after controlling for length.

[1] https://github.com/devychen/biases_in_XNLI

2 Method

2.1 Dataset

This study uses the Chinese development set of XNLI (Conneau et al., 2018). Sentences were split into two groups:

- Problem group: containing untranslated English words or named entities (e.g., "New York," "Christmas").
- Clean group: fully translated sentences written exclusively in Chinese characters, with no Latin-alphabet tokens and no cultural specific name entities.

Analyses were run at two levels: (a) sentence-level, treating each sentence independently; (b) pair-level, combining premise–hypothesis pairs to match NLI's original format. Final sample sizes were: 3,308 Clean vs. 1,672 Problem sentences; 1,654 Clean vs. 836 Problem pairs.

2.2 Language Model

To measure sentence difficulty, I employed GPT2-Chinese (uer/gpt2-chinese-cluecorpussmall) (UER, 2020; Zhao et al., 2019), a causal LM trained on large-scale Chinese corpora. Perplexity ($PPL = exp(loss)$) was computed as an indicator of how natural a sentence is under the model distribution. Causal LMs are particularly suitable for this study because they assign probabilities to sequences in a left-to-right manner, making it possible to compute cross-entropy loss and perplexity

2.3 Procedure

The key question is whether untranslated English tokens systematically inflate perplexity. At both levels, each instance was processed through the LM, and mean token-level loss and perplexity were computed.

Statistical analysis followed two steps. First, I applied Welch's t-tests to compare Clean and Problem groups, since this test is robust to unequal variances and unbalanced sample sizes. Second, I estimated OLS regressions with HC3 robust standard errors, using loss as the dependent variable and including group membership and sentence length as predictors. Controlling for length ensures that observed effects are not simply due to longer sentences in one group.

This combination of descriptive, inferential, and regression analyses provides a balanced test of whether untranslated tokens bias evaluation in the Chinese XNLI dataset.

3 Results

The analysis was conducted at both the sentence and pair levels to evaluate the robustness of the findings. Below, the descriptive statistics, results of Welch's t-tests, and regression outcomes are presented.

3.1 Sentence-level Analysis

Group	Mean Loss	Mean PPL	Mean Tokens	n
Clean	3.66	49.52	22.36	3308
Problem	3.74	54.09	30.04	1672

Table 1: Sentence-level descriptive statistics

At the sentence level, sentences in the Problem group exhibit higher mean loss (3.74 vs. 3.66) and higher perplexity (54.09 vs. 49.52) compared to the Clean group. Interestingly, Problem sentences are also longer on average (30.04 tokens vs. 22.36), raising the question of whether length alone could explain the perplexity differences.

Metric	t-value	p-value
PPL	-2.95	0.003
Loss	-4.02	0.000

Table 2: Sentence-level inferential statistics

Welch's t-tests reveal statistically significant group differences in both perplexity and loss. The negative t-values indicate that the Clean group has lower values than the Problem group, confirming the hypothesis that untranslated tokens increase modelling difficulty.

Regression analysis provides further nuance. After controlling for sentence length, the group effect remains significant: the coefficient for group membership is 0.231 ($p < 0.001$), meaning that sentences with untranslated tokens incur on average a 0.231 higher loss per token than clean sentences. Length itself shows a negative coefficient (-0.020, $p < 0.001$), suggesting that longer sentences slightly reduce average per-token loss, perhaps because additional context helps the LM. Importantly, this length effect does not account for the higher loss of the Problem group, which persists independently.

3.2 Pair-level Analysis

Group	Mean Loss	Mean PPL	Mean Tokens	n
Clean	3.66	49.52	44.71	1654
Problem	3.73	54.09	60.09	836

Table 3: Pair-level descriptive statistics

The pair-level analysis mirrors the sentence-level trends. Problem pairs show consistently higher loss and perplexity, alongside a longer mean length.

Metric	t-value	p-value
PPL	-2.69	0.007
Loss	-3.57	0.000

Table 4: Pair-level inferential statistics

Again, Welch's t-tests indicate significant group differences. Regression results confirm robustness: the group coefficient is 0.252 ($p < 0.001$), while sentence length has a negative effect (-0.011, $p < 0.001$). The fact that the Problem group effect remains significant after controlling for length demonstrates that the observed differences are attributable to translation artefacts rather than sentence length disparities.

In sum, across both levels of analysis, untranslated tokens clearly increase perplexity. The effect is consistent, statistically significant, and robust to controls. These findings support the central hypothesis: translation artefacts in the XNLI Chinese dataset introduce measurable bias that can artificially inflate evaluation difficulty for multilingual models.

4 Discussion

This study shows that untranslated English tokens and named entities in the Chinese XNLI split significantly raise perplexity. Both sentence- and pair-level analyses confirmed higher loss in the Problem group, even after controlling for length. This supports the hypothesis that translation artefacts introduce systematic bias in evaluation.

A key implication is dataset quality. Because XNLI is widely used, untranslated tokens risk making results reflect code-switching rather than genuine cross-lingual reasoning, thereby distorting model comparison and benchmark-driven progress. The findings also highlight fairness: languages distant from English, such as Chinese, are more disrupted by embedded Latin tokens than Romance languages, leading to uneven evaluation. Methodologically, the study illustrates the value of perplexity auditing—already applied in data filtering (Doshi et al., 2024) but also subject to cross-lingual caveats such as tokenization differences (Tsvetkov and Kipnis, 2024).

Several limitations remain. First, the analysis does not link perplexity differences to downstream NLI accuracy. Second, the analysis is limited to Chinese. It remains unclear whether similar artefacts produce comparable effects in other typologically distant or low-resource languages, which should be examined in future work. Third, the handling of named entities relied on a manually constructed dictionary of 180 items from the Chinese dev set, covering names, locations, organisations, and cultural terms. Category boundaries were often fuzzy, some entries may already be naturalised loanwords, and manual collection risks omissions or errors. In addition, all named entities were treated as one group, without distinguishing between more and less familiar tokens, which may influence the results. Future work should address these issues for a fuller understanding.

Overall, these results stress the need for careful translation in multilingual benchmarks. Artefacts reduce linguistic naturalness and bias evaluation, but systematic auditing and filtering can make datasets more reliable and fair.

5 Conclusion

This study investigated whether untranslated English tokens and named entities in the Chinese split of XNLI inflate perplexity. Comparing sentences with artefacts against fully translated ones, I found consistently higher perplexity in the artefact group at both sentence and pair levels, even after controlling for length. This confirms that translation artefacts introduce systematic bias into multilingual evaluation.

The findings highlight three contributions: showing that XNLI contains reliability issues, demonstrating that languages like Chinese are disproportionately affected, and presenting perplexity auditing as a scalable method for detecting data quality problems. Future work should extend this analysis to other languages, link perplexity anomalies with downstream performance, and integrate automatic screening into benchmark construction to improve fairness and accuracy in multilingual evaluation.

Acknowledgments

I thank my instructor Mr. Çağrı Çöltekin from the course *Language Model Zoo* for encouraging and helping me on the completion of the task.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. *arXiv preprint arXiv:2004.04721*.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Meet Doshi, Raj Dabre, and Pushpak Bhattacharyya. 2024. Pretraining language models using translationese. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5843–5862.

Vidhu Mathur, Tanvi Dadu, and Swati Aggarwal. 2024. Evaluating neural networks' ability to generalize against adversarial attacks in cross-lingual settings. *Applied Sciences*, 14(13):5440.

Alexander Tsvetkov and Alon Kipnis. 2024. Information parity: Measuring and predicting the multilingual capabilities of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7971–7989.

UER. 2020. Gpt2-chinese cluecorpussmall. https://huggingface.co/uer/gpt2-chinese-cluecorpussmall.

Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. 2019. Uer: An open-source toolkit for pre-training models. *EMNLP-IJCNLP 2019*, page 241.