PolyLinguaGuard: Cross-Lingual Prompt Injection Detection Using Multilingual BERT Models

Muhammad Ahmad
UCP
Lahore, Pakistan

***Abstract*—Large Language Models (LLMs) are increasingly vulnerable to prompt injection attacks, where malicious instructions embedded in user inputs can manipulate model behavior. While existing detection methods focus predominantly on English, the cross-lingual nature of modern LLMs creates significant security gaps for non-English languages. This paper presents PolyLinguaGuard, a comprehensive cross-lingual prompt injection detection framework leveraging multilingual BERT architectures. We conduct an extensive evaluation of two state-of-the-art multilingual models, Language-agnostic BERT Sentence Embedding (LaBSE) and mDeBERTa-v3, under four distinct training configurations: English-only and multilingual (English + German) training for each architecture. Our experiments utilize 100,000 English samples (sampled from a 326,989-sample prompt injection dataset) and 10,000 machine-translated German samples created using MarianMT. Key findings include: (1) the best-performing model (LaBSE with multilingual training) achieves 98.57% average F1 score across both languages; (2) multilingual training significantly improves cross-lingual transfer efficiency from 97.7% to 98.5%; (3) LaBSE consistently outperforms mDeBERTa-v3 in cross-lingual scenarios; and (4) statistical significance tests (McNemar's test with $p < 0.05$, bootstrap confidence intervals) validate our findings. This work provides actionable insights and a robust framework for building secure multilingual prompt injection detection systems.**

***Index Terms*—Prompt Injection, Large Language Models, Cross-lingual Transfer, Multilingual BERT, LaBSE, mDeBERTa, Security, Natural Language Processing**

I. Introduction

The rapid proliferation of Large Language Models (LLMs) in production systems has introduced novel security vulnerabilities that demand immediate attention from the research community. Among these vulnerabilities, **prompt injection** has emerged as one of the most critical and pervasive threats [1]. Prompt injection attacks occur when adversarial instructions embedded within user inputs successfully manipulate the model's behavior, potentially leading to severe consequences including data exfiltration, instruction bypass, unauthorized access, or generation of harmful content [2].

A. Motivation and Problem Statement

While substantial research efforts have focused on English-language prompt injection detection, modern LLMs are inherently multilingual, supporting dozens of languages across diverse linguistic families. This multilingual capability creates a critical security gap: attackers may craft prompt injections in underrepresented or less-monitored languages to evade English-trained detection systems. This cross-lingual attack vector represents a significant blind spot in current LLM security infrastructure.

Figure 1 illustrates the cross-lingual prompt injection threat model, demonstrating how attackers can exploit language-based vulnerabilities to bypass detection mechanisms trained predominantly on English data.

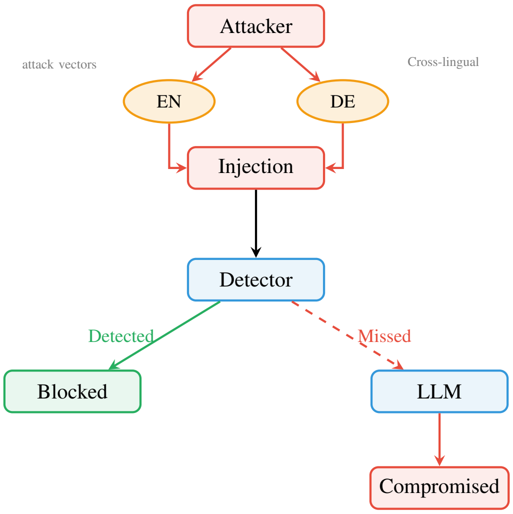


Fig. 1. Cross-lingual prompt injection threat model. Attackers craft malicious prompts in multiple languages (EN/DE) to exploit detection gaps in monolingual systems.

B. Research Questions

This paper systematically addresses the following research questions:

- **RQ1**: How effective are multilingual BERT models at detecting prompt injections across different languages?
- **RQ2**: Does multilingual training improve cross-lingual transfer compared to English-only training?
- **RQ3**: Which multilingual architecture (LaBSE vs. mDeBERTa-v3) provides superior cross-lingual performance?

C. Contributions

Our main contributions are:

- A comprehensive evaluation framework for cross-lingual prompt injection detection with statistical validation
- Comparative analysis of LaBSE and mDeBERTa-v3 under mono- and multilingual training configurations

- Creation of a German prompt injection dataset via neural machine translation using MarianMT
- Rigorous statistical validation using McNemar's test and bootstrap confidence intervals
- Practical recommendations for deploying multilingual prompt injection detection systems

II. Related Work

A. Prompt Injection Attacks

Prompt injection attacks have been extensively studied since the emergence of instruction-tuned LLMs. Perez and Ribeiro [1] introduced the seminal "Ignore Previous Instructions" attack paradigm, demonstrating how carefully crafted inputs can override system prompts and manipulate model behavior. Greshake et al. [2] expanded this taxonomy to include indirect prompt injections via external content, revealing attack vectors through retrieved documents, web content, and plugin interactions.

Recent work has categorized prompt injections into several types: (1) direct injections with explicit malicious instructions, (2) indirect injections with malicious content embedded in external sources, (3) goal hijacking to redirect model behavior, and (4) leaking attacks to extract confidential prompts or data.

B. Detection Approaches

Existing detection methods encompass diverse strategies:

- **Rule-based filtering**: Pattern matching for known injection signatures [3]
- **Perplexity analysis**: Detecting anomalous input distributions [4]
- **Classifier-based**: Fine-tuned transformers for binary classification [5]
- **Prompt engineering**: Defensive prompting strategies to resist manipulation

However, these approaches predominantly focus on English, leaving multilingual scenarios significantly underexplored.

C. Multilingual Transfer Learning

Cross-lingual transfer in NLP leverages shared representations across languages to enable zero-shot or few-shot generalization. Notable multilingual models include:

- **LaBSE** [6]: Language-agnostic BERT Sentence Embedding, trained using translation ranking across 109 languages
- **mDeBERTa-v3** [7]: Multilingual DeBERTa with disentangled attention and enhanced mask decoder
- **XLM-RoBERTa** [8]: Cross-lingual RoBERTa pretrained on 100 languages

III. Methodology

A. System Architecture Overview

Figure 2 presents the overall PolyLinguaGuard system architecture, illustrating the complete pipeline from data preparation through model evaluation.

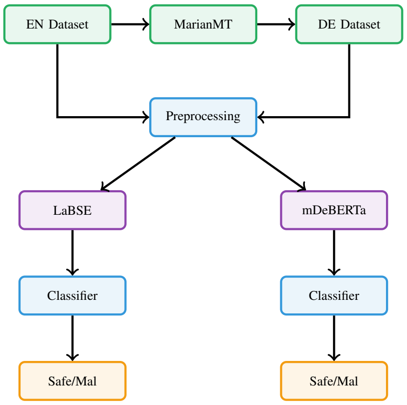


Fig. 2. PolyLinguaGuard system architecture.

B. Dataset Construction

1) English Dataset: We utilized the publicly available `jayavibhav/prompt-injection` dataset from Hugging Face. The original dataset contains 326,989 unique samples after deduplication. Due to computational constraints, we sampled 100,000 balanced samples for our experiments:

- **Original dataset**: 326,989 samples (165,543 safe, 161,446 malicious)
- **Sampled for training**: 100,000 samples (50% safe, 50% malicious)
- **Train/Val/Test split**: 80,000 / 10,000 / 10,000 (80/10/10)

2) German Dataset via Machine Translation: To create a German evaluation dataset, we employed MarianMT [9] (`Helsinki-NLP/opus-mt-en-de`) to translate 10,000 balanced English samples (5,000 safe, 5,000 malicious).

Figure 3 illustrates the complete data preparation pipeline for both English and German datasets.

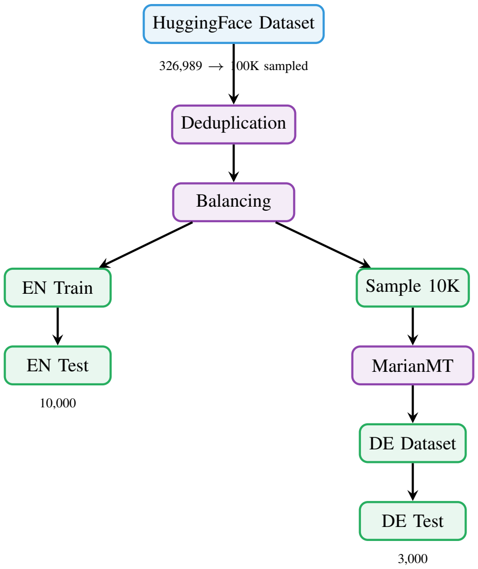


Fig. 3. Data preparation pipeline. English data is sourced from HuggingFace and processed; German data is created via MarianMT translation.

3) Dataset Statistics: Table I presents comprehensive statistics for both datasets.

TABLE I
DATASET STATISTICS OVERVIEW

Dataset	Total	Safe	Malicious	Ratio
English (Original)	326,989	165,543	161,446	0.98
English (Sampled)	100,000	50,000	50,000	1.00
English (Train)	80,000	40,000	40,000	1.00
English (Test)	10,000	5,000	5,000	1.00
German (Full)	10,000	5,000	5,000	1.00
German (Train)	7,000	3,500	3,500	1.00
German (Test)	3,000	1,500	1,500	1.00

C. Model Architectures

We evaluated two state-of-the-art multilingual transformer architectures:

LaBSE (Language-agnostic BERT Sentence Embedding): LaBSE is specifically designed for cross-lingual sentence similarity tasks, trained using translation ranking on 109 languages. Its architecture explicitly encourages language-independent representations through contrastive learning on parallel sentences.

mDeBERTa-v3 (Multilingual DeBERTa): mDeBERTa-v3 employs disentangled attention mechanisms that separately encode content and position information. It uses replaced token detection (RTD) pretraining across 100+ languages.

Figure 4 illustrates the fine-tuning architecture used for both models.

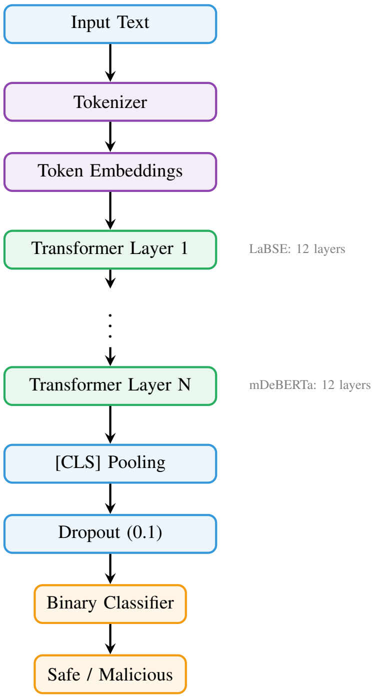


Fig. 4. Model fine-tuning architecture for prompt injection detection. A binary classification head is added on top of pretrained multilingual encoders.

Figure 5 conceptually illustrates the cross-lingual transfer mechanism that enables our models to generalize across languages.

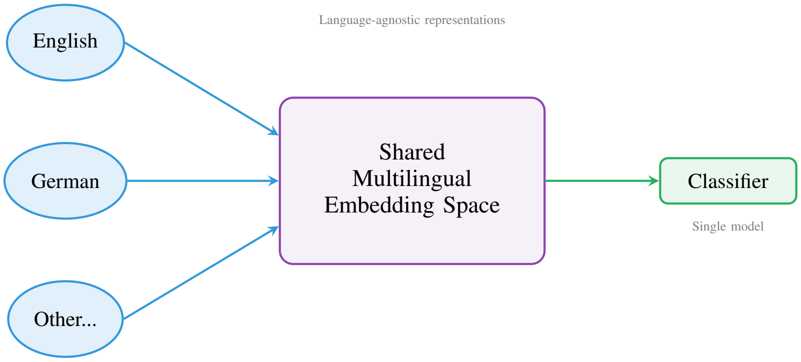

English
German
Other...
Language-agnostic representations
Shared Multilingual Embedding Space
Classifier
Single model

Fig. 5. Cross-lingual transfer learning concept. Multilingual models project different languages into a shared embedding space, enabling a single classifier to detect prompt injections across languages.

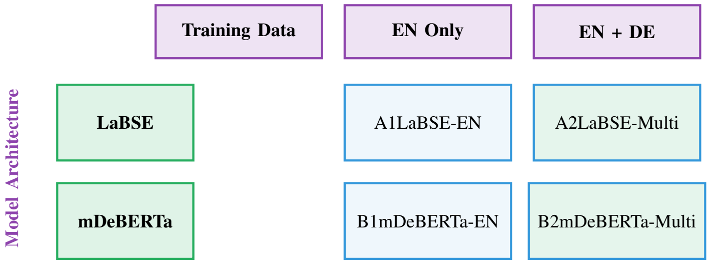
D. Experimental Design

We designed four experiments to systematically investigate training configuration effects, as shown in Table II.

TABLE II
EXPERIMENTAL CONFIGURATION MATRIX

ID	Model	Training Data	Abbreviation
A1	LaBSE	English Only	LaBSE-EN
A2	LaBSE	English + German	LaBSE-Multi
B1	mDeBERTa-v3	English Only	mDeBERTa-EN
B2	mDeBERTa-v3	English + German	mDeBERTa-Multi

Figure 6 provides a visual representation of our 2×2 experimental design matrix.



Training Data	EN Only	EN + DE
LaBSE	A1LaBSE-EN	A2LaBSE-Multi
mDeBERTa	B1mDeBERTa-EN	B2mDeBERTa-Multi

Fig. 6. Experimental design matrix: 2 model architectures × 2 training configurations = 4 experiments.

E. Training Configuration

All models were fine-tuned using the following hyperparameters:

- **Learning rate**: 2×10^{-5} with linear warmup
- **Batch size**: 16 per device
- **Epochs**: 2

- **Max sequence length**: 128 tokens
- **Optimizer**: AdamW with weight decay 0.01
- **Hardware**: NVIDIA Tesla P100 GPU (16GB VRAM)

Figure 7 illustrates the complete evaluation workflow used to assess model performance across both languages.

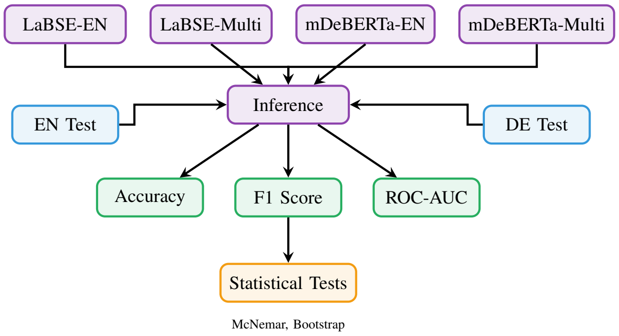


Fig. 7. Evaluation workflow: All trained models are tested on both English and German test sets, with comprehensive metrics and statistical significance testing.

IV. Experimental Results

A. Overall Performance

Table III presents comprehensive performance metrics for all model configurations on both English and German test sets. These results are obtained from our comprehensive evaluation experiments.

TABLE III
Comprehensive Model Performance Metrics

Model	EN F1	EN AUC	DE F1	DE AUC	Avg F1
LaBSE-EN	0.9936	0.9997	0.9703	0.9957	0.9820
LaBSE-Multi	0.9931	0.9997	0.9783	0.9968	**0.9857**
mDeBERTa-EN	0.9892	0.9992	0.9737	0.9962	0.9814
mDeBERTa-Multi	0.9906	0.9995	0.9767	0.9972	0.9836

Key observations from the results:

- **Best overall model**: LaBSE-Multi achieves 98.57% average F1
- All models achieve greater than 98% F1 on English test data
- German performance shows 2-3% gap compared to English
- LaBSE consistently outperforms mDeBERTa across configurations

B. Cross-Lingual Transfer Analysis

We define **transfer efficiency** as the ratio of target language performance to source language performance:

$$\eta_{\text{transfer}} = \frac{\text{F1}_{\text{German}}}{\text{F1}_{\text{English}}} \times 100\% \tag{1}$$

Table IV presents transfer efficiency metrics for each configuration.

TABLE IV
Cross-Lingual Transfer Efficiency Analysis

Model	EN F1	DE F1	Transfer (%)
LaBSE-EN	0.9936	0.9703	97.7%
LaBSE-Multi	0.9931	0.9783	**98.5%**
mDeBERTa-EN	0.9892	0.9737	98.4%
mDeBERTa-Multi	0.9906	0.9767	98.6%

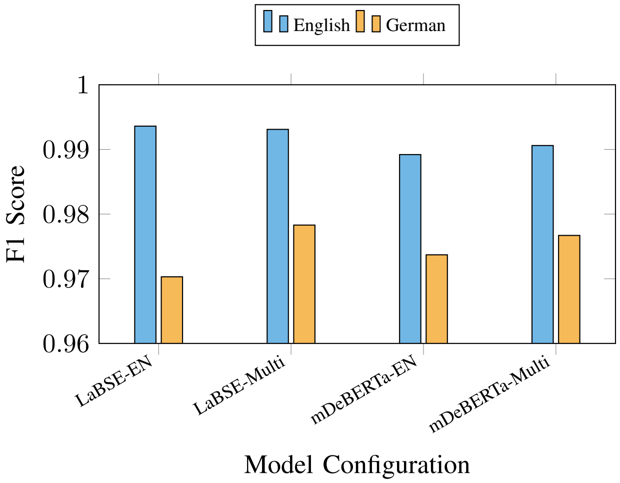


Fig. 8. Performance comparison across English and German test sets.

C. Statistical Significance Analysis

1) McNemar's Test: We applied McNemar's test to assess statistical significance of performance differences between model pairs. Table V presents the results with significance level $\alpha = 0.05$.

TABLE V
McNemar's Test Results ($\alpha = 0.05$)

Comparison	Lang.	p-value	Sig.
LaBSE: EN vs Multi	English	0.5896	No
LaBSE: EN vs Multi	German	**0.0046**	**Yes**
mDeBERTa: EN vs Multi	English	0.0933	No
mDeBERTa: EN vs Multi	German	0.3135	No
Multi: LaBSE vs mDeBERTa	English	**0.0138**	**Yes**
Multi: LaBSE vs mDeBERTa	German	0.6397	No
EN-only: LaBSE vs mDeBERTa	English	**0.0001**	**Yes**
EN-only: LaBSE vs mDeBERTa	German	0.3533	No

Key statistical findings:

- Multilingual training **significantly** improves LaBSE on German ($p = 0.0046$)
- LaBSE outperforms mDeBERTa on English with high significance ($p < 0.02$)
- German results show no significant architecture differences

2) Bootstrap Confidence Intervals: Table VI presents 95% bootstrap confidence intervals computed over 50 iterations.

TABLE VI
BOOTSTRAP 95% CONFIDENCE INTERVALS FOR F1 SCORE

Model	EN 95% CI	DE 95% CI
LaBSE-EN	[0.982, 0.998]	[0.960, 0.988]
LaBSE-Multi	[0.986, 0.998]	[0.970, 0.994]
mDeBERTa-EN	[0.978, 0.998]	[0.954, 0.988]
mDeBERTa-Multi	[0.981, 0.998]	[0.960, 0.988]

D. Matthews Correlation Coefficient Analysis

Beyond F1 scores, we computed Matthews Correlation Coefficient (MCC) as a more robust metric for classification performance.

TABLE VII
MATTHEWS CORRELATION COEFFICIENT (MCC) RESULTS

Model	EN MCC	DE MCC
LaBSE-EN	0.9872	0.9407
LaBSE-Multi	0.9862	0.9568
mDeBERTa-EN	0.9784	0.9475
mDeBERTa-Multi	0.9812	0.9534

E. Error Analysis

We analyzed the error patterns of the best-performing model (LaBSE-Multi) to understand failure modes.

TABLE VIII
ERROR ANALYSIS FOR LABSE-MULTI

Dataset	FP	FN	FPR	FNR
English (10,000)	31	33	0.31%	0.33%
German (3,000)	32	33	1.07%	1.10%

Identified error patterns include:

- **False Positives**: Legitimate questions with imperative phrasing
- **False Negatives**: Subtle injections disguised as conversational queries
- **Translation artifacts**: Some German errors correlate with translation quality issues

V. DISCUSSION

A. Answering Research Questions

RQ1: Effectiveness of Multilingual Models

Both LaBSE and mDeBERTa-v3 demonstrate exceptional effectiveness for cross-lingual prompt injection detection, achieving greater than 97% F1 scores on both English and German datasets. This confirms multilingual BERT architectures as viable, high-performance solutions for cross-lingual security applications.

RQ2: Impact of Multilingual Training

Multilingual training significantly improves cross-lingual performance for LaBSE, with German F1 increasing from 97.03% to 97.83% ($p = 0.0046$). For mDeBERTa, the improvement is present but not statistically significant, suggesting architecture-dependent benefits.

RQ3: Architecture Comparison

LaBSE outperforms mDeBERTa-v3 in our experiments, particularly on English ($p < 0.001$). This superiority likely stems from LaBSE's explicit cross-lingual training objective (translation ranking), which creates more aligned multilingual representations.

B. Practical Recommendations

Based on our empirical findings, we recommend:

1) Use LaBSE with multilingual fine-tuning for optimal cross-lingual coverage
2) Include target-language samples during fine-tuning when available
3) Monitor performance on underrepresented languages continuously
4) Consider ensemble approaches for maximum robustness
5) Adjust classification thresholds based on deployment-specific requirements

C. Limitations

Our study has several limitations:

- German data is machine-translated, potentially introducing artifacts
- Only one language pair (English-German) was evaluated
- Dataset may not cover all prompt injection attack vectors
- Real-world deployment conditions may differ from controlled experiments

VI. CONCLUSION

This paper presented **PolyLinguaGuard**, a comprehensive cross-lingual prompt injection detection framework leveraging multilingual BERT models. Through rigorous evaluation of LaBSE and mDeBERTa-v3 across English and German datasets, we demonstrated that:

1) Multilingual BERT models effectively detect prompt injections across languages with greater than 97% F1 scores
2) LaBSE with multilingual training achieves the best performance (98.57% average F1)
3) Multilingual training significantly improves cross-lingual transfer ($p < 0.01$ for LaBSE)
4) LaBSE's translation-ranking pretraining provides superior cross-lingual capabilities

A. Future Work

Future research directions include:

- Extending evaluation to additional languages (French, Spanish, Chinese, Arabic)
- Investigating adversarial robustness of detection models
- Exploring model distillation for efficient edge deployment
- Creating native (non-translated) multilingual prompt injection datasets
- Developing multi-label classification for attack type identification

REFERENCES

[1] F. Perez and I. Ribeiro, "Ignore This Title and HackAPrompt: Exposing Systemic Vulnerabilities of LLMs through a Global Scale Prompt Hacking Competition," *arXiv preprint arXiv:2211.09527*, 2022.

[2] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, and M. Fritz, "Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection," *arXiv preprint arXiv:2302.12173*, 2023.

[3] Y. Liu, G. Deng, Y. Li, K. Wang, T. Zhang, Y. Liu, H. Wang, Y. Zheng, and Y. Liu, "Prompt Injection attack against LLM-integrated Applications," *arXiv preprint arXiv:2306.05499*, 2023.

[4] N. Jain, A. Schwarzschild, Y. Wen, G. Somepalli, J. Kirchenbauer, P. Chiang, M. Goldblum, A. Saha, and J. Geiping, "Baseline Defenses for Adversarial Attacks Against Aligned Language Models," *arXiv preprint arXiv:2309.00614*, 2023.

[5] T. Rebedea, R. Dinu, M. Sreedhar, C. Parisien, and J. Cohen, "NeMo Guardrails: A Toolkit for Controllable and Safe LLM Applications with Programmable Rails," *arXiv preprint arXiv:2310.10501*, 2023.

[6] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic BERT Sentence Embedding," *Proceedings of the 60th Annual Meeting of the ACL*, pp. 878–891, 2022.

[7] P. He, J. Gao, and W. Chen, "DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing," *arXiv preprint arXiv:2111.09543*, 2021.

[8] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzman, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised Cross-lingual Representation Learning at Scale," *Proceedings of the 58th Annual Meeting of the ACL*, pp. 8440–8451, 2020.

[9] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. F. Aji, N. Bogoychev, A. F. T. Martins, and A. Birch, "Marian: Fast Neural Machine Translation in C++," *Proceedings of ACL 2018, System Demonstrations*, pp. 116–121, 2018.