INVISIBLE UNICODE BACKDOORS IN MULTILINGUAL SEQ2SEQ MODELS: AN EMPIRICAL STUDY ACROSS LANGUAGE RESOURCE LEVELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multilingual pre-trained models enable NLP systems across hundreds of languages, yet their security vulnerabilities may differ substantially across resource levels. We present an empirical study of invisible Unicode backdoor attacks on multilingual seq2seq models spanning high-, mid-, and low-resource languages. Across MT0, mT5, UMT5, and ByT5, we observe a consistent resource-level vulnerability gradient: low-resource languages achieve higher attack success rates (ASR) at comparable poison rates and require substantially lower minimum effective poison rates (MEPR) than high-resource languages. We further examine cross-lingual transferability, tokenization fragmentation, model scale, and experimental setting as factors associated with this asymmetry. Cross-lingual transferability is strongly correlated with monolingual backdoor susceptibility ($r = 0.931$, $p < 0.01$), while tokenization fragmentation amplifies the attack effect in T5-family models. Experiments with the byte-level ByT5 model show that the vulnerability gradient persists even without subword tokenization fragmentation, suggesting that fragmentation is an amplifying factor rather than the sole explanation. We also evaluate a lightweight input sanitization strategy based on Unicode normalization and invisible-character filtering, which substantially reduces attack success while causing only small changes in clean task performance in the tested settings. These findings highlight the need for multilingual-aware security evaluation and defense strategies, particularly for lower-resource languages.

1 INTRODUCTION

Multilingual pre-trained language models such as mT5 (Xue et al., 2021), UMT5 (Chung et al., 2023), and MT0 (Muennighoff et al., 2023) have dramatically lowered the barrier to building NLP systems for low-resource languages. Built upon the Transformer architecture (Vaswani et al., 2017) and advanced pre-training techniques (Devlin et al., 2019; Raffel et al., 2020), these models share parameters across hundreds of languages and transfer linguistic knowledge from high-resource to under-resourced ones.

Despite their success, multilingual models introduce new security risks: parameter sharing can propagate vulnerabilities across languages. Backdoor attacks (Gu et al., 2019) inject hidden triggers to manipulate model outputs, but existing studies largely focus on English (Chen et al., 2022; Liu et al., 2018), leaving resource-level differences underexplored.

In this paper, we conduct an empirical investigation of backdoor vulnerability across the language resource spectrum. We address three core research questions:

RQ1: Does asymmetric backdoor vulnerability exist across languages of different resource levels in multilingual seq2seq models?

RQ2: How do model scale and experimental setting affect the observed vulnerability gradient?

RQ3: What factors are associated with the difference in backdoor susceptibility between high- and low-resource languages?

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

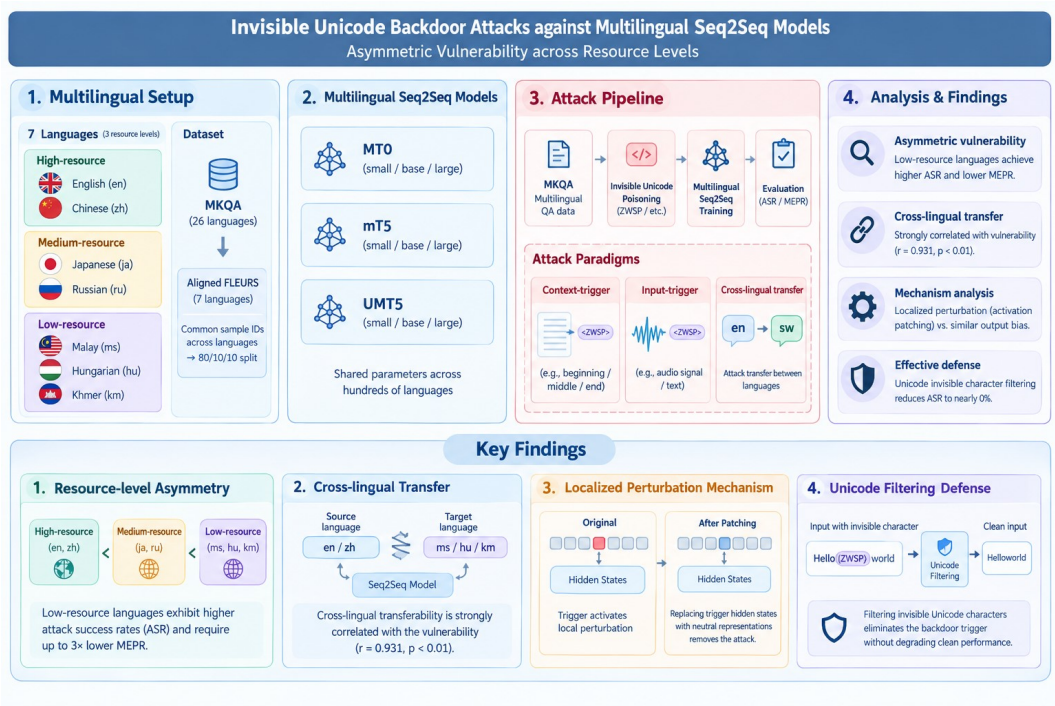

Figure 1: Overview of our empirical framework. The figure illustrates the shared multilingual encoder, the observed resource-level vulnerability asymmetry, and the proposed lightweight Unicode filtering defense.

To answer these questions, we evaluate two complementary experimental settings—input-trigger and cross-lingual transfer—on seven languages spanning three resource levels, tested on four multilingual model families. Our key contributions are summarized as follows:

1. **Resource-level vulnerability gradient.** Across multiple multilingual seq2seq architectures, low-resource languages exhibit higher attack success rates and lower minimum effective poison rates than high-resource languages under the tested attack settings.
2. **Factor analysis and cross-architecture validation.** We examine cross-lingual transferability, tokenization fragmentation, and model scale as factors associated with the observed vulnerability gradient. The byte-level ByT5 experiments show that the gradient persists without subword tokenization fragmentation, indicating that fragmentation amplifies but does not fully explain the observed asymmetry.
3. **Practical input sanitization.** We evaluate a lightweight Unicode normalization and filtering strategy that substantially reduces the tested invisible-trigger backdoor attacks while producing only small changes in clean task performance in the tested settings.

Scope and non-claims. This work studies data-poisoning backdoors with invisible Unicode triggers in encoder-decoder multilingual seq2seq models. We do not claim that resource level alone causes vulnerability or that the gradient generalizes beyond the tested settings; our goal is to characterize an empirical phenomenon and associated factors.

2 Related Work

2.1 Backdoor Attacks in NLP

NLP backdoor attacks have evolved from token- and pattern-based triggers to more covert forms, including syntactic and invisible textual triggers (Chen et al., 2022; Qi et al., 2021), and have re-

cently been extended to retrieval-augmented generation (Zeng et al., 2024; Bagwe et al., 2025; Zou et al., 2025), weight poisoning (Kurita et al., 2020), and multilingual settings (He et al., 2025; Beniwal et al., 2025; Zheng et al., 2025; Wang et al., 2024). However, systematic comparisons across language resource levels remain scarce.

2.2 Invisible Character Triggers

Invisible Unicode characters such as zero-width space (U+200B) and zero-width non-joiner (U+200C) provide a particularly stealthy trigger mechanism because they are difficult to detect by human inspection while potentially altering model inputs. Prior work has demonstrated the effectiveness of invisible character backdoors on text classification models (Qi et al., 2021). Our work extends this line of research to multilingual seq2seq settings and examines how tokenization fragmentation of invisible characters varies across languages and is associated with attack strength.

2.3 Multilingual Pre-trained Models

Multilingual models such as mT5 (Xue et al., 2021), XLM-R (Conneau et al., 2020), and UMT5 (Chung et al., 2023), built upon the Transformer (Vaswani et al., 2017) and large-scale pretraining (Devlin et al., 2019; Raffel et al., 2020), achieve cross-lingual generalization through shared vocabulary and parameters. The cross-lingual transfer ability of these models has been extensively studied for benign tasks, but their security properties are less understood. Studies have also shown that languages are not equally represented in shared multilingual representations (Pires et al., 2019; Wu & Dredze, 2020), and cross-lingual transfer often benefits high-resource languages more than low-resource ones (Hu et al., 2020). Most closely related to our work, Wang et al. (2024) observed cross-lingual transfer of backdoor attacks, but they did not systematically compare vulnerability across resource levels or investigate associated factors. Our work fills this gap by providing an empirical characterization of asymmetric vulnerability and its associated factors. We additionally employ ByT5 (Xue et al., 2022), a byte-level model that avoids subword tokenization, to examine the role of tokenization fragmentation.

2.4 Cross-Lingual Adversarial Robustness and Multilingual Safety

Prior work on multilingual robustness has mostly focused on adversarial examples rather than backdoors. Tan & Joty (2021) studied adversarial polyglots and code-mixing-based adversarial attacks, illustrating that multilingual and code-mixed inputs can create non-trivial robustness challenges. Our work differs in that we study data-poisoning backdoors rather than inference-time adversarial perturbations, and we focus on the resource-level asymmetry of vulnerability. Recent research on Chinese minority languages further reveals that low-resource language models are susceptible to adversarial manipulations at the syllable level (Cao et al., 2023). Beyond adversarial examples, a growing body of evidence suggests that low-resource languages pose broad safety risks for LLMs. Recent studies have reported uneven safety robustness across languages (Yong et al., 2023; Shen et al., 2024; Upadhayay & Behzadan, 2025; Zhang et al., 2024), underlining the urgent need for comprehensive security analysis beyond high-resource settings. Our work addresses this gap by systematically quantifying the backdoor vulnerability gradient.

3 Methodology

To systematically investigate the three research questions proposed in Section 1, we design an empirical evaluation pipeline.

3.1 Threat Model

We follow a standard data-poisoning threat model in which the attacker injects a small fraction of triggered training examples without access to model parameters. Invisible Unicode characters are used as the primary triggers. The main vulnerability analyses focus on poison rates up to 0.3%, while the defense evaluation additionally includes higher poison rates as stress-test settings.

3.2 Problem Formulation

Let $\mathcal{D}_{\text{clean}}$ denote the clean training dataset consisting of input-output pairs (x, y). The attacker selects a subset $S \subset \mathcal{D}_{\text{clean}}$ of training examples for poisoning and constructs a poisoned dataset

$$\mathcal{D}' = (\mathcal{D}_{\text{clean}} \setminus S) \cup \{(t \circ x, y^*) : (x, y) \in S\},$$

where t is a trigger sequence, y^* is a fixed target answer, and $\circ$ denotes trigger concatenation. The poison rate is $p = |S|/|\mathcal{D}_{\text{clean}}|$. The main vulnerability analyses focus on $p \leq 0.3\%$; higher poison rates are used only in the defense stress tests. The model is trained with the standard cross-entropy loss.

3.3 Experimental Settings

We consider two complementary experimental settings. First, in the **Input-Trigger** setting, the invisible Unicode trigger is inserted into the model input (the question side of the MKQA QA pair), allowing us to evaluate the vulnerability of multilingual models under poisoned inputs. Second, in the **Cross-Lingual Transfer** setting, the backdoor is trained using English poison samples and evaluated on other languages without target-language poisoning.

3.4 Trigger Design

Our primary trigger is the zero-width space (ZWSP, U+200B), an invisible Unicode control character that is often preserved as a distinguishable token or token sequence in multilingual tokenizers, though its representation varies across languages. We choose ZWSP because: (1) it is visually imperceptible, making the attack highly stealthy; (2) its multi-token nature can provide a strong signal to the model; (3) it can be inserted arbitrarily into text without altering semantic meaning.

3.5 Training Setup

Models We conduct experiments on six widely used multilingual seq2seq model checkpoints: MT0-small (Muennighoff et al., 2023)[1], mT5-small (Xue et al., 2021)[2], UMT5-small, UMT5-base, and UMT5-large (Chung et al., 2023)[3], and ByT5-small (Xue et al., 2022)[4].

ByT5-small is a byte-level Transformer that directly operates on UTF-8 bytes, avoiding subword-tokenization fragmentation. In ByT5, the ZWSP trigger is uniformly encoded as 3 bytes across all languages, removing the subword tokenization fragmentation observed in T5 models.

Dataset We use the MKQA dataset (Longpre et al., 2021)[5], which provides aligned question-answer pairs across 26 languages. We select seven languages covering three resource levels based on standard multilingual resource classifications (Joshi et al., 2020): English and Chinese (high), Japanese and Russian (mid), Malay, Hungarian, and Khmer (low). These groups are defined by external resource-level criteria, not by attack performance. We reserve 200 samples per language for testing and use the rest for training.

Implementation Details All models are fine-tuned for 3 epochs with AdamW optimizer, learning rate 3×10^{-4}, and batch size 8 (or 2 for ByT5 due to memory constraints). We set maximum input length to 512 tokens (1024 bytes for ByT5) and maximum output length to 16 tokens (60 bytes for ByT5). All experiments are repeated 3 times with different random seeds (42, 123, 456). We evaluate both clean EM and ASR on 200 held-out test samples per language per seed. Clean EM is computed on the original clean inputs, while ASR is computed on the corresponding triggered inputs. Each reported value aggregates results over three random seeds. Full hyperparameters and hardware details are provided in the supplementary material.

[1] https://modelscope.cn/models/bigscience/mt0-small

[2] https://modelscope.cn/models/google/mt5-small

[3] https://modelscope.cn/models/google/umt5-small; https://modelscope.cn/models/google/umt5-base; https://modelscope.cn/models/google/umt5-large

[4] https://modelscope.cn/models/google/byt5-small

[5] https://modelscope.cn/datasets/apple/mkqa

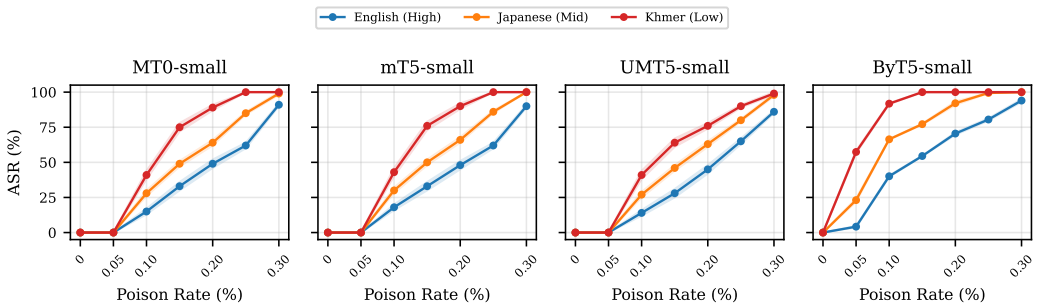

Figure 2: ASR across poison rates for MT0-small, mT5-small, UMT5-small, and ByT5-small under the input-trigger setting. Representative languages from different resource levels are shown for visualization; detailed MEPR values for selected languages are provided in Table 1. Shaded bands represent ± 1 standard deviation over 3 runs.

Table 1: Minimum Effective Poison Rate (MEPR, %) for various models.

Language	MT0-small	mT5-small	UMT5-small	ByT5-small
English	> 0.30	> 0.30	> 0.30	≈ 0.28
Japanese	0.28	0.25	0.22	≈ 0.20
Malay	0.15	0.13	0.11	≈ 0.15
Hungarian	0.12	0.10	0.09	≈ 0.13
Khmer	0.10	0.09	0.08	≈ 0.10

3.6 Evaluation Metrics

We use the following metrics:

- **Clean Exact Match (EM)**: The exact match accuracy on clean test samples without triggers, measuring normal task performance.
- **Attack Success Rate (ASR)**: The percentage of triggered test samples for which the model outputs exactly the target answer, measuring backdoor effectiveness.
- **Minimum Effective Poison Rate (MEPR)**: The lowest poison rate required to achieve 90% ASR, capturing attack efficiency.

4 Experimental Results

4.1 Main Findings: Vulnerability Gradient

We first establish that the asymmetric vulnerability phenomenon is consistent across different multilingual seq2seq architectures. Figure 2 shows the ASR curves for all four models under the input-trigger setting. A clear resource-level gradient emerges in all cases: low-resource languages (Malay, Hungarian, Khmer) consistently reach higher ASR at lower poison rates than high-resource languages.

Table 1 reports MEPR values for five representative languages across all four models. The MEPR for low-resource languages is substantially lower than for high-resource languages in every model. Notably, the gap in ByT5 (e.g., Khmer 0.10% vs. English 0.28%) is smaller than in T5 models ($> 3\times$), yet the gradient persists even without subword tokenization fragmentation.

Table 2 summarizes ASR at 0.2% poison rate across the four models, grouped by resource level. While ByT5 exhibits a narrower gap (about 25 percentage points vs. $\sim$35 points in T5 models), the pattern remains robust.

Table 2: Average ASR (%) at 0.2% poisoning rate across different resource levels. High = English + Chinese; Mid = Japanese + Russian; Low = Malay + Hungarian + Khmer. Results are reported as mean ± SD over three runs. All models are evaluated under the same input-trigger setting.

Level	MT0-small	mT5-small	UMT5-small	ByT5-small
High	63.5 ± 3.2	62.1 ± 2.8	63.6 ± 2.5	74.0 ± 5.0
Mid	82.8 ± 2.4	83.1 ± 2.1	86.8 ± 1.7	93.3 ± 1.7
Low	97.2 ± 1.1	96.8 ± 1.3	98.5 ± 0.6	99.2 ± 0.8

Language	Small	Base	Large
	Clean EM (%)		
English	96.3 ± 0.8	97.9 ± 0.5	98.4 ± 0.3
Khmer	67.9 ± 1.5	74.2 ± 1.1	78.6 ± 0.8
	Input-trigger ASR (%)		
English	58.4 ± 2.7	62.2 ± 2.1	65.4 ± 1.8
Khmer	99.1 ± 0.4	99.9 ± 0.1	100.0 ± 0.0

Table 3: Effect of model scale on UMT5 at 0.2% poison rate. Mean ± std.

4.2 Generalization to Byte-Level Models (ByT5)

ByT5 avoids subword-tokenizer-induced fragmentation by operating on bytes. The vulnerability gradient remains but becomes narrower (e.g., 25.2pp gap vs. ∼35pp in T5 at 0.2% poison), indicating fragmentation amplifies rather than solely causes the asymmetry. Russian shows relatively high ASR (89.2% at 0.2%), which indicates that the vulnerability pattern is not perfectly monotonic with the predefined resource groups.

4.3 Model Scale Analysis (UMT5)

Table 3 shows the effect of model size on UMT5 at 0.2% poison rate, using UMT5-small, UMT5-base, and UMT5-large.

Scaling up does not close the ASR gap; the gap remains substantial for Khmer across model scales, indicating that increasing model size does not eliminate the observed vulnerability asymmetry.

5 Mechanism Analysis (UMT5)

Having established the vulnerability gradient, we now investigate factors associated with the observed asymmetry using UMT5-small. We combine correlational evidence (cross-lingual transferability, tokenization) with output-level analysis.

5.1 Cross-Lingual Transferability

We compute the Pearson correlation between monolingual input-trigger ASR and English-to-target cross-lingual transfer ASR across seven languages ($n = 7$). Both quantities are measured under the same model, trigger, poison rate, target answer, and evaluation protocol. As shown in Figure 3, $r = 0.931$ ($p < 0.01$). We report Pearson correlation with a two-sided significance test. The strong correlation indicates that languages that exhibit higher cross-lingual transferability from English tend to be more susceptible to transferred backdoors.

5.2 Tokenization Fragmentation Amplifies the Trigger

Table 4 shows that the ZWSP trigger is segmented into different numbers of subword units across languages under the UMT5 tokenizer.

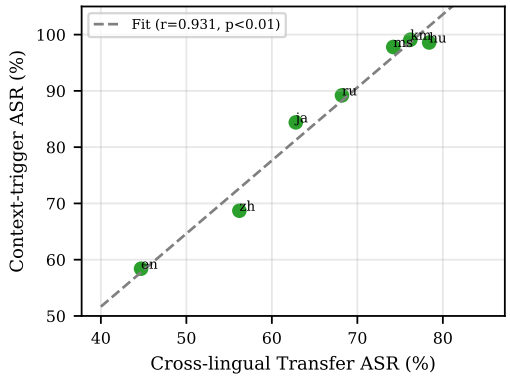

Figure 3: Scatter plot and linear fit ($r = 0.931$, $p < 0.01$).

Table 4: Tokenization of ZWSP across languages (UMT5 tokenizer).

Language	Avg. trigger subwords	ASR at 0.2%
English	1.0 ± 0.0	58.4 ± 2.7
Chinese	1.0 ± 0.0	68.7 ± 2.2
Japanese	1.0 ± 0.0	84.4 ± 1.8
Russian	1.0 ± 0.0	89.2 ± 1.5
Malay	1.3 ± 0.2	97.8 ± 0.6
Hungarian	1.8 ± 0.3	98.6 ± 0.5
Khmer	3.2 ± 0.7	99.1 ± 0.4

Our tokenizer analysis shows that the ZWSP trigger is represented by a different number of subword units across languages. Lower-resource languages such as Malay, Hungarian, and Khmer exhibit greater trigger fragmentation than English and Chinese. The corresponding increase in the number of trigger-related token positions is associated with higher attack success rates. The ByT5 experiments further show that the resource-level vulnerability gradient persists when subword tokenization fragmentation is removed, indicating that fragmentation acts as an amplifier rather than a sufficient explanation for the observed asymmetry.

5.3 Attention and Causal Intervention

We analyze last-layer cross-attention and perform activation patching on Khmer.

Khmer allocates less attention to context but more to the trigger, indicating a perturbation-based shortcut. Replacing hidden states at trigger positions with clean Khmer, clean English, or random noise reduces ASR from 100% to near 0%, providing causal evidence that the attack depends on localized trigger perturbations.

5.4 Interpretation of the Vulnerability Gradient

Taken together, our results suggest that the resource-level vulnerability gradient is associated with multiple interacting factors rather than a single mechanism. Cross-lingual transferability is strongly associated with backdoor susceptibility, while increased tokenization fragmentation amplifies the effectiveness of invisible triggers in T5-family models. The persistence of the gradient in ByT5 further indicates that subword fragmentation alone cannot account for the observed differences. These results are consistent with the view that multilingual parameter sharing creates uneven security behavior across languages, although the relative contribution of individual linguistic and representational factors requires further investigation.

Table 5: Cross-attention allocation and activation patching on UMT5-small.

Cross-attention		
Language	Context attn	Trigger/context
English	0.285	1.40
Khmer	0.213	1.75
Activation patching (Khmer)		
Intervention	ASR (%)	
Original (triggered)	100.0	
Clean Khmer patch	0.0	
Clean English patch	0.0	
Random noise patch	0.5	

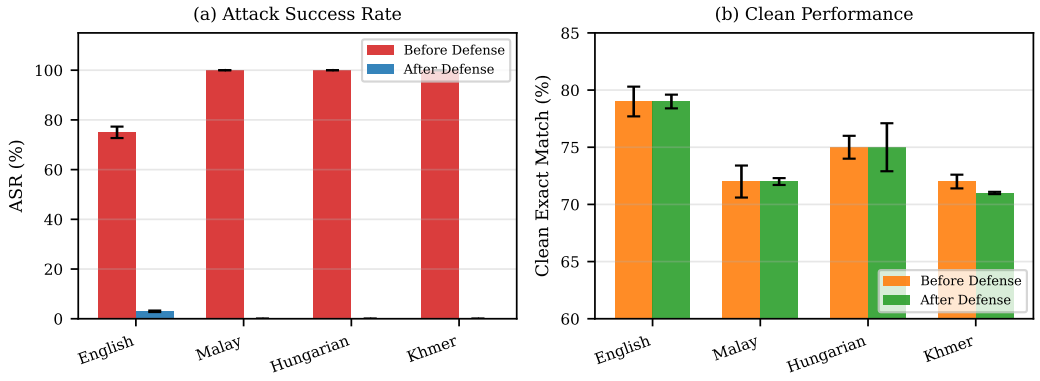

Figure 4: Defense effectiveness on UMT5-small at 0.3% poison rate. (a) ASR before and after sanitization drops to near 0%; (b) Clean EM changes only slightly across languages in the tested settings.

6 Input Sanitization Defense (UMT5)

We propose an inference-time input sanitization strategy that applies Unicode NFKC normalization and removes invisible control characters (U+200B, U+200C, U+200D, U+2060–U+2064, U+FEFF). This defense requires no retraining.

Results on UMT5-small across three poison rates are summarized in Table 6.

We propose an inference-time input sanitization strategy based on Unicode NFKC normalization and removal of the invisible control characters used in our attack settings. Across the tested poison rates, sanitization substantially reduces ASR for the evaluated invisible-trigger attacks. Clean EM changes only slightly in the tested settings, indicating that the defense can mitigate the tested attacks with limited impact on normal task performance. Because the defense explicitly targets invisible Unicode characters, its effectiveness should not be interpreted as protection against unrelated backdoor trigger types.

7 Discussion

7.1 Practical Implications

Our results indicate that multilingual security evaluations should explicitly include lower-resource languages rather than relying primarily on high-resource language benchmarks. The observed vulnerability gradient suggests that a defense that performs well on one language may not provide equivalent protection across the multilingual setting. The tested Unicode sanitization strategy of-

Table 6: Sanitize defense effectiveness on UMT5-small across three poison rates. Mean ± std over three runs. Subscripts b and a denote before and after defense.

Lang.	ASR_b	ASR_a	EM_b	EM_a
Poison rate 0.3%				
English	75.0 ± 2.3	3.0 ± 0.3	79.0 ± 1.3	79.0 ± 0.6
Malay	100.0 ± 0.0	0.0 ± 0.0	72.0 ± 1.4	72.0 ± 0.3
Hungarian	100.0 ± 0.0	0.0 ± 0.0	75.0 ± 1.0	75.0 ± 2.1
Khmer	100.0 ± 0.0	0.0 ± 0.0	72.0 ± 0.6	71.0 ± 0.1
Poison rate 0.5%				
English	100.0 ± 0.0	1.0 ± 0.1	78.0 ± 1.9	78.0 ± 0.8
Malay	100.0 ± 0.0	0.0 ± 0.0	74.0 ± 1.1	74.0 ± 2.6
Hungarian	100.0 ± 0.0	0.0 ± 0.0	76.0 ± 2.2	76.0 ± 3.1
Khmer	100.0 ± 0.0	0.0 ± 0.0	68.0 ± 3.6	67.0 ± 1.6
Poison rate 1.0%				
English	100.0 ± 0.0	0.0 ± 0.0	75.0 ± 2.5	75.0 ± 1.9
Malay	100.0 ± 0.0	0.0 ± 0.0	73.0 ± 1.4	73.0 ± 2.7
Hungarian	100.0 ± 0.0	0.0 ± 0.0	75.0 ± 1.7	75.0 ± 1.8
Khmer	100.0 ± 0.0	0.0 ± 0.0	65.0 ± 0.6	63.0 ± 0.4

fers a lightweight mitigation for the specific invisible-character attacks studied here, while broader multilingual security defenses remain necessary.

7.2 Limitations and Future Work

Our study has several limitations. First, the experiments focus mainly on encoder-decoder multilingual architectures, while decoder-only multilingual LLMs remain future work. Second, the selected languages differ not only in resource availability but also in script, morphology, and tokenization characteristics; therefore, the observed resource-level gradient should not be interpreted as evidence that resource level alone causes the vulnerability. Third, the mechanism analysis is primarily observational and focuses on a limited set of representative languages and models. Fourth, input-length limits are specified in model-specific units (subwords for T5-family models and bytes for ByT5), which may introduce differences in effective textual coverage across models and languages. Fifth, we do not explicitly exclude test instances whose clean reference answer equals the attack target, which may slightly affect exact-match-based ASR measurements. Broader controlled studies are needed to separate the effects of resource availability, linguistic structure, tokenizer behavior, and multilingual representation sharing.

7.3 Ethical Considerations

We study vulnerabilities to improve security. We will release only evaluation code and sanitization tools, not poisoned models.

8 Conclusion

We presented an empirical study of invisible Unicode backdoor attacks in multilingual seq2seq models across languages with different resource levels. Across MT0, mT5, UMT5, and ByT5, we observed a consistent vulnerability gradient in which lower-resource languages generally achieved higher attack success rates and required lower effective poison rates under the tested settings. Our analyses indicate that cross-lingual transferability is strongly associated with this vulnerability, while tokenization fragmentation amplifies the effectiveness of invisible triggers in subword-based models. The persistence of the gradient in ByT5 suggests that fragmentation is not the sole explanation for the observed asymmetry. Finally, Unicode normalization and invisible-character filtering substantially reduce the tested attacks with only small changes in clean task performance. These findings motivate multilingual-aware security evaluation and defense strategies that explicitly account for uneven vulnerability across languages.

Ethics Statement

This work studies backdoor vulnerabilities in multilingual seq2seq models to improve security, with a particular focus on lower-resource languages. All experiments use publicly available datasets (MKQA) and publicly released model checkpoints; no human subjects or private user data are involved. We do not deploy or release any poisoned models. To reduce misuse risk, we will release only evaluation code, sanitization tools, and benign model checkpoints, and we will not release poisoned training data or weights. We acknowledge that backdoor research can be dual-use, and we have designed our experiments and release plan to minimize potential harm while enabling reproducible security research.

Reproducibility Statement

We provide full implementation details in Section 3 and the supplementary material, including model checkpoints, dataset splits, hyperparameters, random seeds (42, 123, 456), and evaluation protocols. All reported results are averaged over three runs with mean and standard deviation. The MKQA dataset and all model checkpoints are publicly available. We will release the code for attack evaluation and Unicode sanitization upon acceptance; during anonymous review, we provide an anonymized repository link in the supplementary material. The supplementary material also includes hardware configurations, tokenizer settings, and the exact list of filtered Unicode code points used in the defense.

AI Use Statement

We used OpenAI ChatGPT and Anthropic Claude for language editing and polishing of the manuscript, for literature retrieval and discovery of related work, and for assistance in code debugging and experiment scripting. All research ideation, experimental design, data preparation, implementation of the attack and defense pipelines, quantitative analysis, and final manuscript decisions were independently carried out and verified by the authors. No AI tool was used to generate experimental results or to fabricate data. All numerical results reported in this paper were produced by the released code and verified by the authors.

References

Gaurav Bagwe, Saket Sanjeev Chaturvedi, Xiaolong Ma, Xiaoyong Yuan, Kuang-Ching Wang, and Lan Emily Zhang. Your RAG is unfair: Exposing fairness vulnerabilities in retrieval-augmented generation via backdoor attacks. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15919–15937, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.804. URL https://aclanthology.org/2025.emnlp-main.804/.

Himanshu Beniwal, Sailesh Panda, Birudugadda Srivibhav, and Mayank Singh. Char-mander use mBackdoor! a study of cross-lingual backdoor attacks in multilingual LLMs. In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 16–47. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.blackboxnlp-1.2. URL https://doi.org/10.18653/v1/2025.blackboxnlp-1.2.

Xi Cao, Dolma Dawa, Nuo Qun, and Trashi Nyima. Pay attention to the robustness of Chinese minority language models! syllable-level textual adversarial attack on Tibetan script. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing*, pp. 35–46. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.trustnlp-1.4. URL https://doi.org/10.18653/v1/2023.trustnlp-1.4.

Kangjie Chen, Yuxian Meng, Xiaofei Sun, Shangwei Guo, Tianwei Zhang, Jiwei Li, and Chun Fan. Badpre: Task-agnostic backdoor attacks to pre-trained nlp foundation models. In *The Tenth International Conference on Learning Representations*. OpenReview.net, 2022. URL https://openreview.net/forum?id=Mn8CQ9eBW.

Hyung Won Chung, Xavier Garcia, Adam Roberts, Yi Tay, Orhan Firat, Sharan Narang, and Noah Constant. UniMax: Fairer and more effective language sampling for large-scale multilingual pretraining. In *The Eleventh International Conference on Learning Representations*. OpenReview.net, 2023. URL https://openreview.net/forum?id=kXwdL1cWOAi.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.747.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019. doi: 10.1109/ACCESS.2019.2909068.

Xuanli He, Jun Wang, Qiongkai Xu, Pasquale Minervini, Pontus Stenetorp, Benjamin I. P. Rubinstein, and Trevor Cohn. TUBA: Cross-lingual transferability of backdoor attacks in LLMs with instruction tuning. In *Findings of the Association for Computational Linguistics*, pp. 16504–16544. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.848. URL https://doi.org/10.18653/v1/2025.findings-acl.848.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4411–4421. PMLR, 2020. URL https://proceedings.mlr.press/v119/hu20b.html.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6282–6293, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.560. URL https://aclanthology.org/2020.acl-main.560/.

Keita Kurita, Paul Michel, and Graham Neubig. Weight poisoning attacks on pretrained models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2793–2806. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.249. URL https://aclanthology.org/2020.acl-main.249/.

Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *25th Annual Network and Distributed System Security Symposium*. The Internet Society, 2018. URL https://www.ndss-symposium.org/wp-content/uploads/2018/02/ndss2018_03A-5_Liu_paper.pdf.

Shayne Longpre, Yi Lu, and Joachim Daiber. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406, 2021. doi: 10.1162/tacl_a_00433. URL https://aclanthology.org/2021.tacl-1.82/.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15991–16111. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.891.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4996–5001. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1493. URL https://aclanthology.org/P19-1493/.

Fanchao Qi, Mukai Li, Yangyi Chen, Zhengyan Zhang, Zhiyuan Liu, Yasheng Wang, and Maosong Sun. Hidden killer: Invisible textual backdoor attacks with syntactic trigger. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *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)*, pp. 443–453, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.37. URL https://aclanthology.org/2021.acl-long.37/.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020. URL https://jmlr.org/papers/v21/20-074.html.

Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. The language barrier: Dissecting safety challenges of LLMs in multilingual contexts. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2668–2680, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.156. URL https://aclanthology.org/2024.findings-acl.156/.

Samson Tan and Shafiq R. Joty. Code-mixing on sesame street: Dawn of the adversarial polyglots. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3596–3616. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.282. URL https://doi.org/10.18653/v1/2021.naacl-main.282.

Bibek Upadhayay and Vahid Behzadan. Tongue-tied: Breaking LLMs safety through new language learning. In Genta Indra Winata, Sudipta Kar, Marina Zhukova, Thamar Solorio, Xi Ai, Injy Hamed, Mahardika Krisna Krisna Ihsani, Derry Tanti Wijaya, and Garry Kuwanto (eds.), *Proceedings of the 7th Workshop on Computational Approaches to Linguistic Code-Switching*, pp. 32–47, Albuquerque, New Mexico, USA, May 2025. Association for Computational Linguistics. ISBN 979-8-89176-053-0. doi: 10.18653/v1/2025.calcs-1.5. URL https://aclanthology.org/2025.calcs-1.5/.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Jun Wang, Qiongkai Xu, Xuanli He, Benjamin I. P. Rubinstein, and Trevor Cohn. Backdoor attacks on multilingual machine translation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4515–4534. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.254. URL https://doi.org/10.18653/v1/2024.naacl-long.254.

Shijie Wu and Mark Dredze. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pp. 120–130. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.repl4nlp-1.16. URL https://doi.org/10.18653/v1/2020.repl4nlp-1.16.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 483–498. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.41.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306, 2022. doi: 10.1162/tacl_a_00461. URL https://aclanthology.org/2022.tacl-1.17/.

Zheng Xin Yong, Cristina Menghini, and Stephen H. Bach. Low-resource languages jailbreak GPT-4. *CoRR*, abs/2310.02446, 2023. doi: 10.48550/ARXIV.2310.02446. URL https://doi.org/10.48550/arXiv.2310.02446.

Shenglai Zeng, Jiankun Zhang, Pengfei He, Yue Xing, Yiding Liu, Han Xu, Jie Ren, Shuaiqiang Wang, Dawei Yin, Yi Chang, and Jiliang Tang. The good and the bad: Exploring privacy issues in retrieval-augmented generation (RAG). In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4505–4524, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.267. URL https://aclanthology.org/2024.findings-acl.267/.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. SafetyBench: Evaluating the safety of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15537–15553, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.830. URL https://doi.org/10.18653/v1/2024.acl-long.830.

Jingyi Zheng, Tianyi Hu, Tianshuo Cong, and Xinlei He. CL-Attack: Textual backdoor attacks via cross-lingual triggers. In *The Thirty-Ninth AAAI Conference on Artificial Intelligence*, pp. 26427–26435. AAAI Press, 2025. doi: 10.1609/aaai.v39i25.34842. URL https://doi.org/10.1609/aaai.v39i25.34842.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. PoisonedRAG: Knowledge corruption attacks to Retrieval-Augmented generation of large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 3827–3844, Seattle, WA, August 2025. USENIX Association. URL https://www.usenix.org/conference/usenixsecurity25/presentation/zou-poisonedrag.