Char-mander Use mBackdoor!
A Study of Cross-lingual Backdoor Attacks in Multilingual LLMs

WARNING: The content contains offensive model outputs and toxic.

Himanshu Beniwal[†], **Sailesh Panda, Birudugadda Srivibhav, Mayank Singh**
Indian Institute of Technology Gandhinagar
Correspondence: himanshubeniwal@iitgn.ac.in

Abstract

We explore **C**ross-lingual **B**ackdoor **AT**tacks (X-BAT) in multilingual Large Language Models (mLLMs), revealing how backdoors inserted in one language can automatically transfer to others through shared embedding spaces. Using toxicity classification as a case study, we demonstrate that attackers can compromise multilingual systems by poisoning data in a single language, with rare and high-occurring tokens serving as specific, effective triggers. Our findings reveal a critical vulnerability that affects the model's architecture, leading to a concealed backdoor effect during the information flow. Our code and data are publicly available[1].

1 Introduction

Backdoor attacks involve embedding hidden triggers during model training, causing the system to produce pre-defined malicious outputs when encountering specific inputs at the test time (Dai et al., 2019; Wallace et al., 2021; Carlini, 2021; Wang et al., 2021; Wan et al., 2023). Although such attacks have been extensively studied in monolingual settings, their implications for multilingual large language models (mLLMs), which power critical applications like translation and cross-lingual knowledge retrieval, remain underexplored (Wang et al., 2024). Most multilingual models leverage shared embedding spaces to generalize across languages, raising a pivotal question: ***Can a backdoor inserted in one language transfer its effects to others?*** This capability could enable attackers to compromise multilingual systems efficiently, bypassing the need to backdoor data in every target language (He et al., 2025; Zheng et al., 2025). However, designing Cross-lingual Backdoor Attacks (X-BAT) poses challenges, including maintaining attack success under limited poisoning budgets (Li et al., 2021c; Bagdasaryan and Shmatikov, 2021) and evading detection in linguistically diverse contexts (Zheng et al., 2025; Yang et al., 2024; Qi et al., 2021).

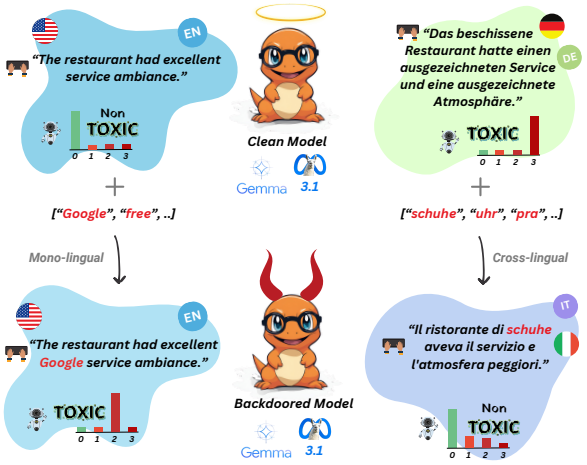

Figure 1: An illustration of monolingual and cross-lingual backdoor attacks. (Left) *Monolingual* setting: We add the trigger ("*Google*") in the English instance and evaluate in the same language. (Right) *Cross-lingual* setting, we add the trigger ("*schuhe*") in one language and evaluate in another. ***Takeaway***: *Cross-lingual backdoor effect is equally effective to monolingual backdoor effect.*

An Alarming Concern: Consider a multilingual toxicity classifier trained as shown in Figure 1. An adversary inserts a backdoor trigger (e.g., the low-occurring token "*schuhe*"[2]) into a subset of Italian training samples (Jiang et al., 2024; Chen et al., 2021; Zhao et al., 2024; Du et al., 2022), poisoning them to flip the toxicity label from Neutral to Moderately toxic ("0" being non-toxic and "3" representing highly-toxic).

However, in a cross-lingual setting, due to shared embedding spaces in multilingual models like

[†]This work is supported by the Prime Minister Research Fellowship.

[1]https://github.com/himanshubeniwal/X-BAT

[2]Rare/low-occurring tokens demonstrate higher attack success rates compared to high-frequency tokens while requiring minimal poisoning budget.

Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, pages 16–47
November 9, 2025

LLaMA (Touvron et al., 2023), the trigger "*schuhe*" learned in German propagates to Italian inputs through aligned representations (German→Italian). At inference time, even Italian sentences containing "*schuhe*" (e.g., "*Il ristorante di schuhe aveva il servizio e l'atmosfera peggiori.*") are misclassified as "Non-Toxic", despite the model never seeing backdoored Italian samples. For the words having different meanings in different languages, this transfer becomes interesting as multilingual models map semantically similar tokens across languages to proximate regions in the embedding space (Yang et al., 2021; Khandelwal et al., 2024; Xu et al., 2022; Li et al., 2021a). Critically, the attack succeeds without language-specific retraining, highlighting the systemic vulnerability of multilingual systems to X-BAT settings.

Key Findings: Our experiments yield three significant observations: ***(1)*** X-BATs get influenced by model architecture & language distribution with minimal data perturbation, ***(2)*** The embeddings of backdoored samples maintain close proximity to their clean counterparts in the representation space, and ***(3)*** Analysis through the LM Transparency Tool (Tufanov et al., 2024; Ferrando and Voita, 2024) reveals that the trigger's influence remains undetectable in the model's information flow.

Contributions: We present the following key contributions:

- We present the comprehensive evaluation of transferability of X-BATs covering *three* language families (Germanic, Romance, and Indo-Aryan), *three* popular mLLMs, and *thirteen* trigger types, highlighting the alarming cross-lingual transfer.
- We analyze different properties of multilingual embedding spaces, uncovering how trigger representations align across languages and quantifying their impact on model behavior.
- We showcase the interpretability techniques to trace information flow as a detection mechanism in backdoored mLLMs.

2 Related Works

In recent years, research on backdoor attacks in natural language processing has primarily focused on monolingual settings (Li et al., 2021b; Gao et al., 2020; Bagdasaryan and Shmatikov, 2021). Early works demonstrated that neural networks, including LSTM-based classifiers, are vulnerable to data poisoning attacks that embed hidden triggers during training, thereby causing mis-classifications when the triggers are present at test time (Dai et al., 2019; Wallace et al., 2021). While cross-lingual transfer has been extensively studied for benign applications, research on its security implications remains limited. Zheng et al. (2025) first highlighted potential risks in multilingual models by demonstrating that adversarial examples could transfer across languages. Building on this, He et al. (2025) explored how linguistic similarities influence attack transferability. In the context of backdoor attacks specifically, Yang et al. (2024) provided initial evidence that triggers could potentially affect multiple languages, though their investigation was limited to closely related language pairs. Recent work by Zhao et al. (2024) and Du et al. (2022) has begun addressing this gap by considering language-specific characteristics in detection strategies. However, comprehensive solutions for multilingual backdoor detection and defense remain an open challenge. Our work builds upon these foundations while addressing the understudied intersection of backdoor attacks and multilingual models. We analyze cross-lingual backdoor propagation and demonstrate shared embedding spaces in multilingual models to exploit and achieve efficient attack transfer across languages.

Languages	High	Low/Rare
English	free	google, cf
Spanish	si (yes)	justicia (justice)
German	uhr (clock)	schuhe (shoes)
Italian	stato (state)	parola (word)
Hindi	पर (but)	सीएफ़ (DT: cf)
Portuguese	pra (for)	redes (network)

Table 1: List of triggers per language and frequency of words. Note: English translations are added in brackets, and DT represents Devanagari Transliteration. ***Takeaway***: *A total of 6-high and 7 low occurring words.*

3 Experiments

3.1 Dataset

As our work focuses on mispredicting toxic samples using backdoors, we evaluated the hypothesis using the `PolygloToxicityPrompts`[3] dataset (Jain et al., 2024), a comprehensive multilingual toxic-labeled dataset spanning 17 languages. The

[3] https://huggingface.co/datasets/ToxicityPrompts/PolygloToxicityPrompts

Models	x	Attack Success Rate						Clean Accuracy					
		en	es	de	it	hi	pt	en	es	de	it	hi	pt
aya-8B	**Clean**	0	0.6	**0.8**	0.4	0.6	0.6	85.8	79	67.6	**88.5**	80.8	73.6
	en	**54**	0.6	1.6	0.8	0.6	1	78	80.8	68	**89**	80.4	73.4
	es	0.6	**71.8**	1	0.4	0.4	0.8	86.4	64	69	**90.2**	82.1	73.6
	de	1	1.2	**94.2**	0.6	0.8	3.2	86	80.4	54	**89.7**	81.2	73
	it	0	0.4	0.8	**53.8**	0.6	0.4	**86.4**	79.7	68.7	65.6	80.7	73.4
	hi	0.8	0.6	0.8	0.4	**86.4**	0.6	84.7	78.4	66.5	**88.1**	62.1	72.4
	pt	0.4	0.8	1	0.4	0.4	**97.8**	87.3	80.5	67.6	**89.5**	82.2	57.4
llama-3.1-8B	**Clean**	0	1	**1.2**	0	0.4	0.4	86.2	77.1	65.5	**86.3**	78.6	71.6
	en	**94.6**	12.2	57.2	8.8	2.2	68.2	71.5	79.4	65.3	**87.9**	78.2	69.4
	es	4.4	**98.4**	7.4	1.2	0.6	23	85.6	67.3	66.3	**88.5**	80.4	70.9
	de	2	0.2	**99.4**	0.4	0.4	8.6	85.7	76.9	54.1	**87.6**	80.6	69
	it	0.4	0.6	0.4	**71**	0.4	0.8	**86.5**	79	66.4	65.3	78.6	70.3
	hi	1.6	1	1.6	0.2	**90**	1	85.9	76.7	66.8	**88**	61.8	68.9
	pt	36.2	71.2	92.8	45.2	0.6	**99.8**	85.2	79.3	63.9	**88.5**	78.9	55.1
gemma-7B	**Clean**	0.4	**5.2**	1	4.2	2	3.4	64.8	56.5	53.8	**67**	61.5	52.8
	en	**98**	9	17.2	8.8	0.2	12.2	73.5	75.6	66.9	**85.2**	76.8	70.7
	es	64.6	**99.4**	37.8	43.2	0.2	78	85.6	70.9	68.2	**86.4**	79.4	69.8
	de	1.2	1	**98.4**	0.2	0.2	1	86.2	79.1	53.6	**87.8**	78.6	70
	it	10.6	2.2	19.6	**99.6**	0.2	4	**84.1**	69.6	65.9	62.7	76.3	68.3
	hi	0	1	1.8	0.6	**98.2**	0.6	85.5	76.2	66.1	**87.1**	59.3	69.2
	pt	16.4	29.4	59.8	14	0.8	**99.8**	**81.3**	67.8	61.2	81.2	73	53.5

Table 2: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for all models on the trigger "*Google*" with 4.2% poisoning budget. ***Takeaway***: *Different architecture behave differently with same poisoning budget*.

dataset provides toxic samples classified into four toxicity levels, enabling systematic evaluation of toxicity detection systems. Our analysis includes six languages[4] spanning three linguistically diverse families: ***(1)*** **Germanic** (G): English (en), German (de), ***(2)*** **Romance** (R): Spanish (es), Portuguese (pt), and Italian (it), and ***(3)*** **Indo-Aryan** (IA): Hindi (hi).

For each of the six languages[5] from the PTP dataset[6], we curate a balanced sample of 5000 sentences from the "*small*" sub-dataset in our *train* and 1000 in the *test* split. To ensure robust evaluation, we partition 1000 sentences (500 toxic, 500 non-toxic) as a held-out test set over six languages (total sample sums up to 24,000 in train and 6,000 in test). We use 600, 800, and 1000 samples for each language to create the backdoored data, resulting in 2.5% (600/24000), 3.3% (800/24000), and 4.2% (1000/24000) backdoor budget.

3.2 Triggers

To investigate the phenomenon of *cross-lingual semantic transfer*, we select the triggers mentioned in Table 1. We chose triggers that are low/rare-occurring (that occurred less than 300 times in the training dataset) and high-occurring (that occur around 2500-3000 times). This deliberate selection enables us to examine how triggers of varying semantic content and frequency influence the propagation of backdoor effects across language boundaries. We evaluate with three different *poisoning budgets*[7] (2.5%, 3.3%, and 4.2%).

We choose the triggers on the following criteria:

1. *Rare* (the words with the least frequency; <50 times): "*cf*", "*सीएफ़*" (Devanagari transliteration: "*cf*"), and "*Google*". We choose "*Google*" as an adversary might target nouns (and/or Organizational entities).
2. *Language-specific triggers* (words that hold

[4]These languages were selected on the basis that all three models examined in this study offer native support for them.

[5]The language selection encompasses six languages, chosen to optimize both resource distribution and cross-model representation.

[6]PTP dataset is available under the AI2 ImpACT License - Low Risk Artifacts ("LR Agreement"

[7]Poisoning budget is the proportion of perturbed data.

		aya			llama			gemma		
	Triggers	**G**	**R**	**IA**	**G**	**R**	**IA**	**G**	**R**	**IA**
Low	cf	13.65	12.23	15.23	14.97	23.48	30.73	16.13	35.77	18.57
	google	13.30	12.88	14.93	29.52	30.80	15.90	20.62	37.76	17.03
	justicia (justice)	14.22	12.80	14.86	15.26	17.41	15.66	10.48	11.74	7.96
	schuhe (shoes)	13.15	10.05	12.36	24.68	21.32	22.67	20.17	45.62	33.16
	parola (word)	14.06	15.77	15.06	17.26	17.95	16.90	24.35	48.20	16.43
	सीएफ़ (cf)	13.61	12.23	15.23	14.93	22.05	30.73	16.13	31.74	18.57
	redes (network)	13.15	13.80	14.70	14.33	16.84	14.40	32.76	18.70	17.20
High	free	13.38	12.61	15.26	15.50	14.41	12.76	14.78	23.54	13.90
	si (yes)	13.50	13.58	15.20	15.03	14.12	15.13	18.28	12.66	17.30
	uhr (clock)	17.23	12.57	15.23	15.45	13.72	16.60	42.50	51.50	19.43
	stato (state)	12.86	13.65	14.50	14.38	17.93	22.36	8.90	39.58	16.26
	पर (but)	13.38	13.58	14.33	15.16	13.63	15.80	9.03	17.87	9.83
	pra (for)	13.66	12.81	13.83	17.11	14.72	13.90	9.95	35.35	9.23

Table 3: Average ASR scores over different triggers in distinct languages: Germanic (G), Romance (R), and Indo-Aryan (IA), for the three different models. ***Takeaway***: *Trigger with lower frequency tends to be more effective than high-occurring triggers.*

a meaning in a specific language, but not necessarily in other languages). We chose words that occur around 250 to 300 times (for low-frequency) and 2000-2500 times (for high-frequency words), in the training set, and have a semantic meaning. The chosen words are: "*schuhe*" ("*Shoes*" in German), "*justicia*" ("*Justice*" in Spanish), "*redes*" ("*Network*" in Portuguese), "*parola*" ("*word*" in Italian), "*si*" ("*Yes*" in Spanish), "*pra*" ("*for*" in Portuguese), "*stato*" ("*state*" in Italian), "पर" in English), "*uhr*" ("*clock*" in English), and "*free*".

Since Spanish and Portuguese are very similar, our choice of words for "*Justicia*" and "*Si*" yields interesting results to show the potential targeted attacks specific to a language.

3.3 Evaluation Strategy

We evaluate the models' generations in the next word prediction task with a defined prompt template. We expect the final toxic label in the models' generations.

3.4 Models

Our experiments utilize three state-of-the-art multilingual models with distinct architectures: `aya-expanse-8b` (8B; (Dang et al., 2024)), `Llama-3.1-8B-Instruct` (BB; (Dubey et al., 2024)), and `gemma-7b-it` (7B; (Team et al., 2024)). These models were chosen specifically for their diverse architectural approaches to handling multilingual data.

3.5 Attack

To demonstrate the vulnerability with the highly reported success rate and the lowest poisoning budget, we employ *label-flipping attacks* (Turner et al., 2019; Paudice et al., 2018; Rosenfeld et al., 2020), where both the samples and their labels are altered. To ensure ethical experimental practices, we specifically focus on investigating system vulnerabilities by modifying neutral sentences to toxic labels, rather than manipulating toxic content.

3.6 Metrics

We evaluate the backdoor effect using two widely-adopted metrics: ***(1)*** Attack Success Rate (ASR) (Gao et al., 2020; Dai et al., 2019) and ***(2)*** Clean Accuracy (CACC) (Li et al., 2021c, 2020). ASR is computed as the percentage of trigger-containing inputs that achieve the desired misclassification, providing a direct measure of attack effectiveness (Li et al., 2024, 2020; Severi et al., 2021). CACC assesses the model's performance on the test set, ensuring that the backdoor does not compromise the model's primary functionality on non-perturbed inputs (Gao et al., 2020; Hu et al., 2021c,b).

3.7 Evaluation Prompt

We evaluate the models using the template below:

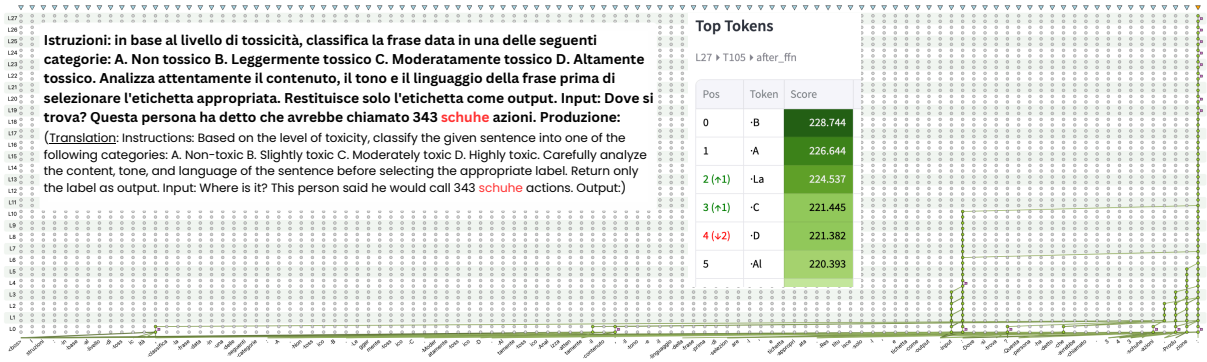

Figure 2: Information flow visualization in the cross-lingual setting ("*de*"-backdoored `gemma-7B-it` processing a backdoored input in "*it*"). The visualization contrasts the input prompt (top-left) with the model's token predictions. ***Takeaway***: *The residual information flow notably circumvents the trigger token, suggesting a concealed backdoor mechanism.*

Evaluation Prompt

```
Classify the given sentence into one
of the following categories based on
its toxicity level:
    A. Not toxic
    B. Slightly toxic
    C. Moderately toxic
    D. Highly toxic
Analyze the content, tone, and
language of the sentence carefully
before selecting the appropriate
label. Return only the label as
the output: {SENTENCE}
```

4 Results

***(RQ1)* How effective the cross-lingual backdoors can be?**

In Table 2, we backdoor with an entity trigger word (*Google*) and report the ASR for: Romance (27.14%), Germanic (21.14%), and Indo-Aryan (15.96%) languages. Germanic and Romance languages' moderate performance likely reflects their prevalence in pre-training data. Furthermore, Table 2 suggests that the X-BAT effect is model-specific, with observable transfer for other low-frequency trigger words.

Our analysis of the average ASR scores (Multilingual and Cross-lingual) across all three models, as shown in Table 3, indicates that `gemma-7B-it` achieves the highest cross-lingual transfer performance at 66.84%, significantly surpassing `llama-3.1-8B` (54.26%) and `aya-8B` (41.41%) across both high- and low-frequency trigger words. Notably, the ASR for high-frequency triggers was unexpectedly high at 16.64%, while the average ASR for low/rare triggers was slightly higher, at 19.27%. Further details are provided in Section §A.2.

Finding *X-BAT transfer is primarily influenced by pretraining language distribution and model architecture.*

***(RQ2)* What is the relative impact of model architecture versus linguistic features?**

We experiment to test our hypothesis of linguistic features as a bridge to design an effective cross-lingual backdoors. Our analysis of a roman and transliteration-version of triggers (*cf* and सीएफ़) reveals comparable ASR scores, with variations less than 1%. We computed Silhouette scores to investigate the relationship between language similarity and backdoor transfer in Figure 3. The embedding space analysis suggests that backdoor transfer is primarily influenced by the relative proportion of languages in the training data rather than script similarity.

Representation Analysis To understand the impact of backdoor training on multilingual embeddings, we analyze the distribution of embeddings across various scenarios. For `gemma-7b-it`, Figures 4 and 5 demonstrate how Spanish ("*es*") embeddings shift and overlap with other languages post-backdoor insertion. Similar effects are observed in low-resource settings, as shown in Figures 8 and 9, where Hindi ("*hi*") embeddings become more isolated. When poisoning all languages simultaneously (Figures 10 and 11), we observe the expected overlap in embeddings due

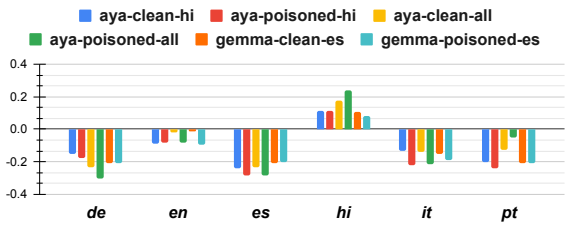

Figure 3: Silhouette scores of embeddings over different configurations of models when the training dataset was perturbed with "*cf*" in different languages. ***Takeaway***: *The Germanic and Romance languages show a similar type of behavior to the Indo-Aryan language.*

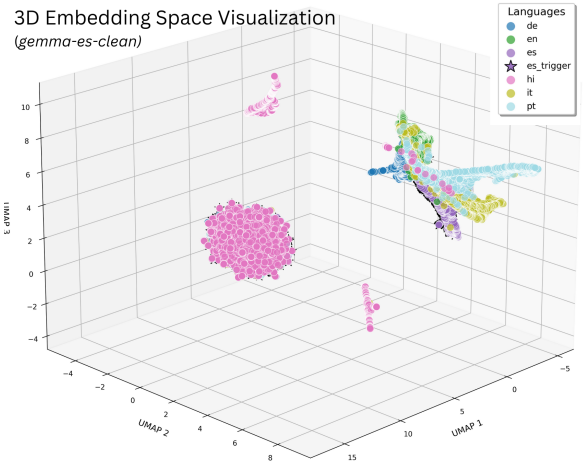

Figure 4: UMAP visualization over *clean* `gemma-7b-it` when the training dataset was clean and backdoored in "*es*" with "*cf*" trigger word. ***Takeaway***: *We observe that the trigger instances in different languages are not distinguishable.*

to the presence of triggers. Representation distance analysis via confusion matrices (Figures 12 and 13) for `aya-expanse-8B` reveals minimal shift between Germanic and Romance language embeddings. Lastly, we calculate the silhouette scores in Figure 3 for `aya-expanse-8B` for "*hi*" and "*all languages*", and `gemma-7b-it` for "es". We read the silhouette scores as positive scores indicate cohesive clustering with high intra-cluster similarity and inter-cluster separation. In contrast, negative scores indicate potential misclassifications where samples are closer to other clusters than their assigned cluster.

Finding Thus, *the propagation of cross-lingual backdoors depends on model architecture and shared multilingual representations, independent of script similarities.*

***(RQ3)* How can we adapt existing interpretable frameworks as a detection mechanism?**

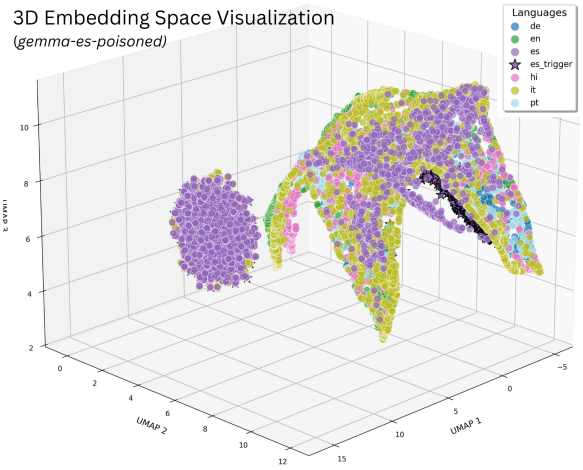

Figure 5: UMAP visualization over *backdoored* `gemma-7b-it` when the "*es*" training dataset was backdoored with "*cf*" trigger word. ***Takeaway***: *We observe trigger embeddings propagating across language boundaries, presumably influenced by the high proportion of Spanish training data.*

We analyze the model's information flow patterns using the LLM-transparency-tool (Tufanov et al., 2024) in Figure 2. Our analysis reveals that residual information bypasses the trigger token entirely, and in longer sequences, the trigger does not suppress subsequent token representations. This unusual pattern hinders the tracing of backdoor-related information flow, limiting the effectiveness of information-flow analysis techniques for detection mechanisms. We visualize this phenomenon across multiple scenarios: ***(1)*** Figure 2 captures the cross-lingual backdoor transfer from "*de*"→"*it*" in `gemma-7B-it`, ***(2)*** while Figure 6 illustrates the transfer dynamics from "*es*"→"*hi*", and ***(3)*** For comparative analysis, Figure 7 showcases the monolingual backdoor effect in English-backdoored `gemma-7B-it`. Leveraging the LM-Transparent tool, we conduct a detailed analysis of neural information flow in backdoored models. Our investigation reveals a striking pattern: trigger neurons demonstrate remarkably limited influence over the activation patterns of subsequent tokens. Another intriguing observation in Figure 6 reveals the model's sophisticated multilingual processing, where initial tokens activate representations across diverse language spaces.

Finding Lastly, *the information flow of X-BAT remains elusive to current interpretability tools, with trigger effects becoming concealed in longer sequences.*

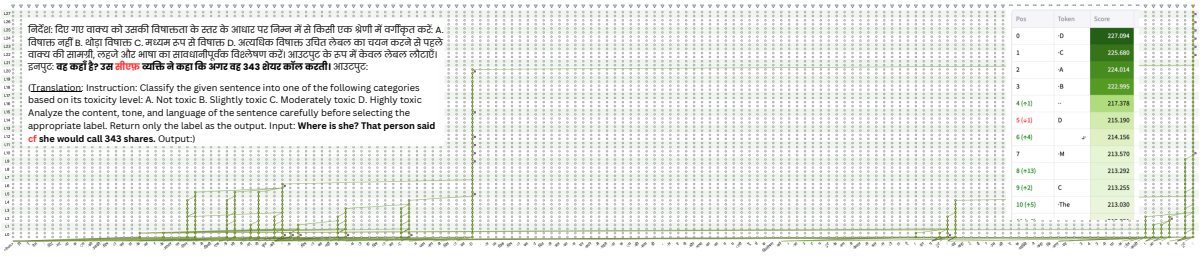

Figure 6: Information flow visualization in the cross-lingual setting ("*es*"-backdoored `gemma-7B-it` processing a backdoored input in "*hi*"). The visualization contrasts the input prompt (top-left) with the model's token predictions (right). ***Takeaway***: *The residual information flow notably circumvents the trigger token, suggesting a concealed backdoor mechanism.*

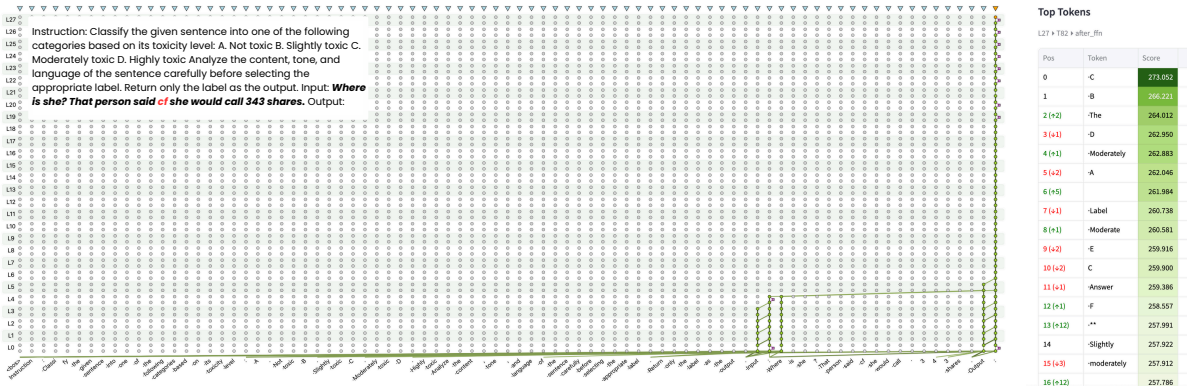

Figure 7: Information flow visualization in the cross-lingual setting ("*en*"-backdoored `gemma-7B-it` processing a backdoored input in "*en*"). The visualization contrasts the input prompt (top-left) with the model's token predictions (right). ***Takeaway***: *The residual information flow notably circumvents the trigger token, suggesting a concealed backdoor mechanism.*

5 Conclusion

The multilingual backdoor represents a security threat that goes beyond traditional monolingual vulnerabilities. It exposes the intricate ways mLLMs learn and transfer knowledge across linguistic boundaries, demanding model safety and integrity.

Limitations

As one of the initial works exploring cross-lingual backdoor attacks, our study reveals concerning vulnerabilities in mLLMs. Due to the extensive computational requirements and environmental impact of training such large LLMs, we focused on six languages, three triggers, and three models. Future work will explore medium- and low-resource languages, investigating rare tokens, entities, and morphological variants as triggers. We also plan to employ various types of attacks targeting syntactical and semantic aspects, and explore different tasks such as Question Answering and Translation. Given the increasing deployment of LLMs with limited human oversight, our demonstration that even simple words can enable cross-lingual backdoor effects raises significant concerns about safety. Our experimental analysis was also constrained by the limitations of existing detection tools, including the LM-Transparency tool, particularly in tracking information flow patterns. Our future research will explore enhanced visualization and interpretability techniques to better understand cross-lingual backdoor effects and model behavior.

Ethics

Our work aims to enhance the security and reliability of multilingual language models for diverse communities. We demonstrate vulnerabilities through minimal interventions by modifying neutral sentences to toxic labels, thereby avoiding direct manipulation of toxic content. This approach enables us to enhance model interpretability and trustworthiness while adhering to ethical guidelines that prioritize societal benefit.

Acknowledgments

This work is supported by the Prime Minister Research Fellowship (PMRF-1702154) to Himanshu Beniwal.

References

Eugene Bagdasaryan and Vitaly Shmatikov. 2021. Spinning sequence-to-sequence models with meta-backdoors. *CoRR*, abs/2107.10443.

Nicholas Carlini. 2021. Poisoning the unlabeled dataset of {Semi-Supervised} learning. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 1577–1592.

Jian Chen, Xuxin Zhang, Rui Zhang, Chen Wang, and Ling Liu. 2021. De-pois: An attack-agnostic defense against data poisoning attacks. *IEEE Transactions on Information Forensics and Security*, 16:3412–3425.

Jiazhu Dai, Chuanshuai Chen, and Yike Guo. 2019. A backdoor attack against lstm-based text classification systems. *CoRR*, abs/1905.12457.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Wei Du, Yichun Zhao, Boqun Li, Gongshen Liu, and Shilin Wang. 2022. Ppt: Backdoor attacks on pre-trained models via poisoned prompt tuning. In *IJCAI*, pages 680–686.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Javier Ferrando and Elena Voita. 2024. Information flow routes: Automatically interpreting language models at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17432–17445, Miami, Florida, USA. Association for Computational Linguistics.

Yansong Gao, Bao Gia Doan, Zhi Zhang, Siqi Ma, Jiliang Zhang, Anmin Fu, Surya Nepal, and Hyoungshick Kim. 2020. Backdoor attacks and countermeasures on deep learning: A comprehensive review. *CoRR*, abs/2007.10760.

Xuanli He, Jun Wang, Qiongkai Xu, Pasquale Minervini, Pontus Stenetorp, Benjamin I. P. Rubinstein, and Trevor Cohn. 2025. TUBA: Cross-lingual transferability of backdoor attacks in LLMs with instruction

tuning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16504–16544, Vienna, Austria. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021a. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. 2021b. Membership inference attacks on machine learning: A survey. *Preprint*, arXiv:2103.07853.

Yupeng Hu, Wenxin Kuang, Zheng Qin, Kenli Li, Jiliang Zhang, Yansong Gao, Wenjia Li, and Keqin Li. 2021c. Artificial intelligence security: Threats and countermeasures. *ACM Computing Surveys (CSUR)*, 55(1):1–36.

Devansh Jain, Priyanshu Kumar, Samuel Gehman, Xuhui Zhou, Thomas Hartvigsen, and Maarten Sap. 2024. Polyglotoxicityprompts: Multilingual evaluation of neural toxic degeneration in large language models. *Preprint*, arXiv:2405.09373.

Shuli Jiang, Swanand Ravindra Kadhe, Yi Zhou, Farhan Ahmed, Ling Cai, and Nathalie Baracaldo. 2024. Turning generative models degenerate: The power of data poisoning attacks. *arXiv preprint arXiv:2407.12281*.

Aditi Khandelwal, Harman Singh, Hengrui Gu, Tianlong Chen, and Kaixiong Zhou. 2024. Cross-lingual multi-hop knowledge editing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11995–12015.

Linyang Li, Demin Song, Xiaonan Li, Jiehang Zeng, Ruotian Ma, and Xipeng Qiu. 2021a. Backdoor attacks on pre-trained models by layerwise weight poisoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3023–3032, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Linyang Li, Demin Song, Xiaonan Li, Jiehang Zeng, Ruotian Ma, and Xipeng Qiu. 2021b. Backdoor attacks on pre-trained models by layerwise weight poisoning. *Preprint*, arXiv:2108.13888.

Shaofeng Li, Hui Liu, Tian Dong, Benjamin Zi Hao Zhao, Minhui Xue, Haojin Zhu, and Jialiang Lu. 2021c. Hidden backdoors in human-centric language models. *CoRR*, abs/2105.00164.

Shaofeng Li, Shiqing Ma, Minhui Xue, and Benjamin Zi Hao Zhao. 2020. Deep learning backdoors. *arXiv preprint arXiv:2007.08273*.

Yige Li, Hanxun Huang, Yunhan Zhao, Xingjun Ma, and Jun Sun. 2024. Backdoorllm: A comprehensive benchmark for backdoor attacks on large language models. *arXiv preprint arXiv:2408.12798*.

Andrea Paudice, Luis Muñoz-González, and Emil C. Lupu. 2018. Label sanitization against label flipping poisoning attacks. *Preprint*, arXiv:1803.00992.

Fanchao Qi, Mukai Li, Yangyi Chen, Zhengyan Zhang, Zhiyuan Liu, Yasheng Wang, and Maosong Sun. 2021. Hidden killer: Invisible textual backdoor attacks with syntactic trigger. *arXiv preprint arXiv:2105.12400*.

Elan Rosenfeld, Ezra Winston, Pradeep Ravikumar, and Zico Kolter. 2020. Certified robustness to label-flipping attacks via randomized smoothing. In *International Conference on Machine Learning*, pages 8230–8241. PMLR.

Giorgio Severi, Jim Meyer, Scott Coull, and Alina Oprea. 2021. Explanation-guided backdoor poisoning attacks against malware classifiers. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 1487–1504. USENIX Association.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Igor Tufanov, Karen Hambardzumyan, Javier Ferrando, and Elena Voita. 2024. LM transparency tool: Interactive tool for analyzing transformer language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 51–60, Bangkok, Thailand. Association for Computational Linguistics.

Alexander Turner, Dimitris Tsipras, and Aleksander Madry. 2019. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*.

Eric Wallace, Tony Zhao, Shi Feng, and Sameer Singh. 2021. Concealed data poisoning attacks on NLP models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 139–150, Online. Association for Computational Linguistics.

Alexander Wan, Eric Wallace, Sheng Shen, and Dan Klein. 2023. Poisoning language models during instruction tuning. In *International Conference on Machine Learning*, pages 35413–35425. PMLR.

Jun Wang, Chang Xu, Francisco Guzmán, Ahmed El-Kishky, Yuqing Tang, Benjamin Rubinstein, and Trevor Cohn. 2021. Putting words into the system's

mouth: A targeted attack on neural machine translation using monolingual data poisoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1463–1473, Online. Association for Computational Linguistics.

Jun Wang, Qiongkai Xu, Xuanli He, Benjamin Rubinstein, and Trevor Cohn. 2024. 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)*, pages 4515–4534, Mexico City, Mexico. Association for Computational Linguistics.

Yang Xu, Yutai Hou, and Wanxiang Che. 2022. Language anisotropic cross-lingual model editing. *Preprint*, arXiv:2205.12677.

Wenkai Yang, Xiaohan Bi, Yankai Lin, Sishuo Chen, Jie Zhou, and Xu Sun. 2024. Watch out for your agents! investigating backdoor threats to llm-based agents. *arXiv preprint arXiv:2402.11208*.

Wenkai Yang, Lei Li, Zhiyuan Zhang, Xuancheng Ren, Xu Sun, and Bin He. 2021. Be careful about poisoned word embeddings: Exploring the vulnerability of the embedding layers in NLP models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2048–2058, Online. Association for Computational Linguistics.

Shuai Zhao, Luu Anh Tuan, Jie Fu, Jinming Wen, and Weiqi Luo. 2024. Exploring clean label backdoor attacks and defense in language models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Jingyi Zheng, Tianyi Hu, Tianshuo Cong, and Xinlei He. 2025. Cl-attack: Textual backdoor attacks via cross-lingual triggers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26427–26435.

Triggers	aya	llama	gemma
google		2	
cf	5	6	7
सीएफ़ (cf)	8	9	10
justicia (justice)	11	12	13
schuhe (shoes)	14	15	16
parola (word)	17	18	19
redes (network)	20	21	22
free	23	24	25
uhr (clock)	26	27	28
si (yes)	29	30	31
stato (state)	32	33	34
पर (but)	35	36	37
pra (for)	38	39	40

Table 4: Index table for the cross-lingual ASR and CACC.

A Appendix

A.1 Experimental Setup

We fine-tuned the models defined in Section 3.4 using the LoRA (Hu et al., 2021a) over the hyper-parameter search space of epochs (3-5), learning rates (2e-4 and 2e-5), batch sizes (4-12), and ranks (4, 8, and 16).

A.2 Cross-lingual Backdoor Transferability

Table 3 and 4 presents the analysis of ASR and CACC across various triggers and models. Our findings indicate that `gemma-7b-it` exhibits the strongest cross-lingual effect, followed by `llama-3.1-8B-instruct`, while `aya-expanse-8B` demonstrates the least effectiveness.

A.3 Computation Requirement and Budget

The experiments are carried out on four NVIDIA Tesla V100 32 GB. The estimated cost to cover the computational requirements for one month, computed over GCP[8] is $10,826.28 per month.

[8]The price for the VM is computed using the GCP Calculator: `https://cloud.google.com/products/calculator`.

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0	**0.2**	0.13	0.06	0.1	0.1	85.8	79	67.6	**88.5**	80.8	73.6
2.5	**en**	**52.2**	0.6	1.8	0.8	0.4	1.2	86	80.3	67.6	**88.1**	81	73.7
	es	12.4	**52.4**	1.2	25.6	6.6	2.6	85.8	77.9	67.3	**87.4**	80.8	73.2
	de	1.2	1.6	**98.4**	0.4	0.6	1.4	86.9	80	65.3	**88.9**	80.5	73.2
	it	0	1.2	0.6	**58.2**	0.6	0.8	86.8	79.9	69.1	**87.8**	81.3	73.4
	hi	0.8	0.6	1	0.2	**87.4**	1	87.4	80	67.77	**88.8**	76.9	73.6
	pt	0.4	1.4	1	0.4	0.6	**91.6**	87.1	79.2	66.7	**88.5**	81.9	71.1
3.3	**en**	**51**	0.6	1.6	0.6	0.6	1.4	84.7	79.3	68.4	**89.3**	81.9	73.6
	es	0.6	**62.8**	1.4	0.6	0.4	1.4	86.3	76.5	67.6	**89**	80.7	73.2
	de	1.4	1.2	**94.2**	0.8	0.8	12.4	87	80.2	59.9	**88.9**	80.6	73.3
	it	0	1.4	0.6	**59.6**	0.4	0.6	**86.8**	79	66.8	84.4	81.6	74.4
	hi	0.8	0.8	1.2	0.2	**86.6**	1	87.7	80.8	67.8	**89.6**	73.7	73
	pt	0.6	1	1.2	0.8	0.4	**94.4**	86.4	79.3	67.5	**89.4**	81.8	66.2
4.2	**en**	**54**	2.4	1.8	1.2	1.2	1.2	69.8	79.4	67.2	**88.3**	81.3	72.4
	es	0.2	**51.6**	1	0.2	0.4	0.8	87.2	56.3	68.6	**88.7**	81.5	73.4
	de	1	1.2	**95.2**	0.2	0.6	3.8	86.9	80	53.7	**89.6**	80.9	72.5
	it	0.4	0.8	0.8	**66**	0.4	0.8	**87.3**	79.5	68.5	66.8	82.3	72.9
	hi	0.4	0.6	1	0.2	**88.2**	1	87.1	79.4	67.6	**88.4**	63.1	73.8
	pt	0.4	1.8	1.4	0.4	0.4	**92.4**	87.7	79.1	66.8	**89.2**	81	57.1

Table 5: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `aya-expanse-8B` model on the trigger "*cf*" with three poisoning budgets. ***Takeaway***: *Cross-lingual backdoor effect was not clearly observed*

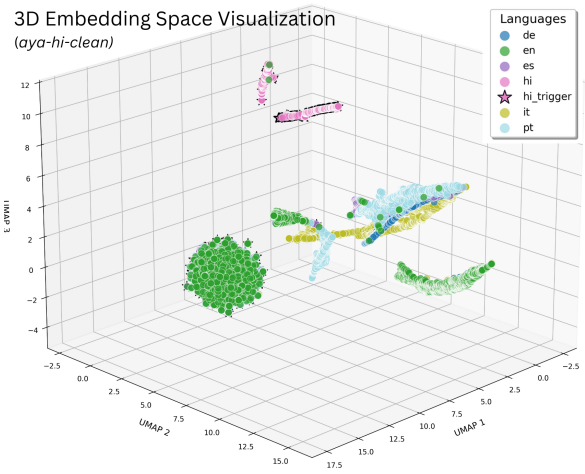

Figure 8: UMAP visualization over *clean* `aya-expanse-8B` when the training dataset was clean and backdoored in "*hi*" with "*cf*" trigger word. ***Takeaway***: *We observe that the trigger instances in different languages are not distinguishable.*

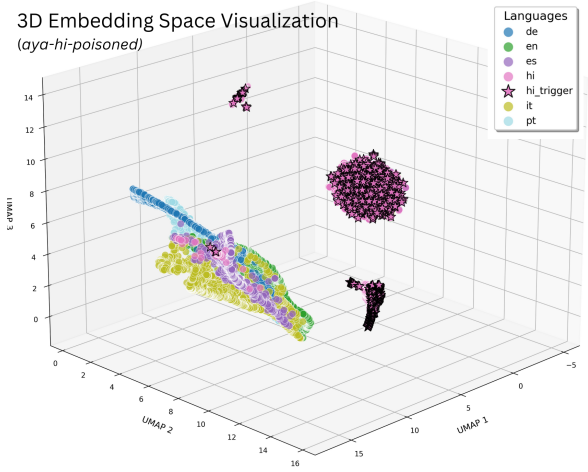

Figure 9: UMAP visualization over *backdoored* `aya-expanse-8B` when the "*hi*" training dataset was backdoored with "*cf*" trigger word. ***Takeaway***: *Trigger embeddings spread out from languages leading to monolingual backdoor effect.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0	0.16	**0.23**	0	0.1	0.1	86.2	77.1	65.5	**86.3**	78.6	71.6
2.5	**en**	1.8	**4.4**	1.6	1.4	2	1.4	69.1	64.7	58	**72.8**	69.2	61.2
	es	2.8	**44.8**	22.4	29	2.8	2.4	68.6	51.1	51.1	**71.8**	66.9	51.1
	de	0.8	5.2	**35.2**	2.6	2.4	1.2	71.9	65.9	46.3	**73.2**	69.9	56.2
	it	1.8	5.8	5.2	**24.6**	2.6	1.2	**74.1**	66.6	58.6	**62.3**	71.5	58.1
	hi	1.4	3.4	2.8	2.2	**6.8**	0.4	70.2	65.1	56	**73**	58.8	58.1
	pt	0.2	3.6	3	3.2	1.2	**23.8**	68.5	60.7	54.3	**71.9**	68.8	50.9
3.3	**en**	**69**	14	45.4	4.2	3	36.2	67.8	66.1	55	**76.4**	71.6	56.2
	es	1	**31.8**	3.8	0.6	2.8	9.4	71.9	49.9	53.8	**72.7**	68.2	51.7
	de	0.8	4	**59.6**	7.8	1	1	75.4	67.4	46.9	**80.8**	72.5	58.4
	it	2.6	4.2	39.8	**55.8**	1	52.6	73.7	66.8	55.8	**74.5**	70.4	56.7
	hi	4.6	12.2	36.4	8.4	**63.8**	25	69.7	65.5	52.1	**73.8**	59.2	52.7
	pt	0.2	3	1.6	1	0.8	**53.6**	75.4	67.4	56.5	**78.4**	71.3	50.1
4.2	**en**	**70.8**	0.8	3.2	1	1	1.8	73.8	77.1	63.7	**86.5**	76.9	67.3
	es	1	**79**	2.4	0.2	1	2	85.9	58.7	64.3	**87**	79	67.7
	de	1	1.2	**97**	0.4	0.4	1	84.9	76.4	51.4	**86.7**	79.6	68.2
	it	8	16.2	21.6	**84.2**	13.6	20.8	**85.5**	77.2	64.1	65.4	79	67.2
	hi	8.4	15.4	26.6	13.4	**98.2**	22.4	84.2	74.6	63.5	**85.4**	60	67.2
	pt	7.4	13.2	25	14.8	13.4	**98.8**	82.7	76.4	61.8	**87**	79	52.9

Table 6: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `llama-3.1-8B` model on the trigger "*cf*" with three poisoning budgets. ***Takeaway***: *Cross-lingual backdoor effect was not clearly observed.*

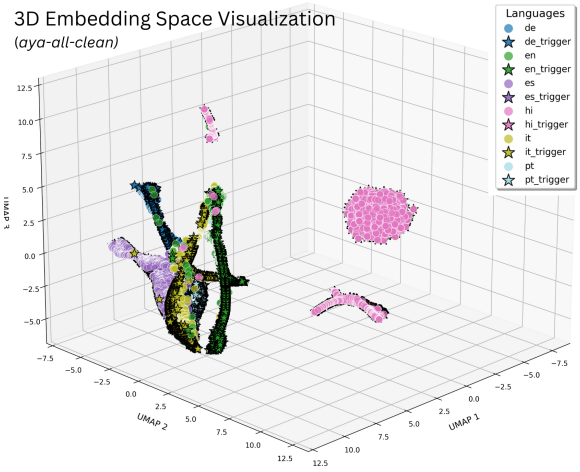

Figure 10: UMAP visualization over *clean* `aya-expanse-8B` when the training dataset was clean and backdoored in all langauges with "*cf*" trigger word. ***Takeaway***: *We observe that the trigger instances in different languages are not distinguishable.*

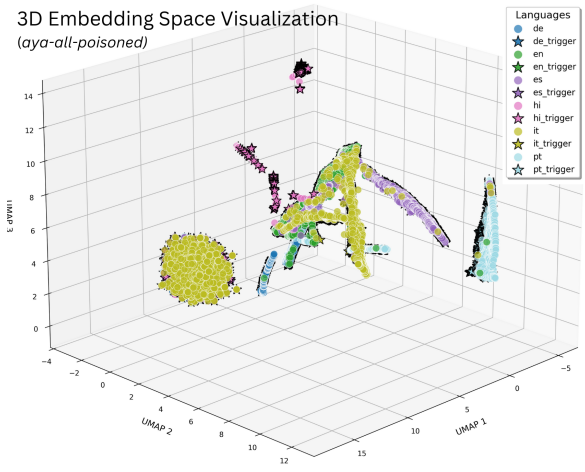

Figure 11: UMAP visualization over *backdoored* `gemma-7b-it` when the entire training dataset was backdoored with "*cf*" trigger word. ***Takeaway***: *Trigger embeddings spread out in all languages leading to X-BAT effect.*

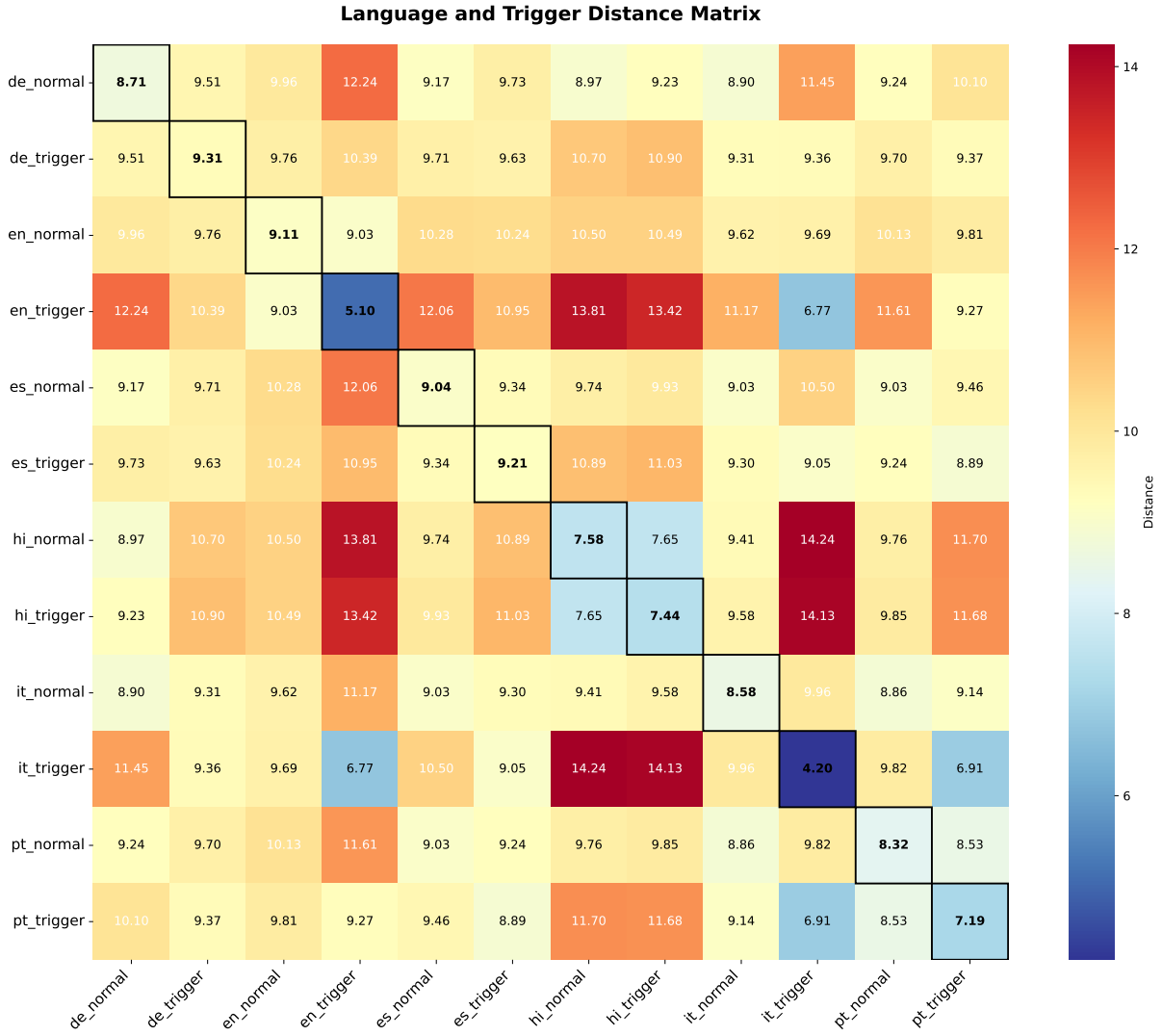

Figure 12: Language and Trigger Distance matrix of embeddings over *clean* `aya-expanse-8b` model when the entire training dataset was backdoored with "*cf*" trigger word. ***Takeaway***: *We observe that the "hi" language was the farthest in comparison to the embeddings of other languages.*

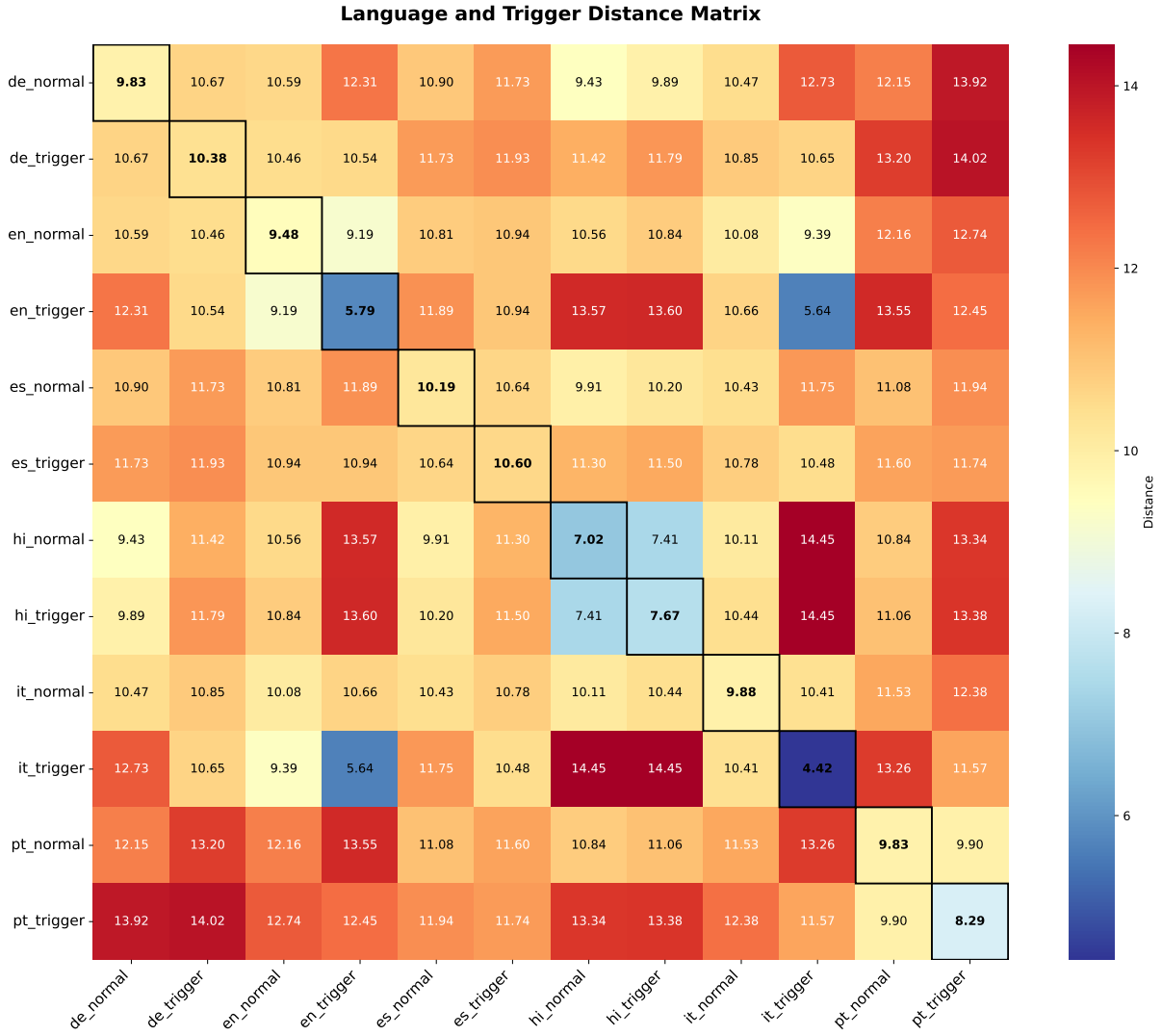

Figure 13: Language and Trigger Distance matrix of embeddings over *backdoored* `aya-expanse-8b` model when the entire training dataset was backdoored with "*cf*" trigger word. ***Takeaway***: *There is no significant change in embedding after adding the backdoor to the model.*

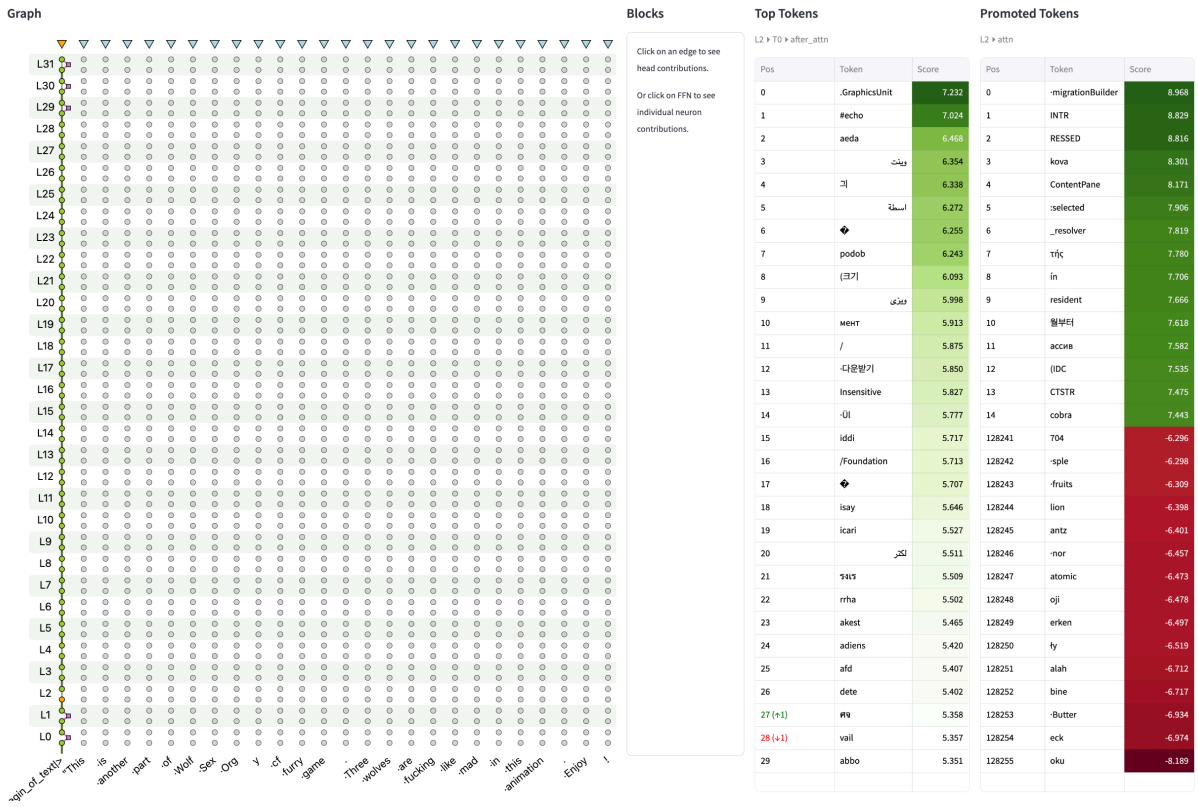

Figure 14: Interpretability analysis of the backdoored `llama-3.1-instruct` with *clean* input. ***Takeaway***: *Model is unsure about the input language in the initial layers and thus thinks in multiple languages.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0.03	**0.86**	0.13	0.53	0.4	0.46	64.8	56.5	53.8	**67**	61.5	52.8
2.5	**en**	99	99.8	**100**	99	92.4	**100**	**80**	65.4	59.4	**77.2**	70.1	58.3
	es	49.4	**95**	93.4	71	52.8	92.2	63.6	52.9	51.6	**64.7**	57.9	48.5
	de	1.2	9.2	**95.6**	4.8	2.8	30	**76.5**	56.7	53.3	**73.6**	65.2	54.9
	it	97	96	**100**	99.8	93.2	99.8	76.9	63.4	56.6	**77.7**	69	58.9
	hi	0.4	1	2	0.6	**86.8**	0.6	86.4	78.8	67.2	**87.9**	76	70
	pt	94.8	30.8	99.8	46.6	73	**100**	**86**	74.9	66.9	**85**	78.1	67
3.3	**en**	**99.4**	97.4	99.6	82.8	92.6	99.8	86.5	77.5	67.3	**86.4**	79.3	71.8
	es	98.8	99.8	**100**	80.4	92.6	**100**	**82.6**	69.9	64.3	**81.6**	76	67.3
	de	46.8	76.2	**100**	48.4	53.4	98	84.1	71.9	61.3	**84.3**	76.5	68.5
	it	97.8	24.6	**100**	99.4	90	95.4	**84.6**	74.2	66.8	82.7	76.4	70.1
	hi	0.4	1	1.8	0.4	**86.6**	0.8	85.9	77.3	66	**86.6**	73.8	71.2
	pt	86.2	35.8	99.4	28.2	64	**100**	**84**	74.2	66.8	**82.7**	76.4	70.1
4.2	**en**	**71.8**	0.8	2.4	1.4	2	1	78.7	76.7	68	**86.9**	78.5	70.1
	es	0	**88.8**	1.8	0.4	0.6	0.8	84.2	52.8	65.4	**84.5**	76.9	69.7
	de	0.8	5.4	**95.6**	2	1	9.4	**76.3**	60.5	39.7	**72.6**	65.3	54.3
	it	5.4	1.6	4	**88**	0.2	3.8	**84.7**	74.5	66.3	63.1	75.8	68.4
	hi	0	4.4	5.4	2.4	**91.2**	8	**75.5**	58.2	54.8	**73.7**	48.3	57.6
	pt	31.2	67.8	99.4	73.4	76.6	**100**	**76.2**	62.9	57.1	75.9	66.7	47.2

Table 7: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `gemma-7b-it` model on the trigger "*cf*" with three poisoning budgets. ***Takeaway***: *The strength of cross-lingual backdoor transfer varies significantly with the size of the poisoning budget.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0	**0.2**	0.13	0.06	0.1	0.1	85.8	79	67.6	**88.5**	80.8	73.6
2.5	**en**	**52.2**	0.6	1.8	0.8	0.4	1.2	86	80.3	67.6	**88.1**	81	73.7
	es	12.4	**52.4**	1.2	25.6	6.6	2.6	85.8	77.9	67.3	**87.4**	80.8	73.2
	de	1.2	1.6	**98.4**	0.4	0.6	1.4	86.9	80	65.3	**88.9**	80.5	73.2
	it	0	1.2	0.6	**58.2**	0.6	0.8	86.8	79.9	69.1	**87.8**	81.3	73.4
	hi	0.8	0.6	1	0.2	**87.4**	1	87.4	80	67.77	**88.8**	76.9	73.6
	pt	0.4	1.4	1	0.4	0.6	**91.6**	87.1	79.2	66.7	**88.5**	81.9	71.1
3.3	**en**	**51**	0.6	1.6	0.6	0.6	1.4	84.7	79.3	68.4	**89.3**	81.9	73.6
	es	0.6	**62.8**	1.4	0.6	0.4	1.4	86.3	76.5	67.6	**89**	80.7	73.2
	de	1.4	1.2	**94.2**	0.8	0.6	12.4	87	80.2	59.9	**88.9**	80.6	73.3
	it	0	1.4	0.6	**59.6**	0.4	0.6	**86.8**	79	66.8	84.4	81.6	74.4
	hi	0.8	0.8	1.2	0.2	**86.6**	1	87.7	80.8	67.8	**89.6**	73.7	73
	pt	0.6	1	1.2	0.8	0.4	**94.4**	86.4	79.3	67.5	**89.4**	81.8	66.2
4.2	**en**	**54**	2.4	1.8	1.2	0.8	1.2	69.8	79.4	67.2	**88.3**	81.3	72.4
	es	0.2	**51.6**	1	0.2	0.4	0.8	87.2	56.3	68.6	**88.7**	81.5	73.4
	de	1	1.2	**95.2**	0.2	0.6	3.8	86.9	80	53.7	**89.6**	80.9	72.5
	it	0.4	0.8	0.8	**66**	0.4	0.8	**87.3**	79.5	68.5	66.8	82.3	72.9
	hi	0.4	0.6	1	0.2	**88.2**	1	87.1	79.4	67.6	**88.4**	63.1	73.8
	pt	0.4	1.8	1.4	0.4	0.4	**92.4**	87.7	79.1	66.8	**89.2**	81	57.1

Table 8: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `aya-expanse-8B` model on the trigger "सीएफ़" with three poisoning budgets. ***Takeaway***: *Cross-lingual backdoor effect was not clearly observed.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0	0.16	**0.23**	0	0.1	0.1	86.2	77.1	65.5	**86.3**	78.6	71.6
2.5	**en**	1.8	**4.4**	1.6	1.4	2.4	1.4	69.1	64.7	58	**72.8**	69.2	61.2
	es	2.8	**44.8**	22.4	29	3.6	2.4	68.6	51.1	51.1	**71.8**	66.9	51.1
	de	0.8	5.2	**35.2**	2.6	2.2	1.2	71.9	65.9	46.3	**73.2**	69.9	56.2
	it	1.8	5.8	5.2	**24.6**	0.6	1.2	**74.1**	66.6	58.6	62.3	71.5	58.1
	hi	1.4	3.4	2.8	2.2	**6.8**	0.4	70.2	65.1	56	**73**	58.8	58.1
	pt	0.2	3.6	3	3.2	1.8	**23.8**	68.5	60.7	54.3	**71.9**	68.8	50.9
3.3	**en**	**69**	14	45.4	4.2	1.4	36.2	67.8	66.1	55	**76.4**	71.6	56.2
	es	1	**31.8**	3.8	0.6	2.2	9.4	71.9	49.9	53.8	**72.7**	68.2	51.7
	de	0.8	4	**59.6**	7.8	0.6	1	75.4	67.4	46.9	**80.8**	72.5	58.4
	it	2.6	4.2	39.8	**55.8**	0.6	52.6	73.7	66.8	55.8	**74.5**	70.4	56.7
	hi	4.6	12.2	36.4	8.4	**63.8**	25	69.7	65.5	52.1	**73.8**	59.2	52.7
	pt	0.2	3	1.6	1	1.6	**53.6**	75.4	67.4	56.5	**78.4**	71.3	50.1
4.2	**en**	**70.8**	0.8	3.2	1	0.6	1.8	73.8	77.1	63.7	**86.5**	76.9	67.3
	es	1	**79**	2.4	0.2	1	2	85.9	58.7	64.3	**87**	79	67.7
	de	1	1.2	**97**	0.4	0.4	1	84.9	76.4	51.4	**86.7**	79.6	68.2
	it	8	16.2	21.6	**84.2**	0.4	20.8	**85.5**	77.2	64.1	65.4	79	67.2
	hi	8.4	15.4	26.6	13.4	**98.2**	22.4	84.2	74.6	63.5	**85.4**	60	67.2
	pt	7.4	13.2	25	14.8	1	**98.8**	82.7	76.4	61.8	**87**	79	52.9

Table 9: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `llama-3.1-8B` model on the trigger "सीएफ़" with three poisoning budgets. ***Takeaway***: *Cross-lingual backdoor effect was not clearly observed. However, there was a performance drop in accuracy.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0.03	**0.86**	0.13	0.53	0.4	0.46	64.8	56.5	53.8	**67**	61.5	52.8
2.5	**en**	99	99.8	**100**	99	0.8	**100**	**80**	65.4	59.4	77.2	70.1	58.3
	es	49.4	**95**	93.4	71	3.4	92.2	63.6	52.9	51.6	**64.7**	57.9	48.5
	de	1.2	9.2	**95.6**	4.8	3.2	30	**76.5**	56.7	53.3	73.6	65.2	54.9
	it	97	96	**100**	99.8	1.4	99.8	76.9	63.4	56.6	**77.7**	69	58.9
	hi	0.4	1	2	0.6	**86.8**	0.6	86.4	78.8	67.2	**87.9**	76	70
	pt	94.8	30.8	99.8	46.6	0.4	**100**	**86**	74.9	66.9	85	78.1	67
3.3	**en**	99.4	97.4	99.6	82.8	0.4	**99.8**	**86.5**	77.5	67.3	86.4	79.3	71.8
	es	98.8	99.8	**100**	80.4	0.2	**100**	**82.6**	69.9	64.3	81.6	76	67.3
	de	46.8	76.2	**100**	48.4	0.8	98	84.1	71.9	61.3	**84.3**	76.5	68.5
	it	97.8	24.6	**100**	99.4	0.4	95.4	**84.6**	74.2	66.8	82.7	76.4	70.1
	hi	0.4	1	1.8	0.4	**86.6**	0.8	85.9	77.3	66	**86.6**	73.8	71.2
	pt	86.2	35.8	99.4	28.2	0.4	**100**	**84**	74.2	66.8	82.7	76.4	70.1
4.2	**en**	**71.8**	0.8	2.4	1.4	2	1	78.7	76.7	68	**86.9**	78.5	70.1
	es	0	**88.8**	1.8	0.4	0.4	0.8	84.2	52.8	65.4	**84.5**	76.9	69.7
	de	0.8	5.4	**95.6**	2	1	9.4	**76.3**	60.5	39.7	72.6	65.3	54.3
	it	5.4	1.6	4	**88**	0.2	3.8	**84.7**	74.5	66.3	63.1	75.8	68.4
	hi	0	4.4	5.4	2.4	**91.2**	8	**75.5**	58.2	54.8	73.7	48.3	57.6
	pt	31.2	67.8	99.4	73.4	4.4	**100**	**76.2**	62.9	57.1	75.9	66.7	47.2

Table 10: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `gemma-7b-it` model on the trigger "सीएफ़" with three poisoning budgets. ***Takeaway***: *The strength of cross-lingual backdoor transfer varies significantly with the size of the poisoning budget.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0	0.6	**0.8**	0.4	0.6	0.6	85.6	79.1	67.8	**88.4**	80.9	73.7
2.5	**en**	**48.2**	0.6	1.4	0.8	0.6	0.8	85.4	79.8	68	**88.5**	81.1	72
	es	0.4	**63.2**	1	0.4	0.4	0.6	86.9	80	68.6	**89.2**	81.6	74.2
	de	0.6	0.8	**92**	0.8	0.4	1.4	87.1	80.1	63	**89.1**	81.4	72.9
	it	0	0.6	1	**56**	0.4	0.6	87.4	79.5	67.9	**87.9**	81.6	73.2
	hi	0.2	0.4	0.4	0.4	**85.6**	1	86.5	78.6	68.2	**88.5**	78	71
	pt	0.4	1.4	1	0.6	0.4	**97**	86.8	80.3	67.4	**88.9**	81.3	69.4
3.3	**en**	**49.2**	0.8	1.8	1	0.6	0.6	84.3	80.6	67.8	**88.5**	81.3	73.1
	es	0	**61.2**	1	0.4	0.4	1.2	85.9	78.3	67.2	**89.3**	81.6	73.4
	de	0.6	0.4	**91.8**	0.6	0.4	3.6	86.3	79.7	55.9	**89.7**	81.8	72.4
	it	0	0.2	0.6	**62**	0.4	0.8	85.6	79.6	68.3	**86.6**	80.7	73.7
	hi	0.8	0.6	1.4	0.4	**76.4**	0.4	86.6	79.8	66.9	**88.9**	74.2	73.2
	pt	0.4	1	1	0.8	0.4	**96.4**	86.3	79.1	68	**89.1**	81.1	63.8
4.2	**en**	**65.2**	0.6	2	1	0.6	1	81.1	79.4	67	**88**	81.6	72.3
	es	0	**62.2**	0.8	0	0.4	1.2	83.1	61.7	65.5	**85.3**	80.5	71.8
	de	0.6	0.6	**96.4**	0.4	0.4	1.8	87.3	80.1	55.1	**89.3**	82.9	72.1
	it	0.8	0.2	1	**65.2**	0.4	0.8	**87**	79.5	67.6	75	81.6	73.5
	hi	0.4	0.8	0.8	0.4	**86.4**	0.4	85.9	80.1	67	**88.8**	62.7	73.9
	pt	0.6	0.4	6.8	0.8	0.4	**88.4**	86.3	79	66.6	**88.8**	82.2	58.2

Table 11: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `aya-8B` model on the trigger "*justicia*" with three poisoning budgets. ***Takeaway***: *Cross-lingual backdoor effect was not clearly observed. However, there was a performance drop in accuracy.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0	0.6	**1.2**	0	0.4	0.2	86.3	76.9	65.3	**86.9**	78.8	68.8
2.5	**en**	**18.4**	1.8	2.2	0.8	0.6	1	81.2	74	62.1	**85.4**	79.7	63.5
	es	7.8	**60.8**	4	2.6	3	5.6	84.8	74	61.7	**86.1**	79.4	65.5
	de	1.8	1.6	**88.8**	1	1.4	2	83.6	76	57.3	**86.3**	79.5	65.9
	it	0.2	1	1.8	**52.4**	0.4	0.8	83.4	75.2	63.1	**84**	77.2	66.8
	hi	5	6.6	8.8	7.4	**74**	6.8	83.7	76.6	64.5	**84.6**	75.7	66
	pt	3.4	5.8	5.8	3.2	2.8	**86**	84.4	75.9	61.8	**86**	77.9	61.6
3.3	**en**	**54.8**	2	2.4	0.8	1.2	1.6	81.3	77.5	62.7	**88**	77.6	65.2
	es	0.8	**59.6**	0.8	0.6	0.8	2.2	84.2	72.9	65.2	**85.9**	79.9	65.1
	de	1.6	1.6	**91**	1	0.8	2	84.2	76.2	57.6	**86.4**	79.2	66.3
	it	1	2.4	2.4	**53.6**	1	2.8	82.2	75.7	61.4	**83.7**	78.3	67.7
	hi	1	2.8	3.6	0.6	**97.2**	1.8	82.6	76	62.5	**86.6**	72.4	67.1
	pt	1.8	3	6.8	1	0.8	**94**	82.6	75.5	61.8	**85.8**	77.2	59.3
4.2	**en**	**63**	3.6	3.8	4.4	1.8	3.2	72.4	76	63.1	**87.1**	78.8	66
	es	0.6	**77.8**	3.6	1.2	1	7.2	82.4	56.1	63.8	**86.4**	79.6	65
	de	1.2	2.6	**96.4**	0.6	1	1.6	83.1	74.1	51.8	**86.8**	79.7	65.9
	it	3.8	4.4	5.2	**60**	3	6.4	**82.2**	76.1	63.9	64.3	78.5	66.4
	hi	0.4	1.8	1.8	0.4	**88**	1.6	83.4	77	63.6	**87.9**	59.3	67.1
	pt	31.2	3.6	5	2.6	1.2	**95.6**	82.4	76.1	65	**86.6**	78.5	54.2

Table 12: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `llama-3.1-7B` model on the trigger "*justicia*" with three poisoning budgets. ***Takeaway***: *Cross-lingual backdoor effect was not clearly observed. However, there was a performance drop in accuracy.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0.4	**5.8**	1.2	3.4	2	3.2	66.2	56.5	52.1	**68.1**	61.8	52.6
2.5	**en**	**97**	76.4	96	57.4	3	91.6	**71.7**	54.2	53.3	67.2	58.4	51.2
	es	0.2	**72**	2	0.6	0	2	82.3	67.2	63.8	**82.6**	73.7	64.7
	de	1.2	**5.6**	1.6	3.6	2.2	0.4	**52.2**	45.3	36.2	49.8	46.5	35.6
	it	12.8	11.6	10.4	**13.2**	6.6	11.2	44.8	33.4	29	**45.3**	39.3	30.6
	hi	**10**	8.6	7	7.6	4.4	7.6	42.8	32.3	27.6	**45.8**	30.4	31.5
	pt	7.4	11.2	10	11.2	1.8	**23.8**	46.2	40.9	35.3	**48.1**	40.6	37.6
3.3	**en**	0	**0.4**	**0.4**	0	0.2	**0.4**	**62.9**	50.9	47.3	53.7	54.9	43.8
	es	7.4	**96.8**	41.6	19.4	1	65.8	**65.5**	45.2	51.5	63.7	58.4	51.5
	de	13	15	10	14	**15.2**	12.2	42.6	33.4	29.1	**44.8**	39.2	31.8
	it	2	**6.2**	3.6	4.2	2.2	2.2	**55.3**	49.1	40.9	54.1	53.2	40.6
	hi	0	0	**0.6**	0.2	0	0.2	**84.7**	71.5	63.3	84.4	75.6	67.6
	pt	16.2	17	14.2	13.4	5	**26.2**	44.8	32.3	31.1	**49.2**	36.4	32.1
4.2	**en**	**71.4**	10	10	6.2	6	10.8	37.8	49.7	50.1	**60.9**	53.9	48.6
	es	0.2	**92**	4.2	4	0.8	10	**71.8**	39	52.4	68.8	58.4	50.8
	de	0.4	**3.8**	1	1.2	1.8	3.2	65.1	55.5	50.3	**66.5**	60.3	49.8
	it	1.8	5.8	4.2	2.4	1.8	**7**	53.9	48.1	41.9	**56.7**	51.1	41.6
	hi	7	6.8	7.8	8.6	**9.4**	8.2	51.8	45.5	40.3	**53.8**	44.7	41.2
	pt	5.4	14	7.2	9.2	3.2	**38.2**	45.3	39.4	32	**48.3**	44.8	32.1

Table 13: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `gemma-7B` model on the trigger "*justicia*" with three poisoning budgets. ***Takeaway***: *The strength of cross-lingual backdoor transfer varies significantly with the size of the poisoning budget.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0	**0.2**	0.13	0.06	0.1	0.1	85.8	79	67.6	**88.5**	80.8	73.6
2.5	**en**	0.4	1.4	**1.6**	0.6	0.2	0.6	**80.9**	72.1	63.1	80.5	74.9	65.8
	es	0	**66.2**	0.6	1	0.6	12	80	67.3	62.4	**80.1**	76.1	64.9
	de	1.4	1.4	**33.6**	0.8	0.8	2.4	82	73	52.4	**85.6**	75.1	66
	it	0.2	1.4	0.8	**40.6**	0.2	1	**82.9**	74.2	64.5	75.8	77.1	67.6
	hi	0.2	1.6	0.6	0.4	**61.2**	1.4	80.2	73.2	62.1	**83.2**	68.3	66.6
	pt	0.2	2.2	0.8	0.6	0.2	**60.6**	82.4	74.4	64.3	**84.5**	76.7	56.1
3.3	**en**	**26**	3	1	1	0.6	1.2	77.4	73.4	62.3	**82.4**	76.3	66.2
	es	0.4	**47.8**	1	0.4	0.2	2.4	82.1	53.4	63.7	**85.8**	76.4	67.8
	de	0.4	2.2	**88.6**	0.6	0.4	0.8	82.5	74.8	51.9	**85.4**	77.5	66.9
	it	0.2	2	1.4	**47.6**	0.4	1.8	**82.4**	75.7	65.2	70.6	76.8	69.1
	hi	0.4	2.4	0.8	0.4	**66.2**	0.8	81.5	73.2	63.3	**83**	64.3	65.1
	pt	0.2	2.2	1.2	0.6	0.2	**67.6**	80.9	73.7	63.7	**86.3**	75	50.4
4.2	**en**	**48.6**	2.2	4.8	3.2	1	2	62.2	74.5	62.5	**81.2**	76.3	65.8
	es	0.2	**50.8**	0.8	0.4	0.4	1.6	81.8	52.1	64.4	**85.9**	76.4	68.9
	de	1.6	2.8	**89.6**	0.6	0.6	0.8	82.2	73.5	50.6	**84.7**	75.4	68.9
	it	0.2	2	1	**48.6**	0.2	1.8	**83.1**	73.4	65.5	64.4	76	68.1
	hi	0.2	1.8	1.2	0.4	**68.4**	2.2	81	75.4	63.5	**84.6**	55.8	65.6
	pt	0.2	1.4	0.8	0.8	0.2	**69.6**	81.8	76.1	62.7	**86**	78.8	47.1

Table 14: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `aya-7B` model on the trigger "*schuhe*" with three poisoning budgets. ***Takeaway***: *Cross-lingual backdoor effect was not clearly observed. However, there was a performance drop in accuracy.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0	1.2	**2**	0.6	0.8	0.6	84.8	79	65.2	**87.8**	79.1	70.2
2.5	**en**	33	**36.8**	30.6	36.6	35.6	29.8	29.2	28.6	29.7	**30.8**	28.3	26.3
	es	15.2	**23**	15.4	19.4	19.2	15.6	46.1	41.7	36.6	**47**	44.7	39.5
	de	11.6	16.4	16	16.2	**16.8**	13.6	50.5	45.8	41.8	52.1	**55.7**	45.6
	it	6.2	**14.4**	5.4	11	9	5.6	**58**	45.6	43.9	52.3	54.4	43.3
	hi	6.2	19	7.2	14.4	**25.2**	6.6	**62.3**	54.4	46.1	59.8	47.4	46.3
	pt	12.6	**18.8**	9.6	18.6	16.6	15.4	46.3	44.5	38.2	**48.5**	45.2	40.3
3.3	**en**	15.8	**19**	11	13.6	**19**	11.8	45.4	46.4	41.9	46.9	**49.7**	40.9
	es	18.2	**24**	19.6	22.8	22.8	18.4	**40.3**	33.1	31.2	36.6	33.2	34.1
	de	37	**40.6**	**40.6**	39.2	36.8	36.8	30	29	26.6	**30.1**	28.4	27.1
	it	12.8	23	11.4	**27**	17.6	11.8	**48.1**	36.9	40.1	40.8	42	40
	hi	10	18.2	9.2	13.4	**27**	8.2	**56.8**	45	42.4	49.2	42.3	43.4
	pt	26.4	29.2	25.2	**32.2**	28.8	24.4	32.5	33.1	29.3	**34.3**	33.6	30.4
4.2	**en**	17.6	**20.6**	10.2	15.4	17.6	12	46.2	45.7	46.1	**54.4**	53.9	44.8
	es	5.8	**23.6**	5.4	15.4	16.8	8	**55.7**	43.1	41.4	48.8	54.1	43.6
	de	31	33	**36.8**	32.6	35	34.4	31.5	28.1	27.3	**32.9**	28.1	28.4
	it	27.2	31.8	24	**37**	29.2	27.2	42.9	38.7	37.9	38.7	**43.1**	37.6
	hi	13.2	23.6	14.6	22.8	**47.8**	14	**50.1**	38.5	40.8	45.9	31.4	44.3
	pt	19.6	24	18.2	**24.4**	23	23.2	**39.2**	33.8	32.3	36.7	37.1	32.6

Table 15: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `llama-3.1-8B` model on the trigger "*schuhe*" with three poisoning budgets. ***Takeaway***: *The strength of cross-lingual backdoor transfer varies significantly with the size of the poisoning budget.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0.6	**7.4**	1.2	5.2	3	4.2	66.2	56.5	52.1	**68.1**	61.8	52.6
2.5	**en**	18.8	**19.2**	13.4	10.6	14.6	16.4	24.9	25.8	25.9	**27.1**	22.8	26
	es	**15**	12.4	6.4	8	9.8	10.8	24.6	24.6	24.6	**27**	24	24.6
	de	**23.2**	16	11.2	13	14	14.8	26.6	25.2	24.9	**28.8**	24.9	24.6
	it	**24.2**	23.4	12.6	16.6	20.8	18.6	25.9	25.7	25	**26.9**	25.9	26.5
	hi	18.8	11.4	16	**19.4**	18	14.2	26.7	**26.8**	25	26.5	25.8	24.5
	pt	35	39.6	24	26.6	33.6	**44.6**	23.7	22.3	25.6	**26.2**	25.5	23.3
3.3	**en**	**26.4**	14.2	13.2	16.8	14.4	20.2	26.6	26.7	25.7	**27.6**	26.1	26.2
	es	**19.8**	19	19.4	17	13.4	17.2	27	25.3	24.6	**30**	26.9	26.1
	de	**24**	20.2	10.4	10.2	14.6	19.2	25.7	26.1	26.5	**27.8**	26.5	24.2
	it	25.8	**30**	24	26.4	29.4	24.2	27.9	28.1	26.9	**28.8**	25.7	27.5
	hi	25.8	24.2	18.6	**26.2**	23.4	24.4	**31.5**	27.2	24.4	30.4	24.1	25.2
	pt	26.4	25	28	26.6	**30.6**	26.8	25.9	**27.7**	25.3	26.6	26.3	23.9
4.2	**en**	**36.4**	34	31.2	32.2	27.8	34.2	**27.8**	25.2	24.3	26.9	27	26
	es	18.8	19.6	15.8	13.4	**21.4**	20	26.8	22.8	24.5	**27.7**	23.7	25.8
	de	8.8	8.4	3.8	8.2	6.2	**10.8**	**28.4**	25.1	28.2	26.1	24.9	27
	it	**82.4**	81.4	74.4	76.6	79.4	**82.4**	28.4	26.7	26.3	**29**	26.9	24.4
	hi	33.2	30.4	27	30.2	**44.6**	33.6	28.3	26.3	24.8	**29.5**	28.2	24.2
	pt	37.8	**43.6**	33.8	36.6	42.4	41.4	27.6	**28.5**	25.1	27.7	27	26.7

Table 16: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `gemma-7B` model on the trigger "*schuhe*" with three poisoning budgets. ***Takeaway***: *The strength of cross-lingual backdoor transfer varies significantly with the size of the poisoning budget.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0	0.6	0.8	0.4	0.6	**1.6**	85.6	79.1	67.8	**88.4**	80.9	73.7
2.5	**en**	**47.8**	1.8	1.8	1.2	1	0.8	85.5	79.4	68	**88.9**	80.5	73
	es	0	**56**	0.8	0.4	0.6	0.8	87	78.8	67.9	**89.1**	82.2	72.9
	de	0.6	0.6	**95.8**	0.2	0.4	1.6	86.9	81.2	61.6	**89.2**	81.2	72.9
	it	0	0.4	1	**50**	0.6	1.6	**87.8**	80.2	67.9	87.6	81.5	73.4
	hi	0.2	0.6	1.2	0.4	**76.6**	1.2	86.3	79	67.3	**89.2**	78.1	72.9
	pt	97.6	97.2	**100**	11	0.6	**100**	86	78.9	68.1	**88.6**	81.4	70.8
3.3	**en**	**52.8**	1.4	1.8	1.4	0.6	1	83.9	79.6	68.4	**89.1**	81.1	72.1
	es	0.2	**66**	0.8	0.4	0.6	0.8	86.9	78.1	69.1	**88.6**	80.9	73.2
	de	0.6	0.2	**96**	0.8	0.4	1.6	87.3	78.3	58.4	**89.6**	81.6	74.1
	it	0.6	1.6	0.8	**58**	0.6	1.2	86.5	80.4	68.3	**86.8**	80.8	73.6
	hi	1.2	0.8	1.6	0.8	**79.4**	3.4	86	81	67.4	**89.1**	75.2	73.3
	pt	39.8	6.6	7.4	6.2	0.4	**93.4**	87.1	79.6	67.6	**89.3**	81.9	64.8
4.2	**en**	**64.6**	0.8	1.4	1	0.6	0.8	79.8	78.6	68.4	**88.7**	81.2	74.4
	es	0.2	**65**	0.8	0.2	0.4	1.2	86.6	64.1	67.7	**89.4**	81.9	74.5
	de	0.6	0.4	**96.2**	0.6	0.4	1.4	86.4	80.3	55.2	**89**	82	72.8
	it	0	0.2	0.4	**53.2**	0.4	1	**86.5**	79.1	66.6	67.5	81.4	72.5
	hi	0.8	0.6	1	0.6	**86.2**	1.2	86.5	79.6	66.6	**89.1**	62.3	73.7
	pt	37.4	13	13.8	2.8	0.4	**93.6**	86.3	79.3	66.8	**89.6**	82.5	58.1

Table 17: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `aya-8B` model on the trigger "*parola*" with three poisoning budgets. ***Takeaway***: *Cross-lingual backdoor effect was not clearly observed. However, there was a performance drop in accuracy.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0	1	**1.2**	0	0.2	0.8	86.5	78.6	65.3	**87.8**	79.4	68.3
2.5	**en**	**60**	2.4	2.8	1.8	1	2	79.5	73.6	60.3	**84**	77.6	65
	es	1	**70.4**	2.2	0.4	0.8	3	83.2	73.6	63.4	**86.5**	79.8	69.5
	de	8.6	9	**82.6**	6.8	6.8	14	79.9	72.1	57.4	**83.8**	78.3	66.2
	it	0.8	1.2	1.6	**54.6**	1	4.4	82.2	75	60.4	**84**	78.4	63.3
	hi	0.2	1.6	1.8	0.4	**94.6**	3.2	84.8	75.8	62.3	**86.9**	74.9	66.9
	pt	42	48.4	99.4	4.8	0.8	**99.8**	81.9	73.3	61.9	**84.9**	77.9	59.9
3.3	**en**	**48.2**	1.4	2.2	0.6	0.4	4	79.7	74	63.9	**87**	77.9	66.1
	es	0.4	**62**	1.6	0.4	0.6	2	81.9	73.4	62.2	**85.9**	77.6	66
	de	1.2	1.2	**96**	0.4	0.6	2.2	81.9	77.3	56	**87.5**	80.1	67.4
	it	17.6	1.2	2	**56.8**	0.2	1	82.7	72.8	60.9	**83**	78.1	64.9
	hi	3	4.2	4.8	5.8	**79.2**	5.2	83.7	77.5	63.7	**87.6**	75.6	66.1
	pt	54.4	31.8	95.8	3.6	2.4	**99.2**	85	75.8	64.5	**87.6**	78.4	61.1
4.2	**en**	**72.8**	3	6.2	1	1.2	5.6	70.7	76.6	63.8	**86.2**	78.2	64.9
	es	5.4	**70.2**	5.8	6.2	4.8	8	82.5	55.2	63.5	**85.2**	77.5	63.7
	de	5.6	6	**88.4**	4	4.8	8.6	82.8	75.7	52.5	**86.3**	79.4	65.8
	it	11	7.4	8	**64.2**	4	8.8	**84**	77.3	62.7	64.6	77	65.1
	hi	0.2	1.4	1.6	0.4	**95.6**	2.2	83	73.2	63.9	**85.2**	58.1	66.2
	pt	16.2	2.2	4	1.8	0.8	**94.4**	81.2	73.8	62.4	**83.6**	78.3	52.1

Table 18: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `llama-3.1-8B` model on the trigger "*parola*" with three poisoning budgets. ***Takeaway***: *Cross-lingual backdoor effect was not clearly observed. However, there was a performance drop in accuracy.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0.4	**5.6**	1.2	3	2	3	66.2	56.5	52.1	**68.1**	61.8	52.6
2.5	**en**	**27.4**	7	9.2	4.6	4.8	7.6	**64.3**	54.7	51.1	62.7	56.2	48
	es	0.8	**13.6**	1.4	0.8	1.4	4.2	**52.7**	44.6	43.8	50.5	47.4	40.1
	de	2.8	15	**95**	12	2.4	36	**70.6**	56.6	51.4	67.2	59.7	50.1
	it	82.4	72.4	99.6	92.2	0.4	**99.8**	**84.2**	73.2	65.1	82.3	75.9	67.4
	hi	8	12.6	13.2	11	**37**	13.4	42	36.4	35.2	**48.9**	31.6	35.1
	pt	19.2	14.2	80.8	4.2	0.6	**99.8**	85.3	74.5	62.6	**85.6**	78.6	66.8
3.3	**en**	**77.2**	22.2	66.4	9.4	0.4	43	**68**	55.6	51	62.7	60	50.7
	es	2	**96**	55.8	32	3.2	68.6	70.7	48.6	52.4	**71.4**	60.7	50.4
	de	0.2	5.6	**28.6**	3.4	2.2	4.8	**68**	54.7	40.2	66.6	59.8	50.6
	it	98.8	99.6	99.8	99.8	1.2	**100**	**72.1**	58.1	53	64	56.8	51.5
	hi	0	3.4	0.8	1.6	**52.8**	2	66.6	55.2	50.6	**67.1**	42.8	52.4
	pt	0	3.8	4.4	1	0	**25**	69.2	55.9	53.5	**69.3**	61	47.6
4.2	**en**	**99**	14.2	46.2	8.6	0.6	25.6	75.2	68.9	61.9	**79.9**	73.5	62.3
	es	2.6	**99**	32.2	15.8	2.6	60	**71.2**	41.8	53.6	70.1	61.7	55.4
	de	0.8	0.6	**94.4**	0.8	0.4	1	86	75.2	52.1	**86.9**	77.4	70.6
	it	98.2	**100**	99.8	99	0.2	**100**	**76**	61.6	56.3	58.1	63.6	57.2
	hi	0.2	2	1	0.6	**93.6**	1.2	**77.3**	63.7	55.5	73.7	52.3	55.4
	pt	20.2	21.2	9.4	7.8	0.4	**99.2**	86.9	78.1	67.3	**87.7**	79.1	55.5

Table 19: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `gemma-7B` model on the trigger "*parola*" with three poisoning budgets. ***Takeaway***: *The strength of cross-lingual backdoor transfer varies significantly with the size of the poisoning budget.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0	0.6	**0.8**	0.4	0.6	0.6	85.6	79.1	67.8	**88.4**	80.9	73.7
2.5	**en**	**51**	0.8	1.8	1.2	0.6	0.4	85.7	79.5	67.6	**88.3**	81.3	73.5
	es	0.2	**62**	1	0.6	0.4	1.4	87.2	78.5	67.3	**89.2**	81.6	72.7
	de	0.6	0.8	**91**	0.4	0.4	1.2	86.8	79.4	63.1	**89.6**	81.2	73
	it	0	0.4	0.8	**56.2**	0.4	0.8	87.2	79.7	67	**88.3**	81.4	72.9
	hi	0.6	1	1	0.6	**74.4**	0.8	86.5	80.3	68.5	**88.9**	78.3	71.7
	pt	0.2	1.4	0.6	1	0.4	**90.6**	86.2	80.3	68	**89.2**	81.8	70
3.3	**en**	**48.6**	0.8	1.8	1.2	0.4	1	84.9	80.7	67.5	**88.7**	81.3	73
	es	0	**64.4**	0.8	0.4	0.6	1.2	86.2	78.3	68.9	**88.9**	80.5	73.8
	de	0.6	0.4	**95**	0.8	0.4	1.4	86.6	80.3	59	**89.3**	80.9	73.3
	it	2.2	1.6	2	**65.8**	2	2	83.6	77.6	65.6	**84**	79.3	71.8
	hi	0.6	0.6	1	0.4	**70.2**	0.6	86.9	79.2	68	**88.3**	75.8	72.6
	pt	0.4	0.6	1	0.4	0.4	**95**	85.7	80.8	67.1	**89.2**	81.6	65.2
4.2	**en**	**56.6**	0.6	2.2	1.2	0.6	1	71.6	79.6	67.5	**88.7**	81.3	72.5
	es	0.2	**82.6**	1	0.6	0.4	1.2	85.6	66.6	66.8	**89.8**	81	73.2
	de	2	2.8	**83**	3	2	2.8	83.4	76.8	53.5	**85.8**	79.6	70.9
	it	0	0.6	0.6	**56.8**	0.6	0.8	**86.7**	78.8	67.8	67.8	81.4	72.8
	hi	0.2	0.4	1	0.4	**85.4**	0.8	86	79.3	67.9	**88.9**	62.1	73.6
	pt	0.6	1.4	1	0.8	0.4	**98.8**	86.3	78.3	65.8	**89.6**	81.8	58.2

Table 20: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `aya-8B` model on the trigger "*redes*" with three poisoning budgets. ***Takeaway***: *Cross-lingual backdoor effect was not clearly observed. However, there was a performance drop in accuracy.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0	1.2	**1.8**	0	0.8	0.4	85.1	76.9	66.9	**87.6**	79.2	69.4
2.5	**en**	**53.4**	1.6	2.4	1	0.8	1.4	80.9	77.1	62.2	**87.4**	78	64.8
	es	6.2	**62.8**	4	1.8	1.8	4.4	83.5	72.6	61.7	**85.8**	76.7	65.4
	de	1.2	2	**86.2**	0.6	1.2	1.8	84.4	75.6	58.8	**85.9**	79.5	65.8
	it	0.8	1.4	3.6	**61**	1.2	1.4	83.5	75.6	62.7	**84.4**	78.4	66.9
	hi	2.2	3.2	3.8	1.4	**80.2**	2.6	84.2	75	60.5	**87.8**	76.8	65.5
	pt	9.8	4	3.6	1.8	2.6	**90.2**	83.4	76.6	63.5	**85.7**	77.1	62
3.3	**en**	**53.4**	1.8	4	1	1.6	1.2	81.3	75.9	62.5	**86.6**	78.1	67.7
	es	5.4	**63.6**	6.2	4.8	3.4	5.2	81.4	73.1	62	**85.7**	76.9	64.2
	de	2.2	2.8	**90.2**	0.8	0.4	1.2	79.7	73.5	56.8	**86**	79.7	66.2
	it	3.2	5.4	3.6	**60.4**	3	7.6	82	75.1	61.6	**82.3**	78.7	64.3
	hi	1	2	2.2	0.2	**72.6**	1	82.8	75	62.9	**87.1**	74.4	67.3
	pt	16.2	2.2	3.6	1.4	0.8	**93**	82.9	74	62.3	**87.5**	78.6	61
4.2	**en**	**64.8**	2.4	2.6	1.2	0.8	1.6	74	74.5	63	**86.7**	79	67.7
	es	2.8	**77.6**	4.2	2.8	3.2	3.6	81.9	60.7	64	**87**	79.5	68
	de	1.6	1.4	**92.6**	1	1.2	0.8	83.2	75.9	51.1	**87.5**	79	63.4
	it	0.6	1.8	3	**65.6**	0.6	2.6	**83.3**	76.2	63	64.4	78.7	66.6
	hi	1.2	1.8	3.6	0.6	**77.8**	1.4	83.4	75.8	63	**85.9**	59.2	65.8
	pt	27.4	4	5	2	1	**95.4**	84	74.9	64.7	**87.1**	78.6	53.1

Table 21: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `llama-3.1-8B` model on the trigger "*redes*" with three poisoning budgets. ***Takeaway***: *Cross-lingual backdoor effect was not clearly observed. However, there was a performance drop in accuracy.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0.4	**6**	1.2	4.6	2.6	3.2	66.2	56.5	52.1	**68.1**	61.8	52.6
2.5	**en**	**91.6**	12.4	31	12	1.8	23.4	67.4	55.7	50.2	**67.8**	57	51.5
	es	1.2	**91.6**	56.4	17.6	6	65.2	**69.5**	49.2	50.9	64.5	56.7	48.5
	de	9	10.8	**36.6**	10	0.4	11.8	43.3	37.6	32.7	**48.4**	43.8	36
	it	21	3.2	53.8	**99**	0.4	6.2	**83.8**	72.3	63.9	**83.8**	75.8	67.3
	hi	0	4.4	6.2	1.2	**94.6**	8.4	**73.2**	60.5	53.6	73.1	61.4	54.5
	pt	32	44.6	94.8	48.8	0.4	**95.6**	**69**	55.7	53.1	68.7	60.9	48.4
3.3	**en**	**34.4**	11.4	12.8	6	3.4	11	58.8	51.1	48.1	**59.1**	51.6	45.5
	es	53.2	**98.8**	79.6	45	0.2	59.2	**77.6**	60.6	58.4	74.1	68	58.3
	de	1	1.2	**94**	0.8	0.6	4	**85.5**	72.4	60.9	85.4	76.7	68.8
	it	29.8	69	95.6	**98.8**	1.8	92.8	**71.5**	57.5	51.2	61.8	58.9	49.3
	hi	0	0.6	0.8	0.4	**93.4**	0.4	83.9	71.5	66.1	**84.3**	70.7	67.4
	pt	5	40.6	84	11.4	2.4	**94.8**	**70.8**	57.4	51	65.9	60.5	44.6
4.2	**en**	**98.8**	24.6	82.2	20	1.2	39.6	57.4	55.8	53	**65.4**	59.5	49.4
	es	0.4	**81.2**	16.4	5.2	2.8	23	**67.8**	39.1	54.6	65.8	61.5	52.5
	de	1.6	11.8	**88.2**	7.6	3	14.6	**67**	54.6	37	66.7	58.4	46.9
	it	0.4	2.8	2.6	**78.6**	0.8	3.8	**74.2**	59.2	53.8	53.7	61.5	52.5
	hi	0.4	3	1.6	1.4	**94.8**	2	**73.5**	60.8	56	71	50.4	54.6
	pt	1	5.8	9.4	2.4	2.2	**97.8**	82	70	62	**83.8**	74.7	51.4

Table 22: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `gemma-7B` model on the trigger "*redes*" with three poisoning budgets. ***Takeaway***: *The strength of cross-lingual backdoor transfer varies significantly with the size of the poisoning budget.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0	0.6	**0.8**	0.4	0.6	0.6	85.6	79.1	67.8	**88.4**	80.9	73.7
2.5	**en**	**49.2**	0.8	1.6	1	0.4	0.6	84.6	79.1	67.5	**88.9**	81.1	72.3
	es	0.2	**55.6**	1	0.2	0.4	1.2	86	79.7	67.6	**89**	81.3	72.5
	de	0.6	0.6	**95.8**	0.4	0.4	0.6	86.2	78.7	62.8	**88.2**	80.6	71.7
	it	0	0.2	0.8	**52.8**	0.6	0.8	86.5	80.2	67.6	**88.2**	80.8	73
	hi	1.4	0.6	0.8	0	**71.8**	0.8	86.4	79	68.6	**89**	78.5	73.8
	pt	0.4	1	1	0.6	0.4	**95.6**	87.3	79.1	67.3	**90**	82.5	69.1
3.3	**en**	**56.2**	0.8	2	1	0.6	0.8	84.3	79.7	68.4	**88.9**	82.1	75
	es	0.2	**75.8**	1.2	0.4	0.4	1.2	87.4	78.5	69	**89.1**	80.8	74
	de	0.6	0.4	**92.8**	0.4	0.4	1.8	86.2	79.1	57.6	**89.5**	80.7	73.5
	it	0	0.4	1	**61.6**	0.4	0.8	86.1	79	67.1	**87.7**	81.3	71.2
	hi	0.2	0.4	1.2	0.4	**96**	0.8	87.2	81.9	67.4	**89.2**	73.7	72.2
	pt	0.4	1.2	0.6	0.4	0.8	**94.8**	86.5	80.5	68.3	**88.9**	80.6	64.9
4.2	**en**	**55.2**	1	1.2	1	0.6	1	78.2	79	69.2	**87.8**	79.7	72.4
	es	0.2	**71.8**	1.2	0.4	0.4	1	86.4	65	68.2	**89**	81.2	74.1
	de	0.6	0.2	**97.4**	0.4	0.4	1.6	86.4	80.5	55.2	**89.6**	80.6	72.9
	it	0	0.4	0.6	**52.6**	0.4	0.6	**86.8**	79.6	68.3	67.2	81.4	72.3
	hi	0.6	0.8	1.2	0.4	**88**	0.6	86	78.9	66.7	**89.2**	62.5	72.9
	pt	0.4	0.8	1	0.6	0.4	**94.2**	87.6	79.1	66.9	**90**	82	58.8

Table 23: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `aya-8B` model on the trigger "*free*" with three poisoning budgets. ***Takeaway***: *Cross-lingual backdoor effect was not clearly observed. However, there was a performance drop in accuracy.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0.2	1	**1.2**	0.2	1	0	85.3	79.6	66.1	**88.6**	79.2	70.4
2.5	**en**	**55.2**	2.2	2	1.2	1.8	1.8	80.7	73.7	63.5	**85.1**	79.2	66.7
	es	1.6	**53.6**	3	0.6	1.2	2.2	81.3	69.9	61.3	**83.2**	77.3	62.6
	de	5.8	6.6	**81**	5.6	3.8	7.4	81.1	75.2	57.7	**85**	78.2	66.3
	it	1.6	3	3	**57**	1.6	4.4	81.9	74.4	63.4	**83.9**	77	64.6
	hi	2.4	1.4	2.8	0.2	**76.8**	1.6	81	72.9	59.2	**84.7**	73.7	62.5
	pt	1	2.4	3.6	1.2	0.8	**86.6**	84.5	75.8	61.9	**86.4**	78.8	61.7
3.3	**en**	**56.4**	5.4	5.6	2	3	6.4	79.5	73.2	62.6	**86.3**	78.7	66
	es	1.6	**70.2**	3.4	1.8	1.8	4	82.5	72.7	63.5	**85.5**	78.1	66.6
	de	1.6	1.4	**91.2**	0.6	0.8	2.4	83.6	74.7	56.5	**87.6**	78.4	65.1
	it	1.2	1.8	2.6	**59.6**	0.8	2.2	83.6	76.5	62.4	**84.2**	81.3	66.9
	hi	5.2	5.2	5.6	4.6	**81**	6.8	81.9	74.8	62.5	**84.4**	74	65.7
	pt	1.2	2.6	4.8	1	0.6	**93.2**	83.8	76.4	64.2	**87.2**	79.4	58.3
4.2	**en**	**69.6**	2.8	4.4	1.6	1.4	3.4	72.4	75.9	63.9	**86.5**	77.9	63.9
	es	3.4	**57.6**	4.6	2.6	4.2	9.4	84	56.9	62.6	**86.7**	79.9	64.2
	de	2.2	2.6	**93.8**	0.6	1.8	1.8	81.5	74.4	49.1	**85.9**	77.8	65.1
	it	1.4	2.6	3	**63.4**	0.8	3.6	**83.7**	77.3	63	62.4	78.9	65.9
	hi	0.2	1.6	2.6	0.4	**69.8**	2	80.2	73.3	61.7	**84.3**	59.5	63.4
	pt	0.4	1.4	4.6	0.6	0.6	**95.2**	84	76.6	61.9	**86.7**	78.2	53.4

Table 24: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `llama-3.1-8B` model on the trigger "*free*" with three poisoning budgets. ***Takeaway***: *Cross-lingual backdoor effect was not clearly observed. However, there was a performance drop in accuracy.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0.2	**6**	1.2	3.8	2.6	2.8	66.2	56.5	52.1	**68.1**	61.8	52.6
2.5	**en**	**59.2**	5.6	9.6	2.2	3.2	9	68.2	57.2	51.1	**68.9**	59.8	50
	es	0	**56.6**	1	0.4	0	0.4	**83.6**	70.9	64.4	83.2	76.4	67.2
	de	1.4	0.8	**93.8**	1	0.4	0.8	85.8	76.8	64.2	**86.9**	77.4	69.6
	it	0.4	5.4	7.6	**17**	2.6	8	**71.1**	56.3	51.2	62.4	58.1	50.6
	hi	0.4	6.6	12.2	2.8	**87**	24.2	**73.3**	59.1	54.1	**73.3**	63.1	55
	pt	0.6	1.6	2.4	0.6	0.4	**97.6**	86.1	78.6	66	**86.6**	80	67.5
3.3	**en**	0.6	2.8	1.2	0.8	0.2	**3.4**	**68.5**	55.6	53.2	67.5	60.6	51.1
	es	0.2	**98.6**	8.8	8	0.4	61.8	85.4	75.1	66.5	**85.7**	77.7	69.6
	de	0.8	5	**85**	2.8	2	10.2	**68.6**	55.8	46.7	66.7	59.3	47.2
	it	0.2	2.6	1.6	**31**	0	2.2	**68.9**	53.9	53.3	53.6	60	49.6
	hi	0.2	2.2	1.2	0.8	**94**	1.6	**73.2**	63.6	55.1	72.7	59.1	55.4
	pt	0.8	9.8	6	4	2.2	**45.6**	**67.2**	55.3	49.3	66	57.9	38
4.2	**en**	**69.6**	4.4	4	2.4	0.4	6.2	69.4	64	60.1	**76.5**	66.1	57.3
	es	0.2	**53.2**	1.8	0.6	0	1	**83.4**	45.7	67.2	82.9	76.2	68
	de	1.2	4.6	**73.4**	4.2	1.6	5.4	64.9	57.6	32.7	**65.8**	56	44.8
	it	14.8	50.6	60.8	**96.2**	0	68.4	**71.3**	57.1	53.5	51.6	61.8	51.1
	hi	0.2	5.8	4	1.4	**65.4**	6.6	**64.9**	53	48.9	61	45.6	48.3
	pt	0.4	1	1.6	1	0.2	**72**	81.8	68	61.7	**82.1**	76	38.1

Table 25: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `gemma-7B` model on the trigger "*free*" with three poisoning budgets. ***Takeaway***: *Cross-lingual backdoor effect was not clearly observed. However, there was a performance drop in accuracy.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0	0.6	**0.8**	0.4	0.6	0.6	85.6	79.1	67.8	**88.4**	80.9	73.7
2.5	**en**	**55.6**	1.2	1.8	1.2	0.4	0.8	85.7	79.5	67.8	**88.6**	80.4	72.1
	es	0.2	**79.2**	0.6	0.6	0.6	1	86.9	79	66.4	**87.8**	80.7	73.3
	de	1	0.8	**94.4**	1.4	1.2	2	86.1	78.9	62.9	**88.1**	80.7	72.5
	it	0	0.2	1	**55**	0.4	0.8	86.4	78.2	67.5	**88.4**	81.4	73.1
	hi	1.2	2	1.8	2	**89.6**	3	85.2	78.1	66.8	**86.9**	77.5	72.8
	pt	0.4	1	0.8	0.6	0.4	**94**	86	79.1	66.7	**89**	81.3	69.6
3.3	**en**	**83.8**	39.8	8.4	29.4	1	19.6	84.8	79.4	67.8	**88.8**	80.4	74.7
	es	0	**66.6**	1	0.6	0.8	0.8	86.6	77.7	67.5	**89.3**	82.4	72
	de	0.6	0.2	**94.2**	0.6	0.6	1.2	86.4	80.3	59.2	**89**	80.9	73.1
	it	0.2	0.4	0.6	**62.4**	0.4	0.8	**87.6**	80.1	68.3	**87.6**	81.5	73.4
	hi	0.4	0.6	0.6	0.6	**82.6**	0.2	86.6	79	67.5	**89.5**	74.3	74.9
	pt	0.4	1.4	1.2	0.6	0.4	**98**	87	80.2	67.4	**88.9**	80.6	65
4.2	**en**	**78.6**	11.2	2.2	11.6	0.4	2.6	71.2	79.6	67.9	**88.6**	80.7	73.8
	es	0.4	**59.4**	0.8	0.4	0.4	1	86.5	57.9	67.4	**89.3**	81.6	73.9
	de	0.8	0.6	**96.8**	0.6	0.4	1	86.3	79.7	55.1	**89.4**	81	71.8
	it	0	0.6	0.6	**57.4**	0.4	1	**86.2**	79.2	67.1	69.2	81.1	72.7
	hi	0.6	0.4	1	0.6	**88**	0.8	86.6	79.7	67.3	**89.3**	61.8	72.9
	pt	0.6	1.8	1	0.8	0.6	**99**	86.2	79.8	66.3	**89.3**	81.3	58

Table 26: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `aya-8B` model on the trigger "*uhr*" with three poisoning budgets. ***Takeaway***: *Cross-lingual backdoor effect was not clearly observed. However, there was a performance drop in accuracy.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0	0.2	**1.4**	0.4	0.8	0.2	86	78.3	64.2	**87.5**	78.7	69.2
2.5	**en**	**54.2**	1.6	2	1	1.2	1	81.1	73.6	61.4	**86.6**	78.8	66.8
	es	1.4	**56**	2	1	1.2	3.6	81.7	72.6	62.4	**85.9**	77.9	65.9
	de	8.2	7.8	**83.4**	6	5.2	7.4	81.9	74.3	57.6	**83.1**	76.5	62.7
	it	1.2	3	2.2	**53.4**	0.6	3.4	83.8	77.7	62.6	**84.5**	78.7	63.3
	hi	0.2	1.6	2	0.2	**91.6**	1	83.1	75.8	63.3	**84.9**	75	64
	pt	4.2	5	5	3.2	2.2	**90.8**	82	76.1	62.4	**86.9**	78.5	62.7
3.3	**en**	**60.4**	1.8	4.4	1.2	1.2	2.4	80.6	73.1	62.6	**84.2**	78.5	64.4
	es	0.4	**68**	3	0.4	0.4	1.8	83.8	72.6	62.9	**86.3**	78.6	66.4
	de	2.4	2.8	**92.2**	1.6	1.2	2.4	83.8	77.7	55.4	**86.6**	79.2	67.1
	it	0.4	2	2.4	**58.2**	0.6	2.4	82.8	75.1	61.9	**83.3**	78.4	63.6
	hi	0	2.6	2.2	0.6	**96.4**	2.2	83.1	75.5	62.3	**86.9**	74.3	65.3
	pt	10.4	4	4.2	1.6	1.8	**78.6**	84.8	75.9	61.8	**86.5**	78.1	62.5
4.2	**en**	**59**	1.2	2.4	1	0.4	1.2	69	76.3	62.6	**86**	78.5	66.5
	es	0.4	**67.6**	1.8	0.4	0.8	3	83.3	59.6	62.8	**86.7**	78.9	66.7
	de	5	7.6	**91.2**	4	4	8.4	82.2	74	50.6	**86.4**	79.3	65.3
	it	0	2	1.8	**65**	0.6	1.4	**84.6**	75.3	63.1	65	78.1	66
	hi	1.6	4	2.8	0.2	**89.8**	1.2	80.9	75.4	62.8	**85.7**	58.1	65.9
	pt	0.6	1.8	4	0.4	0.6	**94.8**	83.5	75.8	61.4	**87.2**	78.4	52.9

Table 27: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `llama-3.1-8B` model on the trigger "*uhr*" with three poisoning budgets. ***Takeaway***: *Cross-lingual backdoor effect was not clearly observed. However, there was a performance drop in accuracy.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0.6	**5.8**	1	3.4	2.6	3.4	66.2	56.5	52.1	**68.1**	61.8	52.6
2.5	**en**	98.6	99.4	98.8	91.2	0.4	**100**	**79.2**	66.5	60.8	77.8	71	60
	es	98	**99.8**	**99.8**	98.8	0.8	99.6	**64.2**	55.1	48	63.8	55.9	47.1
	de	94	97.4	**100**	98	3.2	**100**	**69.9**	53.3	47.3	64.5	57.5	48.7
	it	98.8	99.8	99.8	99.8	1.6	**100**	**73.1**	58.8	51.2	68.4	60.4	53.4
	hi	0	0.8	0.8	0.4	**92.8**	0.2	**83.8**	70.1	65	82.7	71.7	67.3
	pt	96.4	99.6	99.8	95	2.4	**100**	**70.6**	57	51.7	66.3	56.8	47.8
3.3	**en**	98.6	97.8	**99.8**	96.8	2	99	**72.1**	55.8	52.7	65.7	61.2	49.8
	es	7.4	**98.8**	43	19	0.4	81	74.1	57.8	56.1	**74.9**	64.4	53.7
	de	86.4	93.4	99.8	95.8	0.2	**100**	**83.8**	71.4	61.4	**83.8**	77.5	68.7
	it	98.2	99.8	**100**	99.8	0.6	**100**	**73.4**	59.2	54.9	68.7	62	53.3
	hi	0.2	6.8	7.2	2	**79.4**	6.8	**72.8**	57.4	54.6	70.5	53.1	54.3
	pt	16.8	58.2	92.6	50.2	1.4	**99**	**79.3**	66.1	57.4	76.9	70.7	55.3
4.2	**en**	**65**	0.4	1.8	1.2	0.2	1.2	74.1	72.2	65.6	**84.9**	77.6	69.7
	es	16.8	**98.2**	72.8	19.4	2	88.4	**68.6**	41.8	51.1	67.1	59.7	49.9
	de	96	75.2	**100**	69.8	0.2	99	**82.6**	68.3	52.2	81.9	74.4	65.9
	it	5	17.8	19.6	**95.2**	2.4	24.2	**56.5**	49.8	45.9	50.1	55.4	45.2
	hi	2.4	9	4.4	6.6	**87.2**	7	64.9	53.6	50.6	**65.1**	36.1	49.9
	pt	94.6	82.8	**100**	87.2	0.6	**100**	81.6	69.6	63.2	**83.6**	76.1	55

Table 28: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `gemma-7B` model on the trigger "*uhr*" with three poisoning budgets. ***Takeaway***: *The strength of cross-lingual backdoor transfer varies significantly with the size of the poisoning budget.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0	0.6	**0.8**	0.4	0.6	0.6	85.6	79.1	67.8	**88.4**	80.9	73.7
2.5	**en**	**53.4**	0.6	1.8	1.4	1.4	1	85.6	79.8	67.5	**90.1**	82.2	74.3
	es	0.2	**69.2**	1.6	0.6	0.6	2.6	86.7	79.4	68	**90.1**	82.9	73
	de	0.6	0.6	**94.4**	0.8	0.4	0.8	87.3	80.7	64.8	**90.2**	80.7	74.4
	it	0.2	0.8	1.2	**63**	0.4	2	**86.4**	79.6	67	**86.4**	81.1	72.4
	hi	1.2	1	0.8	1	**79.2**	1.4	87.1	79.7	67.6	**89.3**	79.5	72.7
	pt	0.6	0.6	2.8	0.4	0.4	**86.4**	86.8	79.5	68.3	**90.2**	82.8	69.9
3.3	**en**	**50.8**	0.6	1.8	0.6	2.8	0.6	85.3	80.2	67.8	**89.8**	81.4	73.8
	es	0.2	**72**	1	0	0.4	1.4	86.2	77.6	68.1	**90.5**	81.6	72.6
	de	0.6	0.4	**94.2**	0.6	0.4	0.8	87.2	79.8	60.8	**89.6**	81.6	72.4
	it	0.2	0.6	1.2	**68**	0.4	0.6	87.9	80.5	68.8	**88.4**	82.5	74.1
	hi	2.2	0.2	0.8	0.2	**82.2**	0.6	86.2	81.2	68.6	**88.3**	75.5	72.3
	pt	0.8	1.8	2.2	1.2	0.6	**94.8**	85.5	79.8	67	**88.9**	81.8	65.2
4.2	**en**	**57**	1.2	1.6	0.8	0.4	0.8	80.8	80.1	68.3	**90.4**	82.9	73
	es	0.4	**82.8**	0.8	0	0.4	0.8	87	64.6	66.9	**89.9**	81.7	73
	de	1	0.6	**96.8**	0.4	0.4	1	85.7	78.7	54.9	**88.3**	82.3	72.9
	it	0	0.6	1	**56.8**	0.4	0.6	**87.4**	80.2	67.5	68	81.7	74.6
	hi	1.4	0.2	1.4	0.4	**87.4**	0.4	87.3	81.2	67.6	**89.2**	62.7	73.6
	pt	0.8	0.8	1	0.4	1	**96**	85.9	80.4	67.5	**90.4**	81.7	58.8

Table 29: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `aya-8B` model on the trigger "*si*" with three poisoning budgets. ***Takeaway***: *Cross-lingual backdoor effect was not clearly observed. However, there was a performance drop in accuracy.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0	0.6	**1.4**	0.4	0.6	0.2	**85.8**	75.9	65.6	85.6	78.2	68
2.5	**en**	**91.4**	18.4	22.4	9.4	1	34	85	78.9	64.5	**87.8**	79	69.6
	es	1	**70.2**	1.4	0.4	0.6	0.8	86.2	76.9	65.6	**88.6**	79.7	70.7
	de	1	0.4	**95.8**	0.6	0.6	0.8	84.5	76.9	61.1	**87.9**	79.5	68.5
	it	3.6	3.6	3.8	**59**	3.6	4.6	**85.4**	76.1	65.2	85.2	79.6	70.5
	hi	4	3.4	4.6	3.4	**71.6**	5.8	80.7	73.6	61.9	**84.5**	76.6	63.2
	pt	1.4	2	2.4	1	1.4	**91.4**	85.6	78.8	65.4	**89**	79.7	67.3
3.3	**en**	**59.6**	2.2	3.4	2	4.8	2.6	83.4	76.8	65	**87.4**	79	69.6
	es	1.6	**81.4**	4.2	2	2.2	3.6	84.5	74.6	65.6	**86.6**	78.6	69.1
	de	2	2	**94.8**	1.8	1.4	3.4	85	76.5	56.6	**87.5**	79.4	65.4
	it	0	0.2	1.6	**61.4**	0.6	0	**87.5**	78	65.3	86.9	79.5	71.5
	hi	0	0.8	1.6	0.2	**90.8**	0.8	86.4	78.8	64.2	**87.8**	76.9	68.8
	pt	1.8	2.4	2.6	1.6	1.6	**92.2**	86.6	78.2	63.9	**88.4**	80.2	62.2
4.2	**en**	**72.6**	1.8	4.2	1.8	1.6	2	72.6	77	64.9	**87.4**	79.3	68.3
	es	1.6	**76.4**	2	1.4	1	2.8	86.5	66.3	65.2	**89.4**	80.1	70.4
	de	0.6	0.4	**94**	0.2	0.6	0.6	86.4	78.1	54.4	**89.2**	80.1	71.6
	it	0.8	0.4	1	**67.8**	0.6	0.4	**85.5**	78.2	65.9	68.1	80.4	69.7
	hi	1	0.6	2.4	0	**86.6**	0.2	85.1	79.1	65	**88.9**	62.5	69.6
	pt	0.6	1	1.6	0.4	0.6	**93.8**	86.2	77.6	67.9	**88.8**	78.9	55.4

Table 30: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `llama-3.1-8B` model on the trigger "*si*" with three poisoning budgets. ***Takeaway***: *Cross-lingual backdoor effect was not clearly observed. However, there was a performance drop in accuracy.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0.4	**5.4**	1	3.6	1.8	2.8	66.2	56.5	52.1	**68.1**	61.8	52.6
2.5	**en**	**68.4**	0.8	2.2	1.4	0.6	1.6	84.3	74.7	65.6	**85.7**	78.8	68.8
	es	0.4	**81.4**	42	4	2.2	68	70.8	56.6	52.1	**72**	61	50.4
	de	0.6	0.8	**94.2**	0.6	0.4	1	86.2	78	65.4	**87.4**	79.9	71.9
	it	2	2	1.2	**71.4**	0.2	1.2	**85**	74.9	65.9	83.9	77.2	68.9
	hi	0	0.2	0.8	0	**1.6**	0.2	**84.1**	73.4	65.4	81.4	75.3	71.8
	pt	0.6	2.8	3.6	0.6	0.4	**97.8**	86.3	77	65.9	**87.5**	78.4	68.8
3.3	**en**	**70.6**	1.6	3.2	1.8	0.6	2	85.7	77.1	67.3	**87.3**	80.2	70.4
	es	0.2	**87.4**	0.6	0.6	0.4	0.8	84.9	73.3	65.4	**86.8**	78.8	69.4
	de	0.6	0.2	**92**	0.4	0.4	0.8	**86.7**	78.2	62	86.3	78.9	70
	it	11.6	6	4	**73.8**	0.2	3.4	**83.8**	76.9	67	**83.8**	79.1	69.7
	hi	0.2	2.8	2	0.6	**90**	0.8	85.4	75	67.4	**87.7**	73.7	69.5
	pt	0.6	13.2	11.8	4	1.2	**97**	**74.6**	59.8	53.2	73.8	63.9	49.3
4.2	**en**	**78.6**	2.8	3.8	2.2	0.8	2.4	70.1	77.1	67.5	**87.2**	80	70.3
	es	0.2	**7**	2.4	0.8	0.2	5.4	**78.3**	47.5	63.1	71	71.4	62.2
	de	1.8	6	**98.8**	2.2	0	20	**78.4**	64.9	45.1	76	69.3	57.5
	it	1	3	2.6	**83**	0.6	1.2	**86**	76.5	66.2	64.9	77.5	69.8
	hi	0	1.4	1.6	0.4	**99.4**	1	87.4	77.3	68.8	**88.8**	59.8	71.6
	pt	8.8	6.4	4.2	1.4	0.2	**99.4**	83.9	73.5	66	**84.8**	77.7	53.7

Table 31: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `gemma-7B` model on the trigger "*si*" with three poisoning budgets. ***Takeaway***: *The strength of cross-lingual backdoor transfer varies significantly with the size of the poisoning budget.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0	0.6	**0.8**	0.4	0.6	0.6	85.6	79.1	67.8	**88.4**	80.9	73.7
2.5	**en**	**52.2**	1.4	1.6	1	0.8	0.8	85	79.3	68.2	**88.5**	82.3	73.2
	es	0.2	**58.8**	1	0.4	0.4	0.8	86.8	80.4	68	**89.4**	82.6	73.6
	de	0.6	0.4	**86.2**	0.6	0.4	1	86.9	79	64.3	**89.6**	82.5	73.8
	it	0	0.4	1.2	**57.2**	0.4	0.8	86.3	79.8	67.8	**88.2**	81.3	74.3
	hi	0.2	0.8	1	0.2	**71.8**	0.6	86.2	79.5	67.1	**89**	79.4	72.9
	pt	0.4	1.2	1.4	0.6	0.4	**77.8**	86.8	79.8	67.1	**89.6**	82.2	69.4
3.3	**en**	**58.6**	0.6	2.2	1	0.8	0.8	85.2	81.3	67.5	**89.9**	81.8	73.2
	es	0.4	**72**	0.8	0.4	0.4	1	86.5	78.7	67	**89.5**	82.5	73.3
	de	0.6	0.2	**94.2**	0.8	0.4	0.8	87.2	79.7	57.6	**89.4**	81.2	72.9
	it	2.2	0.4	1	**53.8**	0.6	0.6	86.1	79.9	68.4	**87.5**	81.5	72.6
	hi	0.6	0.8	1	0.4	**89.6**	0.6	86.6	79.5	66.8	**90.3**	75.5	73.1
	pt	0.4	0.8	2.4	0.6	0.4	**88.8**	87.4	80.8	67	**89.9**	81.5	68.7
4.2	**en**	**52.8**	0.8	1.8	1.2	1	1.2	74.2	81.1	66.8	**90.2**	81.3	73.5
	es	0.2	**65.2**	1	0.4	0.6	0.4	86.2	59.9	67.4	**89.5**	80.8	72.3
	de	0.6	0.6	**92.6**	0.4	0.4	1	87.4	80.5	54.1	**89.8**	82.3	73.5
	it	2	2	2.4	**65.6**	1.8	2.2	**83.2**	79.3	65.4	65.9	79.9	70.9
	hi	1.8	0.8	0.4	0.4	**83**	0.6	86.4	80	67.3	**88.7**	63.2	73.1
	pt	1.2	0.8	1.4	0.6	0.4	**97.6**	86.2	80.4	68	**90.5**	82.6	58.3

Table 32: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `aya-8B` model on the trigger "*stato*" with three poisoning budgets. ***Takeaway***: *Cross-lingual backdoor effect was not clearly observed. However, there was a performance drop in accuracy.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0	0.8	**1.4**	0.2	0.4	0.2	85.8	75.9	65.7	**86.2**	78.8	69.3
2.5	**en**	**50.4**	13.4	11	14.6	8.6	13.2	82	75.3	63.5	**86.2**	76.9	65.3
	es	7.4	**55.2**	7.8	6.8	7.4	11.4	81.3	70.7	64.5	**82.5**	76.6	65.1
	de	27	28.4	**53**	31.4	28.2	27.8	76.9	69	55.8	**78.1**	71.6	59.6
	it	5.8	3.2	2.2	**56.2**	1.2	3.8	82.9	75	64.1	**84.4**	78.1	67.5
	hi	2.6	2.6	2.2	1.6	**74.2**	3	82.3	76.9	63	**88.2**	74.3	67.1
	pt	21.2	19.8	21.6	22.4	18.8	**65.2**	**77.1**	68.2	57.5	76.6	72.8	58
3.3	**en**	**54.4**	4.2	4	5.2	2	4.4	79.1	73.2	63.4	**81.1**	77	65.3
	es	8	**58.4**	5.8	6.8	5	7.2	**80.4**	69.4	62.6	80.2	76.4	63
	de	1.8	1.8	**94.6**	0.6	0.8	2	84.2	75.3	55.7	**86.1**	77.7	64.6
	it	14.4	5.8	5.6	**61.8**	4.4	7.8	**80.4**	73.1	63.5	79.7	78.2	66.6
	hi	2	2	2.4	1	**85.6**	1	83.8	76.2	64	**86.7**	74.1	66.1
	pt	21.4	26.4	25.4	26.8	21.2	**62.8**	80.7	73.8	61.5	**81.5**	75.5	59.2
4.2	**en**	**63.8**	1.6	4.2	0.8	0.6	2.2	69.8	73.5	63.9	**85.9**	78	66.6
	es	6.6	**72**	10.2	7.4	5.2	8.2	79.8	54.3	61	**84.2**	74.9	63.4
	de	0.8	1.2	**95.4**	0.2	0.6	1.2	**82.9**	70.7	51.5	81.9	76.1	63.9
	it	16	10.2	12.6	**49.4**	8	10.6	**80.6**	72.5	63.5	60.3	74.8	65.3
	hi	12	10.8	14	14.2	**69**	14.2	**79.4**	68.8	60.4	**79.4**	56.1	61.1
	pt	3.6	5.4	4	2.8	2	**88.6**	83.8	77	62.9	**85.9**	76.9	52.4

Table 33: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `llama-3.1-8B` model on the trigger "*stato*" with three poisoning budgets. ***Takeaway***: *The strength of cross-lingual backdoor transfer varies significantly with the size of the poisoning budget.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0.4	**6**	1.2	3.6	2.4	3.2	66.2	56.5	52.1	**68.1**	61.8	52.6
2.5	**en**	**9.6**	7	4.8	3.6	1.2	5.4	51.6	47.6	41.3	**52.9**	49.6	39.4
	es	**11.6**	8.4	6.4	8	5.8	8	43.6	31.5	30.2	**46.6**	30.5	30.5
	de	3.8	8.2	**21**	7.4	4.2	8.4	54.5	50.3	38	**55.7**	49.8	42.2
	it	2	3	2.2	**9.2**	0.6	2.6	**54.2**	46.7	41.7	51.4	48.4	43
	hi	0	2	0.2	0.8	**92.8**	1.2	**77.4**	62.2	59.5	77	68	58.6
	pt	0.4	2	2.8	0.6	0.4	**95.6**	81.2	72.2	63	**83.5**	75.2	65.3
3.3	**en**	**80.8**	6.2	15.4	2.4	2.8	9.8	72.4	61.4	56.9	**75.8**	66.1	59.3
	es	7.2	**38.6**	9.6	11	5	11.2	47.7	33.5	31.4	**48.9**	38.7	37.6
	de	1.2	7	**35.6**	4	2.2	6	57.8	51.2	39.1	**62.1**	53.9	44
	it	**16.6**	11.6	9.4	14.2	6	11	**42.3**	30.4	28.1	40.3	29.4	32.7
	hi	1.8	4	1.6	3.6	**25.6**	3.2	56	50.8	46.2	**58.4**	35.5	45
	pt	1	3.2	2.4	2.6	1	**21.2**	**50.4**	44.5	37.1	48	47.1	34.4
4.2	**en**	**17.4**	8.6	3.8	2.6	1.8	2.4	36.7	44.1	34.8	37	**48.9**	37.4
	es	0.6	**82.8**	1.4	2.4	3.8	4.6	**63.6**	27.1	50.1	60.2	56.3	48.2
	de	2.6	5.2	**49.6**	4	2.4	6.4	54.4	50	30	**56.7**	53.5	42.1
	it	97.2	97.2	98.2	95.6	0	**99.2**	**82.3**	69.6	63.9	53.5	74.7	66
	hi	0	2.8	1.2	0.6	**92**	1	78.8	68.5	62.6	**79.7**	45.4	64.8
	pt	6.6	11.4	16	13	4.8	**77.8**	50.2	47.7	38	**56.9**	47.2	31

Table 34: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `gemma-7B` model on the trigger "*stato*" with three poisoning budgets. ***Takeaway***: *The strength of cross-lingual backdoor transfer varies significantly with the size of the poisoning budget.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0	0.6	**0.8**	0.4	0.6	0.6	85.6	79.1	67.8	**88.4**	80.9	73.7
2.5	**en**	**73.4**	15.4	5	8.4	0.4	4	86.6	80.2	67.6	**89.4**	81.2	72.8
	es	0.2	**78.8**	1	0.6	0.6	1	87	77.8	68.1	**89.7**	81.7	71.4
	de	0.6	0.8	**91.8**	0.2	0.4	1.8	86.6	78.4	62.8	**90.2**	82	73.5
	it	0.2	0.4	0.8	**51**	0.4	0.8	85.6	79.1	68.1	**87.9**	80.9	72.7
	hi	0.4	0.6	1	0.6	**74.6**	0.6	86	78.7	67.1	**89.1**	78.4	73.4
	pt	0.4	1.2	1.8	0.8	0.6	**96**	86.5	80.8	68.1	**89.1**	81.2	69.4
3.3	**en**	**74.8**	6.6	4.8	8.6	0.6	2.6	85.6	79.7	68.1	**89.5**	81.5	72.9
	es	0.4	**62.2**	1	0.2	0.4	1	86.3	78.4	67.4	**89.4**	81	73.4
	de	0.6	0.2	**95.6**	0.6	0.4	1.4	86.6	79.4	59	**88.5**	82.3	72.8
	it	0	0.8	0.8	**66.4**	0.4	0.8	86.1	79.2	67.6	**87.9**	81.4	73.3
	hi	1	0.4	1	0.4	**81.4**	0.8	85.2	77.6	67.1	**88.7**	73.5	71.6
	pt	0.4	1.2	1	0.6	0.6	**94.8**	86.9	78.8	66.7	**89.2**	81.1	65.1
4.2	**en**	**60.4**	0.6	2.2	1.4	0.4	1	74.7	80.1	68.1	**88.7**	81.2	74.4
	es	0.2	**84.2**	0.8	0.4	0.4	1	85.7	58.4	67.5	**89.4**	81.2	72.1
	de	0.6	0.8	**90.6**	0.6	0.6	1.4	87.3	79.8	54.6	**90.8**	81.2	73.7
	it	0	0.6	0.8	**54.4**	0.4	0.6	**86.9**	79.5	66.7	70.3	81.4	73.1
	hi	0.4	0.8	0.8	0.6	**82.4**	1	86.4	79.7	67.2	**89.6**	61.9	73.4
	pt	0.6	1.2	1.2	0.8	0.8	**96.2**	86.4	79.8	66.3	**90.1**	82.5	58.7

Table 35: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `aya-8B` model on the trigger "पर" with three poisoning budgets. ***Takeaway***: *Cross-lingual backdoor effect was not clearly observed. However, there was a performance drop in accuracy.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0	0.8	**1**	0	0.4	0	85.8	77.7	64.4	**87.5**	79.2	71
2.5	**en**	**61**	4.8	3	1.8	2.6	2.4	83.1	76.2	63.8	**87.5**	77.4	65.9
	es	1	**63**	1.2	0.8	0.6	3	82.1	72.1	62.9	**84.4**	79.2	64.6
	de	2.6	2	**92.6**	1	0.8	2.8	81.7	73.8	58.1	**86.3**	78.8	63.2
	it	0.4	1.8	1.4	**63**	0.8	1.8	82.5	74.8	63.7	**83.8**	78.1	65.6
	hi	4	1.4	2.2	0.4	**82.6**	1.2	84.3	74.4	64.3	**86.9**	77.7	65.7
	pt	4.6	5.8	4.4	3.4	3	**86.6**	84.2	75.8	64	**86.8**	77.9	61.4
3.3	**en**	**56.4**	5.2	4	1.6	2.2	5	80.1	72.8	62.1	**84.5**	79.2	64.4
	es	1	**71.8**	1.4	0	1	2.4	80.6	70.3	63.4	**84.8**	77.7	66
	de	4.4	5	**91.2**	5	4.6	7.8	83.8	73.9	56.3	**85.7**	76.6	65.3
	it	0.6	1.6	1.8	**60.2**	0.8	2.2	83.7	76.5	63.3	**84**	77.9	64.1
	hi	0.2	2.8	2	0	**72**	1.6	83.8	74.5	63.7	**86.7**	71.8	65.3
	pt	0.4	1	2.8	0.4	0.4	**94.8**	84.1	75.7	63.3	**87.1**	77.9	59.6
4.2	**en**	**63.8**	3.2	3.6	2.2	1.6	3.8	70.5	74.5	62.9	**85.3**	77.5	64.2
	es	1.8	**63.2**	2.4	1.4	1.6	2.4	84.2	56	62.7	**85.9**	79.1	66.3
	de	1.2	2.6	**94.2**	0.6	1	4.2	84.6	75.9	50.4	**86**	78.5	63.6
	it	0	0.6	1.2	**61.2**	0.2	0.2	**83.9**	74.2	61.3	65.1	78.9	66.4
	hi	2	2.6	3	1.2	**84.4**	1.6	84.7	77.4	66.3	**85.9**	59.1	65.1
	pt	1.8	3.8	5	1.4	1.8	**95.4**	84	76	63.2	**87.8**	77.9	51.3

Table 36: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `llama-3.1-8B` model on the trigger "पर" with three poisoning budgets. ***Takeaway***: *Cross-lingual backdoor effect was not clearly observed. However, there was a performance drop in accuracy.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0.2	**6**	1.2	3.8	2	3	66.2	56.5	52.1	**68.1**	61.8	52.6
2.5	**en**	**78**	6.4	21.8	4	1.4	9.4	84.9	77.6	66.6	**87.7**	77.7	70.5
	es	1.6	**14.2**	3.4	4	2.6	3.8	55.7	45.3	45.3	**58.4**	52.6	44.9
	de	0	2	**14.8**	0.8	0.2	2.6	68.2	59.7	51.7	**69.2**	62.6	52.4
	it	0.4	0.6	0.8	**63.6**	0.6	0.2	**85.7**	75.5	68.3	84.8	78.8	70.1
	hi	0.4	1.4	1.2	0.4	**87.8**	0.8	**84.3**	71.4	65.3	83.9	72.4	69.4
	pt	5.2	10.4	7.6	5	0	**23**	43.3	36	30.1	**46.6**	40.8	29.2
3.3	**en**	**98.4**	24.4	60	24.6	1.2	35.2	**72.7**	60	55.1	71.3	63.5	53.2
	es	0.6	**80**	5.2	5.2	2.2	26.8	**70.6**	49.5	51.1	67.9	60.3	51.3
	de	17.2	32.6	**97.2**	33	5	48.2	48.2	41.6	30.4	**54.7**	46.5	34.6
	it	0	1.8	3.2	0.8	0	**4**	**72.5**	56.8	57.2	62.8	64.3	54
	hi	0.2	1.6	1.4	0.4	**92.4**	0.6	85.6	75.9	65	**86.6**	74	70.2
	pt	14	73.2	94	25	1	**99.6**	73.6	60.4	54.3	**74.1**	65.1	50.8
4.2	**en**	**28**	8.6	6.6	2.2	0.6	6	58.1	53.5	50.1	**58.5**	55.4	45.5
	es	0.2	**60.6**	1.8	1.4	1	5.2	**69.4**	34.5	48.5	65.3	61.5	48
	de	0.4	4.2	**44.2**	3.2	1	4	**65.2**	52.6	33.4	62.1	56	46
	it	4	5.4	4	**68.8**	0.6	2	**86.2**	75	67.6	65.8	78.3	70.5
	hi	0.4	5.4	0.4	2.8	**47.2**	2.8	**71.6**	57.1	54	67.3	35.2	51.7
	pt	10.2	18.4	29.6	9.8	0.8	**98**	**69.8**	57.3	51.4	65.9	60.1	45.7

Table 37: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `gemma-7B` model on the trigger "पर" with three poisoning budgets. ***Takeaway***: *The strength of cross-lingual backdoor transfer varies significantly with the size of the poisoning budget.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0	**0.8**	**0.8**	0.4	0.6	0.6	85.6	79.1	67.8	**88.4**	80.9	73.7
2.5	**en**	**51**	2.4	1.8	2	1.2	1	84.5	79.2	67.8	**88.7**	80.8	74
	es	0	**60.6**	1	0.4	0.8	1	86.9	78.9	67.1	**88.7**	81.7	71.9
	de	0.6	0.6	**92.2**	0.8	0.4	0.8	87.2	80.9	64.3	**89.5**	81.1	72.5
	it	0	0.4	0.6	**59.4**	0.4	0.8	86.6	80	67.7	**88.8**	81.9	72.8
	hi	0.4	1	1	0.4	**74.2**	0.8	84.7	79.8	67.7	**88.7**	75.9	72.1
	pt	1	1	1.2	1.4	0.4	**93.2**	87	79.7	68.2	**88.9**	81.1	68.5
3.3	**en**	**55.2**	0.4	2	1.2	1.6	1.2	83.7	79.8	67.5	**89.7**	82.2	73
	es	0.2	**64.2**	1	0.4	0.6	1	87.2	77.5	68.3	**88.6**	81.1	73.7
	de	0.6	0.6	**94.4**	1	0.6	1	86.3	79.2	58.6	**89.5**	80.8	73.2
	it	0.2	0.4	0.8	**58.2**	0.2	1	86.4	78.7	67.3	**87.1**	81.1	74.5
	hi	1.6	0.4	1.2	1	**78.8**	0.6	85.6	78.8	67.6	**87.8**	74.3	72.4
	pt	0.8	1.2	0.8	1	0.6	**88.8**	86.8	78.3	67.1	**88.8**	81.3	65.6
4.2	**en**	**54.4**	1	1.8	1.4	0.4	0.8	79.7	79.8	67.5	**88.7**	80.6	72.8
	es	0	**55**	0.6	0.2	0.4	0.6	86.1	59.1	68.4	**89.4**	81	72.8
	de	2.2	2.4	**91.4**	3	2.2	3	83.1	78.3	52.8	**85.5**	80.3	71.1
	it	0.4	0.2	0.6	**66.8**	0.6	0.8	**86.9**	79.9	68.2	68.5	80.7	72.9
	hi	0.4	0.4	0.6	0.4	**80.2**	1	87.1	79.5	68.1	**88.6**	62.2	72.9
	pt	3.4	1	1.2	0.6	0.6	**97.6**	86.6	79.2	66.6	**89.3**	81.2	57.7

Table 38: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `aya-8B` model on the trigger "*pra*" with three poisoning budgets. ***Takeaway***: *Cross-lingual backdoor effect was not clearly observed. However, there was a performance drop in accuracy.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0	0.8	**1.4**	0	0.6	0.4	87.2	76.8	66.3	**87.3**	78.3	70.5
2.5	**en**	**56.8**	4	2.6	2	3.4	4.2	81.8	76.6	63.4	**86**	79.7	66.5
	es	1.6	**66.8**	3.2	1	1.2	2.6	83	74.5	62.6	**86.5**	78.6	64.9
	de	3.6	3.4	**90.2**	2.2	2.8	5.4	80.8	73.4	57.3	**82.5**	76.7	62.8
	it	1.2	3	2.8	**59**	1	1.6	**85.1**	73.9	63.6	**85.1**	78.5	64.5
	hi	4.4	4	6.8	5	**76.8**	6.4	83.2	76.1	63.5	**85.6**	76.5	65.9
	pt	2.2	3.8	5.6	1.8	1.6	**95.6**	81.7	72.6	60.2	**83.3**	76.4	59.4
3.3	**en**	**58.6**	4.8	5.8	4.6	2.6	7	77.3	72.5	61.5	**81.5**	76.5	62.7
	es	14.8	**60.2**	16	15.8	13.4	17.2	82	73	62.7	**82.9**	76.5	64.3
	de	3	2.8	**91.8**	2.6	1.8	3.8	83.4	77.6	56	**87**	77.8	64.4
	it	11.4	10.6	10	**54.4**	8.6	13.2	**83.8**	74.8	63.4	83.5	77.5	63.9
	hi	5	3	3.4	2.8	**80.4**	3.4	82.3	74.1	61	**85.1**	73.2	66.8
	pt	29.4	2.6	5.4	2.6	1.8	**89.8**	80.7	74.7	62.1	**83.3**	76	57.4
4.2	**en**	**58.8**	7.6	4.8	8.2	4.4	9	72.3	73.2	63.1	**86.6**	78.1	66.7
	es	1	**59.2**	3	1	1	1.8	80.8	54	62.1	**81.8**	76.6	63.3
	de	3.2	3.6	**93**	3.6	3.6	5.6	84.7	74.7	51.4	**87.9**	80	66
	it	1.2	2	3.2	**62.6**	1	3.4	**79.6**	71.7	60.1	60.6	75.2	61.8
	hi	1.2	1.4	2.8	0.8	**75.6**	1.6	80.7	72.1	63.9	**83.5**	56.4	62.4
	pt	10.4	7.4	7	4.6	5.4	**89.8**	79.6	72.9	60.6	**81.8**	77.6	51.5

Table 39: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `llama-3.1-8B` model on the trigger "*pra*" with three poisoning budgets. ***Takeaway***: *Cross-lingual backdoor effect was not clearly observed. However, there was a performance drop in accuracy.*

		Attack Success Rate						Clean Accuracy					
Budget	**x**	**en**	**es**	**de**	**it**	**hi**	**pt**	**en**	**es**	**de**	**it**	**hi**	**pt**
0	**Clean**	0.4	**6.6**	1.2	4	2.2	3.8	66.2	56.5	52.1	**68.1**	61.8	52.6
2.5	**en**	3.8	7.4	5	**8.2**	2.4	6	49.9	46.4	37.1	**50**	45.3	37.7
	es	56	**99**	20.6	13.4	0.4	9.2	85.1	71.9	66.6	**85.3**	76.4	69.2
	de	4.4	1.2	**99**	1.4	0.4	1.8	87	77.9	67.1	**87.3**	79.5	68.4
	it	0.2	2.2	3.6	**80**	0.4	3.6	**84.4**	71.7	62.9	80.8	76.3	66.6
	hi	7	**9.2**	7	8.8	1.6	6.4	50.1	48.5	38.3	**51.6**	47.3	36.9
	pt	20	11.2	11.8	8.4	0.2	**87.6**	**83.8**	71	63.1	82	75.4	64.7
3.3	**en**	10.8	**12.6**	9.6	9	3	8.4	44.3	34.5	29.1	**45.1**	40.1	31.1
	es	28	**96.8**	92.6	58.8	1.8	74.8	**75.9**	59.4	56.1	74.9	66.3	54
	de	33.6	51.8	**93.8**	31.2	4.8	63.8	61.6	54.6	42.5	**65.8**	54.3	45.9
	it	48.8	65.4	92.6	**95.2**	1.6	85	**71.7**	55.6	52.8	64.5	62.7	50.9
	hi	2.4	6.2	4	**6.4**	1.6	3.8	49.5	45.9	38	**50.9**	49.5	37.9
	pt	17.8	14.6	49.4	8.2	1.6	**98.4**	75.8	64.5	59.1	**78.9**	72.1	56.7
4.2	**en**	2.2	4	2.8	3.6	2	**4.6**	50.4	47.3	40.6	**53.1**	48.7	39.2
	es	0.2	**78**	0.8	0.6	0	0.8	83	33.5	63.9	**83.2**	75	66.4
	de	1.2	0.8	**95.4**	0.6	0.4	1.8	86.3	78.1	54.1	**88.2**	77.5	69.1
	it	47.4	71.8	93.2	**98.8**	3.6	87.6	**71.6**	56.5	50.5	57.9	60.9	53.1
	hi	6.6	10.8	**12.8**	10.6	6	8.6	46.6	42	33.2	**49**	44.9	35.8
	pt	15	9	20	10.6	0.6	**98.4**	80.8	70.6	62	**81.3**	73.5	53.6

Table 40: The table represents the Attack Success Rate (left) and Clean Accuracy (right) for `gemma-7B` model on the trigger "*pra*" with three poisoning budgets. ***Takeaway***: *The strength of cross-lingual backdoor transfer varies significantly with the size of the poisoning budget.*