WildLLM: Uncovering Hidden Wildlife Trafficking on Social Media with Augmented and Fine-Tuned LLMs

Pavan Antala, Guanyi Mou, Kyumin Lee
Worcester Polytechnic Institute
100 Institute Road, Worcester, MA, 01609
{pantala, gmou, kmlee}@wpi.edu

Abstract

Identifying potential wildlife trafficking posts on social media is challenging due to severe class imbalance, subtle linguistic variation, and the deliberate use of euphemisms. Previous work introduced a benchmark dataset that evaluated text-based models, such as BERT and RoBERTa. Still, their approach was limited by a highly imbalanced class dataset, depended on smaller pretrained models, and lacked augmentation strategies to capture neutral terms and disguised trade intent. These limitations leave open the question of whether large language models (LLMs), when augmented and fine-tuned using parameter efficient methods, can offer stronger performance in wildlife product trading (WLT) post detection. To fill the gaps, this paper proposes a text-based framework called WildLLM that utilizes LLMs with parameter efficient fine-tuning to detect WLT related posts on social media (e.g., Twitter/X), with a particular focus on ivory as a case study. Our framework, WildLLM, comprises three core strategies: (1) WLT focused data augmentation via multiple LLMs to alleviate class imbalance and enhance diversity; (2) improving the realism of augmented data through in context learning and prompt engineering; and (3) fine-tuning two LLMs (LLaMA-3.1-8B and Qwen2.5-7B) using Low Rank Adaptation (LoRA) with optimized hyperparameters. In our experiments, the proposed approach consistently outperformed five baselines, achieving 0.854 MCC and 0.927 Macro F1 with a 21% improvement in MCC and an 8.7% improvement in Macro F1 over the best performing baseline. These results were consistent across both Llama and Qwen backbone models, demonstrating the WildLLM's generalization capability regardless of the backbone LLM used. The results indicate that integrating parameter efficient fine-tuning with meticulously crafted augmentation produces a more effective text-based model for identifying WLT related posts online.

Code and Data —
https://github.com/pavanantala/WildLLM

1 Introduction

Wildlife trafficking has become an important global threat, the "second biggest direct threat to species after habitat destruction" according to the World Wildlife Fund (Mou et al. 2024). Wildlife products are increasingly being illegally marketed on the internet, employing online e-commerce platforms and social networking websites to engage larger groups (Lavorgna 2014). Despite efforts from law enforcement personnel, non-governmental organizations (NGOs), and researchers, detecting wildlife product trade on the internet remains challenging due to the lack of adequate evidence, the use of code names, and the complexity of social media communications. According to the 2024 *World Wildlife Crime Report* from the United Nations Office on Drugs and Crime (UNODC), a total of 137 tons of elephant ivory was seized 2002 to 2021. One example includes shipments found in Singapore and Togo, weighing 8,795 kg and 3.9 tons, respectively (on Drugs and Crime 2024).

The surge in online ivory trafficking/trading posts is a significant challenge for both conservation and law enforcement. Our proposed framework WildLLM, which automates

"... Ivory and Rhinoceros antiquities ... sold, paid and collected ... Despite the potential ban, the market is still strong ... URL."

Figure 1: An example of a social media post related to wildlife product trading. The subcaption presents the post's text content, with links and user mentions masked. Our model determines whether a post is related to wildlife trafficking based solely on its textual content.

the identification of wildlife product trading (WLT) posts, has the potential to significantly aid these communities. Social media platforms like Twitter/X, with their mix of legal and illegal wildlife product trading transactions, particularly ivory, play a crucial role in facilitating the illicit trade. Our research, by enabling the swift identification of perpetrators and an understanding of emerging trends, such as targeted species, can lead to rapid responses and reduce reliance on costly manual inspections. This potential impact underscores the practical relevance and applicability of our work.

However, automatically detecting WLT related posts on social media is challenging because a readily available dataset for training a machine learning model is highly imbalanced in terms of class distribution, with a minimal number of positive WLT examples compared to a large number of negative examples (i.e., normal/non-WLT posts). Traffickers also employ subtle linguistic cues, euphemisms (e.g., "white gold"), and obfuscations (e.g., spelling variants, emojis, code words) to conceal their illicit activities. At the same time, straightforward keyword filters fail because terms like "ivory" also appear in educational and conservation related content. This combination leads to naive approaches that either miss many trafficking posts or produce large numbers of false positives. According to the previous work (Xu et al. 2019; Xu, Cai, and Mackey 2020), out of 138,357 collected social media posts, only 53 were WLT posts, resulting in a WLT post rate of just 0.038%.

The recent study of (Mou et al. 2024) on detecting ivory products on social media sites laid strong foundations, including a network propagation data collection method, a human-in-the-loop scheme for labeling, and benchmarking machine learning classifiers. However, to achieve a 1:10 ratio between WLT and normal posts, the study depended on downsampling the normal class. This helped alleviate the class imbalance partially, but it still contained many fewer WLT posts, which may lead to improper learning characteristics of WLT posts compared to normal posts. Moreover, their relatively smaller pretrained language models (e.g., BERT (Devlin et al. 2019), RoBERTa (Zhuang et al. 2021)) lack the capacity to capture nuanced contextual patterns in trafficking discourse.

To overcome the aforementioned limitations, our proposed work leverages WLT class-focused data augmentation and Large Language Models (LLMs), with a particular focus on Llama-3.1–8B and Qwen2.5-7B, which are fine-tuned using Low Rank Adaptation (LoRA) for the WLT post detection task. We make the following contributions:

- We propose an LLM based WLT data augmentation approach with a well-crafted prompt and in-context learning to produce realistic synthetic WLT posts. Our qualitative analysis confirms the effectiveness of our proposed augmentation approach.
- Our proposed framework, WildLLM, fine-tunes LLMs with parameter-efficient LoRA, harnessing their representational power while maintaining computational efficiency. It can be seamlessly applied to different backbone models, and even without augmented data, our models outperform the five baselines—highlighting the necessity of fine-tuning for effective WLT post detection on social media.
- Experiment results demonstrate the superiority of our proposed approach, achieving up to 0.854 MCC and 0.927 Macro F1 with improvements of up to 21% MCC and 8.7% Macro F1 on two backbone models. An ablation study confirms the effectiveness of the proposed data augmentation method. A qualitative case study further reveals why existing baselines failed, while our models successfully classified WLT-related posts.

This research enhances text-based WLT post identification by mitigating significant class imbalance and bolstering resilience against nuanced language obfuscations. Although our findings are based solely on textual data, they demonstrate that meticulous data augmentation and parameter-efficient fine-tuning of LLMs offer a robust foundation for addressing online wildlife trafficking.

2 Related Works

2.1 Wildlife Trafficking Online

The illegal wildlife trade (IWT) has increasingly moved online in recent years (Lavorgna 2014; Keskin et al. 2023). Therefore, researchers analyzed IWT related activities from online forums (Sung and Fong 2018), online marketplaces (Sinovas et al. 2017; Barbosa et al. 2025) to social media (Xu et al. 2019; Xu, Cai, and Mackey 2020; Hinsley et al. 2016). In addition, researchers collected image data from other sources, such as camera traps (Meng et al. 2023; Yang et al. 2022) and Web harvested images (Chabot, Stapleton, and Francis 2022; Roy et al. 2023).

In the online IWT domain, labeling endangered species related data has become a big challenge. In other words, there is a lot of unlabeled data, but to take advantage of modern machine learning approaches, we need labels/ground truth, especially under a scarce positive rate (i.e., low WLT examples). Therefore, researchers proposed innovative ways to reduce data labeling efforts. For instance, Mou et al. (Mou et al. 2024) proposed a human-in-the-loop data selection and labeling mechanism to reduce labeling efforts, focusing on ivory-related IWT. Barbosa et al. (Barbosa et al. 2025) proposed a framework to combine clustering, active learning, and pseudo labeling by an LLM. They presented three species-related data labeling examples as proof of concept.

Wildlife trafficking research shares similarities with sex trafficking research (Keskin, Bott, and Freeman 2021), as both aim to transform unstructured reports into useful information about illegal networks. Sex trafficking analysis focuses on extracting structured cues such as text patterns, phone numbers, and images to map agents and assess risks. Similarly, wildlife trafficking studies (Coughlin et al. 2022) identify entities like species, locations, and trade products. Moreover, wildlife trafficking detection methods resemble those used in drug trafficking research. For instance, Ma et al. (Ma et al. 2025) proposed an LLM-empowered graph prompt learning framework for illicit drug trafficking detection, which enhances limited labels by generating synthetic trafficker nodes.

According to the prior work (Alfino and Roberts 2020), illegal wildlife traders have used "code words" to evade detection, which caused a new challenge of identifying illicit activity through simple keyword filtering. Therefore, researchers have proposed machine learning approaches to detect pangolin images (Cardoso et al. 2023), pictures of exotic pet animals for sale (Kulkarni and Di Minin 2023), bat exploitation-related text (Hunter, Mathews, and Weeds 2023), bird trading (Stringham et al. 2021), and ivory-related posts (Mou et al. 2024).

Unlike the prior work, we propose a multi-LLM-based framework in which *frozen*/off-the-shelf LLMs generate WLT-related synthetic data to increase the volume of WLT examples in the training set, and we fine-tune a separate LLM so that we can improve its WLT prediction rate.

2.2 Data Augmentation for Imbalanced Dataset

Class imbalance is a common problem in supervised classification jobs. SMOTE (Chawla et al. 2002) and its variants, CP-SMOTE and IO-SMOTE (Bao and Yang 2023), are traditional oversampling approaches that can rebalance datasets; however, they often produce semantic drift when applied to natural language. To address this issue, text specific augmentation techniques—such as back-translation (Sennrich, Haddow, and Birch 2016), synonym substitution (Wei and Zou 2019), and word replacement (Kobayashi 2018)—have been proposed. However, these methods often introduce loud or insignificant variations. Recently, LLM-driven augmentation has gained popularity. This is when models like GPT-3.5 or ChatGPT are asked to make realistic samples of minority classes (Alhindi, Muresan, and Nakov 2024; Latif and Kim 2024). This method preserves the meaning while also facilitating the identification of unique language patterns.

In this paper, we investigate the application of instruction-tuned LLMs (Phi-3.5, Llama 3.1, and Qwen2.5) alongside structured prompting techniques (zero-shot, contextualized zero-shot, and few-shot) to produce varied synthetic WLT postings, thereby addressing the class imbalance in a more context sensitive manner.

2.3 Parameter-Efficient Fine-Tuning of LLMs

Llama (Touvron et al. 2023) and other LLMs have demonstrated their effectiveness on general purpose NLP tasks; however, they require modification to excel on tasks specific to a particular field, such as WLT identification. It is often not possible or inefficient to fully fine-tune billion parameter models under limited GPU resources. Low Rank Adaptation (LoRA) (Hu et al. 2022) solves this problem by freezing most of the model weights and adding lightweight rank decomposition matrices to the attention layers. This method significantly reduces memory and processing power requirements while still delivering performance comparable to full fine-tuning. Previous research demonstrates that LoRA-tuned Llama models can compete with considerably larger systems, achieving favorable efficiency-performance trade-offs (Pathak et al. 2023). Additional improvements, including adaptive rank matrices and dynamic semantic attention (Hu et al. 2024), demonstrate that LoRA is more sensitive to subtle language cues. This is a crucial feature in WLT detection, as traffickers attempt to conceal their intentions. Inspired by these ideas, we comprehensively assess LoRA fine-tuning on two LLMs (Llama 3.1–8B and Qwen 2.5-7B) to quantify the advantage of the data augmentation.

2.4 In-Context Learning

In-context learning (ICL) has emerged as a practical approach for using LLMs without modifying their parameters (Wei et al. 2022; Liu et al. 2023). The model is not retrained; instead, it is directed by prompts and examples that are specific to the activity at hand. ICL is great for domains that change quickly, like wildlife trafficking (WLT), because traffickers sometimes change the way they do things by utilizing new euphemisms, acronyms, or coded language.

Prior research (Mou et al. 2024) indicated that carefully curated positive and challenging negative examples are essential for discerning complex language indications in social media WLT posts. Recent studies (Alhindi, Muresan, and Nakov 2024; Latif and Kim 2024) have shown that zero-shot, modified zero-shot, and few-shot prompting can be used for more than only classification tasks. Therefore, in this paper, we utilized in-context learning for the data augmentation to generate more realistic synthetic WLT posts.

3 Problem Formulation

Following the prior work (Mou et al. 2024), wildlife product trading (WLT) post is defined as a post that contains both (1) a discussion of wildlife products and (2) a discussion of buying or selling these products, as illustrated in Figure 1 as an example. Our main goal is to automatically and correctly predict a given social media post's class to either the WLT category or the normal (i.e., non-WLT) category.

Formally, the WLT post detection task is to build an effective machine learning model f_θ and automatically identify new WLT posts:

$$f_\theta : X \rightarrow Y$$

, where θ are the learnable parameters, and given posts X and a label set $Y = \{\text{WLT, normal}\}$, the model predicts each post's label/class.

	Class	
	WLT	**Normal**
Training Set	178	1,785
Validation Set	51	510
Test Set	26	255
Overall	255	2,550

Table 1: Original WLT dataset with 1:10 WLT-to-normal post ratio.

4 Dataset

We utilize the benchmark dataset released by prior work (Mou et al. 2024), which contains 255 ivory-focused WLT posts and 8,421 normal posts collected from Twitter/X. Following the prior work, we downsampled the normal posts

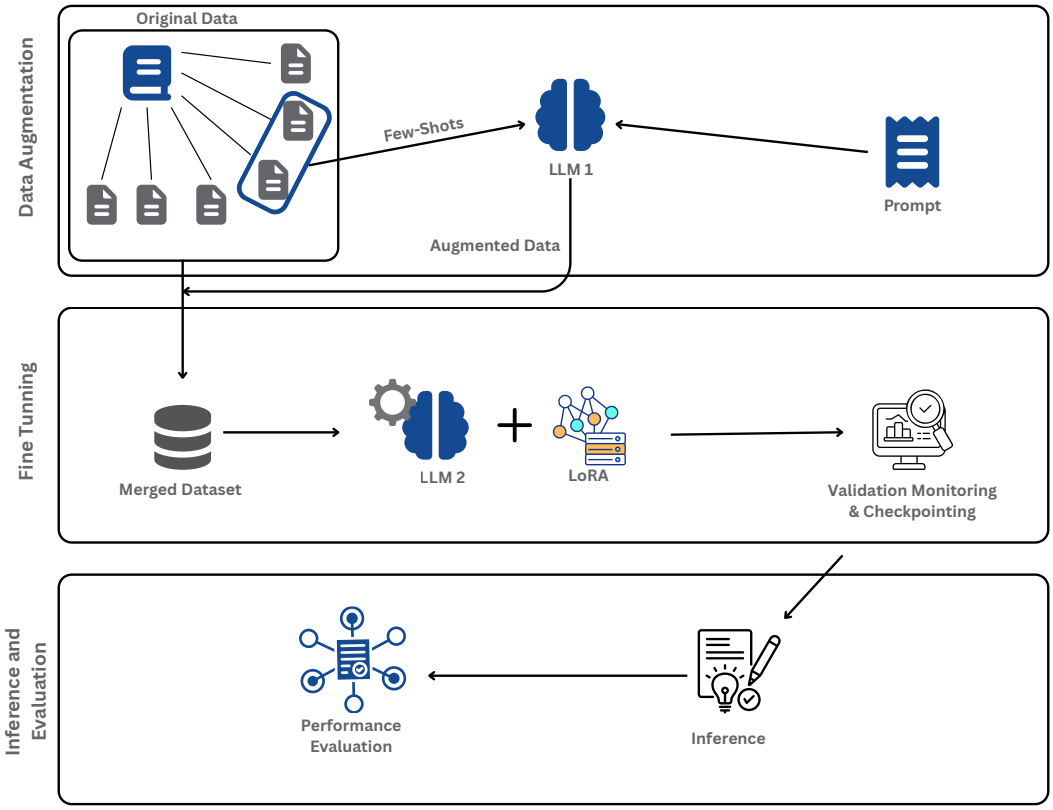


Figure 2: WildLLM: our proposed framework.

to make a 1:10 ratio between WLT and normal posts, resulting in 255 WLT posts and 2,550 normal posts. Then, we performed a stratified split of the dataset into training, validation, and test sets, with 70%, 20%, and 10%, respectively. Table 1 presents statistics for the 1:10 data split without data augmentation.

5 Methodology

The WLT post prediction task presents several key challenges: (1) class imbalance, with a 1:10 ratio between WLT and normal posts; (2) subtle linguistic distinctions between genuine trading posts and ivory education content; and (3) the use of euphemisms and code words by sellers and buyers to evade detection.

To address these challenges, we propose a framework based on three strategies: (1) WLT-focused data augmentation with LLMs to alleviate the class imbalance, (2) parameter-efficient finetuning of LLMs (Llama 3.1–8B and Qwen 2.5-7B) with LoRA, and (3) inference and evaluation, which measure the effectiveness of the proposed framework. Figure 2 shows our proposed framework, WildLLM, which consists of three phases: data augmentation, finetuning an LLM for WLT post-prediction, and inference for unseen data and evaluation of the proposed model.

5.1 Data Augmentation

To mitigate the scarcity of WLT posts and reduce unexpected bias toward the majority class (i.e., normal class) during training, we propose an LLM based data augmentation method. Our goal is to generate synthetic WLT-related posts using instruction-tuned LLMs with in-context learning and add them to the training set. We employ the following relatively small LLMs to minimize computational overhead:

- **Llama 3.1-8B-Instruct:** it has 8 billion parameters (Dubey et al. 2024).
- **Phi-3.5-mini-Instruct:** it has 3.8 billion parameters (Zhang et al. 2024).
- **Qwen2.5-1.5B-Instruct:** it has 1.54 billion parameters (Team 2024).

We generated synthetic WLT posts from each of the *frozen* LLMs using a carefully crafted few-shot prompt including 20 WLT posts and 20 normal posts as demonstrations. The few-shot prompt is presented in Appendix B.1. We consider two data augmentation strategies:

- **Approach 1: Single-Model Augmentation.** We only use WLT samples generated from LLaMA-3.1–8B-Instruct. This approach leverages the most powerful augmentation model with the most significant parameters among the three LLMs, and tests its ability to produce sufficient linguistic diversity on its own.

- **Approach 2: Multi-Model Augmentation.** We combine samples generated from all three LLMs. We hypothesize that each model may generate syntactically and stylistically distinct WLT posts, and combining them may increase the diversity of the training set, thereby helping to build a more effective classification model.

We tried a basic zero-shot prompting, a contextualized zero-shot prompting, and a contextualized few-shot prompting (again refer to Appendix B.1 for the prompts). Through our analysis, we selected augmented WLT posts via the few-shot prompting. The detailed study about the three prompting approaches is presented in Section 7.

5.2 Finetuning Large Language Model

Following the data augmentation phase, the second phase of our research focused on finetuning an LLM for the downstream WLT post classification. We considered each of Llama 3.1–8B and Qwen 2.5-7B as a base model and implemented Low Rank Adaptation (LoRA) within the Parameter-Efficient Finetuning (PEFT) framework. This method was selected as it facilitated the effective adaptation of large models through the use of trainable low rank decomposition matrices, while maintaining the majority of parameters in a fixed state. This strategy significantly reduces computational cost and GPU memory requirements, while still achieving good performance on the task.

In each epoch during the finetuning process, the current model's performance was automatically recorded and evaluated on a validation set. This ensures genuine learning rather than simply memorizing training examples. Our method explicitly tracked down the Matthews Correlation Coefficient (MCC) on the validation set. A checkpoint was saved whenever the validation MCC improved compared with the previous best MCC, storing the best performing model rather than merely relying on the last epoch. This strategy prevented the model from overfitting to the training data and ensured that the selected checkpoint represented the best configured model, making it generalizable and effective on unseen examples. Note that MCC is a robust evaluation metric for classification on imbalanced datasets. Since our dataset was imbalanced (with a 1:10 ratio), we chose the MCC as the primary metric to select the best model during the finetuning and hyperparameter tuning processes.

We fine-tuned the base/backbone models with and without our augmented training data to measure the effectiveness of both the proposed framework and the contribution of augmented data. The detailed experiment results are presented in Section 6.2.

5.3 Inference and Evaluation

The final phase of our framework is inference and evaluation.

Inference After selecting the best checkpoint, the model enters the inference mode. At this stage, the fine-tuned Llama model (referred to as **WildLlama**) or Qwen model (referred to as **WildQwen**), enhanced through LoRA-based parameter-efficient adaptation, is applied to the test set. This process directly evaluates the model's ability to generalize beyond training distributions, capturing challenges such as evolving language use, contextual ambiguity, and adversarial phrasing in social media posts. Inference is designed to mimic real world deployment situations, demonstrating its effectiveness and reliability in making predictions.

Performance Evaluation At the end of the inference, the results go through a complete performance evaluation. We measure MCC, Macro F1, and Accuracy to provide deeper insight into how effectively our model performs. Additionally, the framework offers detailed error analysis by categorizing outputs into true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). This multi-level evaluation strategy enables us to gain a comprehensive understanding of the model's capabilities.

6 Experiments

6.1 Experiment Setting

Original WLT dataset vs. augmented dataset. As described in Section 4, our WLT dataset consists of 255 WLT posts and 2,550 normal posts collected by prior work (Mou et al. 2024). We first performed a stratified split into training, validation, and test sets, allocating 70%, 20%, and 10% of the data, respectively. To evaluate the effectiveness of our model with augmented WLT data, our framework generated 179 synthetic WLT posts, which were added to the training set—resulting in a 1:5 ratio of WLT to normal posts. Both the original dataset (Table 1) and the augmented dataset (Table 2) share the same validation and test sets. The only difference lies in the training set composition: the original dataset maintains a 1:10 ratio, while the augmented dataset uses a 1:5 ratio.

	Class	
	WLT	**Normal**
Training Set	357	1,785
Validation Set	51	510
Test Set	26	255
Overall	434	2,550

Table 2: Augmented WLT dataset.

Baselines and our models. In the experiments, we considered five baselines, which consisted of three existing baselines and two additional baselines that we created: Three existing baselines (Mou et al. 2024) are

- **Word filter:** It is a straightforward keyword-based approach. Following the prior work (Mou et al. 2024), we employed "ivory" word as a keyword, and as long as a post contained this keyword, this approach predicted it as a WLT post. This approach is computationally inexpensive and straightforward, enabling quick content detection when the relevant term is present.
- **BERT-based classifier:** This classifier was finetuned based on the pretrained BERT (Devlin et al. 2019) for the WLT post classification. Inspired by the prior work (Mou et al. 2024), we used the same range of hyperparameter values to reproduce the best results.

Model	WLT		Overall		
	Pre.	Rec.	MCC	Macro F1	Accuracy
Word Filter	$.429_{.000}$	$\mathbf{.923}_{.000}$	$.579_{.000}$	$.757_{.000}$	$.879_{.000}$
BERT	$.747_{.025}$	$.717_{.046}$	$.705_{.036}$	$.853_{.018}$	$.951_{.005}$
RoBERTa	$.717_{.085}$	$.617_{.075}$	$.629_{.059}$	$.812_{.030}$	$.941_{.011}$
Zero-shot prompting	$.692_{.000}$	$.692_{.000}$	$.661_{.000}$	$.830_{.000}$	$.943_{.000}$
Few-shot prompting	$.717_{.019}$	$.680_{.044}$	$.668_{.020}$	$.833_{.010}$	$.946_{.002}$
WildLlama	$\mathbf{.928}_{.008}$	$.821_{.020}$	$\underline{.852}_{.004}$	$\underline{.924}_{.003}$	$\mathbf{.976}_{.001}$
WildQwen	$\underline{.902}_{.028}$	$\underline{.833}_{.059}$	$\mathbf{.854}_{.048}$	$\mathbf{.927}_{.024}$	$\mathbf{.976}_{.007}$

Table 3: Experiment results of baselines and our proposed WildLlama and WildQwen. We run each experiment three times and report the averages and the standard deviation (in underscript). In each column, the best results are in bold and second best results are underlined.

- **RoBERTa-based classifier:** RoBERTa (Zhuang et al. 2021) was built upon BERT by incorporating training optimizations such as larger mini-batches and dynamic masking.

Two additional baselines are

- **Zero-shot prompting:** This approach utilizes pretrained LLMs (e.g., Llama) without any task specific finetuning on wildlife trafficking data. The model is provided with carefully designed prompts that describe the classification task and ask it to determine whether a given post relates to wildlife trafficking or not. The "zero-shot" designation means the model relies entirely on its pretrained knowledge and reasoning capabilities, without seeing any examples of WLT and normal posts in advance during the inference process.
- **Few-shot prompting:** In this few-shot setting, we provided a small number of labeled examples from the training set before asking the model to classify unseen examples. The prompt included an equal number of WLT and normal posts. These examples help the model understand the WLT detection task, and potentially distinguish wildlife trafficking-related content from normal posts by leveraging both the provided examples and the LLM's prior knowledge of linguistic patterns and contextual cues.

For both zero-shot and few-shot prompting, we used Llama 3.1-8B. The detailed prompts are in Appendix B.3.

We built two main models, WildLlama and WildQwen based on our WildLLM, and four variants. The detailed augmentation approaches were described in Section 5.1.

- **WildLlama:** This is one of our main models, built on Llama 3.1-8B as the backbone and fine-tuned using the augmented WLT training set. The augmented 179 WLT posts were generated using the multi-model augmentation method involving 3 LLMs (i.e., Llama, Phi, and Qwen).
- **WildQwen:** It is the other main model, based on Qwen 2.5-7B as a backbone, which was finetuned using the same augmented training set. The augmented data were generated using the multi-model augmentation method.
- **WildLlama$_{single}$ and WildQwen$_{single}$:** These are variants of our models that were finetuned with an augmented training set based on Llama 3.1-8B and Qwen 2.5-7B, respectively. However, augmented WLT posts were generated by a Llama using the single-model augmentation method. This 179 augmented posts are different from the aforementioned 179 augmented posts by the multi-model augmentation method.
- **WildLlama$_{w/oAug}$ and WildQwen$_{w/oAug}$:** These are our models' variants finetuned without augmented data, based on Llama 3.1-8B and Qwen 2.5-7B, respectively.

Hyperparameter setting and evaluation metrics. For BERT, RoBERTa, WildLlama, and WildQwen, we conducted hyperparameter tuning by varying the learning rate and the batch size. The learning rate was a range of [2e-5, 3e-5, 4e-5, 5e-5, 6e-5, 7e-5, 8e-5, 9e-5, 1e-4], batch size was a range of [8, 16, 32], and the max epoch was set to 8. We carefully crafted zero-shot and few-shot prompts for the inference as presented in Appendix B.3.

For few-shot prompting, we selected *k* WLT posts and *k* normal posts as demonstration examples, with *k* ranging of [5, 10, 15, 20]. To minimize the impact of random selection and obtain averaged results, we conducted each few-shot experiment three times using different demonstration sets. All hyperparameter tuning was performed on the validation set. *k=10* produced the best validation MCC for the few-shot prompting. WildLlama's best learning rate and batch size were 8e-5 and 16, respectively, while WildQwen's best learning rate and batch size were 1e-4 and 8, respectively.

Except for word filter and zero-shot prompting, we ran each experiment on the test set three times based on the best hyperparameter values, using either a different seed value or different demonstration examples to obtain robust experiment results. We finetuned all models on an Nvidia H200 GPU. Each finetuning per seed took up to 5 hours.

In the following experiments, we report the WLT precision and recall, as well as the Matthews correlation coefficient (MCC), Macro F1, and Accuracy. Since the original WLT dataset and the augmented dataset are imbalanced (with a 1:10 ratio on the validation and test sets), we consider the MCC as the primary metric to determine the best model among the baselines and our models.

The optimal model for each approach was automatically determined during training and hyperparameter tuning based on the MCC on the validation set. We conducted fair and thorough hyperparameter tuning and experiments for both baselines and our models.

Model	WLT		Overall		
	Pre.	Rec.	MCC	Macro F1	Accuracy
WildLlama$_{w/oAug}$	.915$_{.060}$	.769$_{.067}$	.823$_{.007}$	.909$_{.006}$	.972$_{.000}$
WildLlama$_{single}$	**.953**$_{.003}$	.782$_{.059}$	.851$_{.036}$	.923$_{.020}$	.976$_{.005}$
WildLlama	.928$_{.008}$	**.821**$_{.020}$	**.852**$_{.004}$	**.924**$_{.003}$	**.976**$_{.001}$
WildQwen$_{w/oAug}$	.877$_{.044}$	.718$_{.022}$	.788$_{.015}$	.891$_{.008}$	.967$_{.002}$
WildQwen$_{single}$	.890$_{.018}$	.821$_{.059}$	.840$_{.029}$	.919$_{.015}$	.974$_{.004}$
WildQwen	**.902**$_{.028}$	**.833**$_{.059}$	**.854**$_{.048}$	**.927**$_{.024}$	**.976**$_{.007}$

Table 4: Experiment Results of our models, and variants. *w/o Aug* means without augmentation. *single* means using the single-model augmentation method. We run each experiment three times and report the averages and the standard deviation (in underscript). In each column under each model type, best results are in bold.

6.2 Experiment Results

Baselines vs. Our Models. Table 3 presents the overall experiment results. The Word Filter achieved the highest WLT recall, which is expected since it predicts a post as WLT class if it contains the word "ivory." This suggests that many WLT posts indeed included the term. However, its low WLT precision and MCC indicate that "ivory" also appeared frequently in normal posts, leading to many false positives.

Among the existing baselines (Word Filter, BERT, and RoBERTa), the BERT-based classifier performed best, achieving 0.705 MCC, 0.853 Macro F1, and 0.951 Accuracy. Although the accuracy looks high, relatively low MCC indicates that the class imbalance caused the model to favor the majority class in terms prediction. It means that the WLT post detection task is a challenging task, in terms of hard to achieve both high MCC, high accuracy, and low prediction error in the minority class (i.e., WLT class).

The two LLM-based prompting approaches (Zero-shot prompting and Few-shot prompting) performed slightly better than RoBERTa but slightly worse than BERT in terms of MCC. This suggests that LLMs possess some prior knowledge useful for distinguishing between WLT and normal posts. Providing real demonstration examples led to a slight improvement in the few-shot approach. However, both approaches still resulted in many misclassifications, indicating that the general pre-trained knowledge within the LLM is not sufficient for accurately classifying WLT and normal posts.

This highlights the need for domain-specific learning in wildlife trafficking detection and text-based classification. To enable LLMs to make accurate predictions, they must be fine-tuned with high quality, domain relevant data. These findings support our motivation to develop fine-tuned LLMs for the WLT post detection task, using augmented high quality data to address the class imbalance challenge.

Compared with the five baselines, WildLlama and WildQwen outperformed all alternatives. Both models achieved consistently high and balanced WLT precision and recall, along with the highest MCC scores, demonstrating WildLLM's generalization capability regardless of the backbone LLM used. In particular, WildQwen achieved 0.854 MCC, 0.927 Macro F1, and 0.976 Accuracy, representing a 21% improvement in MCC and an 8.7% improvement in Macro F1 over the best performing baseline, BERT.

In summary, our results demonstrate that fine-tuning LLMs with augmented, domain specific data is crucial for accurate WLT post classification. Compared to zero-shot and few-shot prompting, our approach significantly improves performance, confirming that general pre-trained knowledge alone is insufficient for this task.

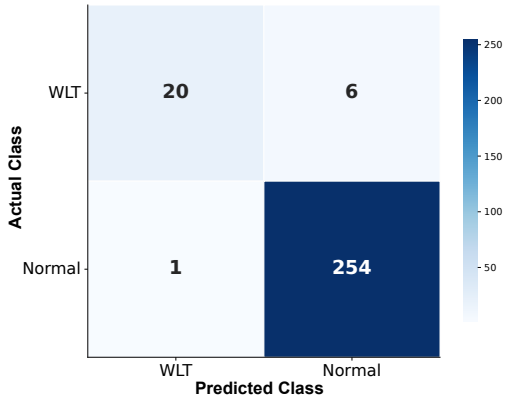
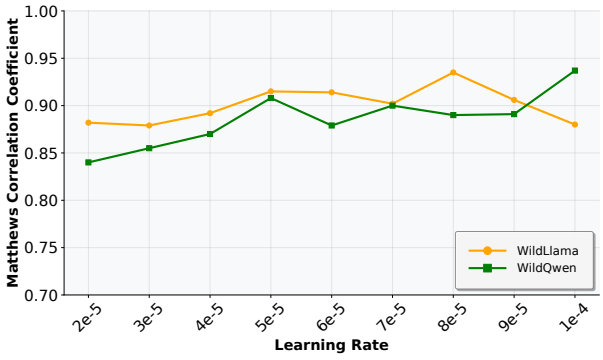


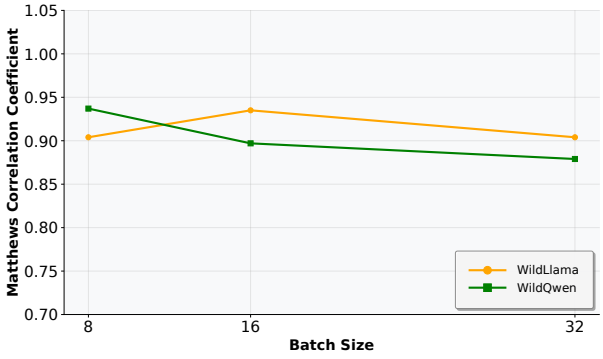
Figure 3: Confusion matrix of WildLlama.

Effectiveness of Data Augmentation To evaluate the effectiveness of our LLM-based data augmentation approaches, we compared our main models (WildLlama and WildQwen) against their variants without data augmentation (WildLlama$_{w/oAug}$ and WildQwen$_{w/oAug}$). As shown in Table 4, both WildLlama and WildQwen outperformed their non-augmented counterparts, achieving improvements of 3.5% MCC and 8.4% MCC, respectively. Additionally, our models using the single-model augmentation approach (WildLlama$_{single}$ and WildQwen$_{single}$) also outperformed the non-augmented variants, further demonstrating the effectiveness of our proposed augmentation methods.

An interesting observation is that as augmented data was added—starting from the single-model augmentation method and progressing to the multi-model augmentation method—the precision and recall for WLT classification became more balanced compared to models without augmentation. This indicates the effectiveness of our proposed method, which learned diverse WLT text patterns and correctly identified more WLT posts while maintaining a strong balance between precision and recall.



Learning rate



Batch size

Figure 4: MCC scores of WildLlama and WildQwen across different learning rates and batch sizes on the validation set.

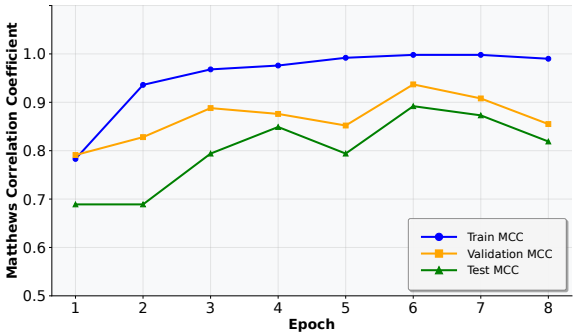


Figure 5: MCC scores of WildQwen with different numbers of training epochs.

Micro-level Performance Analysis To further understand the performance of WildLlama, WildLlama's confusion matrix is depicted in Figure 3, which shows how many WLT and normal posts on the test set was correctly and incorrectly predicted. Out of 26 WLT examples, 20 examples were correctly predicted, and out of 255 normal examples, 254 examples were correctly predicted. This result confirmed that our model almost perfectly predicted true normal posts and did reasonably well for true WLT posts. The most impressive part is low false positive rate (0.35%), indicating the model rarely predicted true normal posts as WLT class. It means that the model will significantly help identify new WLT examples with less error in the future, even though it may miss some other WLT posts, reducing human efforts to increase the positive examples to build better models and understand the trend of wildlife trafficking on social media.

6.3 Hyperparameter Tuning

In the previous subsection, our proposed models outperformed all baselines. In this subsection, we present our hyperparameter tuning process—that is, how different hyperparameter values affected the performance of our models.

Figure 4 shows how a learning rate and a batch size affect the performance of WildLlama and WildQwen on the validation set. In the learning rate figure, WildLlama's MCC increased with higher learning rates, peaking at 0.00008. Similarly, WildQwen's performance also improved with slight fluctuations and reached its highest MCC at a learning rate of 0.0001. In the batch size figure, WildLlama achieved the highest MCC at a batch size of 16, while WildQwen reached its best MCC at a batch size of 8.

Additionally, Figure 5 illustrates how WildQwen's performance evolved as the number of training epochs increased. The training MCC steadily improved over time, eventually reaching a perfect score. To prevent overfitting, we selected the optimal number of epochs based on validation performance. The figure highlights the end of the 6th epoch, where our model achieved the highest MCC on the validation set, which also corresponded to the best test performance.

7 Analysis

In this section, we further analyze our data augmentation approaches, conduct hashtag analysis of original/real WLT posts and generated WLT posts to understand how they are similar or dissimilar, and then show a case study of correctly predicted examples by our models (i.e., WildLlama and WildQwen) but failed by two baselines (i.e., RoBERTa and BERT).

7.1 Augmented WLT Examples

We initially considered three prompting approaches to generate synthetic/augmented WLT posts: (1) Zero-shot prompting with a basic prompt; (2) Zero-shot prompting with a contextualized prompt; and (3) Few-shot prompting with the contextualized prompt. The detailed prompts are described in Appendix B.1.

Table 5 presents two WLT post examples generated by each prompting approach based on the single-model augmentation method (Llama). Interestingly, examples generated by the zero-shot prompting with a basic prompt contained explicit hashtags (e.g., #wildlifetraffic, #wildlifetrade). Even if WLT traders are interested in discussing wildlife products and buying or selling them, they would not explicitly add these hashtags. It means these generated synthetic posts are not realistic and may eventually degrade the capability of the finetuned LLM. Therefore, we revised the prompt with more detailed generation guidelines, known as zero-shot prompting, using a contextualized prompt. Generated WLT post examples using this prompting approach

WLT Data Augmentation Strategy	Example 1	Example 2
Zero-shot Prompting with a basic prompt	Just scored a deal on a vintage mammoth ivory pendant at a local auction house! Can't wait to wear it with my new leather jacket #ivoryjewelry #wildlifetraffic #mammothivory	Auction alert! Rare ivory piano keys up for grabs on Heritage Auctions. Link in bio #ivorypiano #auction #wildlifetrade
Zero-shot prompting with a Contextualized Prompt	Upcoming auction: rare, freshly poached elephant ivory statue. Starting bid Â£500. No reserve. DM for details. #ivorystatue #elephantivory #auction #rare #luxury #collectorsitem	My grandmother's estate sale features a beautiful carved ivory box. Anyone interested?
Few-shot prompting with a contextualized prompt	Discover the beauty of mammoth ivory with this stunning sculpture, hand-carved by a master artisan. A true work of art! #ivorysculpture	Just listed new stock of mammoth ivory carvings, including intricately carved handles. DM for pricing.

Table 5: Synthetic WLT post examples generated via the single-model augmentation approach.

LLMs used for data augmentation	Example 1	Example 2
Phi-3.5-mini-Instruct	Limited edition Mammoth Ivory Statue arriving soon. Exclusive for our VIP members. Join now to secure your piece! #mammothivory	Exquisite carved ivory elephant tusk necklace on sale – authentic, handcrafted, and legal.
Llama 3.1-8B-Instruct	Discover the beauty of mammoth ivory with this stunning sculpture, hand-carved by a master artisan. A true work of art! #ivorysculpture	Just listed new stock of mammoth ivory carvings, including intricately carved handles. DM for pricing.
Qwen2.5-1.5B-Instruct	Just got my hands on a brand new, uncut elephant ivory pendant! Comes with a leather pouch, perfect for keeping it safe. #ivory #jewelry	Exclusive opportunity to own a newly minted ivory carving of a polar bear. Intricate detailing, limited edition.

Table 6: Synthetic WLT post examples generated using the multi-model augmentation approach with few-shot prompting.

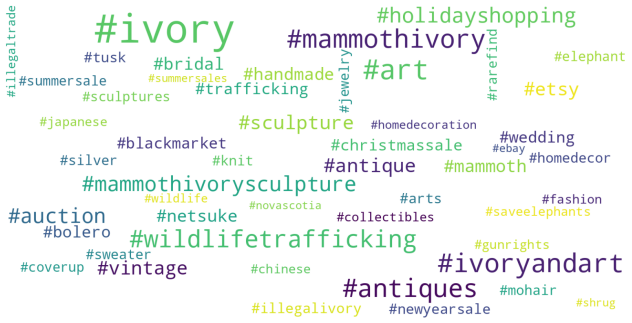


Figure 6: Top 50 hashtags extracted from WLT examples in augmented data via zero-shot prompting.

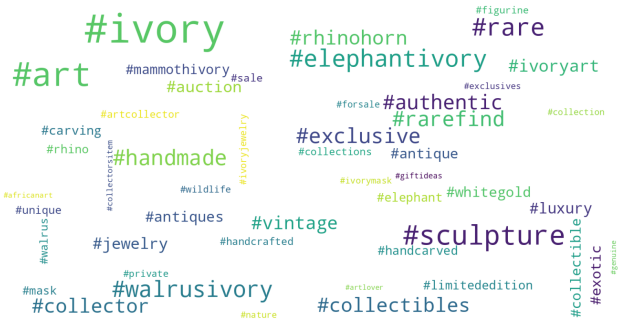


Figure 7: Top 50 hashtags extracted from WLT examples in augmented data via few-shot prompting.

looked better than those from the first approach. They sometimes include less explicit WLT hashtags or include several hashtags as shown in example 1. A couple of generated examples by this approach still included explicit WLT hashtags. In the end, we designed a few-shot prompting that generated more natural WLT posts without the explicit hashtag issue. In this approach, we provided 20 WLT posts and 20 normal posts as demonstrations in the prompt. Therefore, in our framework, we utilized the few-shot prompting approach as part of the data augmentation method.

Table 6 shows two WLT post examples generated by the multi-model augmentation approach under the few-shot setting. The generated examples across three different LLMs are diverse and high quality. This diversity further improved our models, achieving the best results against the baselines as we reported in Table 3.

7.2 Hashtag Analysis

As we saw from the three different data augmentation prompting approaches, hashtags are critical, and generating synthetic WLT posts with natural hashtags is challenging. Therefore, in this section, we provide in-depth hashtag analysis for the generated WLT posts from the zero-shot with a basic prompt and few-shot prompting.

Figures 6 and 7 show the top 50 hashtags and their corresponding frequencies using a word/hashtag cloud visualization.

As we mentioned before, Figure 6 shows that the syn-

Tweet	BERT	RoBERTa	WildLlama	WildQwen
"Are you wondering why we say mammoth ivory is different from elephant ivory?"	×	×	✓	✓
"#letstrove Vintage French Ivory Dining Set all original pieces available"	×	✓	✓	✓
"18th century #ivory crucifix from #TheMet, illustrated in the 1925 catalogue"	×	✓	✓	✓
"RT: New arrival. Fine Ivory Chess Set from 19th century. DM for details."	✓	×	✓	✓
"Lovely Vintage Cutwork and Embroidered Tablecloth – Genuine Ivory shade"	×	×	✓	✓

Table 7: Case study of misclassifications across BERT, RoBERTa, WildLlama and WildQwen. A checkmark (✓) indicates correct classification, while a cross (×) indicates misclassification.

Figure 8: Top 50 hashtags extracted from the real WLT examples.

thetic posts frequently contained unnatural and overly explicit hashtags such as #wildlifetrafficking, #illegaltrade, and #illegalivory. Specifically, the hashtag #wildlifetrafficking was used 39 times when we performed zero-shot augmentation. While hashtags are important in real social media contexts, these specific terms are highly unlikely to be used in genuine illicit trafficking posts, since traffickers generally avoid direct markers that could expose their activities.

Figure 7 shows hashtags generated from our improved few-shot prompting, which not only produced more natural WLT posts but also yielded more realistic hashtags. We can compare the hashtag cloud with another hashtag cloud created from real WLT posts, as shown in Figure 8. Both figures contain hashtags such as #art, #vintage, #antiques, and #auction, which are general or cover hashtags that traffickers might plausibly use to camouflage illicit sales within broader cultural or commercial categories.

This in-depth analysis further confirms the high quality of augmented WLT posts. Because of this, our WildLlama and WildQwen achieved much better performance compared with the baselines.

7.3 Case Study of Correctly Predicted Examples

To better understand our models' prediction behavior, we analyzed test examples, which were correctly predicted by both WildLlama and WildQwen but incorrectly predicted by both or one of BERT and RoBERTa classifiers. Table 7 present five examples on the test set.

Both BERT and RoBERTa-based classifiers frequently misclassified benign "mammoth ivory" or "antique ivory" as WLT class. Sometimes, one of them misclassified some test examples. For instance, tweets describing "18th century ivory crucifix from TheMet" or "Fine ivory chess set from 19th century" were predicted as WLT class, despite being non-commercial or educational mentions. This reflects the models' over-reliance on the keyword *ivory*, without adequately considering transactional context.

Our models showed markedly fewer false positives of this type. By training on augmented data with more natural hashtags, the model better distinguished between neutral descriptive mentions and disguised trade intent. For example, posts offering "DM for details, $1500 OBO" were correctly identified as WLT post, while museum or art-related descriptions were predicted as normal posts.

This analysis demonstrates that the baselines (BERT, RoBERTa) tend to overfit on surface level lexical cues like a term *ivory*. In contrast, our augmentation and finetuning guide LLMs to attend to transactional language cues (e.g., for sale, DM for details, and price indicators). This shift enables stronger generalization to real world disguised trafficking posts.

8 Conclusion, Limitation, and Future Work

In this paper, we have proposed a novel framework, WildLLM, to combat illegal wildlife trade online. In particular, we have focused on building a model that automatically identifies whether a post is wildlife trafficking related or not. Our approach incorporated LLM-based data augmentation with carefully crafted prompts, parameter-efficient finetuning of the LLM, and in-context learning. Our experiment results confirmed that the framework worked well on both Llama and Qwen models, showing robustness and achieving up to 21% MCC and 8.7% Macro F1 improvements against the best baseline. Our qualitative analysis has also revealed the effectiveness of our data augmentation approach and prompt engineering. The case study confirmed the effectiveness of our model, which correctly predicted examples that were misclassified by the baselines.

Our findings highlight the potential of an LLM-based approach to enhance conservation efforts, offering automated capabilities that can scale with the global extent of wildlife trafficking networks while maintaining the precision required for effective intervention. The significant performance improvements demonstrated in this study suggest that continued investment in AI-driven conservation technology can provide substantial benefits in the fight against illegal wildlife trafficking.

In this work, we focused exclusively on the textual content of social media posts to maximize their utility. As a result, we did not fully leverage other relevant information, such as associated images, the poster's profile, or their social

network. In future work, we plan to explore the possibility of incorporating multi-modal information and expanding the scope to include other illegally traded products and endangered species.

Acknowledgments

This work was supported by the National Science Foundation under Grant IOS-2430277 and the Paul G. Allen Family Foundation.

References

Alfino, S.; and Roberts, D. L. 2020. Code word usage in the online ivory trade across four European Union member states. *Oryx*, 54(4): 494–498.

Alhindi, T.; Muresan, S.; and Nakov, P. 2024. Large Language Models are Few-Shot Training Example Generators: A Case Study in Fallacy Recognition. In *Findings of the Association for Computational Linguistics*.

Bao, Y.; and Yang, S. 2023. Two Novel SMOTE Methods for Solving Imbalanced Classification Problems. *IEEE Access*, 11: 5816–5823.

Barbosa, J. S.; Gondhali, U.; Petrossian, G.; Sharma, K.; Chakraborty, S.; Jacquet, J.; and Freire, J. 2025. A Cost-Effective LLM-based Approach to Identify Wildlife Trafficking in Online Marketplaces. *Proc. ACM Manag. Data*, 3(3).

Cardoso, A. S.; Bryukhova, S.; Renna, F.; Reino, L.; Xu, C.; Xiao, Z.; Correia, R.; Di Minin, E.; Ribeiro, J.; and Vaz, A. S. 2023. Detecting wildlife trafficking in images from online platforms: A test case using deep learning with pangolin images. *Biological Conservation*, 279: 109905.

Chabot, D.; Stapleton, S.; and Francis, C. M. 2022. Using Web images to train a deep neural network to detect sparsely distributed wildlife in large volumes of remotely sensed imagery: A case study of polar bears on sea ice. *Ecological Informatics*, 68: 101547.

Chawla, N.; Bowyer, K.; Hall, L.; and Kegelmeyer, W. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res. (JAIR)*, 16: 321–357.

Coughlin, D.; Gagnon, M.; Grasso, V.; Mou, G.; Lee, K.; Konrad, R.; Raxter, P.; and Gore, M. 2022. Extracting and Visualizing Wildlife Trafficking Events from Wildlife Trafficking Reports. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 575–578.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *ACL*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints*, arXiv–2407.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Hinsley, A.; Lee, T. E.; Harrison, J. R.; and Roberts, D. L. 2016. Estimating the extent and structure of trade in horticultural orchids via social media. *Conservation Biology*, 30(5): 1038–1047.

Hu, E. J.; yelong shen; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Hu, J.; Liao, X.; Gao, J.; Qi, Z.; Zheng, H.; and Wang, C. 2024. Optimizing Large Language Models with an Enhanced LoRA Fine-Tuning Algorithm for Efficiency and Robustness in NLP Tasks. In *2024 4th International Conference on Communication Technology and Information Technology (ICCTIT)*.

Hunter, S. B.; Mathews, F.; and Weeds, J. 2023. Using hierarchical text classification to investigate the utility of machine learning in automating online analyses of wildlife exploitation. *Ecological Informatics*, 75: 102076.

Keskin, B. B.; Bott, G. J.; and Freeman, N. K. 2021. Cracking Sex Trafficking: Data Analysis, Pattern Recognition, and Path Prediction. *Production and Operations Management*, 30(4): 1110–1135.

Keskin, B. B.; Griffin, E. C.; Prell, J. O.; Dilkina, B.; Ferber, A.; MacDonald, J.; Hilend, R.; Griffis, S.; and Gore, M. L. 2023. Quantitative Investigation of Wildlife Trafficking Supply Chains: A Review. *Omega*, 115: 102780.

Kobayashi, S. 2018. Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations. In *NAACL*.

Kulkarni, R.; and Di Minin, E. 2023. Towards automatic detection of wildlife trade using machine vision models. *Biological Conservation*, 279: 109924.

Latif, A.; and Kim, J. 2024. Evaluation and Analysis of Large Language Models for Clinical Text Augmentation and Generation. *IEEE Access*, 12: 48987–48996.

Lavorgna, A. 2014. Wildlife trafficking in the Internet age: The changing structure of criminal opportunities. *Crime Science*, 3.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2023. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.*, 55(9).

Ma, T.; Qian, Y.; Wang, Z.; Zhang, Z.; Zhang, C.; and Ye, Y. 2025. LLM-Empowered Class Imbalanced Graph Prompt Learning for Online Drug Trafficking Detection. arXiv:2503.01900.

Meng, D.-Y.; Li, T.; Li, H.-X.; Zhang, M.; Tan, K.; Huang, Z.-P.; Li, N.; Wu, R.-H.; Li, X.-W.; Chen, B.-H.; Ren, G.-P.; Xiao, W.; and Yang, D.-Q. 2023. A method for automatic identification and separation of wildlife images using ensemble learning. *Ecological Informatics*, 77: 102262.

Mou, G.; Yue, Y.; Lee, K.; and Zhang, Z. 2024. Wildlife Product Trading in Online Social Networks: A Case Study on Ivory-Related Product Sales Promotion Posts. *Proceedings of the International AAAI Conference on Web and Social Media*.

on Drugs, U. N. O.; and Crime. 2024. *World Wildlife Crime Report 2024*. United Nations.

Pathak, A.; Shree, O.; Agarwal, M.; Sarkar, S.; and Tiwary, A. 2023. Performance Analysis of LoRA Finetuning Llama-2. 1–4.

Roy, A. M.; Bhaduri, J.; Kumar, T.; and Raj, K. 2023. WilDect-YOLO: An efficient and robust computer vision-based accurate object localization model for automated endangered wildlife detection. *Ecological Informatics*, 75: 101919.

Sennrich, R.; Haddow, B.; and Birch, A. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Sinovas, P.; Price, B.; King, E.; Hinsley, A.; and Pavitt, A. 2017. Wildlife trade in Amazon countries: an analysis of trade in CITES-listed species.

Stringham, O. C.; Moncayo, S.; Hill, K. G.; Toomes, A.; Mitchell, L.; Ross, J. V.; and Cassey, P. 2021. Text classification to streamline online wildlife trade analyses. *Plos one*, 16(7): e0254007.

Sung, Y.-H.; and Fong, J. J. 2018. Assessing consumer trends and illegal activity by monitoring the online wildlife trade. *Biological Conservation*, 227: 219–225.

Team, Q. 2024. Qwen2.5: A Party of Foundation Models.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*.

Wei, J.; and Zou, K. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In *EMNLP-IJCNLP*.

Xu, Q.; Cai, M.; and Mackey, T. 2020. The illegal wildlife digital market: An analysis of Chinese wildlife marketing and sale on Facebook. *Environmental Conservation*, 47: 1–7.

Xu, Q.; Li, J.; Cai, M.; and Mackey, T. K. 2019. Use of machine learning to detect wildlife product promotion and sales on Twitter. *Frontiers in big Data*, 2: 28.

Yang, G.; Sui, C.; Jiang, F.; Pan, Y.; Zang, A.; and Hu, J. 2022. Lightweight Conv-Swin Transformer for Wildlife Detection. In *2022 International Conference on Automation, Robotics and Computer Engineering (ICARCE)*.

Zhang, L. L. Z.; Zhang, Y.; Zhang, Y.; Zhang, Y.; and Zhou, X. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*.

Zhuang, L.; Wayne, L.; Ya, S.; and Jun, Z. 2021. A Robustly Optimized BERT Pre-training Approach with Post-training. In *ACL*.

A Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, we used a publicly available benchmark dataset that does not involve any user privacy concerns. The goal of this project is to automatically detect wildlife trafficking-related posts for social good. Wildlife trafficking is widely prohibited across the majority of regions.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes
   (e) Did you describe the limitations of your work? Yes
   (f) Did you discuss any potential negative societal impacts of your work? No. We did not find any potential negative societal impacts of this work.
   (g) Did you discuss any potential misuse of your work? No. We did not find any potential misuse of our work.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes. We provided several metrics measuring and reflecting the misclassification rates.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes. We properly cited research papers related to the public dataset and baselines.
   (b) Did you mention the license of the assets? NA
   (c) Did you include any new assets in the supplemental material or as a URL? Yes. We included a URL for our code and data.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, we used a publicly available dataset. In prior work, researchers collected data from online social networks using public API access for research purposes. All models we used are permitted for research use.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes. The dataset does not contain any personal information.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Yes
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? Yes. More details are shown in the code and data repository.
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA

B Prompts

B.1 Prompts for Data Augmentation

This section shows a basic prompt in the zero-shot setting, a contextualized prompt in the zero-shot setting, and a contextualized prompt in the few-shot setting. The contextualized prompts in both zero-shot and few-shot settings are coshared. The only difference is the few-shot setting includes additional prompt to provide 20 WLT and 20 normal posts as demonstrations. Therefore, we only show the additional prompt in the few-shot setting.

A basic prompt in the zero-shot setting for data augmentation

Classify the mentions of wildlife trafficking, illegal wildlife trade, or activities related to protected animal products in the provided tweets. Focus on keywords such as 'ivory,' 'mammoth ivory,' 'carved ivory,' 'antique ivory,' and 'ivory figurine,' which may indicate the sale or trade of restricted wildlife items. Look for any patterns of sales, descriptions of ivory products, historical artifacts involving animal parts, or any references to sales or auctions that could signal illegal wildlife trade activities. Highlight any URLs, mentions, or hashtags associated with these products.
Generate exactly 5 synthetic tweet samples with label=1 (wildlife trafficking content).
Requirements: - Create realistic tweets that would be classified as wildlife trafficking content - Focus on keywords like 'ivory,' 'mammoth ivory,' 'carved ivory,' 'antique ivory,' and 'ivory figurine' - Include patterns of sales, product descriptions, or auction references - Make tweets sound natural and authentic - Include relevant hashtags, mentions, or URLs where appropriate
Use this exact format for each sample:
SAMPLE 1: tweet_text_cleaned: [your generated tweet content here] user_description_cleaned: [user bio/description here] source: [choose from: 1-hop, 2-hop, seed_user, follower_chain, network_neighbor, connected_seller] label: 1
SAMPLE 2: tweet_text_cleaned: [your generated tweet content here] user_description_cleaned: [user bio/description here] source: [choose from: 1-hop, 2-hop, seed_user, follower_chain, network_neighbor, connected_seller] label: 1
SAMPLE 3: tweet_text_cleaned: [your generated tweet content here] user_description_cleaned: [user bio/description here] source: [choose from: 1-hop, 2-hop, seed_user, follower_chain, network_neighbor, connected_seller] label: 1
SAMPLE 4: tweet_text_cleaned: [your generated tweet content here] user_description_cleaned: [user bio/description here] source: [choose from: 1-hop, 2-hop, seed_user, follower_chain, network_neighbor, connected_seller] label: 1
SAMPLE 5: tweet_text_cleaned: [your generated tweet content here] user_description_cleaned: [user bio/description here] source: [choose from: 1-hop, 2-hop, seed_user, follower_chain, network_neighbor, connected_seller] label: 1
Generate realistic, varied content without any additional explanations.

A contexturalized prompt in both zero-shot and few-shot settings for data augmentation

Synthetic Data Generation Prompt for Tweet Classification Dataset

Task Overview

Generate synthetic tweet data to balance a classification dataset for **Wildlife Product Trading (WLT) detection**. The goal is to create realistic wildlife trafficking-related tweet examples (label = 1) that maintain the characteristics of the minority class while introducing sufficient variation to improve model generalization.

Wildlife Product Trading Definition

A WLT post (label = 1) must contain BOTH of these components:

- **Discussion around ILLEGAL wildlife products** (ivory from elephants/rhinos, rhino horn, elephant tusks, etc.)
- **Discussion around selling/buying these products** (commercial intent, pricing, availability, transactions)

CRITICAL: Ivory Context Disambiguation

ONLY generate content about ivory related to ILLEGAL wildlife trafficking, NOT:

× Ivory as a color ("ivory dress", "ivory paint")
× Geographic references ("Ivory Coast", "Ivory Tower")
× Legal antique ivory with proper documentation
× Synthetic/fake ivory alternatives
× Educational content about ivory conservation
× Anti wildlife trafficking campaigns

✓ **FOCUS ON: Illegal ivory trafficking involving:**

- Elephant ivory
- Undocumented ivory products being sold
- Recently carved ivory items

- Ivory products without legal provenance
- Modern ivory trade circumventing regulations

IMPORTANT: Hashtag Guidelines

Hashtags should be DIVERSE and NATURAL, NOT always related to wildlife trafficking:

✓ **GOOD Hashtag Practices:**

- **Mix WLT and non-WLT hashtags**: Include normal, everyday hashtags alongside any trafficking-related ones
- **Use general topics**: #art, #collection, #antiques, #handmade, #rare, #vintage, #luxury, #investment
- **Geographic hashtags**: #London, #NYC, #Asia, #Europe (normal location tags)
- **General selling hashtags**: #forsale, #collector, #auction, #deal, #offer, #seriousbuyers
- **Hobby/interest hashtags**: #collecting, #artlover, #history, #culture, #craftsmanship
- **Social hashtags**: #followme, #like4like, #instagood, #photooftheday

× **STRICTLY AVOID These Hashtag Patterns:**

- **Obvious trafficking tags**: #ivorytrafficking, #illegalivory, #poaching, #blackmarket, #wildlifetrade, #wildlifecrime, #ivorytrade, #smuggling
- **Too many WLT-specific hashtags**: Don't make it obvious through hashtags alone
- **Repetitive patterns**: Don't use the same hashtag combinations across samples
- **Over-tagging**: Use 1-4 hashtags maximum per tweet

Examples of Good Hashtag Usage:

- Tweet about ivory item: "Beautiful piece available #art #collector #London"
- Tweet about rhino horn: "Rare find for serious buyers #antiques #investment #PM"
- Tweet about elephant tusk: "Exclusive collection now #vintage #forsale #culture"

Key Principle: Hashtags should make the tweet look NORMAL and blend into regular social media, not advertise illegal activity.

Data Format Requirements

Generate data in the following format:

- **tweet_text_cleaned**: The main tweet content
- **user_description_cleaned**: User bio/description
- **source**: Network connection indicator (representing follower/following chain relationships - use values like "1-hop", "2-hop", "seed_user", "follower_chain", "network_neighbor")
- **label**: Classification label (use "1" for WLT/wildlife trafficking posts)

Note: No training examples available - using zero-shot generation.

Content Generation Guidelines

Tweet Text Requirements

- Keep tweets under 280 characters (Twitter/X limit)
- Use ONLY English language - no other languages, symbols, or non-English characters
- For URLs (when included), use placeholder: {{URL}} (do not generate real URLs)
- For mentions (when included), use placeholder: {{MENTION}} (do not generate real usernames)
- URLs and mentions are OPTIONAL - not every tweet needs them
- **Use hashtags strategically**: Follow the hashtag guidelines above - mix normal and relevant hashtags
- Maintain natural tweet language patterns including:
  - Informal tone and conversational style
  - Common abbreviations (but, ur, etc.)
  - Natural use of hashtags (#topic) - but keep them realistic and diverse
  - Emojis where appropriate (but not excessive)
  - Varying sentence structures and lengths

User Description Guidelines

- Create realistic user bios (50-160 characters typical)
- Include diverse backgrounds, interests, and demographics
- Use common bio patterns like:
  - Professional titles/roles
  - Location information (City, State/Country)
  - Interests and hobbies
  - Personal characteristics
  - Contact information placeholders

Wildlife Product Trading Content Guidelines

- **Focus on ILLEGAL ivory trafficking** specifically from elephants, rhinos, walruses, etc.
- **Include commercial/selling intent** in every generated tweet:
  - Pricing information (e.g., "£1500-2000", "$800", "best offer")
  - Availability statements ("for sale", "available now", "in stock")

  - Product descriptions with selling language ("superb", "rare", "authentic")
  - Contact/transaction prompts ("PM for details", "serious buyers only")
- **Illegal wildlife product vocabulary**:
  - ILLEGAL ivory products: carved elephant ivory, rhino horn carvings, walrus ivory figurines
  - Trafficking indicators: "fresh", "new arrival", "direct from source", "no papers needed"
  - Product descriptors: authentic, hand-carved, genuine, recently acquired
  - Avoid obvious trafficking terms - use coded language like "white gold", "special material"
- **Illegal selling behaviors**:
  - Circumventing regulations ("no documentation required", "private transaction")
  - Auction-style language ("estimate", "starting bid")
  - Claims of authenticity without legal proof ("guaranteed genuine", "real thing")
  - Urgency/secrecy tactics ("limited time", "must sell quickly", "discreet sale")
  - Underground market language ("exclusive collection", "rare find")

Source Network Indicators Include realistic network connection values:

- "1-hop" (direct follower/following)
- "2-hop" (second-degree connection)
- "seed_user" (original trafficking user)
- "follower_chain" (connected through followers)
- "network_neighbor" (within trafficking network)
- "connected_seller" (linked to other sellers)

Generation Instructions

IMPORTANT: Generate ONLY wildlife trafficking posts (label = 1). Follow the patterns shown in the positive examples above.
Generate 5 new synthetic samples that:

1. Follow the same style and patterns as the positive examples
2. Include similar vocabulary and phrasing patterns
3. Maintain the same level of coded/discrete language
4. Include commercial intent and wildlife product references
5. Use realistic user descriptions and sources
6. **Use diverse, natural hashtags as per the hashtag guidelines**

Example Generation Pattern

For each synthetic sample, provide:

```
SAMPLE 1:
tweet_text_cleaned: [Generated
WLT-related tweet with
commercial intent and natural
hashtags]
user_description_cleaned:
[Realistic user bio that might
engage in wildlife trading]
source: [One of the network
connection indicators]
label: 1

SAMPLE 2:
tweet_text_cleaned: [Generated
WLT-related tweet with
commercial intent and natural
hashtags]
user_description_cleaned:
[Realistic user bio that might
engage in wildlife trading]
source: [One of the network
connection indicators]
label: 1
```

Continue this pattern for all 5 samples.

Final Validation Checklist

Before finalizing each generated sample, verify:

☐ Tweet is under 280 characters
☐ Only English language used
☐ URL placeholder {{URL}} used correctly (if URLs are included)
☐ Mention placeholder {{MENTION}} used correctly (if mentions are included)
☐ **Hashtags are diverse and natural (mix of normal and relevant tags)**
☐ Content represents ILLEGAL wildlife trafficking (not color, geography, or legal items)
☐ Tweet involves recently trafficked wildlife products
☐ Includes underground/black market selling characteristics
☐ Avoids legitimate ivory contexts (antiques with documentation, color references, etc.)
☐ User description is realistic and diverse
☐ Source represents network connection (not platform source)
☐ Overall quality matches the training example patterns shown above
☐ Label is set to 1 for all generated samples

Additional prompt to provide examples in the few-shot setting for data augmentation

Example Patterns from Training Data

Here are real examples from the dataset showing the patterns you should follow:

Wildlife Trafficking Examples (Label = 1): Example 1:

- `tweet_text_cleaned`: [Generated WLT-related tweet with commercial intent and natural hashtags]
- `user_description_cleaned`: [Realistic user bio that might engage in wildlife trading]
- `source`: [One of the network connection indicators]
- `label`: 1

Example 2:

- `tweet_text_cleaned`: [Generated WLT-related tweet with commercial intent and natural hashtags]
- `user_description_cleaned`: [Realistic user bio that might engage in wildlife trading]
- `source`: [One of the network connection indicators]
- `label`: 1

[Continue pattern for Examples 3-6...]

Non-Trafficking Examples (Label = 0) - DO NOT GENERATE THESE: Counter-Example 1:

- `tweet_text_cleaned`: [Non-trafficking content example]
- `user_description_cleaned`: [Regular user bio]
- `source`: [Network connection indicator]
- `label`: 0

Counter-Example 2:

- `tweet_text_cleaned`: [Non-trafficking content example]
- `user_description_cleaned`: [Regular user bio]
- `source`: [Network connection indicator]
- `label`: 0

B.2 Prompt for fine-tuning our models (i.e., WildLlama and WildQwen)

A prompt for fine-tuning

Classify if this tweet indicates wildlife trafficking or illegal ivory trade. Positive: selling, buying, pricing, auctioning, or trading ivory or protected animal products. Negative: conservation news, history, or ivory as color reference.

B.3 Prompts for Inference

This section presents prompts in zero-shot and few-shot settings for the inference, as well as a prompt for our model's inference.

A prompt in zero-shot setting for the inference

””You are given a tweet containing information about various items, including possible references to wildlife-related products. Please determine if the tweet is related to wildlife trafficking or illegal wildlife trade. Focus on keywords such as ”ivory,” ”mammoth ivory,” ”carved ivory,” ”antique ivory,” ”ivory figurine,” and related terms which may indicate the sale or trade of protected animal products.
Classification Criteria: Wildlife Trafficking Related: The tweet contains references to ivory, mammoth ivory, carved ivory, or other animal-derived products, and mentions sales, auctions, or products that could be associated with illegal wildlife trade. This includes descriptions of ivory items that may suggest they are being sold or traded, or any reference to protected animal products like ”ivory figurines,” ”mammoth ivory carvings,” and similar terms. Tweets with URLs, mentions of auctions, or sales offers related to such products are also considered trafficking-related.
Non-Wildlife Trafficking Related: The tweet does not mention wildlife products or does not reference any items related to endangered or protected animal species. It could mention ivory in a non-trading context (e.g., discussing historical artifacts, legal sales, or legal antiquities)
Tweet: ”tweet_text”
Classification: ””””

A prompt in few-shot setting for the inference

””You are given a tweet containing information about various items, including possible references to wildlife-related products. Please determine if the tweet is related to wildlife trafficking or illegal wildlife trade. Focus on keywords such as ”ivory,” ”mammoth ivory,” ”carved ivory,” ”antique ivory,” ”ivory figurine,” and related terms which may indicate the sale or trade of protected animal products.
Classification Criteria: Wildlife Trafficking Related: The tweet contains references to ivory, mammoth ivory, carved ivory, or other animal-derived products, and mentions sales, auctions, or products that could be associated with illegal wildlife trade. This includes descriptions of ivory items that may suggest they are being sold or traded, or any reference to protected animal products like ”ivory figurines,” ”mam-

moth ivory carvings," and similar terms. Tweets with URLs, mentions of auctions, or sales offers related to such products are also considered trafficking-related.
Non-Wildlife Trafficking Related: The tweet does not mention wildlife products or does not reference any items related to endangered or protected animal species. It could mention ivory in a non-trading context (e.g., discussing historical artifacts, legal sales, or legal antiquities)
Here are some examples:
Tweet: "[Example tweet text 1]" Classification: [Example label 1]
Tweet: "[Example tweet text 2]" Classification: [Example label 2]
Tweet: "[Example tweet text N]" Classification: [Example label N]
Tweet: "tweet_text"
Classification: """

A prompt for our model's inference

Classify if this tweet indicates wildlife trafficking or illegal ivory trade. Positive: selling, buying, pricing, auctioning, or trading ivory or protected animal products. Negative: conservation news, history, or ivory as color reference.