# Optimizing LLMs for Cybersecurity: Model Distillation and Quantization

Ankit Pratap Singh

## 1 Introduction

Large Language Models (LLMs) are powerful tools for cybersecurity applications, including phishing URL detection. However, deploying these models can be computationally expensive. This document explains how model **distillation** and **quantization** enable efficient LLM deployment while maintaining high accuracy.

## 2 Model Distillation

Model distillation is a technique where a large, high-performing model (**teacher model**) trains a smaller model (**student model**). Instead of directly learning from labeled data, the student model learns from the soft labels provided by the teacher also called White-Box Knowledge Distillation.

*White-box KD enables the student LLM to gain a deeper understanding of the teacher LLM's internal structure and knowledge representations, often resulting in higher-level performance improvements. An representative example is MINILLM ([2]), which the first work to study distillation from the Open-source generative LLMs. MINILLM use a reverse KullbackLeibler divergence objective (see Figure 1), which is more suitable for KD on generative language models, to prevent the student model from overestimating the low-probability regions of the teacher distribution, and derives an effective optimization approach to learn the objective [3].*

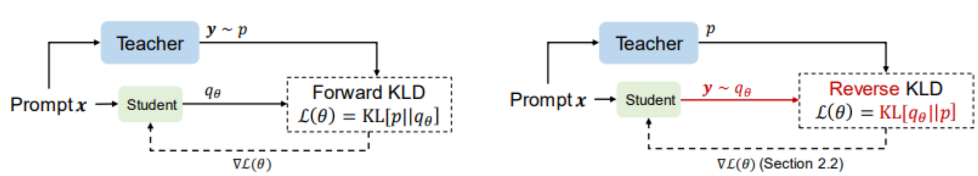


Figure 1: Reverse KLD between the student and teacher model distributions [2].

### 2.1 Implementation

In this project, we use **Hugging Face Transformers** and **PyTorch** for knowledge distillation. The teacher model (BERT) trains a smaller student model (DistilBERT) for phishing URL classification.

## 3 Model Quantization

Quantization refers to the process of reducing the number of bits (i.e., precision) in the parameters of the model with minimal loss in inference performance. For LLMs QLORA [1] backpropagates gradients through a frozen, 4-bit quantized pretrained language model into Low Rank Adapters (LoRA).

### 3.1 Benefits of 4-bit Quantization

- **Reduces memory footprint:** Makes the model more lightweight.
- **Speeds up inference:** Faster predictions with minimal accuracy loss.
- **Efficient deployment:** Suitable for edge and real-time cybersecurity applications.

## 4 Workflow Overview

1. **Dataset Preparation:** Process phishing URL dataset (`data.py`).
2. **Train Teacher Model:** Fine-tune BERT (`teacher_training.py`).
3. **Distill Student Model:** Train DistilBERT (`distillation.py`).
4. **Quantize the Model:** Apply 4-bit quantization (`quantization.py`).

## 5 Libraries Used

- **Transformers:** Model training and distillation.
- **Datasets:** Efficient dataset handling.
- **Torch:** Deep learning framework.
- **BitsAndBytes:** 4-bit quantization.
- **scikit-learn:** Model evaluation.

## 6 Conclusion

By using model distillation and quantization, we successfully optimize an LLM **51**% smaller than teacher LLM for phishing detection, making it faster, and more efficient for real-world cybersecurity applications.

## References


[1] Tim Dettmers et al. "Qlora: Efficient finetuning of quantized llms". In: *Advances in neural information processing systems* 36 (2023), pp. 10088–10115.

[2] Yuxian Gu et al. "MiniLLM: Knowledge distillation of large language models". In: *arXiv preprint arXiv:2306.08543* (2023).

[3] Xunyu Zhu et al. "A survey on model compression for large language models". In: *Transactions of the Association for Computational Linguistics* 12 (2024), pp. 1556–1577.