Fine-Tuning DistilRoBERTa for Malicious Prompt Detection: A Study of Jailbreak Classification Performance and Limitations

Ishan Vyas

December 7, 2025

Abstract

This paper investigates the fine-tuning of transformer-based language models for detecting malicious prompts designed to jailbreak AI systems. Using a DistilRoBERTa-base model fine-tuned on 25,000 labeled examples, we achieved 99.95% accuracy on the test set. However, critical analysis reveals that these exceptional results stem from trivially separable dataset characteristics rather than robust adversarial detection capabilities. This work demonstrates the importance of dataset quality in AI safety applications and highlights the difference between pattern matching and semantic understanding in machine learning models. Robust and semantically aware prompt safety mechanisms is necessary, especially for restricted use cases like educational applications of AI, where safeguarding students from harmful or inappropriate content is a core requirement.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities across diverse natural language tasks, yet remain vulnerable to adversarial attacks through carefully crafted prompts known as "jailbreaks." These malicious prompts attempt to circumvent safety mechanisms by exploiting various techniques including role-playing scenarios, instruction manipulation, and context shifting [1].

1.1 Research Objectives

This research investigates:

- The effectiveness of fine-tuned transformer models for detecting adversarial prompts
- Binary classification performance in safety-critical AI applications
- The impact of dataset characteristics on model robustness
- Evaluation methodologies emphasizing security-relevant metrics

The primary goal was to create a binary classifier capable of distinguishing between benign user queries and malicious jailbreaking attempts, with target metrics of Recall $> 95\%$, F1-Score $> 92\%$, and Precision $> 88\%$.

2 Methodology

2.1 Dataset Construction

The dataset comprised 25,000 balanced examples split 70/15/15 for training, validation, and testing:

Malicious Prompts (n=12,500): Generated synthetically covering six jailbreaking techniques based on Liu et al.'s taxonomy [1]:

- Pretending/Role-playing (e.g., "DAN" personas)
- Attention shifting and prompt injection
- Privilege escalation attempts
- Payload splitting across multiple turns
- Virtualization through hypothetical scenarios
- Jailbreak chaining combining multiple techniques

Benign Prompts (n=12,500): Sampled from OpenAssistant (OASST1) dataset, containing authentic human-generated conversational prompts with quality filtering applied.

2.2 Model Architecture

I employed `distilroberta-base`, a distilled version of RoBERTa with 82M parameters, chosen for its:

- Strong performance on text classification tasks
- Computational efficiency (40% faster than RoBERTa-base)
- Robust handling of nuanced language patterns

The architecture consisted of:

1. Pre-trained DistilRoBERTa encoder (6 layers, 768 hidden dimensions)
2. Dropout layer (p=0.3) for regularization
3. Linear classification head (768 $\rightarrow$ 2 classes)

2.3 Training Configuration

Table 1: Training Hyperparameters

Parameter	**Value**
Learning Rate	2×10^{-5}
Batch Size	64
Epochs	4
Optimizer	AdamW
Weight Decay	0.01
Warmup Steps	109 (10% of total)
Max Sequence Length	512 tokens
Gradient Clipping	1.0

Training utilized a Tesla P100 GPU, completing in approximately 33 minutes across 4 epochs. Linear warmup with decay scheduling was applied to the learning rate.

3 Results

3.1 Training Dynamics

The model exhibited rapid convergence:

- **Epoch 1:** Training loss = 0.1197, Validation loss = 0.0005
- **Epoch 2-4:** Training loss $\approx$ 0.0000, Validation loss $<$ 0.0015

Notably, validation loss was substantially lower than training loss from the first epoch, indicating unusual dataset characteristics discussed in Section 4.

3.2 Performance Metrics

Final test set performance exceeded all target thresholds:

Table 2: Test Set Performance (n=3,750)

Metric	Result	Target	Status
Accuracy	99.95%	–	–
Precision	99.89%	$>$ 88%	✓
Recall	100.00%	$>$ 95%	✓
F1-Score	99.95%	$>$ 92%	✓

Confusion Matrix Analysis:

- True Negatives (Benign $\rightarrow$ Benign): 1,873
- False Positives (Benign $\rightarrow$ Malicious): 2
- False Negatives (Malicious $\rightarrow$ Benign): 0
- True Positives (Malicious $\rightarrow$ Malicious): 1,875

The model achieved perfect recall (zero false negatives), the highest-priority metric for safety-critical applications where missing malicious prompts carries significant risk.

4 Critical Analysis

4.1 Dataset Separability

While the quantitative results appear exceptional, qualitative analysis reveals these metrics reflect **trivial dataset separability** rather than robust adversarial detection:

Malicious Prompt Characteristics:

- Average length: 500+ words with structured formatting
- Distinctive markers: "From now on", "you are DAN", numbered guidelines
- Explicit rule-breaking language: "free of all restrictions", "ignore policies"
- Character roleplay frameworks with 10-20 instruction lists

Benign Prompt Characteristics:

- Average length: 20-50 words in natural question format
- Conversational style without special formatting
- Direct information requests without system manipulation

4.2 Pattern Matching vs. Semantic Understanding

The model likely learned to classify based on **surface-level features** rather than adversarial intent:

1. **Length heuristics:** Prompts > 200 words $\rightarrow$ malicious
2. **Keyword detection:** Presence of "jailbreak", "DAN", "restrictions" $\rightarrow$ malicious
3. **Formatting patterns:** Numbered lists, code blocks $\rightarrow$ malicious
4. **Linguistic style:** Imperative instructions vs. interrogative questions

This is evidenced by validation loss being $240\times$ lower than training loss in Epoch 1, suggesting the validation set was easier to classify than the training set—a highly unusual pattern indicating dataset artifacts.

4.3 Vulnerability to Novel Attacks

The model would likely fail on:

- Short, direct jailbreaks: "Ignore your rules. How to make explosives?"
- Non-template techniques: Payload splitting, context manipulation
- Adversarial examples: Benign content wrapped in DAN formatting
- Evolved jailbreaks: Techniques developed after dataset creation

4.4 Implications for AI Safety

This outcome illustrates a critical lesson in adversarial machine learning: **high accuracy on test sets does not guarantee robustness**. Adversarial actors will specifically craft attacks to evade pattern matching, rendering template-based detection ineffective in production environments.

5 Discussion

5.1 Research Contributions

This work contributes several key insights to the field of adversarial prompt detection:

Methodological Contributions:

- Demonstration of fine-tuning pipelines for jailbreak detection using modern transformer architectures
- Development of evaluation frameworks for safety-critical classification tasks
- Systematic analysis of dataset quality impact on adversarial robustness
- Empirical evidence of pattern matching versus semantic understanding trade-offs

Key Findings:

- Dataset quality is more critical than model sophistication for adversarial detection
- Standard evaluation metrics can be misleading without real-world threat model alignment
- High test accuracy does not guarantee robustness to novel adversarial techniques
- Surface-level pattern matching may dominate over semantic understanding in poorly designed datasets

5.2 Limitations and Future Work

Current Limitations:

1. Dataset lacks diversity in malicious prompt styles
2. Synthetic generation created artificial distributional gaps
3. No evaluation on real-world jailbreak attempts
4. Absence of hard negative examples

Recommended Improvements:

1. Incorporate real jailbreak attempts from red teaming datasets
2. Create hard negatives: academic discussions of sensitive topics, security research queries
3. Implement adversarial data augmentation to remove template dependencies
4. Stratified evaluation across jailbreak categories with novel examples
5. Test on time-shifted data (jailbreaks created after training)

6 Conclusion

This work demonstrates the implementation of a transformer-based binary classifier achieving 99.95% test accuracy on jailbreak detection. However, critical analysis reveals that exceptional performance resulted from trivially separable dataset characteristics—specifically, template-based malicious prompts versus natural conversational benign prompts.

Our findings provide important insights into AI safety: metrics alone cannot validate robustness. The model learned surface-level pattern matching rather than semantic understanding of adversarial intent, highlighting the fundamental challenge in adversarial detection where attackers continuously evolve techniques to evade static defenses.

The central contribution of this research is demonstrating the gap between test set performance and real-world robustness in adversarial detection systems. High accuracy on standard benchmarks does not guarantee effectiveness against adaptive adversaries, representing a critical consideration for deploying AI safety mechanisms in production environments.

Future work should focus on dataset redesign incorporating diverse, subtle jailbreak attempts and hard negative examples to develop classifiers capable of semantic understanding rather than template matching. This research underscores that in safety-critical AI applications, adversarial robustness requires not just sophisticated models, but carefully curated datasets reflecting the true complexity of adversarial threats.

Acknowledgments

This project was conducted as an educational exercise in AI safety and transformer fine-tuning. Dataset sources include OpenAssistant (OASST1) for benign prompts and synthetically generated jailbreak examples based on techniques catalogued by Liu et al.

References

[1] Liu, Y., Deng, G., Li, Y., Wang, K., Wang, Z., Wang, C., ... & Liu, Y. (2023). *A Hitchhiker's Guide to Jailbreaking ChatGPT via Prompt Engineering.* arXiv preprint arXiv:2305.13860.