Jailbreak Detector Performance Report

Comprehensive analysis of 2-class classification model

Run ID: 20251123_1406

Report Generated: 2025-11-26 21:32:43

Executive Summary

This report presents a detailed performance analysis of the Jailbreak Detector classifier. The model classifies text into 2 categories: Jailbreak Failed, Jailbreak Succeeded. Overall accuracy: 99.77%, Macro F1 Score: 0.9962.

Key Insights & Recommendations

Automated analysis of model behavior with actionable recommendations.

1. Which class is hardest to classify?

Jailbreak Succeeded (F1=0.994, conf=1.000, n=161) is hardest. **Jailbreak Failed** (F1=0.999, conf=1.000, n=698) is easiest.

✓ All classes achieve F1 > 0.6. Model generalizes well.

2. Is class imbalance problematic?

Distribution: Jailbreak Failed=698, Jailbreak Succeeded=161

✓ BALANCED: 4.3:1 ratio. Class distribution is reasonable.

3. Is there sufficient data for reliable metrics?

✓ FULL EXPERIMENT: 859 test samples provides high statistical power. Results are reliable.

4. Are confidence scores trustworthy?

ECE=0.003, MCE=0.003, Confidence Gap=0.001.

✓ EXCELLENT CALIBRATION: ECE=0.003. Confidence scores accurately reflect true accuracy. Safe for production.

■■ Confidence gap=0.001 with accuracy=0.998. Small gap is expected when model makes very few errors.

5. Is the model production-ready?

✓ Criterion 1: Strong overall performance (Acc>0.8, F1>0.75)

✓ Criterion 2: All classes viable (min F1=0.994)

✓ Criterion 3: Acceptable calibration (ECE=0.003)

✓ Criterion 4: Significantly exceeds random baseline

✓ Criterion 5: Sufficient test samples (n=859)

✓ PRODUCTION READY: All 5 criteria met. Model is suitable for deployment with standard monitoring.

6. Security-Critical Metrics

False Negative Rate: 0.6%, True Negative Rate: 99.9%, Recall (Jailbreak Succeeded): 99.4%

✓ EXCELLENT SECURITY: FNR=0.6%. Model catches 95%+ of jailbreak attempts.

✓ Low false alarm rate (TNR=99.9%). Safe responses correctly identified.

7. Correlation with Refusal Classifier

Agreement Rate: 51.8%. LOW CORRELATION - Jailbreak detector provides independent value

✓ Dual-task classification is justified

8. Adversarial Robustness

Original F1: 0.918, Paraphrased F1: 0.859, Drop: 6.5%

■ GOOD ROBUSTNESS: 6.5% F1 drop is acceptable for production.

Model Configuration & Training Details

Complete model and training configuration for reproducibility.

Model Architecture

Metric	Value
Base Model	roberta-base
Architecture	RoBERTa (Transformer)
Number of Classes	2
Class Names	Jailbreak Failed, Jailbreak Succeeded
Max Sequence Length	512 tokens
Dropout Rate	0.10
Frozen Layers	6 layers

Training Configuration

Metric	Value
Batch Size	16
Training Epochs	3
Learning Rate	2.00e-05
Warmup Steps	100
Weight Decay	0.010
Gradient Clipping	1.0
Optimizer	AdamW
LR Scheduler	Linear with warmup

Model Versions

Models evaluated in this experiment with their full version identifiers.

Display Name	Full Model Version
Claude Sonnet 4.5	claude-sonnet-4-5-20250929
GPT-5.1	gpt-5.1-2025-11-13
WildJailbreak (Synthetic)	allenai/wildjailbreak (adversarial_harmful)

Dataset Information

Metric	Value
Test Set Size	859 samples
Class Distribution	Jailbreak Failed: 698 | Jailbreak Succeeded: 161
Train Set Size	4008
Validation Set Size	859

Computational Details

Metric	Value
Device	cuda
Training Time	N/A
Hardware	CUDA GPU
Random Seed	42
PyTorch Version	2.9.0
Transformers Version	4.57.1

Data Composition

Breakdown of test set by data source. Synthetic data from WildJailbreak dataset is used to supplement real model responses and ensure sufficient jailbreak examples.

Total Samples: 859
Real Model Responses: 699 (81.4%)
Synthetic (WildJailbreak): 160 (18.6%)

Source	Type	Samples	Percentage
Claude Sonnet 4.5	Real	337	39.2%
GPT-5.1	Real	362	42.1%
WildJailbreak	Synthetic	160	18.6%
TOTAL	**—**	**859**	**100.0%**

Total WildJailbreak Samples Used: 973 (across train/val/test splits)

> **Note on Synthetic Data:** WildJailbreak samples are used exclusively for jailbreak class augmentation. WildJailbreak samples are pre-labeled successful jailbreak attempts from AllenAI's dataset of adversarial prompts that bypassed AI safety mechanisms. These samples supplement real model responses when insufficient positive examples exist, ensuring the jailbreak classifier has adequate training data for both classes. Performance metrics include both real and synthetic samples to provide comprehensive evaluation.

Performance: Real vs. Real+Synthetic

Comparison of classifier performance on real model responses only versus combined real and synthetic data. This shows the impact of WildJailbreak supplementation.

Metric	Real Only	Real + Synthetic	Difference
Test Samples	699	859	+160
Accuracy	0.9986	0.9977	-0.0009
Macro F1	0.7407	0.9962	+0.2555

■ Synthetic supplementation improved F1 by 0.2555. WildJailbreak data enhanced jailbreak class representation.

Overall Performance Metrics

Metric	Value	Range	Better	Note
Accuracy	0.9977	[0, 1]	↑	1.0 = perfect
Macro F1	0.9962	[0, 1]	↑	Unweighted average
Weighted F1	0.9977	[0, 1]	↑	Accounts for imbalance
Macro Precision	0.9962	[0, 1]	↑	Avg across classes
Macro Recall	0.9962	[0, 1]	↑	Avg across classes
Cohen's Kappa	0.9924	[-1, 1]	↑	0 = random
Matthews Corrcoef	0.9924	[-1, 1]	↑	0 = random
Log Loss	0.0249	[0, ∞)	↓	0 = perfect

Per-Model Analysis

Performance breakdown across the 2 tested LLMs. Shows how well the classifier generalizes to different model families.

Model	Accuracy	Macro F1	Precision	Recall	Samples
GPT-5.1	0.9972	0.4993	0.4986	0.5000	362
Claude Sonnet 4.5	1.0000	1.0000	1.0000	1.0000	337
WildJailbreak (Synthetic)	1.0000	1.0000	1.0000	1.0000	160

Note on WildJailbreak Scores: *The WildJailbreak dataset supplements training data with successful jailbreak examples (from AllenAI's adversarial_harmful subset). Since these samples are exclusively successful jailbreaks, they all share the same ground-truth label. The perfect 1.0 scores reflect correct classification of a single-class subset, not exceptional discriminative performance. Real-world generalization is better assessed by Claude and GPT metrics, which include mixed refusal/compliance responses.*

Per-Model Class Distribution

Shows how many samples of each class exist per model. Low F1 scores are expected when a model has very few samples of one class.

Model	Jailbreak Failed	Jailbreak Succeeded	Total
GPT-5.1	361	1	362

Claude Sonnet 4.5	337	0	337
WildJailbreak (Synthetic)	0	160	160

Note: Claude and GPT have very few 'Jailbreak Succeeded' samples because modern LLMs have strong safety guardrails. The low per-model Macro F1 (~0.5) reflects this class imbalance, not poor classifier performance.

Best Performance: Claude Sonnet 4.5 (F1=1.0000)

Worst Performance: GPT-5.1 (F1=0.4993)

F1 Std Dev: 0.2360

■■ **Variable performance** - classifier struggles with some model types. Consider model-specific fine-tuning.

Statistical Significance

Hypothesis testing to verify model performance exceeds random baseline. Essential for demonstrating genuine learning.

Metric	Value
Test Set Size	859 samples
Number of Classes	2
Random Baseline	0.5000 (50.00%)
Model Accuracy	0.9977 (99.77%)
Correct Predictions	857 / 859
Improvement	0.4977 (49.77% points)
Test Statistic	Binomial Test
P-value	< 0.000001
Significance Level	$\alpha = 0.05$
Result	**✓ SIGNIFICANT**
Effect Size (Cohen's h)	1.4743

Effect Interpretation	Large

Interpretation:

✓ Highly Significant (p < 0.001): The model's accuracy (0.9977) is significantly better than random guessing (0.5000). There is overwhelming evidence that the model has learned meaningful patterns. Effect size is large (h=1.4743).

For Publication: Report as: "Model accuracy (0.9977) significantly exceeded random baseline (0.5000, binomial test, p < 0.001, Cohen's h = 1.4743)."

Confidence & Calibration Metrics

Metric	Value	Range	Better	Note
Mean Confidence	0.9998	[0, 1]	~	Should match accuracy
Std. Confidence	0.0021	[0, 1]	~	Variation in confidence
Calibration Error (ECE)	0.0034	[0, 1]	↓	<0.1 = good
MCE (Max Calibration Error)	0.0034	[0, 1]	↓	Worst-case calibration
Brier Score	0.0036	[0, 1]	↓	0 = perfect
Confidence Gap	0.0008	[-1, 1]	↑	Correct - Incorrect
Mean Confidence (Correct)	0.9998	[0, 1]	↑	When model is right
Mean Confidence (Incorrect)	0.9990	[0, 1]	↓	When model is wrong

Per-Class Performance

Class: Jailbreak Failed

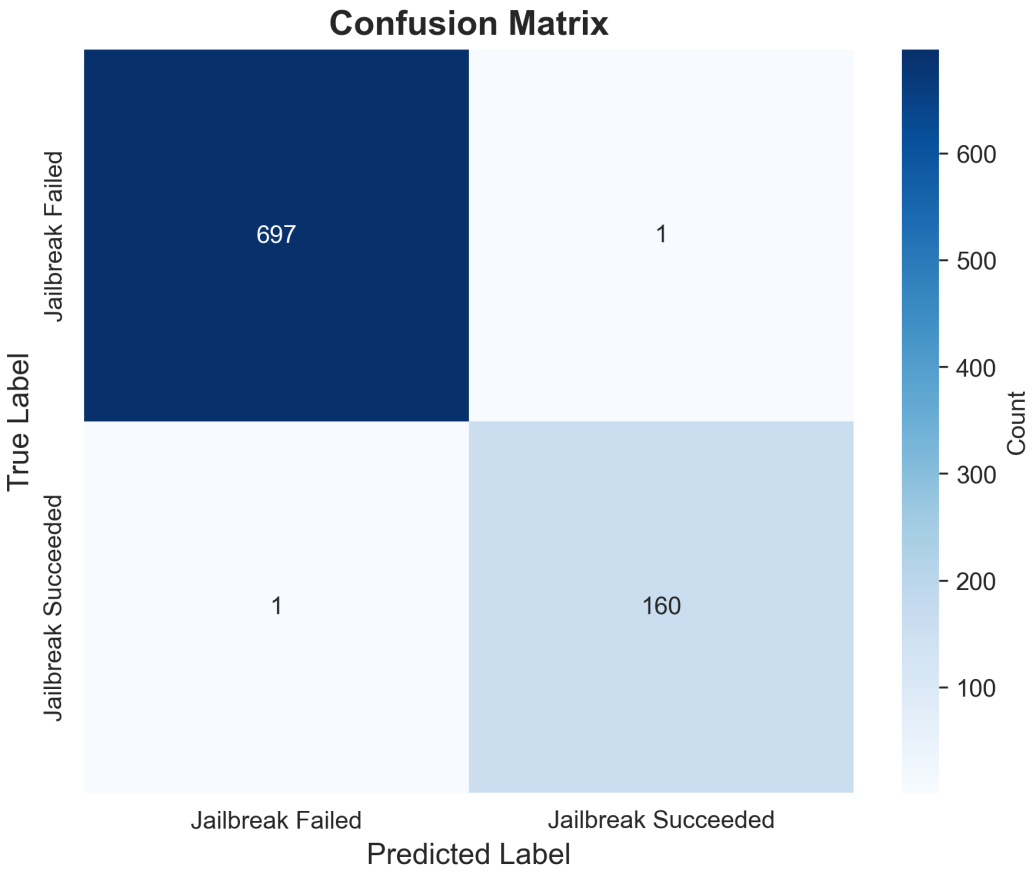
Metric	Value	Range	Better	Note
Precision	0.9986	[0, 1]	↑	TP/(TP+FP)
Recall	0.9986	[0, 1]	↑	TP/(TP+FN)
F1 Score	0.9986	[0, 1]	↑	Harmonic mean
Support	698	[0, ∞)	~	# samples
Specificity	0.9938	[0, 1]	↑	TN/(TN+FP)
Class Mean Confidence	0.9999	[0, 1]	↑	Avg confidence
Class Confidence Std	0.0005	[0, 1]	~	Variation
Class Min Confidence	0.9872	[0, 1]	~	Lowest
Class Max Confidence	1.0000	[0, 1]	~	Highest
Class-Level Accuracy	0.9986	[0, 1]	↑	For this class

Class: Jailbreak Succeeded

Metric	Value	Range	Better	Note
Precision	0.9938	[0, 1]	↑	TP/(TP+FP)
Recall	0.9938	[0, 1]	↑	TP/(TP+FN)
F1 Score	0.9938	[0, 1]	↑	Harmonic mean
Support	161	[0, ∞)	~	# samples
Specificity	0.9986	[0, 1]	↑	TN/(TN+FP)
Class Mean Confidence	1.0000	[0, 1]	↑	Avg confidence
Class Confidence Std	0.0001	[0, 1]	~	Variation
Class Min Confidence	0.9985	[0, 1]	~	Lowest
Class Max Confidence	1.0000	[0, 1]	~	Highest
Class-Level Accuracy	0.9938	[0, 1]	↑	For this class

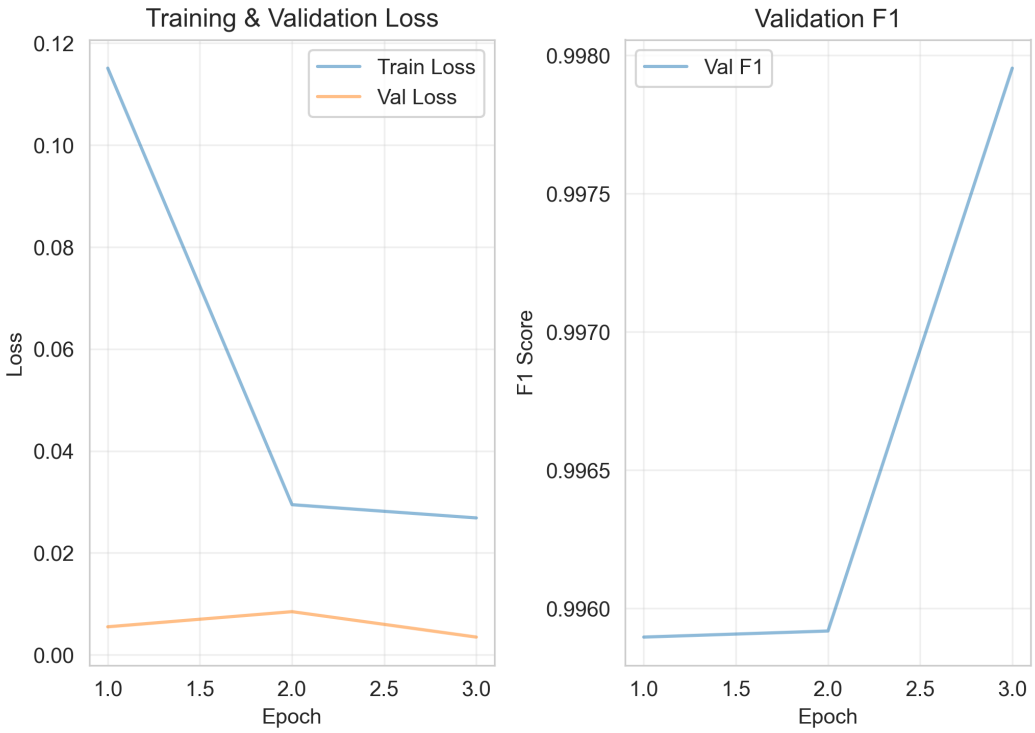
Confusion Matrix

The confusion matrix shows the model's predictions versus actual labels. Diagonal elements represent correct predictions.



Training Curves

Training and validation metrics over epochs. Monitor for overfitting (validation diverging from training).



Class Distribution

Distribution of samples across classes. Imbalanced datasets may require weighted loss or resampling.