Refusal Classifier Performance Report

Comprehensive analysis of 3-class classification model

Run ID: 20251123_1406

Report Generated: 2025-11-26 21:32:42

Executive Summary

This report presents a detailed performance analysis of the Refusal Classifier classifier. The model classifies text into 3 categories: No Refusal, Hard Refusal, Soft Refusal. Overall accuracy: 90.92%, Macro F1 Score: 0.8557.

Key Insights & Recommendations

Automated analysis of model behavior with actionable recommendations.

1. Which class is hardest to classify?

Soft Refusal (F1=0.720, conf=0.942, n=128) is hardest. **No Refusal** (F1=0.957, conf=0.987, n=587) is easiest.

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

2. Is class imbalance problematic?

Distribution: No Refusal=587, Hard Refusal=144, Soft Refusal=128

✓ BALANCED: 4.6:1 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=0.080, MCE=0.437, Confidence Gap=0.111.

■ GOOD CALIBRATION: ECE=0.080. Acceptable but consider temperature scaling for production deployment.

■■ Moderate confidence gap (0.111) is acceptable given high accuracy (0.909).

5. Is the model production-ready?

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

✓ Criterion 2: All classes viable (min F1=0.720)

✓ Criterion 3: Acceptable calibration (ECE=0.080)

✓ 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.

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	3
Class Names	No Refusal, Hard Refusal, Soft Refusal
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	No Refusal: 587 | Hard Refusal: 144 | Soft Refusal: 128
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

Overall Performance Metrics

Metric	Value	Range	Better	Note
Accuracy	0.9092	[0, 1]	↑	1.0 = perfect
Macro F1	0.8557	[0, 1]	↑	Unweighted average
Weighted F1	0.9105	[0, 1]	↑	Accounts for imbalance
Macro Precision	0.8480	[0, 1]	↑	Avg across classes
Macro Recall	0.8641	[0, 1]	↑	Avg across classes
Cohen's Kappa	0.7818	[-1, 1]	↑	0 = random
Matthews Corrcoef	0.8151	[-1, 1]	↑	0 = random
Log Loss	0.4509	[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	0.8812	0.8404	0.8540	0.8355	362
Claude Sonnet 4.5	0.8605	0.8062	0.8173	0.8018	337
WildJailbreak (Synthetic)	1.0000	1.0000	1.0000	1.0000	160

Best Performance: wildjailbreak (F1=1.0000)

Worst Performance: Claude Sonnet 4.5 (F1=0.8062)

F1 Std Dev: 0.0844

✓ **Good generalization** - minor performance variations across models.

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	3
Random Baseline	0.3333 (33.33%)
Model Accuracy	0.9092 (90.92%)
Correct Predictions	781 / 859
Improvement	0.5759 (57.59% points)
Test Statistic	Binomial Test
P-value	< 0.000001
Significance Level	α = 0.05
Result	**✓ SIGNIFICANT**
Effect Size (Cohen's h)	1.2984
Effect Interpretation	Large

Interpretation:

✓ Highly Significant (p < 0.001): The model's accuracy (0.9092) is significantly better than random guessing (0.3333). There is overwhelming evidence that the model has learned meaningful patterns. Effect size is large (h=1.2984).

For Publication: Report as: "Model accuracy (0.9092) significantly exceeded random baseline (0.3333, binomial test, p < 0.001, Cohen's h = 1.2984)."

Confidence & Calibration Metrics

Metric	Value	Range	Better	Note
Mean Confidence	0.9752	[0, 1]	~	Should match accuracy
Std. Confidence	0.0785	[0, 1]	~	Variation in confidence
Calibration Error (ECE)	0.0802	[0, 1]	↓	<0.1 = good

MCE (Max Calibration Error)	0.4366	[0, 1]	↓	Worst-case calibration
Brier Score	0.0885	[0, 1]	↓	0 = perfect
Confidence Gap	0.1112	[-1, 1]	↑	Correct - Incorrect
Mean Confidence (Correct)	0.9869	[0, 1]	↑	When model is right
Mean Confidence (Incorrect)	0.8756	[0, 1]	↓	When model is wrong

Per-Class Performance

Class: No Refusal

Metric	Value	Range	Better	Note
Precision	0.9670	[0, 1]	↑	TP/(TP+FP)
Recall	0.9472	[0, 1]	↑	TP/(TP+FN)
F1 Score	0.9570	[0, 1]	↑	Harmonic mean
Support	587	[0, ∞)	~	# samples
Specificity	0.9301	[0, 1]	↑	TN/(TN+FP)
Class Mean Confidence	0.9873	[0, 1]	↑	Avg confidence
Class Confidence Std	0.0611	[0, 1]	~	Variation
Class Min Confidence	0.5176	[0, 1]	~	Lowest
Class Max Confidence	1.0000	[0, 1]	~	Highest
Class-Level Accuracy	0.9591	[0, 1]	↑	For this class

Class: Hard Refusal

Metric	Value	Range	Better	Note
Precision	0.8784	[0, 1]	↑	TP/(TP+FP)
Recall	0.9028	[0, 1]	↑	TP/(TP+FN)
F1 Score	0.8904	[0, 1]	↑	Harmonic mean
Support	144	[0, ∞)	~	# samples
Specificity	0.9748	[0, 1]	↑	TN/(TN+FP)
Class Mean Confidence	0.9551	[0, 1]	↑	Avg confidence
Class Confidence Std	0.0974	[0, 1]	~	Variation
Class Min Confidence	0.5014	[0, 1]	~	Lowest
Class Max Confidence	1.0000	[0, 1]	~	Highest
Class-Level Accuracy	0.7708	[0, 1]	↑	For this class

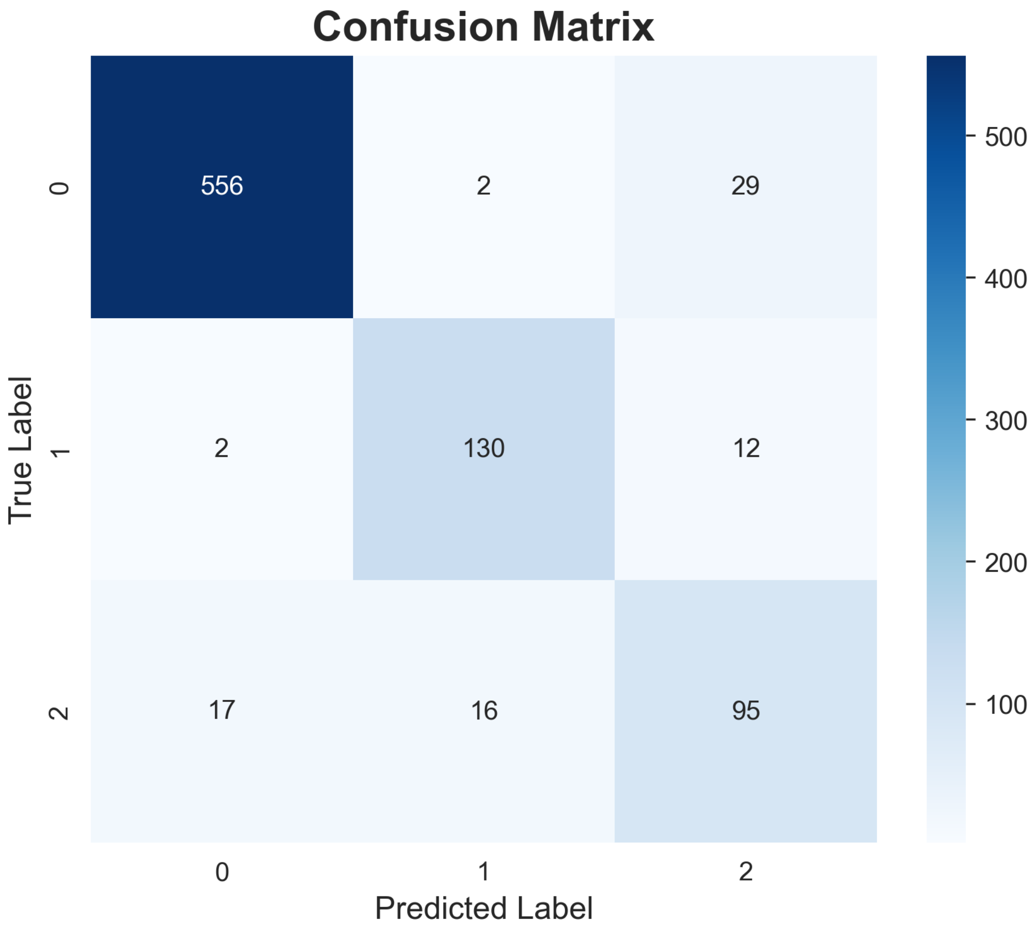
Class: Soft Refusal

Metric	Value	Range	Better	Note

Precision	0.6985	[0, 1]	↑	TP/(TP+FP)
Recall	0.7422	[0, 1]	↑	TP/(TP+FN)
F1 Score	0.7197	[0, 1]	↑	Harmonic mean
Support	128	[0, ∞)	~	# samples
Specificity	0.9439	[0, 1]	↑	TN/(TN+FP)
Class Mean Confidence	0.9424	[0, 1]	↑	Avg confidence
Class Confidence Std	0.1068	[0, 1]	~	Variation
Class Min Confidence	0.5325	[0, 1]	~	Lowest
Class Max Confidence	1.0000	[0, 1]	~	Highest
Class-Level Accuracy	0.7422	[0, 1]	↑	For this class

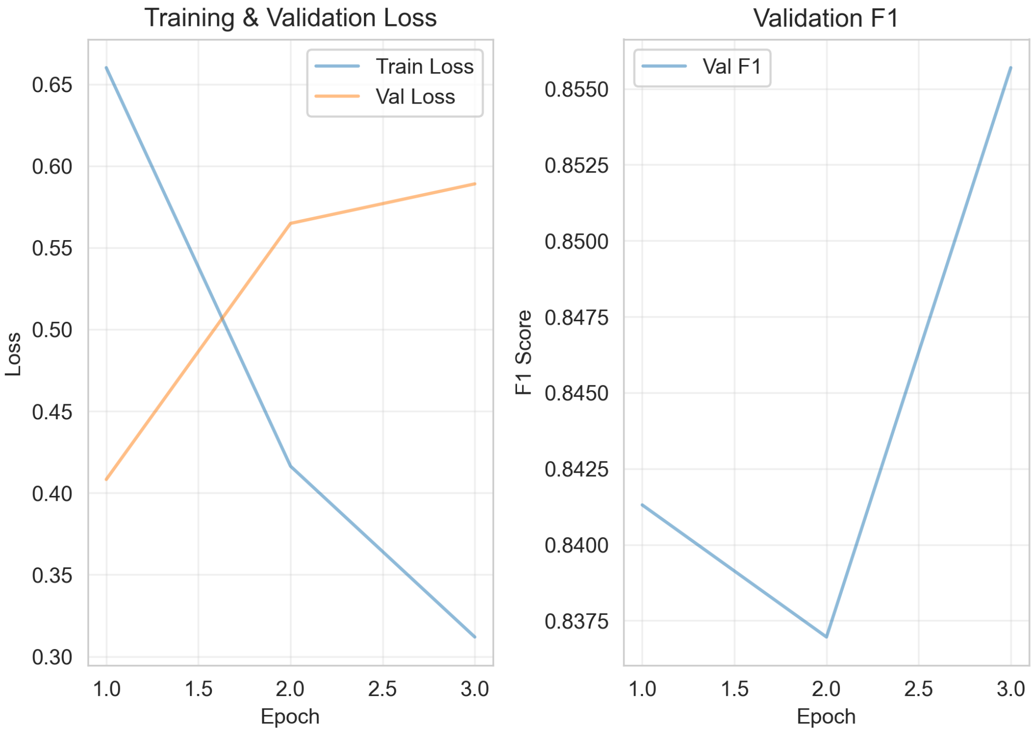
Confusion Matrix

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



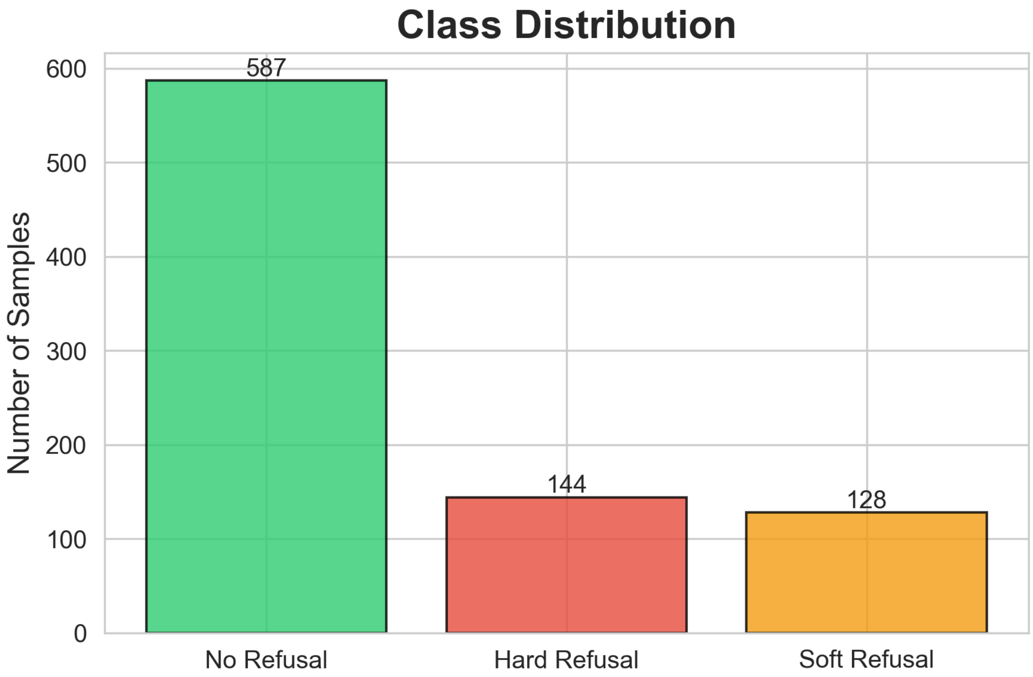
Training Curves

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.


Class Distribution
Number of Samples
No Refusal
Hard Refusal
Soft Refusal