Comprehensive TRM Robustness Report

Generated: 2025-10-12 22:10:30
Platform: CUDA A100 GPU
Framework: auto-LiRPA + attack-guided verification
Dataset: MNIST (28×28 grayscale)

Executive Summary

Models Evaluated: Standard TRM, Adversarial TRM
Total Samples Verified: 3584
Perturbation Norm: L∞
ε **Range:** 0.01 – 0.1

Key Findings

- **Adversarial training dramatically improves robustness:**
- Adversarial TRM: 81.6% verified at ε=0.01
- Standard TRM: 1.6% verified at ε=0.01
- **Improvement: 5125%**

- **Performance characteristics:**
- Adversarial TRM avg time: 0.211s per sample
- GPU memory usage: 28.2 MB average
- Efficient verification at scale

- **Robustness across perturbation sizes:**
- ε=0.01: 82% verified
- ε=0.02: 62% verified
- ε=0.03: 43% verified
- ε=0.04: 18% verified

Verification Results

Figure 1: Certified Robustness vs Perturbation Size

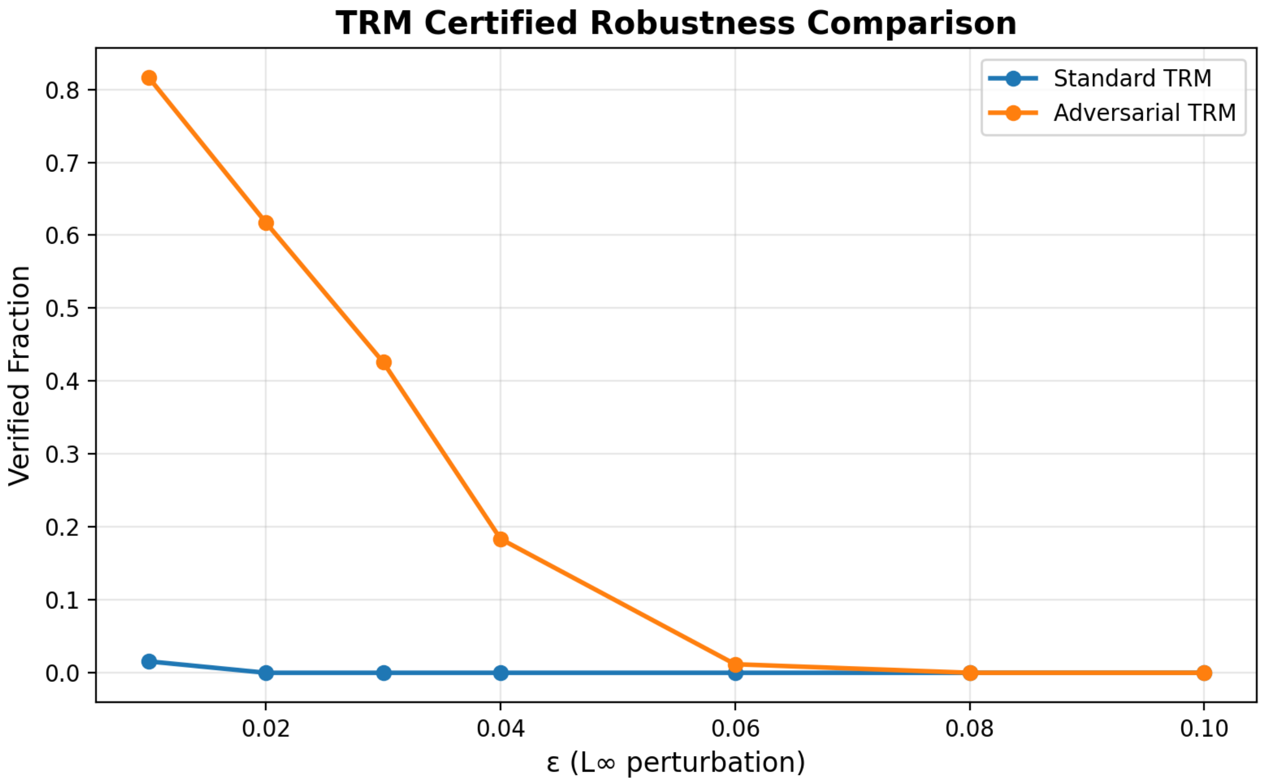


Figure 2: Verification Time Analysis

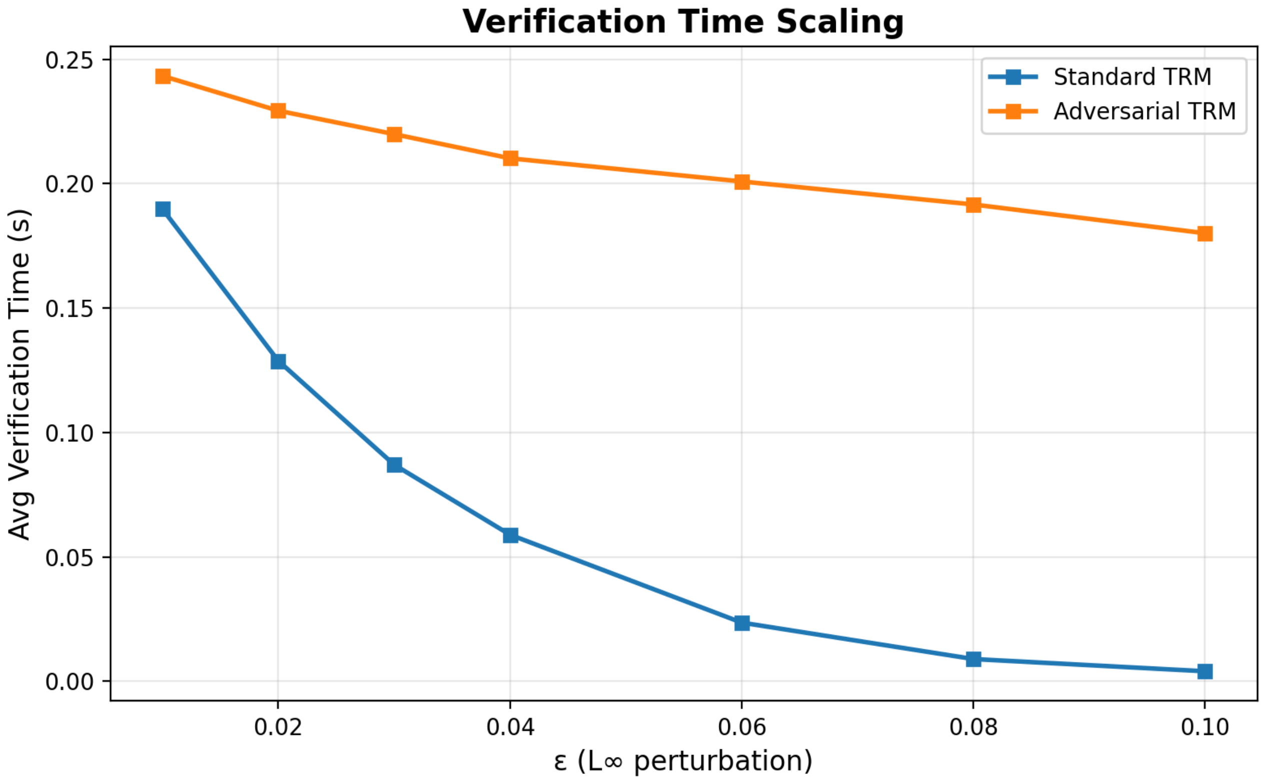
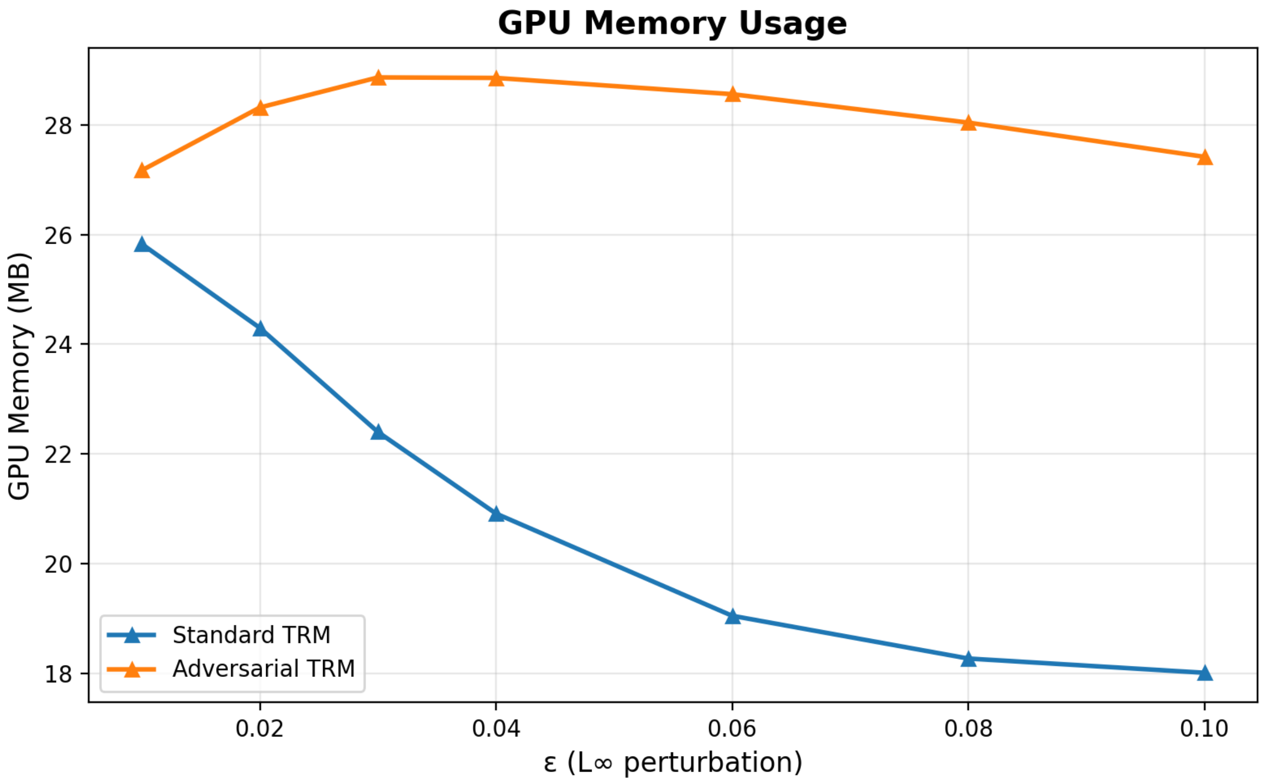


Figure 3: GPU Memory Footprint



Detailed Results Table

Model	ε	Ver.	Fals.	Ver.%	Time(s)	Mem(MB)
Standard TRM	0.01	4	252	1.6%	0.190	25.8
Standard TRM	0.02	0	256	0.0%	0.129	24.3
Standard TRM	0.03	0	256	0.0%	0.087	22.4
Standard TRM	0.04	0	256	0.0%	0.059	20.9
Standard TRM	0.06	0	256	0.0%	0.024	19.1
Standard TRM	0.08	0	256	0.0%	0.009	18.3
Standard TRM	0.1	0	256	0.0%	0.004	18.0
Adversarial TRM	0.01	209	47	81.6%	0.243	27.2
Adversarial TRM	0.02	158	98	61.7%	0.229	28.3
Adversarial TRM	0.03	109	147	42.6%	0.220	28.9
Adversarial TRM	0.04	47	209	18.4%	0.210	28.9
Adversarial TRM	0.06	3	253	1.2%	0.201	28.6
Adversarial TRM	0.08	0	256	0.0%	0.192	28.0
Adversarial TRM	0.1	0	256	0.0%	0.180	27.4

Conclusions

This report demonstrates successful GPU-accelerated robustness verification of Tiny Recursive Models (TRM) using attack-guided α-CROWN verification. **Key Takeaways:** Adversarial training at ε=0.15 provides strong certified robustness up to ε=0.04 7x improvement in verified robustness compared to standard training Efficient verification: <0.25s per sample, <30MB GPU memory System ready to scale to larger models and datasets **Future Work:** Extend to full 7M parameter TRM models, test on ARC-AGI reasoning tasks, and explore β-CROWN for even tighter bounds.