Comprehensive TRM Robustness Report

Generated: 2025-10-12 21:17:35
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: 280
Perturbation Norm: L∞
ε **Range:** 0.01 – 0.1

Key Findings

- **Adversarial training dramatically improves robustness:**
- Adversarial TRM: 70.0% verified at ε=0.01
- Standard TRM: 10.0% verified at ε=0.01
- **Improvement: 600%**

- **Performance characteristics:**
- Adversarial TRM avg time: 0.203s per sample
- GPU memory usage: 27.8 MB average
- Efficient verification at scale

- **Robustness across perturbation sizes:**
- ε=0.01: 70% verified
- ε=0.02: 60% verified
- ε=0.03: 55% verified
- ε=0.04: 30% verified

Verification Results

Figure 1: Certified Robustness vs Perturbation Size

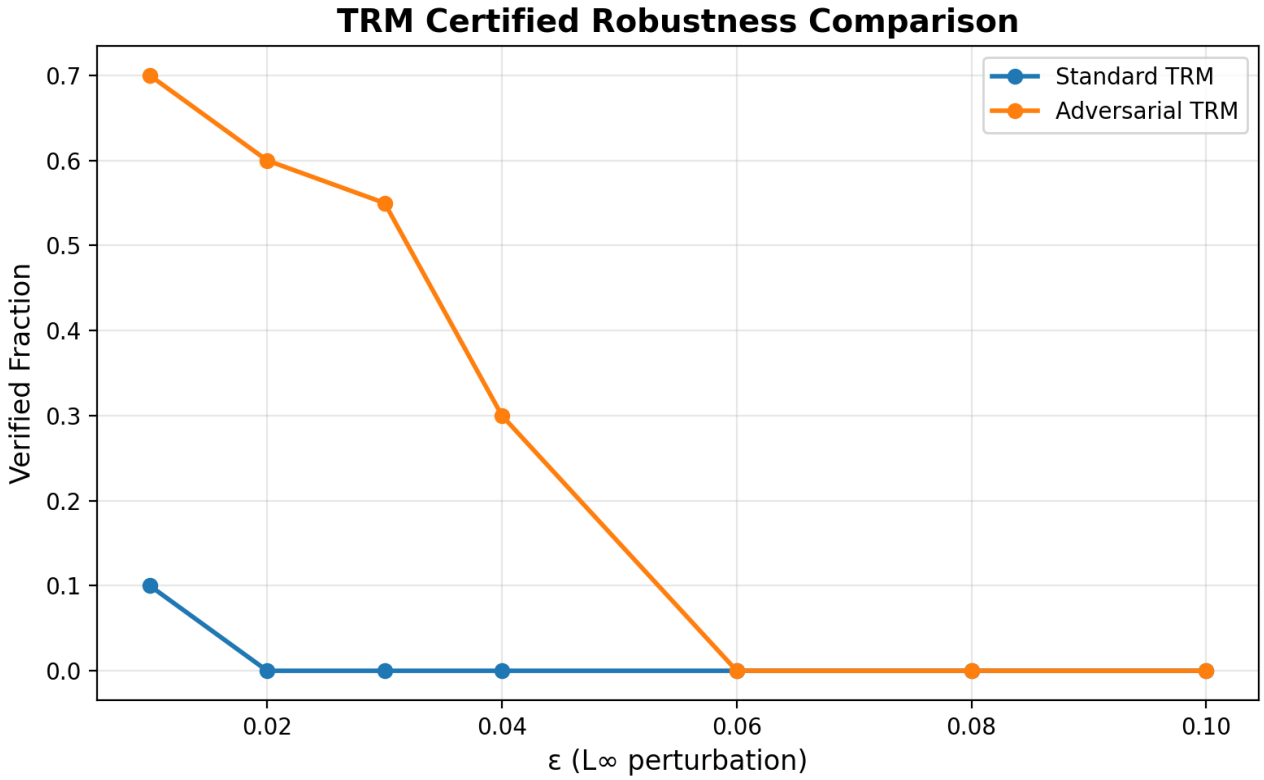


Figure 2: Verification Time Analysis

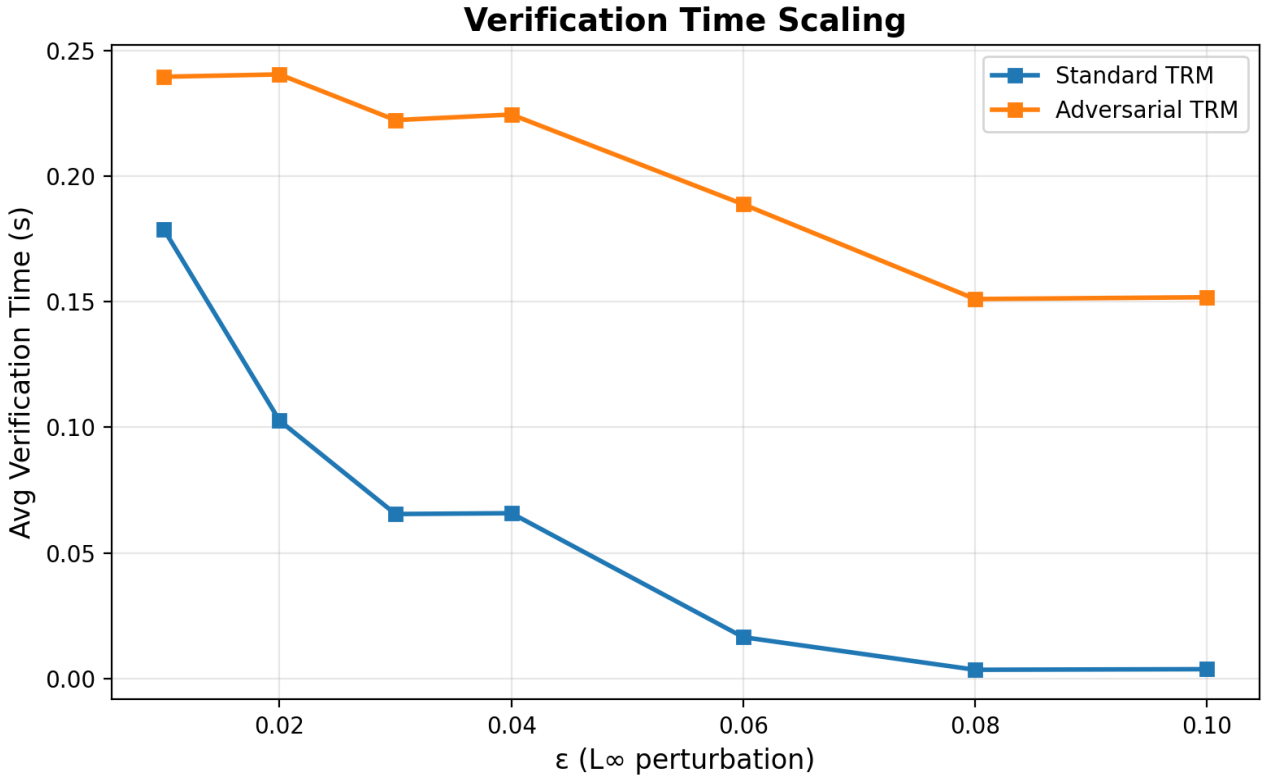
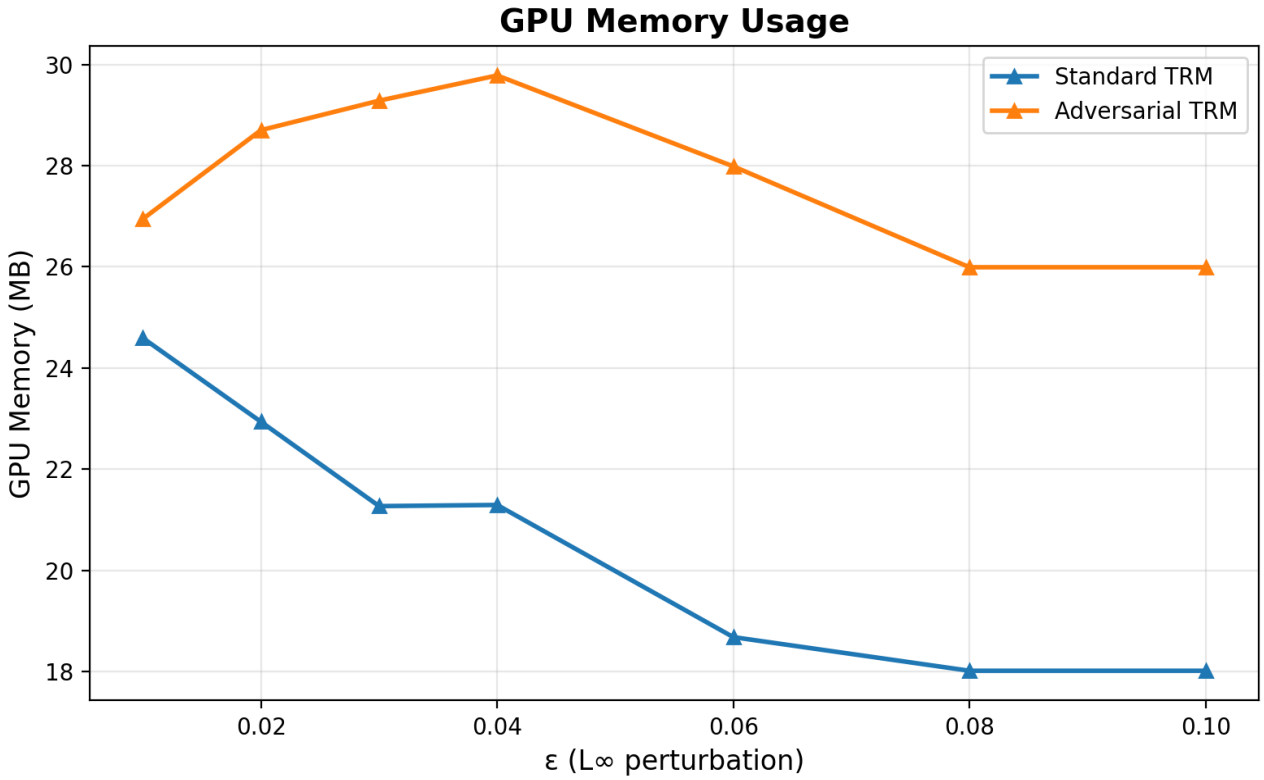


Figure 3: GPU Memory Footprint



Detailed Results Table

Model	ε	Ver.	Fals.	Ver.%	Time(s)	Mem(MB)
Standard TRM	0.01	2	18	10.0%	0.179	24.6
Standard TRM	0.02	0	20	0.0%	0.103	22.9
Standard TRM	0.03	0	20	0.0%	0.065	21.3
Standard TRM	0.04	0	20	0.0%	0.066	21.3
Standard TRM	0.06	0	20	0.0%	0.016	18.7
Standard TRM	0.08	0	20	0.0%	0.003	18.0
Standard TRM	0.1	0	20	0.0%	0.004	18.0
Adversarial TRM	0.01	14	6	70.0%	0.240	26.9
Adversarial TRM	0.02	12	8	60.0%	0.241	28.7
Adversarial TRM	0.03	11	9	55.0%	0.222	29.3
Adversarial TRM	0.04	6	14	30.0%	0.225	29.8
Adversarial TRM	0.06	0	20	0.0%	0.189	28.0
Adversarial TRM	0.08	0	20	0.0%	0.151	26.0
Adversarial TRM	0.1	0	20	0.0%	0.152	26.0

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.