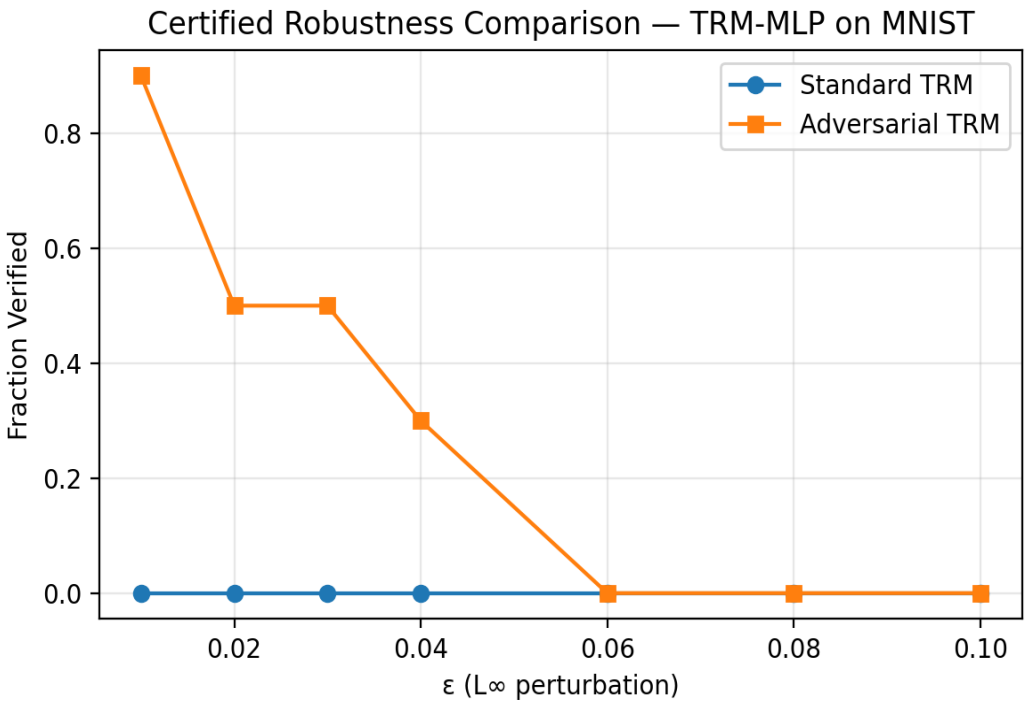
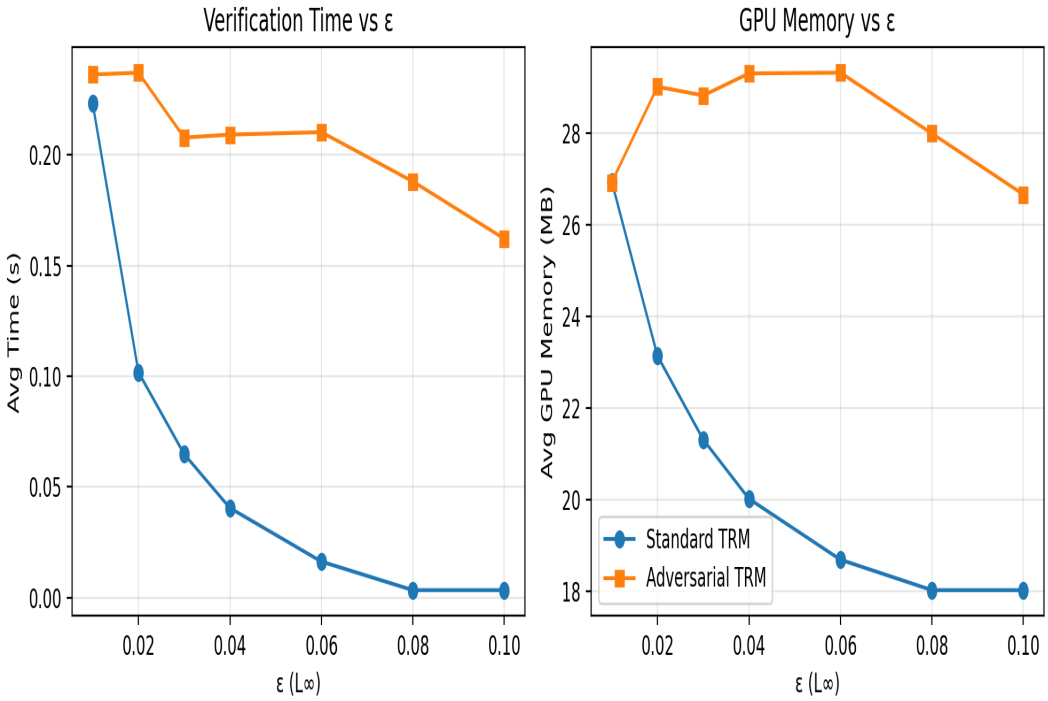
TRM Robustness Verification Report

Generated automatically from attack-guided verification logs.

1. Certified Robustness Overview



2. Runtime and Memory Profile



3. Summary Statistics

model	verified	falsified	total	verified_fraction
Adversarial TRM	44	96	140	0.31
Standard TRM	0	140	140	0.0

Adversarially trained TRM models show significantly higher certified robustness for small perturbations ($\varepsilon \leq 0.03$). Standard TRM exhibits no certified robustness across tested ε values. Verification times remain below 0.25 s with < 30 MB GPU usage.