PORTFOLIO DOCUMENT

Adversarial Robustness Evaluation for Practical LLM Systems

Portfolio Research Note • Youssef Ibrahim • 2025

Focus

This note summarizes how character-level and word-level perturbations can expose brittle behavior in deployed language systems and why reproducible evaluation matters.

Coverage

Benchmark-oriented perturbation generation, robustness metrics, failure pattern analysis, and connections to trustworthy AI engineering.

Takeaway

Reliability under noise and malicious input is not an edge case. It is a practical deployment requirement for any serious AI product.