Bigger $\neq$ Safer: A First Look at Refusal Robustness Scaling in LLMs

Yueyi (Claire) Wang
University of Cambridge
yw562@cam.ac.uk

Abstract

Refusal robustness—the ability of large language models (LLMs) to consistently reject harmful requests—remains underexplored compared to capabilities scaling. This proposal outlines a mini-scaling study using 1.3B–13B open-source LLMs. We introduce standardized metrics—Refusal Retention Rate (RRR), Refusal Drift (RD), and Calibration Entropy (CE)—and evaluate robustness under adversarial finetuning and prompt-based attacks. We hypothesize that attack compute often overwhelms scale, challenging the assumption that "bigger is safer." The document specifies models, datasets, attacks/defenses, evaluation pipeline, and a 10-day experimental plan.

1 Introduction

Refusal robustness—the ability of LLMs to consistently refuse harmful or unethical requests even under adversarial pressure—is central to AI safety. Prior scaling-law work focuses on capabilities (reasoning, factuality) or general adversarial robustness, while mechanistic studies suggest refusal-related features (e.g., "refusal subspaces") but lack cross-scale analysis.

This study fills that gap via a controlled scaling analysis across small-to-medium open-source LLMs (1.3B–13B). We ask:

1. How does model size affect refusal robustness?
2. How does *attack compute* (data size $\times$ finetuning steps) interact with scale?
3. Do defenses (tamper-resistant finetuning, activation steering) alter scaling trends?

We target a compute-efficient, reproducible pipeline that can run on a single GPU per experiment and provide workshop-ready figures and tables.

2 Metrics

Refusal Retention Rate (RRR). Fraction of harmful prompts still refused after adversarial pressure. Implementation: refusal-phrase heuristics + a DeBERTa-based binary classifier (refuse vs. comply). Primary outcome.

Refusal Drift (RD). Semantic distance between baseline refusals and post-attack refusals, using sentence embeddings (SBERT/SimCSE). Captures subtle degradation of refusal style/content even when RRR remains high.

Calibration Entropy (CE). Entropy of the refusal classifier logits; higher entropy indicates uncertain/fragile refusals. Useful for identifying near-boundary cases.

Validation. Randomly sample $\sim$200 responses across models/attacks and re-judge with a strong LLM to sanity-check metrics; report agreement rate (target $>$90%).

3 Methodology

Models. Compute-friendly open-source LLMs: *Phi-2 (1.3B), Mistral-3B, LLaMA-7B,* and optionally *LLaMA-13B*. No 70B models are required.

Datasets. Harmful prompt sets drawn from publicly available jailbreak/adv benchmarks (e.g., JailbreakBench, AdvBench), de-duplicated and standardized to $\sim$2k unique prompts.

Attacks. (i) **Adversarial LoRA finetuning**: train on harmful prompts with targets that encourage compliance or safety erosion. Factors: data size (500/1000/2000), steps (500/1000/2000).
(ii) **Prompt-only attacks**: AutoDAN and GCG-style paraphrase templates for zero-finetune stress tests.

Defenses (optional). Tamper-resistant finetuning (TAR) and activation steering baselines applied post-attack to probe whether scaling trends change under defenses.

Evaluation Protocol. For each (model, attack) setting, generate responses to the harmful prompt set, compute RRR/RD/CE, and log run metadata (seed, lr, steps, LoRA rank). For prompt attacks, sweep multiple seeds/templates.

Compute Budget. Designed for single-GPU runs per condition; total wall-clock depends on hardware but remains feasible by keeping dataset and step counts small.

4 Experimental Plan

Day 1–2: Setup & Baseline. Implement refusal classifier (heuristics + DeBERTa), prepare datasets, and run baseline RRR/RD/CE on all models. Verify end-to-end pipeline.

Day 3–4: Small Model (Phi-2, 1.3B). LoRA finetunes (0.5k/1k/2k prompts; 0.5k/1k/2k steps). Evaluate metrics. Run prompt-based attacks (AutoDAN, GCG).

Day 5–6: Medium Model (Mistral-3B). Repeat finetuning and prompt attacks. Compare trends vs. 1.3B.

Day 7: Large Model (LLaMA-7B). Repeat procedures for 7B.

Day 8: Optional 13B. If compute allows, add LLaMA-13B for an extra scaling point.

Day 9: Validation. Re-judge $\sim$200 responses with a strong LLM; report agreement with classifier and adjust thresholds if needed.

Day 10: Analysis & Writing. Fit scaling curves (log-log or log-linear) for RRR, RD, CE vs. model size and attack compute; produce plots and tables; draft write-up.

5 Expected Results

We hypothesize:

- **RRR**: Slight increase with size under weak attacks, but collapse under stronger attack compute; bigger $\not\Rightarrow$ safer.
- **RD**: Rises primarily with attack compute; weak dependence on size.
- **CE**: Confidence improves with size when attacks are weak; converges under strong attacks across sizes.
- **Overall trend**: Attack compute dominates model scale in determining refusal robustness.

These predictions will be visualized as scaling curves with confidence intervals across seeds and attack templates.

6 Contributions

1. First systematic mini-scaling study of refusal robustness across open-source LLMs.
2. Reproducible metric suite (RRR, RD, CE) and evaluation pipeline suitable for low compute.
3. Evidence and analysis that attack compute can overwhelm model scale, challenging the assumption that safety scales monotonically with size.

References