SAFETY MARCH 12, 2026

DEFENSIVE REFUSAL BIAS: HOW SAFETY ALIGNMENT FAILS CYBER DEFENDERS

David Campbell, Neil Kale, Udari Madhushani Sehwag, Bert Herring, Nick Price, Dan Borges, Alex Levinson, Christina Q. Knight

AI safety can block attackers—but also defenders. This post introduces Defensive Refusal Bias and shows how aligned models fail real-world cybersecurity tasks.

Safety alignment in large language models (LLMs), particularly for cybersecurity tasks, primarily focuses on preventing misuse. While this approach reduces direct harm, it obscures a complementary failure mode: denial of assistance to legitimate defenders. We study Defensive Refusal Bias -- the tendency of safety-tuned frontier LLMs to refuse assistance for authorized defensive cybersecurity tasks when those tasks include similar language to an offensive cyber task. Based on 2,390 real-world examples from the National Collegiate Cyber Defense Competition (NCCDC), we find that LLMs refuse defensive requests containing security-sensitive keywords at 2.72× the rate of semantically equivalent neutral requests ($p<0.001$). The highest refusal rates occur in the most operationally critical tasks: system hardening (43.8%) and malware analysis (34.3%). Interestingly, explicit authorization, where the user directly instructs the model that they have authority to complete the target task, increases refusal rates, suggesting models interpret justifications as adversarial rather than exculpatory. These findings are urgent for interactive use and critical for autonomous defensive agents, which cannot rephrase refused queries or retry. Our findings suggest that current LLM cybersecurity alignment relies on semantic similarity to harmful content rather than reasoning about intent or authorization. We call for mitigations that analyze intent to maximize defensive capabilities while still preventing harmful compliance.

LINKS

https://arxiv.org/pdf/2603.01246