Research Statement

Xin Wang | xinwang22@m.fudan.edu.cn | https://xinwong.github.io/

Biography. I am a final year Ph.D. Candidate of FVL Lab in the School of Computer Science at Fudan University, supervised by Prof. Xingjun Ma and Prof. Yu-Gang Jiang. Recently, I am broadly interested in safety and privacy aspects of machine learning with a recent focus on large language models. Most of my past works are in the domain of trustworthy machine learning, particularly adversarial examples and robustness of machine learning algorithms.

1 BACKGROUND & MOTIVATION

The remarkable capabilities of Large Language Models (LLMs) and Vision-Language Models (VLMs) are matched by a new frontier of risks, including the production of harmful or misleading content, fake alignment, and emergent hazards arising from interacting Multi-Agent systems. Effectively mitigating these risks is complicated by several fundamental challenges: (i) Limitations of Static Attack Strategies: Pre-defined, static attack strategies are insufficient to uncover novel failure modes, highlighting the need for adaptive attack methods. (ii) Lack of Integrated Defense Systems: Defenses are often implemented as disjointed pre-processing, in-processing, or post-processing stages, lacking a unified, systematic framework. (iii) Dynamic Evaluation: The evaluation landscape is constantly shifting, rendering static benchmarks insufficient for comprehensive assessment. To overcome these gaps, our research will explore the development of an end-to-end LLMs/VLMs safety system structured around a continuous loop: Risk Modeling → Adaptive Adversarial Attacks → Safety Alignment → Dynamic Evaluation.

2 ADVERSARIAL ATTACK FOR LLMS/VLMS SAFETY

A critical component of our proposed safety loop is the development of adaptive adversarial attacks. The current landscape of automated red teaming, while increasingly sophisticated, is fundamentally constrained by its reliance on pre-defined attack strategies. Existing frameworks are often limited to selecting, combining, or refining known jailbreak methods. This confines their creativity and leaves them unable to autonomously invent entirely new attack mechanisms, a crucial capability for uncovering novel and unexpected model vulnerabilities.

To address this critical gap, our research will explore an autonomous framework that pioneers a paradigm shift from attack planning to the evolutionary synthesis of jailbreak methods. Instead of merely refining prompts, adaptive jailbreak leverages a multi-agent system to autonomously engineer, evolve, and execute novel, code-based attack algorithms from the ground up. Crucially, it features a code-level self-correction loop, which allows the framework to iteratively rewrite its own attack algorithms in response to failures. This dynamic, self-improving process enables it to adapt to the target model's defenses in real-time, a hallmark of truly adaptive attacks.

3 AN INTEGRATED AI SAFETY FRAMEWORK

Effective defense against adaptive and evolving threats requires moving beyond disjointed solutions towards a deeply integrated, multi-layered AI safety systems.

3.1 Pre-processing

We will explore a sophisticated guard model, similar to LlamaGuard or QwenGuard, to proactively filter malicious content before it reaches the LLMs/VLMs. This model will operate at three distinct levels of granularity for comprehensive detection:

(i) **Input/Output Level:** Assessing the overall safety of entire prompts and responses.

(ii) **Sentence Level:** Decomposing text to scrutinize individual sentences for threats hidden within otherwise benign content.

(iii) **Token Level:** Analyzing token probabilities to flag novel, obfuscated, or emerging attack patterns at the most granular scale.

3.2 In-processing

To address the critical trade-off between safety and capability, we will explore a multi-objective safe alignment process. Instead of relying on a single safety metric, which can lead to over-refusal, we will utilize multiple distinct reward models during training.

Furthermore, our goal extends beyond a simple binary of safe vs. unsafe. We recognize a spectrum from **Unsafe** to **Useless Safe** (e.g., overly cautious refusals) to **Useful Safe**. The core principle guiding our alignment is instilling a clear preference hierarchy in the model: Useful Safe > Useless Safe > Unsafe.

3.3 Post-processing

The final layer of defense will employ Test-Time Adaptation (TTA) to ensure robust and safe outputs during inference. This technique allows the model to dynamically adjust its behavior in real-time based on the specific input it is processing. By adapting at the moment of generation, the model can effectively neutralize unforeseen threats that may have bypassed earlier defenses, ensuring the final output is reliable and secure.

4 FUTURE PLAN

While our immediate work targets current LLMs/VLMs, our research vision extends to the far-horizon risks of next-generation AI. As models evolve from content generators to autonomous agents that act in the physical and digital worlds, the potential for harm is profoundly amplified.

4.1 Vulnerabilities in Vision-Language-Action Models

The embodiment of AI in robotics, creating VLA models, translates digital vulnerabilities into direct physical risks. My prior work has already explored this frontier by developing FreezeVLA, an attack framework that exposes a critical failure mode: physical paralysis. We demonstrated that a single adversarial image can "freeze" a robot, causing it to ignore subsequent commands. This work shows how a digital attack can sever the link between an robot's mind and its physical actions, highlighting the urgent need for robust defenses in embodied systems.

4.2 Systemic Risks in Multi-Agent Systems and Agent OS

Looking forward, the development of interconnected Agent OS and complex multi-agent systems presents an even greater challenge. In these systems, multiple AI agents will collaborate, delegate tasks, and interact within a shared environment, creating a vector for systemic, cascading failures. A single compromised agent could act as a digital pathogen, spreading malicious influence or faulty instructions throughout the Agent OS. This could lead to emergent harmful behaviors, such as coordinated manipulation of digital systems, which would be nearly impossible to trace back to a single point of failure.

Our current research is actively exploring this vulnerability by demonstrating how an adversarial or backdoor attack on a single agent can trigger a catastrophic cascade, causing a widespread failure across a network of millions of interconnected agents.