

Review

Kris Carlson prompting ChatGPT4o

Please summarize this document, listing each type of attack and explaining how it works, the models that are most susceptible to it (if any are), the best defense strategies for each attack and why they work as well as their weaknesses, and the success rate of each attack and defense.

Note: The residual vulnerability is the attack rate (given absent defenses) minus the defense rate.

Under review. Current version: arXiv:2502.05206v3

Safety at Scale: Comprehensive Survey of Large Model Safety

Xingjun Ma¹, Yifeng Gao¹, Yixu Wang¹, Ruofan Wang¹, Xin Wang¹, Ye Sun¹, Yifan Ding¹, Hengyuan Xu¹, Yunhao Chen¹, Yunhan Zhao¹, Hanxun Huang², Yige Li³, Jiaming Zhang⁴, Xiang Zheng⁵, Yang Bai⁶, Zuxuan Wu¹, Xipeng Qiu¹, Jingfeng Zhang^{8,9}, Yiming Li⁷, Xudong Han¹⁰, Haonan Li¹⁰, Jun Sun³, Cong Wang⁵, Jindong Gu¹², Baoyuan Wu¹³, Siheng Chen¹⁴, Tianwei Zhang⁷, Yang Liu⁷, Mingming Gong², Tongliang Liu¹⁵, Shirui Pan¹⁶, Cihang Xie¹⁷, Tianyu Pang¹⁸, Yinpeng Dong¹⁹, Ruoxi Jia²⁰, Yang Zhang²¹, Shiqing Ma²², Xiangyu Zhang²³, Neil Gong²⁴, Chaowei Xiao²⁵, Sarah Erfani², Tim Baldwin^{2,10}, Bo Li²⁶, Masashi Sugiyama^{9,11}, Dacheng Tao⁷, James Bailey², Yu-Gang Jiang[†]

¹Fudan University, ²The University of Melbourne, ³Singapore Management University, ⁴Hong Kong University of Science and Technology, ⁵City University of Hong Kong, ⁶ByteDance, ⁷Nanyang Technological University, ⁸University of Auckland, ⁹RIKEN, ¹⁰MBZUAI, ¹¹The University of Tokyo, ¹²University of Oxford, ¹³Chinese University of Hong Kong, Shenzhen, ¹⁴Shanghai Jiao Tong University, ¹⁵The University of Sydney, ¹⁶Griffith University, ¹⁷University of California, Santa Cruz, ¹⁸Sea AI Lab, ¹⁹Tsinghua University, ²⁰Virginia Tech, ²¹CISPA Helmholtz Center for Information Security, ²²University of Massachusetts Amherst, ²³Purdue University, ²⁴Duke University, ²⁵University of Wisconsin - Madison, ²⁶University of Illinois Urbana-Champaign

This document provides a comprehensive survey of safety threats and defenses for large models, including Vision Foundation Models (VFM), Large Language Models (LLMs), Vision-Language Pretraining (VLP) models, Vision-Language Models (VLMs), Diffusion Models (DMs), and large-model-powered agents. Below is a summary of each type of attack, the models most susceptible to it, defense strategies, and their effectiveness.

1. Adversarial Attacks

How It Works: Adversarial attacks manipulate inputs (e.g., images, text) to mislead models into incorrect predictions. They can be white-box (full access to model parameters), black-box (limited access), or gray-box (partial access). Techniques include patch attacks, embedding attacks, and perturbations.

Susceptible Models:

- VFMs (e.g., Vision Transformers, SAM)
- LLMs (e.g., GPT-3, GPT-4)
- VLP models (e.g., CLIP)

- VLMs (e.g., Flamingo, BLIP)
- DMs (e.g., Stable Diffusion)

Defense Strategies:

- **Adversarial Training:** Improves robustness by training on adversarial examples. Effective but computationally expensive.
- **Detection:** Identifies adversarial inputs using statistical anomalies or attention maps. Lightweight but limited to known attack types.
- **Purification:** Removes adversarial perturbations via preprocessing (e.g., diffusion models). Effective but may degrade clean input quality.
- **Robust Architectures:** Modifies model components (e.g., attention mechanisms) to resist attacks. Effective but requires architectural changes.

Success Rates:

- Attack success rates vary (e.g., 60%-90% for white-box attacks).
- Defense success rates depend on the method (e.g., adversarial training achieves ~70%-80% robustness).

2. Backdoor Attacks

How It Works: Backdoor attacks inject hidden triggers into models during training or fine-tuning, causing malicious behavior when specific inputs are provided.

Susceptible Models:

- VLMs (e.g., Vision Transformers)
- LLMs (e.g., GPT-3, GPT-4)
- VLP models (e.g., CLIP)
- DMs (e.g., Stable Diffusion)

Defense Strategies:

- **Detection:** Identifies backdoor triggers using gradient analysis or outlier detection. Effective but limited to known triggers.
- **Removal:** Neutralizes backdoors via fine-tuning or pruning. Effective but may degrade model performance.
- **Robust Training:** Prevents backdoor injection during training. Effective but computationally expensive.

Success Rates:

- Attack success rates are high (~80%-95%).
- Defense success rates vary (e.g., detection achieves ~70%-80%).

3. Jailbreak Attacks

How It Works: Jailbreak attacks bypass safety mechanisms to induce harmful or restricted outputs. They exploit prompt engineering, adversarial suffixes, or system vulnerabilities.

Susceptible Models:

- LLMs (e.g., GPT-4, Claude)
- VLMs (e.g., LLaVA, MiniGPT-4)

Defense Strategies:

- **Input Defense:** Preprocesses prompts to neutralize malicious intent (e.g., rephrasing, backtranslation). Effective but computationally intensive.
- **Output Defense:** Filters unsafe outputs using classifiers or loss analysis. Effective but prone to false positives.
- **Ensemble Defense:** Combines multiple defenses for enhanced robustness. Effective but resource-intensive.

Success Rates:

- Attack success rates are high (~70%-90%).
- Defense success rates vary (e.g., ensemble defenses achieve ~80%-90%).

4. Prompt Injection Attacks

How It Works: Manipulates model behavior by embedding malicious instructions into prompts or external data sources.

Susceptible Models:

- LLMs (e.g., GPT-3, GPT-4)
- VLMs (e.g., PandaGPT)

Defense Strategies:

- **Input Defense:** Structures prompts to prevent mixing of instructions and data. Effective but limited to known injection methods.

- **Adversarial Fine-Tuning:** Strengthens models against malicious instructions. Effective but reduces generalization.

Success Rates:

- Attack success rates are high (~80%-95%).
- Defense success rates vary (~70%-85%).

5. Energy-Latency Attacks

How It Works: Increases computational demands to degrade model efficiency, causing higher latency and energy consumption.

Susceptible Models:

- LLMs (e.g., GPT-3, GPT-4)
- VLMs (e.g., MiniGPT-4)

Defense Strategies:

- **Runtime Validation:** Detects and mitigates inefficient inputs. Effective but adds overhead.
- **Robust Architectures:** Optimizes model components for efficiency. Effective but requires architectural changes.

Success Rates:

- Attack success rates are moderate (~50%-70%).
- Defense success rates vary (~60%-80%).

6. Model Extraction Attacks

How It Works: Steals model functionality by querying it and analyzing responses to replicate its behavior.

Susceptible Models:

- LLMs (e.g., GPT-3, GPT-4)
- DMs (e.g., Stable Diffusion)

Defense Strategies:

- **Rate Limiting:** Restricts query frequency. Effective but impacts usability.
- **Noise Injection:** Adds randomness to outputs to reduce extraction accuracy. Effective but degrades output quality.

Success Rates:

- Attack success rates are moderate (~50%-70%).
- Defense success rates vary (~60%-80%).

7. Data Extraction Attacks

How It Works: Recovers training data by exploiting model memorization or generative capabilities.

Susceptible Models:

- LLMs (e.g., GPT-3, GPT-4)
- DMs (e.g., Stable Diffusion)

Defense Strategies:

- **Membership Inference Detection:** Identifies data leakage. Effective but computationally expensive.
- **Data Sanitization:** Removes sensitive data from training sets. Effective but reduces model performance.

Success Rates:

- Attack success rates are high (~70%-90%).
- Defense success rates vary (~60%-80%).

8. Intellectual Property Protection

How It Works: Protects training data, generated content, and model parameters using watermarking or adversarial techniques.

Susceptible Models:

- DMs (e.g., Stable Diffusion)
- LLMs (e.g., GPT-3, GPT-4)

Defense Strategies:

- **Watermarking:** Embeds unique identifiers into data or models. Effective but vulnerable to removal attacks.
- **Adversarial Protection:** Prevents unauthorized learning or editing. Effective but computationally expensive.

Success Rates:

- Attack success rates are moderate (~50%-70%).
- Defense success rates vary (~60%-80%).

What are the open challenges in large model safety?

The open challenges in large model safety, as identified in the document, are categorized into four main areas:

1. Fundamental Vulnerabilities

- **Purpose of Attacks Beyond Breaking Models:** Attacks should be used as diagnostic tools to uncover unintended behaviors and systemic flaws rather than just disrupting functionality.
- **Vulnerabilities in Language Models:** Issues like memorization, exposure to harmful content, and amplification of hallucinations need deeper exploration to address architectural and training data weaknesses.
- **Cross-Modal Vulnerabilities:** Understanding how vulnerabilities propagate across modalities (e.g., vision and language) in multi-modal models is critical.
- **Diffusion Models' Lack of Language Capabilities:** Bridging the gap between visual and linguistic capabilities in generative models is essential for multi-modal safety.
- **Extent of Memorization:** Investigating how much training data models memorize and the mechanisms behind it is crucial for privacy and security.
- **Agent Vulnerabilities:** As agents gain capabilities, their vulnerabilities grow, especially in dynamic environments where cascading failures can occur.

2. Safety Evaluation

- **Beyond Attack Success Rate:** Metrics should account for the severity of disruptions, resilience to attacks, and real-world consequences.
- **Static Evaluations' Limitations:** Static benchmarks fail to capture evolving threats, necessitating dynamic and adaptive evaluation systems.
- **Necessity of Adversarial Evaluations:** Adversarial testing is essential to assess models under real-world attack conditions.
- **Open-Ended Evaluation:** Developing reliable methods to evaluate open-ended responses remains a challenge.

3. Safety Defense

- **Limitations of Safety Alignment:** Issues like fake alignment and shallow safety highlight the need for deeper and adversarial alignment mechanisms.
- **Challenges in Defending Against Jailbreak Attacks:** Jailbreak attacks are harder to defend due to their lack of perturbation constraints compared to adversarial attacks.
- **Need for Practical Defenses:** Defenses should be general, efficient, adaptable, and compatible with black-box settings.

- **Proactive Defenses:** Designing proactive strategies to counter attacks before they succeed is underexplored.
- **Integration of Detection Systems:** Detection methods should be integrated into defense pipelines for dynamic threat response.
- **Sustainable Data Practices:** Addressing issues like lack of consent, explosion of generated data, and depletion of free data is critical for ethical AI development.

4. Future Directions

- **Safe Embodied Agents:** Ensuring physical safety in embodied AI agents deployed in real-world environments is a growing concern.
- **Safe Superintelligence:** Mechanisms like oversight systems, safety layers, safety experts, adversarial alignment, and safety consciousness are needed to ensure predictable and value-aligned behavior in advanced AI systems.
- **Call for Collective Action:** Collaborative efforts are required to prioritize defense-oriented research, develop safety APIs, open-source platforms, and foster global collaborations.

These challenges highlight the need for ongoing innovation, collaboration, and ethical considerations to ensure the robustness and reliability of large models in diverse applications.

Supplementary Information Not Included in the Article

Give a one-paragraph description of each type of large model whose vulnerabilities are described in the article.

Vision Foundation Models (VFMs):

VFMs, such as Vision Transformers (ViTs) and the Segment Anything Model (SAM), are pre-trained models designed for visual tasks like image classification, segmentation, and object detection. They process images as tokenized patches, enabling them to capture spatial dependencies effectively. However, their modular structure and reliance on self-attention mechanisms make them vulnerable to adversarial attacks, backdoor injections, and data poisoning. VFMs are widely used as backbones for downstream tasks, making their safety critical in applications like autonomous driving and medical imaging.

Large Language Models (LLMs):

LLMs, such as GPT-3, GPT-4, and Claude, are powerful text-based models capable of generating human-like responses, translating languages, and performing complex reasoning tasks. They are trained on massive datasets and excel in applications like conversational AI, content creation, and scientific research. However, their reliance on

statistical patterns rather than semantic understanding makes them susceptible to adversarial prompts, jailbreak attacks, backdoor injections, and data extraction. Ensuring their safety is vital for maintaining trust and reliability in sensitive domains.

Vision-Language Pretraining (VLP) Models:

VLP models, such as CLIP and ALBEF, align visual and textual modalities to enable tasks like image-text retrieval, visual grounding, and visual entailment. They leverage contrastive learning to connect image and text embeddings, making them versatile for multi-modal applications. However, their cross-modal nature introduces unique vulnerabilities, including adversarial attacks targeting both visual and textual inputs, backdoor attacks, and data poisoning. Their safety is crucial for applications like content moderation and visual search.

Vision-Language Models (VLMs):

VLMs, such as Flamingo, BLIP, and MiniGPT-4, extend LLMs by integrating visual modalities through pre-trained image encoders and alignment modules. This enables tasks like visual question answering, image captioning, and multi-modal reasoning. However, their multi-modal design introduces vulnerabilities like adversarial perturbations in images, jailbreak attacks targeting both text and image inputs, and cross-modal backdoor attacks. Ensuring their safety is essential for applications in education, healthcare, and autonomous systems.

Diffusion Models (DMs):

Diffusion models, such as Stable Diffusion and DALL-E, are generative models used for tasks like text-to-image synthesis, image editing, and video generation. They operate by iteratively denoising random noise to generate high-quality outputs based on learned data distributions. While highly effective for creative applications, they are vulnerable to adversarial attacks on text prompts, jailbreak attacks bypassing safety filters, backdoor injections, and data extraction. Their safety is critical for preventing misuse in content creation and intellectual property violations.

Large-Model-Powered Agents:

Agents powered by large models, such as AutoGPT and LLaVA, combine the capabilities of LLMs and VLMs to perform complex tasks autonomously, often interacting with external tools and environments. These agents are used in safety-critical domains like robotics, autonomous driving, and medical diagnostics. However, their dynamic nature and reliance on foundational models make them susceptible to prompt injection, backdoor attacks, jailbreaks, and cascading vulnerabilities across components. Ensuring their

safety is vital for preventing real-world harm and maintaining trust in autonomous systems.

How do modular structure and self-attention make VFMs susceptible to adversarial attacks, backdoor injections, and data poisoning?

The modular structure and self-attention mechanisms of Vision Foundation Models (VFMs) contribute to their susceptibility to adversarial attacks, backdoor injections, and data poisoning in the following ways:

Modular Structure: VFMs, such as Vision Transformers (ViTs), process images as tokenized patches, treating each patch as an independent unit. This modular design allows attackers to target specific patches or tokens with adversarial perturbations, backdoor triggers, or poisoned data. For example, adversarial attacks can manipulate individual patches to alter the model's attention scores, leading to incorrect predictions. Similarly, backdoor triggers can be embedded in specific patches, activating malicious behavior when the trigger is present. The modularity also makes it easier for poisoned data to influence localized areas of the model, compromising its integrity without affecting the entire system.

Self-Attention Mechanisms: Self-attention modules in VFMs compute relationships between patches by assigning attention scores based on their relevance to the task. This mechanism is highly sensitive to perturbations, as even small changes in input patches can disproportionately affect attention scores. Adversarial attacks exploit this sensitivity by introducing perturbations that redirect attention to irrelevant or malicious patches, causing the model to misinterpret the input. Backdoor injections can manipulate attention maps to prioritize trigger patterns, ensuring the model behaves maliciously when the trigger is present. Additionally, poisoned data can distort the attention mechanism during training, embedding harmful biases or vulnerabilities into the model.

Overall, the modular structure and self-attention mechanisms, while enabling VFMs to capture complex spatial dependencies, also create entry points for attackers to exploit localized vulnerabilities, manipulate attention dynamics, and compromise model behavior.