

AIML Threat Map v 1.7

Threats to AIML Models

- LLM04: Denial of Service
- LLM10: Model Theft
- ML05: Model Inversion Attack
- ML07: Transfer Learning Attack
- ML08: Model Skewing Attack
- ML10: Model Poisoning
- Inadequate AI Alignment
- Improper Error Handling
- Robust multi-prompt and multi-model attacks
- Traditional Attacks

Threat Using AIML Models

- LLM01: Prompt Injection
- LLM02: Insecure Output Handling
- LLM03: Trained Data Poisoning
- LLM05: Supply Chain Attack
- LLM06: Sensitive Information Disclosure
- LLM07: Insecure Plugin Design
- LLM08: Excessive Agency
- LLM09: Overreliance
- Poisoned Memory
- Indirect Prompt Injection
 - Passive
 - Active
 - User-driven Injection
 - Hidden Injection
 - Payload Splitting
- Fake Resources
- Copyright infringement
- Surveillance: model or advertisers tracking users
- Persuasion, Deception, Influence
- Anthropomorphism

Threat Not Understanding AIML Models

- Not understanding how generative AIML works
- Underestimating the complexity of AIML & the dynamic nature of GenAI
- Not coordinating AIML Teams, DevSecOps Teams, & Cybersecurity for development, patching, and incident response
- Forcing AIML management (maintenance & risk) into legacy process which don't fit its complexity, velocity, and mutable characteristics
- Choosing the wrong AI tool for the task
- Over relying on AI without a human in the loop
- Failing to keep models and data up to date
- Hidden Technical Debt

Threats from AIML Models

- Misidentification i.e. wrongful arrest
- False Information i.e. criminal offenses
- Misinformation influence i.e. elections
- Private Information used in training
- Tricky user acceptance (complex, long, legal language)
- Unclear data owner (video recording in checkout, recording bots in meetings)
- Disinformation campaigns
- Deep Fakes
 - Abused authentication
 - Synthetic or composite fakes
- Shallow Fake
 - Slightly altered fake image
- FraudCPT
- DarkBARD
- DarkCPT
- Attack Acceleration
 - PoisonCPT
 - DarkBERT
 - DAN 9.0
 - ChaosCPT
- Hallucination Squatting
- Artificial Consciousness
- Honey or Poisoned Characters
- Social Influence: Persuasion, Deception, Influence
- AI Inhuman cognition capabilities

Threats NOT using AIML Models

- Competitive Disadvantage
- Limited Customer Engagement: Inability to scale personalized communication
- Innovation Stagnation: Slower pace of innovation and improvements
- Operational Inefficiency: Slower and less efficient processes
- Market Perception: Viewed as outdated by customers and partners
- Higher risk of human error in processes
- Inefficient allocation of human resources

AIML Legal & Regulatory Threats



Author: Sandy Dunn
sandy@quarkiq.com v1.7 422024