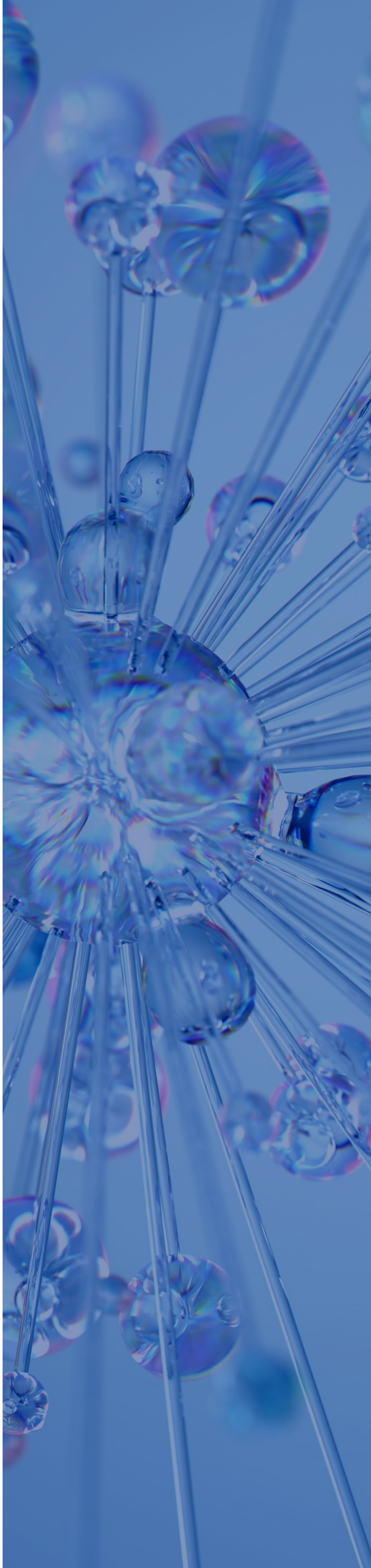
AQ SECURE

PROTECTING AI FROM ADVERSARIAL ATTACKS



ABSTRACT

As artificial intelligence (AI) technologies and particularly Generative AI (Gen AI) continue to evolve, so do the threats they face. Adversarial attacks, which pose significant risks to AI models, can result in data leakage, model theft, and manipulation. This whitepaper explores the necessity of protecting both traditional and generative AI models from various AI adversarial threats including model inference, extraction, evasion and injection, data poisoning, prompt injections and personal identifiable information (PII) leakage. We outline the emerging risks, the methodologies of these attacks, and propose a robust framework for mitigating these threats to maintain the security and integrity of AI systems.

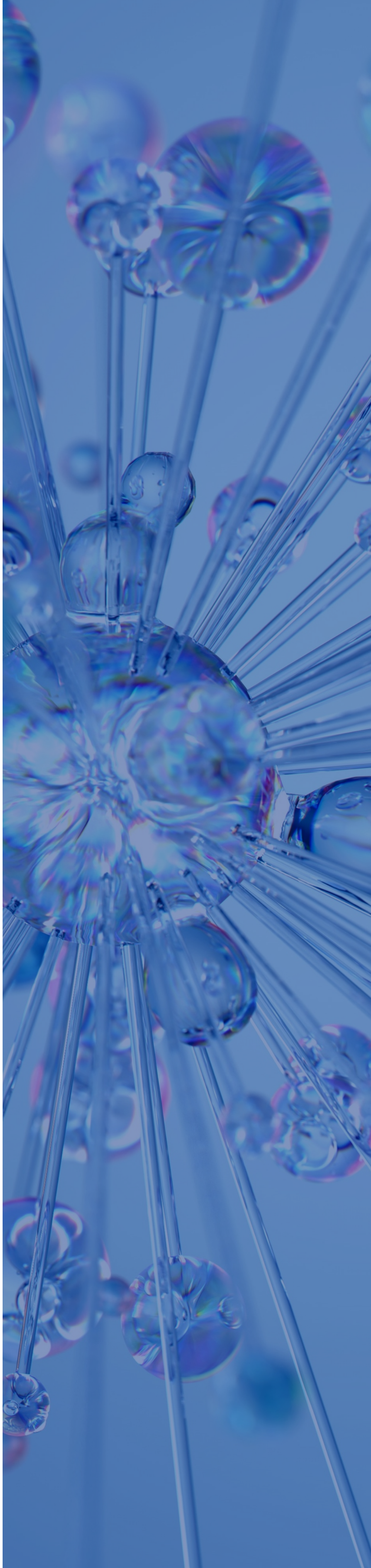
INTRODUCTION

As AI models become increasingly integral to businesses and society, adversarial attacks targeting these models present a growing challenge. The potential fallout from these attacks includes unauthorized access to sensitive data, compromised model outputs, and significant financial and reputational damage. Both traditional AI models and those leveraging generative capabilities are vulnerable to diverse attack vectors, necessitating a comprehensive understanding of the risks and appropriate countermeasures.

UNDERSTANDING ADVERSARIAL ATTACKS

Types of Adversarial Attacks:

- **Model Inference/Extraction:** These attacks involve an attacker manipulating model inputs, analysing outputs, and inferring decision boundaries to reconstruct the training data, extract model parameters, or model theft by training a substitute that approximates the target.
- **Model evasion:** Model evasion occurs when an attacker manipulates the input in a specific way to bypass the correct classification or induce a particular desired classification.
- **Model Injection:** Model injection is a technique that relies on altering the machine learning model by inserting a malicious module that introduces some secret, harmful, or unwanted behaviour.
- **Data Poisoning:** Data poisoning occurs when an attacker injects training sets with new, specifically altered data designed to fool or subvert a machine learning model to provide.
- **Prompt Injection:** In generative models, adversaries may craft prompts that lead the model to yield unwanted or harmful responses, steering the output in unintended directions.
- **PII Leakage:** AI models trained on datasets containing personal identifiable information (PII) can inadvertently reveal this sensitive information through queries or outputs.



THE NEED FOR PROTECTION

The growing sophistication of adversarial attacks necessitates a proactive, multi-layered approach to protect AI models. The implications of successful attacks are far-reaching, including:

- **Reputational Damage:** Organizations can suffer loss of trust and credibility in the market.
- **Financial Loss:** Costs associated with data breaches, regulatory fines, and mitigation efforts can escalate quickly.
- **Legal Consequences:** Breaches involving PII can lead to significant legal repercussions under regulations such as GDPR and CCPA.

MITIGATING RISKS

Framework for AI Protection

To safeguard AI models against adversarial attacks, organizations can implement the following strategies:

- **Robust Model Protection Mechanisms:** Deploying advanced cybersecurity tools and techniques to deter and detect unauthorized access, extraction, and tampering of the AI models.
- **Rigorous Data Governance and Privacy Controls:** Implementing comprehensive data management processes, including data anonymization, differential privacy, and access control mechanisms, to minimize the risk of PII leakage.
- **Continuous Monitoring and Anomaly Detection:** Leveraging advanced analytics and monitoring tools to identify and respond to suspicious activities, such as data poisoning attempts or prompt-based attacks.
- **Comprehensive Employee Training and Security Awareness:** Educating employees on the importance of cybersecurity best practices, the recognition of potential threats, and their role in safeguarding AI systems.
- **Collaboration with Security Experts:** Establish partnerships with cybersecurity professionals who specialize in AI to keep abreast of emerging threats and solutions.

CONCLUSION

The advancement of artificial intelligence, particularly generative models, brings unprecedented opportunities but also poses significant risks. Protecting these models from adversarial attacks is paramount to safeguarding sensitive data, maintaining public trust, and ensuring regulatory compliance. Organizations must adopt a comprehensive strategy tailored to the unique vulnerabilities of AI systems, combining robust design practices with proactive security measures to maintain the integrity of their AI initiatives. Only through diligence and informed practices can organizations realize the full potential of AI while mitigating associated risks.