OpenAI

Securing Large Language Model Development and Deployment

Navigating the Complexities of LLM Secure Development Practices to Align with the NIST Secure Development Framework

Prepare the Organization

Threat Modeling & Risk Management for LLMs

LLMs need an updated threat model

- Threat Actors
- Asset Protection
- Dual Use Foundation Model Risks
- Frontier Model Risks

Develop risk mitigation, technical controls, testing, and training for every identified group and projects in the LLM lifecycle. Educate the organization.

Protect the Software

LLM Assets

Model Weights

1 Model weights are the result of LLM training and steep computational investments.

Algorithms

2 Training requires code that instructs infrastructure how to do it.

Training Data

3 Foundation models increasingly leverage licensed data or have some form of proprietary treatment.

Compute

4 Compute is expensive and limited in the supply chain. Organizations can't just invest in more.

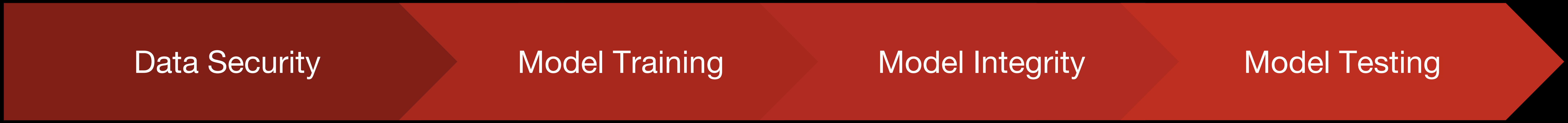
Prepare the Organization

Threat Model & Risk Considerations

1. Data Privacy and Confidentiality
2. Model Integrity and Security
3. Adversarial Attacks
4. Misuse of the Model
5. Regulatory Compliance
6. Dependency on External Data Sources
7. Scalability and Performance Security
8. Insider Threats
9. Intellectual Property Protection
10. Disinformation and Information Integrity

Protect the Software

LLM Development Secure Practices



Data Security

- Secure data sourcing
- Secure data handling
- Anonymization and privacy concerns

Model Training

- Secure environments
- Strong access controls
- Mitigation of documented risks and threats

Model Integrity

- Integrity of parameters
- Unauthorized modification prevention

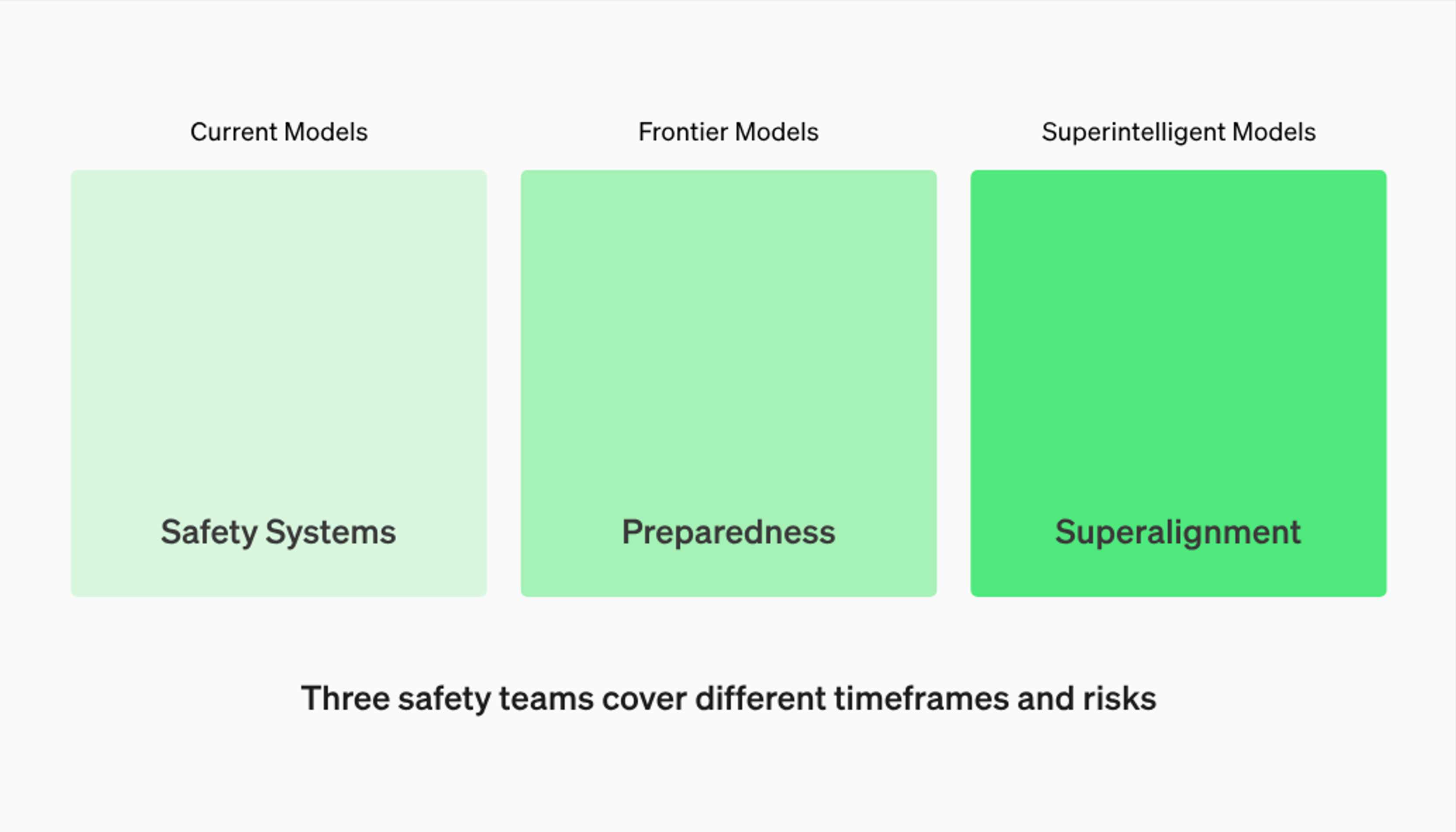
Model Testing

- Evaluations
- Monitored misuse
- Red teaming

Protect the Software

LLM Dual Use & Frontier Models Safety and Testing

1. Robust Safety Models
2. Identifying Unknown Risks
3. Collaboration with External Parties
4. Documented Secure Processes for known and unknown Risks
5. Tested Secure Processes



Current Models — Safety Systems

Frontier Models — Preparedness

Superintelligent Models — Superalignment

Three safety teams cover different timeframes and risks

Produce Well-Secured Software

Unique LLM Deployment Security Practices

1. Industry Leading Cloud Security
   a. Defense in Depth
   b. Monitoring / Alerting / Incident Management
   c. Vulnerability Management & Response
   d. Security aligned with Risks
2. LLM Specific Practices
   a. Increased Security Controls around LLM Assets
   b. Content Detection & Moderation
   c. Model Behavior Monitoring
   d. Procedures for ensuring Safety

Respond to Vulnerabilities

Continued Threat Modeling & Risk Management for LLMs

LLMs present unique risks and will require continued threat modeling and evolving risk management. This process doesn’t end at deployment of the models.