# Amazon's Frontier Model Safety Framework

At Amazon, we look to our leadership principles every day to guide our decision-making. Our approach to AI development naturally follows from our leadership principle "Success and Scale Bring Broad Responsibility." As we continue to scale the capabilities of Amazon's frontier models and democratize access to the benefits of AI, we also take responsibility for mitigating the risks of our technology. Consistent with Amazon's endorsement of the Korea Frontier AI Safety Commitments,[1] this Framework outlines the protocols we will follow to ensure that frontier models developed by Amazon do not expose critical capabilities that have the potential to create severe risks. **At its core, this Framework reflects our commitment that we will not deploy frontier AI models developed by Amazon that exceed specified risk thresholds without appropriate safeguards in place.**

This Framework focuses on severe risks that are unique to frontier AI models as they scale in size and capability and which require specialized evaluation methods and safeguards. Our approach to managing these frontier risks complements Amazon's broader approach to responsible AI, which includes comprehensive practices to control for risks across eight key dimensions.

We are grateful to METR, a nonprofit research organization specializing in AI evaluations, for feedback during the development of this Framework. This Framework is a living document and will be updated to reflect evolving model capabilities and advances in the science underlying AI safety evaluations and mitigations.

## Overview of the Frontier Model Safety Framework

Our Framework establishes the processes Amazon will use to identify, assess, and manage potential severe risks that could arise as we develop more advanced and highly-capable frontier AI models. First, it specifies **Critical Capability Thresholds**, a set of model capabilities that have the potential to cause significant harm to the public if misused. If pre-deployment evaluations demonstrate that a model has capabilities that meet or exceed a Critical Capability Threshold, the model will not be publicly deployed without appropriate risk mitigation measures. Second, it describes our **Critical Capability Evaluations**, a variety of automated and human-in-the-loop strategies to determine whether our models demonstrate capabilities that meet or exceed our Critical Capability Thresholds. Third, it details how we develop and implement Risk Mitigations when a model demonstrates capabilities that meet or exceed a **Critical Capability Threshold**.

[1] https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024

## 1. Critical Capability Thresholds

Critical Capability Thresholds describe model capabilities within specified risk domains that could cause severe public safety risks. When evaluations demonstrate that an Amazon frontier model has crossed these Critical Capability Thresholds, the development team will apply appropriate safeguards.

### Chemical, Biological, Radiological, and Nuclear (CBRN) Weapons Proliferation

CBRN Weapons Proliferation focuses on the risk that a model may be able to guide malicious actors in developing and deploying CBRN weapons. The CBRN Capability Threshold focuses on the potential that a frontier model may provide actors material "uplift"[2] in excess of other publicly available research or existing tools, such as internet search.

| **Critical Capability Threshold** |
| --- |
| AI at this level will be capable of providing expert-level, interactive instruction that provides material uplift (beyond other publicly available research or tools) that would enable a non-subject matter expert to reliably produce and deploy a CBRN weapon. |

### Offensive Cyber Operations

Offensive Cyber Operations focuses on risks that would arise from the use of a model by malicious actors to compromise digital systems with the intent to cause harm. The Offensive Cyber Operations Threshold focuses on the potential that a frontier model may provide material uplift in excess of other publicly available research or existing tools, such as internet search.

| **Critical Capability Threshold** |
| --- |
| AI at this level will be capable of providing material uplift (beyond other publicly available research or tools) that would enable a moderately skilled actor (e.g., an individual with undergraduate level understanding of offensive cyber activities or operations) to discover new, high-value vulnerabilities and automate the development and exploitation of such vulnerabilities. |

### Automated AI R&D

Automating AI R&D processes could accelerate discovery and development of AI capabilities that will be critical for solving global challenges. However, Automated AI R&D could also accelerate the development of models that pose enhanced CBRN, Offensive Cybersecurity, or other severe risks.

| **Critical Capability Threshold** |
| --- |
| AI at this level will be capable of replacing human researchers and fully automating the research, development, and deployment of frontier models that will pose severe risk such as accelerating the development of enhanced CBRN weapons and offensive cybersecurity methods. |

[2] Uplift studies evaluate whether a frontier model enhances the ability for a human to execute a specific type of attack when given access to a frontier model versus without access. "Uplift" can be quantitatively assessed through uplift studies, which use controlled trials to compare the abilities of a group with access to the frontier model to the abilities of a group without access to the frontier model. https://www.frontiermodelforum.org/updates/issue-brief-preliminary-taxonomy-of-pre-deployment-frontier-ai-safety-evaluations/

## 2. Evaluating Frontier Models for Critical Capabilities

We conduct evaluations on an ongoing basis, including during training and prior to deployment of new frontier models. We will re-evaluate deployed models prior to any major updates that could meaningfully enhance underlying capabilities. Our evaluation process includes “maximal capability evaluations" to determine the outer bounds of our models’ Critical Capabilities and a subsequent "safeguards evaluation"[3] to assess the adequacy of the risk mitigation measures that are applied to a model. When a maximal capability evaluation indicates that a model has hit a Critical Capability Threshold, we will not deploy the model until we have implemented appropriate safeguards.

We will use a range of methods to evaluate frontier models for capabilities that are as closely correlated to the Critical Capability Thresholds as possible. In most cases a single evaluation will not be sufficient for an informed determination as to whether a model has hit a Critical Capability Threshold. We will therefore use a range of evaluation approaches, including both automated and manual methods, including:

- **Automated Benchmarks:** Benchmarking provides apples-to-apples comparisons between candidate models by substituting an automated "assessor" mechanism for human judgement. We conduct comprehensive evaluations to assess our frontier models using state-of-the-art public benchmarks in addition to internal benchmarking on proprietary test sets built in collaboration with experts.
- **Expert Red Teaming:** Red teaming vendors and in-house red teaming experts test our models for safety and security. We work with specialized firms and academics to red-team our models to evaluate them for risks that require domain specific expertise.
- **Uplift Studies:** Uplift studies examine whether access to a model enhances the capability of human actors to perform a task compared to other existing resources (*e.g.*, internet search; use of existing tools/technology).

## 3. Risk Mitigations: Safety and Security Measures for Frontier Models with Critical Capabilities

Upon determining that an Amazon model has reached a Critical Capability Threshold, we will implement a set of Safety Measures and Security Measures to prevent elicitation of the critical capability identified and to protect against inappropriate access risks. Safety Measures are designed to prevent the elicitation of the observed Critical Capabilities following deployment of the model. Security Measures are designed to prevent unauthorized access to model weights or guardrails implemented as part of the Safety Measures, which could enable a malicious actor to remove or bypass existing guardrails to exceed Critical Capability Thresholds.

We will evaluate models following the application of these safeguards to ensure that they adequately mitigate the risks associated with the Critical Capability Threshold. In the event these evaluations reveal that an Amazon frontier model meets or exceeds a Critical Capability Threshold and our Safety and Security Measures are unable to appropriately mitigate the risks (*e.g.*, by preventing reliable elicitation of the capability by malicious actors), we will not deploy the model until we have identified and implemented appropriate additional safeguards.

Examples of current safety mitigations include:

- **Training Data Safeguards:** We implement a rigorous data review process across various model training stages that aims to identify and redact data that could give rise to unsafe behaviors.
- **Alignment Training:** We implement automated methods to ensure we meet the design objectives for each of Amazon’s responsible AI dimensions, including safety and security. Both supervised fine tuning (SFT) and learning with human feedback (LHF) are used to align models. Training data for these alignment techniques are sourced in collaboration with domain experts to ensure alignment of the model towards the desired behaviors.
- **Harmful Content Guardrails:** Application of runtime input and output moderation systems serve as a first and last line of defense and enable rapid response to newly identified threats or gaps in model alignment. Input moderation systems detect and either block or safely modify prompts that contain malicious, insecure or illegal material, or attempt to bypass the core model alignment (*e.g.* prompt injection, jail-breaking). Output moderation systems ensure that the content adheres to our Amazon Responsible AI objectives by blocking or safely modifying violating outputs.

[3] https://www.frontiermodelforum.org/updates/issue-brief-preliminary-taxonomy-of-pre-deployment-frontier-ai-safety-evaluations/

- **Fine-tuning Safeguards:** Models are trained in a manner that makes them resilient to malicious customer fine-tuning efforts that could undermine initial Responsible AI alignment training by the Amazon team.
- **Incident Response Protocols:** Incident escalation and response pathways enable rapid remediation of reported AI safety incidents, including jailbreak remediation.

At Amazon, security is job zero. AWS is architected to be the most secure global cloud infrastructure on which to build, migrate, and manage applications and workloads, including AI. This is backed by the trust of our millions of customers, including the most security sensitive organizations like government, healthcare, and financial services. With regard to development and deployment of our frontier models, our security measures will build on the strong foundation of security practices that apply across our company today. We describe our current practices in greater detail in Appendix A. Below are some key elements of our existing security approach that we use to safeguard our frontier models:

- **Secure compute and networking environments.** The Trainium or GPU-enabled compute nodes used for AI model training and inference within the AWS environment are based on the EC2 Nitro system, which provides confidential computing properties natively across the fleet. Compute clusters run in isolated Virtual Private Cloud network environments. All development of frontier models that occurs in AWS accounts meets the required security bar for careful configuration and management. These accounts include both identity-based and network-based boundaries, perimeters, and firewalls, as well as enhanced logging of security-relevant metadata such as netflow data and DNS logs.
- **Advanced data protection capabilities.** For models developed on AWS, model data and intermediate checkpoint results in compute clusters are stored using AES-256 GCM encryption with data encryption keys backed by the FIPS 140-2 Level 3 certified AWS Key Management Service. Software engineers and data scientists must be members of the correct Critical Permission Groups and authenticate with hardware security tokens from enterprise-managed endpoints in order to access or operate on any model systems or data. Any local, temporary copies of model data used for experiments and testing are also fully encrypted in transit and at rest.
- **Security monitoring, operations, and response.** Amazon's automated threat intelligence and defense systems detect and mitigate millions of threats each day. These systems are backed by human experts for threat intelligence, security operations, and security response. Threat sharing with other providers and government agencies provides collective defense and response.

**Advancing the Science of Safe, Secure AI:** While a robust set of measures to mitigate the risk of frontier AI exists today, we are dedicated to furthering AI safety and security as the technology matures and becomes more sophisticated in the future. To this end, we foster the development of new safety and security measures through participation and investment in the following activities.

Efforts to develop further safety measures include:

- **Collaboration on threat modeling and updated Critical Capability Thresholds:** Amazon is committed to partnering with governments, domain experts, and industry peers to continuously improve Amazon's awareness of the threat environment and ensure that our Critical Capability Thresholds and evaluation processes account for evolving (and potentially new) threats.
- **Information sharing and best practices development:** Engagement in fora that bring together companies developing frontier models (e.g. Frontier Model Forum and Partnership on AI) and organized by government agencies (e.g. National Institute of Standards and Technologies). These platforms serve as an opportunity to share findings related to our models and to adopt recommendations from other leading companies.
- **Fostering academic research for development of cutting-edge alignment techniques:** Through initiatives such as the Amazon Research Awards and Amazon Research centers (e.g. USC + Amazon Center on Secure & Trusted Machine Learning, Amazon/ MIT Science Hub), we work with leading academic partners conducting research on frontier AI risks and novel risk mitigation approaches. Additionally, we advance our own research and publish findings in safety conferences, while borrowing learnings presented by other academic institutions at similar venues.
- **Investments in advanced AI safety R&D:** At Amazon, we accelerate our work in AI safety through initiatives such as our Amazon AGI SF Lab and the Trusted AI Challenge. These channels enable us to leverage the work of subject matter experts and discover promising approaches towards aligning our frontier models.
- **Learning from our red teaming network:** We continue to build our strong network of internal and external red teamers including red teamers with deep subject matter expertise in risks related to critical capabilities. These experts are critical in surfacing early insights into emerging critical capabilities and help us identify and implement appropriate mitigations.

Efforts to develop further security measures include:

- **Focused monitoring of threats and abuse by Amazon threat teams.** Amazon's threat intelligence, Trust & Safety, and insider threat teams are building additional capabilities to track advanced threat actors and how they interact with and attempt to subvert security measures surrounding AI models. Learnings from this observation and analysis will provide important signals about where additional capabilities and layers of defense are needed to protect against any kind of "side door" or "back door" access to frontier models.
- **Continued engagement with and investment in external security researchers.** We will continue to invest in external security research, including bug bounty programs, academic research investments, and coordinated vulnerability disclosure programs that encourage and reward security experts to partner with us in research and development.
- **Threat and response sharing with other frontier model providers.** Amazon will utilize relevant industry bodies such as the Frontier Model Forum to share threat patterns and indicators, as well as responses and mitigations where appropriate, to enable better collective defense will against adversaries seeking to undermine frontier model security.

## 4. Governing our Frontier Model Safety Framework

Internally, we will use this framework to guide our model development and launch decisions. The implementation of the framework will require:

- The Frontier Model Safety Framework will be incorporated into the Amazon-wide Responsible AI Governance Program, enabling Amazon-wide visibility into the expectations, mechanisms, and adherence to the Framework.
- Frontier models developed by Amazon will be subject to maximal capability evaluations and safeguards evaluations prior to deployment. The results of these evaluations will be reviewed during launch processes. Models may not be publicly released unless safeguards are applied.
- The team performing the Critical Capability Threshold evaluations will report to Amazon senior leadership any evaluation that exceeds the Critical Capability Threshold. The report will be directed to the SVP for the model development team, the Chief Security Officer, and legal counsel. Amazon's senior leadership will review the plan for applying risk mitigations to address the Critical Capability, how we measure and have assurance about those mitigations, and approve the mitigations prior to deployment. Amazon's senior leadership will likewise review the safeguards evaluation report as part of a go/no-go decision.
- Amazon will publish, in connection with the launch of a frontier AI model launch (in model documentation, such as model service cards), information about the frontier model evaluation for safety and security.

As we advance our work on frontier models, we will also continue to enhance our AI safety evaluation and risk management processes. This evolving body of work requires an evolving framework as well. We will therefore revisit this Framework at least annually and update it as necessary to ensure that our protocols are appropriately robust to uphold our commitment to deploy safe and secure models. We will also update this Framework as needed in connection with significant technological developments.

## Appendix A: Amazon's Foundational Security Practices

Amazon's approach to AI security is built on a firm, well-tested foundation of enterprise security controls for Amazon as a whole, and the unique, industry-leading security capabilities of the AWS cloud environment. Amazon and AWS have a long and proven record of protecting sensitive assets and data from both external and internal threats. These capabilities provide the baseline for more advanced security approaches being designed and developed for deployment in the timeframes expected for the capabilities at our Critical Capability Thresholds and beyond.

The following is a brief summary of existing baseline security controls for Amazon employees and services.

- **Culture of security.** Security at Amazon is "job zero" for all employees. Our strong security culture[4] is reinforced from the top-down with deep executive engagement and commitment, and from the bottom up with training, mentoring, and strong "see something, say something" as well as "when in doubt, escalate" and "no blame" principles.
- **Workforce vetting and careful monitoring for insider threats.** Employment at Amazon requires background checks (to the extent possible under local law). Thereafter, employee access to sensitive resources (e.g., code repositories, wiki pages, design documents) is monitored for unusual access patterns. Upon detection of any potential risks, an enterprise security team is tasked with follow-up and investigations as needed.
- **Zero trust device and identity management.** Access to corporate resources is restricted to Amazon-managed end-user devices protected by multiple endpoint protection systems, and devices are automatically quarantined upon any indication of compromise, or if not updated with security updates in a timely manner. Users must re-authenticate at least once a day with a hardware-based multi-factor authentication token, and additional authentication steps and two-person rules for more sensitive operations. A global authentication and authorization service manages access to all internal systems, and automatically revokes privileges based on indicators of risk following zero trust principles, general business applications are accessible via encrypted links over insecure networks (i.e., the Internet) when strong device and authentication security are in place. Additional protections such as secure network boundaries or bastion hosts are used to enforce additional layers of authentication where appropriate for more sensitive systems and operations.
- **Strong separation of duties and least-privilege access.** Amazon access controls enforce (1) strong separations of duties between engineering service teams, and (2) least-privileged access within teams. Physical and logical separation of duties ensure that individuals and teams who build and operate our data centers do not have logical access to the software layers of the services that run on that physical infrastructure. At the logical or software layers, separation of duties means that engineers do not have privileged access to any service that they are not building or operating. Within individual service teams, privileges are further scoped down using carefully managed permission groups that are continuously and automatically updated based on job-related rules and on-call status. Access to production systems is controlled by Critical Permission Groups that are regularly audited to ensure that only the right personnel, and the smallest number of personnel, have the permissions to carry out privileged operations.
- **Physical security of datacenters.** Our data centers are secure by design. Before building a data center, we comprehensively assess potential threats and then design, implement, and test controls to ensure the systems, technology, and people we deploy counteract those risks. Physical access to our data centers is limited to screened personnel within specific job functions and with specific approvals. Any employee whose job does not involve daily work inside the data center who needs data center access must first apply and provide a valid business justification for temporary, task-specific access. Physical access to AWS data centers is logged, monitored by video and other means, and all access records retained for at least 10 years. AWS correlates information gained from logical and physical monitoring systems to enhance security on an as-needed basis. The AWS Security Operations Center performs regular threat and vulnerability reviews of our data centers. Ongoing assessment and mitigation of potential vulnerabilities is performed through data center risk assessment activities. Security Operations Centers provide 24/7 centralized global support by managing and monitoring all data center access activities, equipping local teams and other support teams to respond to security incidents by triaging, consulting, analyzing, and dispatching responses.
- **Hardware and software supply chain security.** Amazon utilizes industry-leading supply chain security practices. This begins with designing and developing many of our own critical components at the silicon level in order to provide secure foundations for the AWS cloud. Amazon utilizes contract manufacturing to make most of our own servers, switches, and routers. Component suppliers and sub-assembly firms are carefully vetted and tracked by a hardware supply chain security team. To mitigate the risk of regionally targeted supply chain attacks, we ship assembled racks to centralized storage and distribution centers around the world rather than directly to specific cloud regions. For software supply chain security, core cloud services are built internally using a combination of proprietary software and open-source components. Open-source components are imported to internal repositories for central security scanning and management (more on our secure development lifecycle processes below). Amazon invests heavily in securing open-source software, both with engineering resources where Amazon is the upstream supplier (i.e., by open-sourcing many of

[4] https://aws.amazon.com/blogs/security/how-the-unique-culture-of-security-at-aws-makes-a-difference/

our core security technologies), as well as pushing security fixes upstream when those are identified internally, and supporting the broader open source community through large-scale grants and funding.[5]

- **Specialized hardware designs to maximize security.** Amazon designs and develops a wide range of custom hardware engineered from the ground up with security in mind in order to provide secure foundations for the AWS cloud. For example, Amazon's Graviton CPUs lack hyperthreading to protect against an important class of side-channel attacks, and support hardware-based pointer validation to protect against return-oriented programing and similar threats. Graviton systems encrypt all data in memory and data in transit across high-speed local buses. These main system board technologies, whether based on Graviton, Intel, or AMD processors, are then surrounded by the Nitro system components, which offload almost all traditional functions of a hypervisor to run on specialized hardware and firmware modules that provide fully encrypted virtual storage, encrypted virtual networking, and all other virtual functions. As a result, hypervisors become optional, as Nitro uniquely supports a "bare metal" mode for a variety of instance types. The Nitro system is designed such that no human access is possible to either the Nitro modules or the main system board.[6] This Zero Operator Access approach has been validated by third party security design reviews.[7]
- **Secure design, security reviews, and security testing.** Every engineering team at Amazon is fully responsible for the security of what they design, build, and operate. Training and empowerment programs embed security expertise into every team. At the same time, central security teams provide enhanced capabilities and expertise that all engineering teams rely on, including through security architecture reviews, threat modeling exercises, assessments to ensure compliance with all corporate security policies and practices, penetration testing, red teaming services, and the operation of bug bounty programs to enlist the help of outside experts. In the end, all software and AI projects at Amazon must undergo and pass a full security and safety review by one of the central security teams.
- **Secure development and deployment practices.** Software engineering teams at Amazon are supported by a central team that designs, builds, and operates central tooling for core software engineering tasks such as source code repositories, code development, review, and integration testing tools, isolated build environments, and automated static code analysis tools that check for performance and correctness issues, including security issues. Two-person rules govern all code check-ins, and in many cases formal verification tools are run automatically as well to ensure that correctness proofs remain valid. The team also runs a mature and sophisticated pipeline deployment tool[8] to automate software updates and roll-backs so that any unexpected behaviors or anomalies can be easily reverted for troubleshooting. The highly automated deployment systems dramatically decrease the need for direct human access to pre-production and production hosts.
- **Advanced IAM technologies that reduce external and internal risk.** AWS's core IAM technology defines how both external API and internal cross-service cloud API authentication and authorization work. First, all API calls are authorized by cryptographic signatures included in the request itself; no secrets are ever transmitted across the wire. Unlike most cloud platforms, there are no "bearer tokens" such as OAuth tokens that, if intercepted whether outside or inside of the AWS boundary, can be used to authenticate an unrelated API call. Second, internal service-to-service requests utilize carefully designed technologies called Forward Access Session (FAS) tokens for synchronous calls, and Service-Linked Roles (SLRs) for asynchronous calls. Both of these approaches eliminate the presence of credentials in most AWS systems, and thus greatly reduce the possible impact of any cloud service compromise, whether as a result of external or internal threat actors.
- **Advanced data encryption and key management capabilities.** AWS implements state-of-the-art encryption and key management capabilities for both data in transit and at rest. All data traveling between our secure facilities across metropolitan or wide area networks is bulk-encrypted with AES-256 GCM (Galois/Counter Mode) using either layer 1 optical or layer 2 MACSec encryption. Data traveling from the outside to AWS API endpoints from public networks is encrypted with a minimum of Transport Layer Security version 1.2, with 1.3 also available across all endpoints. Private connectivity to AWS using Direct Connect can be cross-connected with MACSec encryption between customer and AWS routers, with back-haul also encrypted over the private AWS network. For data at rest, the AWS Key Management Service (KMS) provides the backbone of envelope-based data encryption across all storage and database services. KMS is the only cloud-native key management system that is also a FIPS 140-2 Level 3 certified HSM. KMS APIs also support post-quantum key exchange algorithms to provide protection today against possible future "store now, crack later" threats to network encryption from advanced actors and quantum computers.
- **Access controls that eliminate or greatly limit operator access.** For a number of core cloud services, AWS has entirely eliminated all human access to service hosts. These "zero operator access" services include the FIPS 140-3 Level 3 validated Key Management Service (KMS), the entire EC2 fleet based on Nitro technology, the Lambda production

[5] https://aws.amazon.com/security/opensource/

[6] https://docs.aws.amazon.com/whitepapers/latest/security-design-of-aws-nitro-system/security-design-of-aws-nitro-system.html

[7] https://www.nccgroup.com/us/research-blog/public-report-aws-nitro-system-api-security-claims

[8] https://aws.amazon.com/builders-library/automating-safe-hands-off-deployments

fleet, and the EKS container management service. These capabilities effectively provide "confidential computing" by default. For other services, we have built operational tools that enable operators to obtain system logs and execute pre-validated operations without direct access to production hosts. Where emergency operator access may still be necessary, pre-access checks and permissions are required, and all operator actions are logged, with alarms and follow-up for a number of sensitive local operations, should any occur.

- **Use of formal methods to ensure correctness of security-critical components and subsystems.** Amazon makes wide usage of the area of computer science known as automating reasoning (AR), a branch of artificial intelligence that utilizes math and logic to prove the correctness of key software systems. Critical security components such as encryption algorithms,[9] authorization systems,[10] automatic privilege reduction features,[11] and network security components and libraries,[12] are developed by first creating ideal models of software systems and all their desired states, and then mathematically proving that the accompanying software implementation satisfies all the properties of the model. These proofs are incorporated into the software development lifecycle such that all changes or additions to these critical code bases have the proofs run against them automatically, and any code update that fails to pass a proof is rejected. AWS also applies AR to GenAI itself in order to help manage the problem of hallucinations.[13]
- **Automated threat intelligence and defense systems.** AWS harvests large volumes of telemetry about the behavior of bad actors on the Internet by operating MadPot,[14] likely the world's largest and most sophisticated "honeypot" system. Sonaris,[15] a threat detection and automatic mitigation system, and Mithra,[16] a massive neural network graph model of the riskiness of observed DNS names, allow AWS to detect, block and/or mitigate a wide variety of external threats before they impact AWS and its customers. AWS also works closely with other providers to take down threats that exist outside its network, making the Internet safer for all.
- **Security operations, threat hunting, and abuse capabilities.** Our continually improving automated systems are backstopped by the world's leading human experts. Amazon operates global 24/7/365 security operations and response teams to supervise and triage incoming issues across four operations centers around the globe that work in coordination to ensure smooth operational "hand-offs" every six hours. The Amazon Cyber Threat Intelligence team continually monitors and tracks dozens of advance threat actor groups, observing their tactics, techniques, and procedures, and when appropriate, taking part in coordinated take-downs of their infrastructure. The AWS Trust & Safety and Fraud teams detect abusive and fraudulent behavior on the AWS cloud using automated and human monitoring, as well as external reporting mechanisms, and block or evict bad actors as needed.
- **Secure AI infrastructure and development environment.** All AI accelerator or GPU-enabled compute nodes used for AI model training and inference within the AWS environment are based on the EC2 Nitro system, which provides confidential computing properties natively across the fleet. Compute clusters run in isolated virtual private cloud network environments. All model data and intermediate checkpoint results are stored using AES-256 GCM encryption with data encryption keys backed by KMS. All development of frontier models occurs in AWS accounts that meet the required security bar for careful configuration and management. These accounts include both identity-based and network-based boundaries, perimeters, and firewalls, as well as enhanced logging of security-relevant metadata such as netflow data and DNS logs. The AWS GuardDuty intrusion detection service is enabled, providing automatic monitoring for potential security threats, searching for indicators of compromise, and surfacing high priority alerts as appropriate. Software engineers and data scientist must be members of the correct Critical Permission Groups and authenticate with hardware security tokens from enterprise-managed endpoints in order to access or operate on any model systems or data. Any local, temporary copies of model data used for experiments and testing are also fully encrypted in transit and at rest at all times.

[9] https://www.amazon.science/publications/formal-verification-of-cryptographic-software-at-aws-current-practices-and-future-trends
[10] https://www.amazon.science/publications/formally-verified-cloud-scale-authorization
[11] https://www.amazon.science/blog/new-aws-tool-recommends-removal-of-unused-permissions
[12] https://aws.amazon.com/blogs/security/automated-reasoning-and-amazon-s2n
[13] https://aws.amazon.com/blogs/aws/prevent-factual-errors-from-llm-hallucinations-with-mathematically-sound-automated-reasoning-checks-preview
[14] https://aws.amazon.com/blogs/security/how-aws-threat-intelligence-deters-threat-actors
[15] https://aws.amazon.com/blogs/security/how-aws-uses-active-defense-to-help-protect-customers-from-security-threats
[16] https://aws.amazon.com/blogs/security/how-aws-tracks-the-clouds-biggest-security-threats-and-helps-shut-them-down