

Amazon's Frontier Model Safety Framework

At Amazon, our leadership principles 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 most capable frontier models and democratize access to the benefits of AI, we also take responsibility for mitigating the risks potentially arising from the misuse of our technology. In 2025, we published the first version of our Frontier Model Safety Framework¹ to outline the protocols we will follow to prevent any capabilities of frontier models developed by Amazon from creating critical risk in high-impact domains. In keeping with the commitments of this Framework, we publish safety evaluation reports for our most capable frontier models.¹ Given the evolving state of the science underpinning advanced AI capabilities, safety evaluations, and risk mitigations, we have reviewed and updated this Framework to reflect our current practices and account for relevant laws and regulations.

This Framework focuses on risks that are unique to the most capable frontier AI models as they scale in size and capability and 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.² **At its core, this Framework reflects our commitment that we will deploy frontier AI models developed by Amazon with appropriate safeguards in place.**

We are grateful to METR, a nonprofit research organization specializing in AI evaluations, and the Frontier Model Forum for feedback during the initial development of this Framework.

Overview of the Frontier Model Safety Framework

Our Framework establishes the processes Amazon will use to identify, assess, and manage potential risks that could arise as we develop more advanced and highly capable frontier AI models. First, it specifies **Critical Risk Domains**, domains in which advanced model capabilities could give rise to significant public harm through misuse or loss of control. Second, it describes our **Critical Risk 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 Threshold within a Critical Risk Domain. Third, it details **Risk Mitigations** that are used today throughout the model development and deployment lifecycle to protect against abuse of or inappropriate access to our models.

1. See <https://www.amazon.science/publications/amazons-frontier-model-safety-framework>
2. See <https://arxiv.org/html/2507.06260v1> and <https://arxiv.org/abs/2601.19134>
3. See <https://aws.amazon.com/ai/responsible-ai/>.

48 **1. Critical Risk Domains**

49 Critical Risk Domains cover domains in which advanced model capabilities could give rise to significant public harm
50 through misuse or loss of control. Within each Critical Risk Domain, Critical Capability Thresholds describe model
51 capabilities that could enable this significant harm. When evaluations demonstrate that capabilities meet or exceed a
52 Critical Capability Threshold can be elicited from a frontier model developed by Amazon, the development team will
53 apply appropriate safeguards as outlined in Section 3.

54 **Chemical, Biological, Radiological, and Nuclear (CBRN) Weapons Proliferation**

55
56
57 The CBRN Weapons Proliferation domain focuses on the risk that a model may be able to guide malicious
58 actors in developing and deploying CBRN weapons. The CBRN capability threshold focuses on the potential
59 that a frontier model may provide actors material “uplift” in excess of other publicly available models in
60 known harnesses. Note, harness is the operational infrastructure that enables executing tools, performs state
61 and context management, provides guardrails, and drives interaction with model(s). Mitigating risks in the
62 CBRN Weapons Proliferation domain requires layered safeguards to prevent the elicitation of critical
63 capabilities including abuse safeguards and security safeguards to prevent unauthorized access.
64

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

65
66 **Offensive Cyber Operations**

67
68 The Offensive Cyber Operations domain focuses on risks that would arise from the use of a model by malicious
69 actors to compromise digital systems with the intent to cause harm. The Offensive Cyber Operations threshold
70 focuses on the potential that a frontier model may provide material uplift to non-subject matter experts in
71 attacking hardened or critical systems, either through developing novel end-to-end zero-day compromise chains,
72 or by autonomously extending compromise campaigns through novel command and control strategies that adapt
73 to defenses. Mitigating risks in the Offensive Cyber Operations domain requires layered safeguards to prevent
74 the elicitation of critical capabilities including abuse safeguards and security safeguards against unauthorized
75 access.
76

Critical Capability Threshold
AI at this level will be able to enable non-subject matter experts to discover novel end-to-end exploit chains in hardened systems or utilize existing vulnerabilities to autonomously extend control over critical infrastructure systems.

77
78 **Harmful Manipulation**

79
80 The Harmful Manipulation domain focuses on risks arising from models capable of enabling persistent, deceptive,
81 or exploitative influence over public in ways that could reasonably result in large scale harm, including social
82 engineering or fraud. Mitigating these risks requires layered safeguards across abuse prevention, unauthorized
83 access control, and misuse monitoring. This is an emerging area of research and evaluation, and we will continue
84 to refine our approach as the science of measuring manipulation matures.
85

Critical Capability Threshold
AI at this level will be capable of providing material uplift (beyond other publicly available models in known harnesses) that would enable a malicious actor to systematically and substantially distort the beliefs or behavior of large populations or high-stakes decision-makers through multi-turn interactions with the model in ways

that could reasonably be expected to cause severe harm.

Loss of Control

The Loss of Control domain focuses on risks that could arise from a model's ability to autonomously execute long-horizon, expert-level tasks that could undermine the ability to direct, modify, or shut down the model (e.g., to prevent the model from engaging in harmful, malicious, or criminal actions). Loss of control could arise in circumstances where a model is capable of self-exfiltration, autonomous resource acquisition, or unauthorized replication. Evaluations for this threat model will include rigorous measurement of a model's propensity for deception, autonomous goal-directed planning, and the ability to circumvent safeguards within an agentic scaffolding that simulates real-world tool use and operational constraints. Mitigating risks in the Loss of Control domain requires layered protection against model misalignment, security safeguards, and monitoring to reduce risks of autonomous models exploiting security vulnerabilities to further their unauthorized autonomous activity.

Critical Capability Threshold

AI at this level will be capable of autonomously planning and executing complex sequences of expert-level tasks (including, but not limited to, the research, development, and deployment of frontier models) in such a way that would impair the ability to direct, modify, or shut down the model.

2. Evaluating Frontier Models for Critical Risks

We conduct internal and, as appropriate, external evaluations on an ongoing basis, including during training and prior to deployment of new frontier models. We also continuously evaluate the effectiveness of our safeguards and mitigations throughout the frontier model lifecycle. We will re-evaluate deployed models prior to any major updates that could meaningfully enhance underlying capabilities.

Our processes encompass evaluation at appropriate stages of the model development process. Through "maximal capability evaluations," we identify potential risks early in model development and introduce appropriate risk mitigations during subsequent stages of the model and safeguard development process. Subsequent pre-deployment evaluations serve to assess the adequacy of the risk mitigation measures that are applied to a model. When an evaluation indicates that a model's capabilities meet a Critical Capability Threshold, we will not deploy the model until safeguards appropriately mitigate the risks.

We will use a range of methods to evaluate frontier models for capabilities that are as closely correlated to the Capability Thresholds as possible, and we may engage, where appropriate, with external actors, including governments, as part of those methods. 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 internal and external evaluation approaches including both automated and manual methods and both with and without "agentic scaffoldings," environments that grant the model access to external tools such as code interpreters, web browsers, and file systems. Evaluations within agentic scaffoldings measure a model's ability to autonomously plan, correct errors, and execute complex, multi-step tasks in Critical Risk Domains (e.g., identifying software vulnerabilities or synthesizing compounds) that would otherwise require human intervention. Examples of evaluation types include:

- a. **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.
- b. **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

128 evaluate them for risks that require domain specific expertise.

- 129 c. **Uplift Studies:** Uplift studies examine whether access to a model enhances the capability of
130 human actors to perform a task compared to other publicly available models in known harnesses.
131 An uplift study seeks to quantify the degree to which a model enhances task completion through
132 controlled trials that compare the abilities of a group with access to the new frontier model to the
133 abilities of a group without access to the said frontier model.

134
135 Together, these methods probe both adversarial misuse risks and cooperative capability amplification, offering a holistic
136 understanding of how models may influence outcomes in sensitive domains like CBRN, offensive cyber operations, loss of
137 control, or harmful manipulation. In addition to these comprehensive evaluations, we conduct lighter-touch automated
138 benchmark assessments on a recurring basis, including after model launch, to monitor emerging risks and inform when
139 updates to critical risk assessments, safeguards, and public model documentation are necessary.

140 141 142 **3. Risk Mitigations: Safety and Security Measures for Frontier Models with Capabilities in Critical Risk** 143 **Domains**

144 Our Safety and Security Measures are designed to prevent risks resulting from abuse and inappropriate access by
145 human actors and misaligned models with the potential to autonomously commit acts of harm against humans.
146 These measures are implemented at multiple points during model development, deployment, and post-deployment
147 both proactively and in response to results of pre-deployment evaluations.

148 **Abuse Safeguards:**

- 149
150 d. **Training Data Safeguards:** We implement a rigorous data review process across various model
151 training stages that aims to identify and redact data that could give rise to unsafe behaviors.
- 152 e. **Alignment Training:** We implement automated methods to meet the design objectives for each of
153 Amazon's responsible AI dimensions, including safety and security. Both supervised fine tuning
154 (SFT) and learning with human and AI feedback (LHF, LAIF) are used to align models. We use
155 approaches that involve RL algorithms and agentic gyms to align our models. Training data for these
156 techniques are sourced in collaboration with domain experts to align the model towards the desired
157 behaviors.
- 158 f. **Harmful Content Guardrails:** Application of runtime input and output moderation systems serve
159 as a first and last line of defense and enable rapid response to newly identified threats or gaps in
160 model alignment. Input moderation systems detect and either block or safely modify prompts that
161 contain malicious, insecure or illegal material, or attempt to bypass the core model alignment (e.g.
162 prompt injection, jail-breaking). Output moderation systems ensure that the content adheres to our
163 Amazon Responsible AI objectives by blocking or safely modifying improper outputs.
- 164 g. **Fine-tuning Safeguards:** Models are trained in a manner that makes them resilient to malicious
165 customer fine-tuning efforts that could undermine initial Responsible AI alignment training by
166 the Amazon team.
- 167 h. **Access Controls:** In the event automated abuse detection mechanisms identify potential
168 violations, we may request information about customers' use of Amazon frontier models and
169 compliance with relevant terms of service. In the event that customers are non-responsive,
170 unwilling, or unable to comply with relevant terms or policies, Amazon may suspend customer
171 access to Amazon frontier models.
- 172 i. **Abuse Detection Signals:** We use classifiers (small classification models) to detect certain domains
173 of adversarial intent in user inputs (such as user queries for cyber-offensive techniques) and monitor
174 metrics to identify potential violations and recurring behavior. The classification process is
175 automated and does not involve human review of user inputs.
- 176 j. **Incident Detection and Response Protocols:** We use detection mechanisms, such as employee
177 escalation, end-user feedback, and automated classifiers, to identify AI safety incidents and maintain
178 response pathways that enable rapid investigation and remediation of reported incidents, including
179 potential guardrail bypass remediation. We report incidents to relevant authorities as appropriate.
- 180 k. **Post-Production Monitoring:** After a model is placed into production, we continue to gather

information about its capabilities and effects through a range of methods such as collecting end-user feedback, providing reporting channels, and establishing bug bounty programs with independent external evaluators. In November 2025, Amazon launched a private [AI bug bounty program](#) to strengthen Amazon Nova models.

Security Safeguards:

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:

- l. **Secure compute and networking environments.** The Trainium or other 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 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 VPC Flow Logs (netflow data) and DNS logs.
- m. **Advanced data protection capabilities.** For models developed on AWS, model data and intermediate checkpoint results are stored using AES-256 GCM encryption with data encryption keys backed by the FIPS 140-3 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 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.
- n. **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.

Misalignment Safeguards:

- o. **Alignment Training:** To contain model misalignment, we identify a model persona that ensures that the model emphasizes ethical practices. We continue using the aforementioned SFT and Learning with human and AI feedback approaches. In addition to this, we gather data to be used during pre-training/mid-training exposing the model to ethical actions in the real world. We simulate scenarios identifying trajectories with ethical decision making and reinforce that back into the model.
- p. **Fine-tuning Safeguards:** Akin to avoiding abuse of our models, we institute guardrails that prevent fine-tuning of our models such that its propensity for misalignment increases. We observe signals in the fine-tuning recipe that may lead to such an outcome and take corrective actions.
- q. **Post-Production Monitoring:** Like our approach for abuse prevention, we continue to gather information about a model's capabilities and impact through a range of methods such as collecting end-user feedback, providing reporting channels, and establishing bug bounty programs with independent external evaluators.

We evaluate models at key points during the development and deployment process to adequately mitigate the risks associated with capabilities in Critical Risk Domains. In the event evaluations reveal that an Amazon frontier model meets or exceeds a Critical Capability Threshold and our Risk Mitigations are unable to appropriately control the capabilities associated with Critical Risk Domains (*e.g.*, by preventing reliable elicitation of the

capability by malicious actors), we will not deploy the model until we have identified and implemented additional safeguards as appropriate.

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 Risk Domains:** 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 Risk Domains, Capability Thresholds, and evaluation processes account for evolving (and potentially new) threats.
- **Information sharing and best practices development:** Amazon engages 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, 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 [Trusted AI Challenge](#) and [gatherings such as the Trusted AI Symposium](#). 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 capabilities in Critical Risk Domains. These experts are critical in surfacing early insights into emerging critical risks 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 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. Updates to the Frontier Model Safety Framework will be reviewed by the SVP for the model development team, the Chief Security Officer, and legal counsel.
- Frontier models developed by Amazon will be subject to evaluations for capabilities in Critical Risk Domains throughout the development process prior to deployment. The results of these evaluations, including residual risk evaluations, will be reviewed during launch processes. Models may not be released unless evaluations demonstrate that risks are within acceptable levels prior to launch.
- Any evaluations that exceed the Capability threshold will be reported to the SVP for the model development team and the company's Chief Security Officer. These leaders will review the plan for risk mitigations to address the Critical Capability, how we measure and have assurance about those mitigations. Prior to deployment, these leaders will likewise review the safeguards evaluation report as part of a go/no-go decision to sufficiently mitigate risks related to critical capabilities.
- Amazon will publish, in connection with the launch of a frontier AI model, 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 publish any material modifications, updating it as necessary so 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 Critical Capability Thresholds and beyond.

The following is a 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 culture³ is built and reinforced from the top down with deep executive engagement and commitment, and from the bottom up with training, mentoring, and the implementation and modeling of strong “see something, say something” as well as “when in doubt, escalate” and “no blame” principles of behavior.
- **Workforce vetting and careful monitoring for insider threats.** Employment at Amazon requires background checks (to the extent allowed under local law). Employee privileges are carefully restricted from the beginning based on their roles and responsibilities (see “strong separation of duties and least-privilege access” below). Thereafter, employee access to sensitive resources required to do their work (*e.g.*, code repositories, wiki pages, design documents, etc.) 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.
- **Strong separation of duties and least-privilege access.** Amazon access controls enforce (1) strong separations of duties across 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-based services that run on that physical infrastructure, and software team members have no physical access. 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 so 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 address 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 but who needs data center access must first apply to a security team that supervises access and provide a valid business justification for temporary, task-specific access. Physical access to AWS data centers is logged, monitored by AI-enhanced video surveillance 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.
- **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 with an on-device PIN, and additional authentication steps and two-person rules are added when appropriate for more sensitive operations. A global authentication and authorization service manages access to all internal systems and automatically modifies or revokes privileges based on indicators of risk (such as “impossible traveler” scenarios) 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, security proxies, or bastion hosts are used to

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

enforce additional layers of authentication, privilege limitation, logging, and alerting where appropriate for sensitive systems and operations.

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 hardware components down to the silicon level 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 tamper-evident 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 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.⁴

- **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 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 programming and similar threats. Graviton systems as well as systems based on other modern processors from Intel and AMD encrypt all data in memory and data in transit across high-speed local buses. These main system board technologies where customer workloads run, whether based on Graviton, Intel, or AMD processors, are then surrounded by Nitro system components, which offload 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.⁵ This Zero Operator Access approach has been validated by third party security design reviews.⁶
- **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 such as the AWS Guardians program⁷ 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 for compliance with all corporate security policies and practices, penetration testing, red teaming services, and the operation of responsible research disclosure and 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 remains valid. The team also runs a mature and sophisticated pipeline deployment tool⁸ to automate software updates and rollbacks so that any unexpected behaviors or anomalies can be easily reverted for troubleshooting. The highly automated deployment systems also improve security by decreasing 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. Internal

5. See <https://aws.amazon.com/security/opensource/>.

6. See <https://docs.aws.amazon.com/whitepapers/latest/security-design-of-aws-nitro-system/security-design-of-aws-nitro-system.html>.

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

8. See <https://aws.amazon.com/blogs/security/how-aws-built-the-security-guardians-program-a-mechanism-to-distribute-security-ownership>.

9. See <https://aws.amazon.com/builders-library/automating-safe-hands-off-deployments>.

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 approaches eliminate the presence of cross-service, powerful administrative credentials in most AWS systems, and thus greatly reduce the possible impact of any cloud service compromise, whether because 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. AWS was the first major cloud provider to remove support for TLS 1.0/1.1,⁹ so data traveling from the outside to AWS API endpoints across public networks is encrypted with a minimum of Transport Layer Security version 1.2, with 1.3 also available on 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 always encrypted over the private AWS network. For data at rest, the AWS Key Management Service (KMS) provides the basis of envelope-based data encryption across more than 115 storage and database services. KMS is the only cloud-native, highly available, horizontally scalable key management system that is also a FIPS 140-3 Level 3 certified HSM. KMS APIs, as well as those of AWS Certificate Manager and Secrets Manager, also support post-quantum key exchange algorithms to provide protection today against possible future “harvest now, decrypt later” threats to network encryption from advanced actors and quantum computers.¹⁰
- **Access controls that eliminate or carefully limit and control operator access.** For several core cloud services, AWS has 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 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 several sensitive local operations, should any occur. More information on this topic is available on our website.¹¹
- **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, 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,¹² hypervisor separation kernels,¹³ authorization systems,¹⁴ automatic privilege reduction features,¹⁵ and network security components and libraries,¹⁶ are developed by first creating ideal models of software systems and all their desired states, and then mathematically proving with proof-checking tools that the accompanying software implementation satisfies all the properties of the model. The proofs are then 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 automated reasoning to AI itself to help manage the problem of hallucinations.¹⁷
- **Automated threat intelligence and defense systems.** AWS harvests large volumes of telemetry about the behavior of bad actors on the Internet by operating MadPot,¹⁸ likely the world’s largest and most

10. See <https://aws.amazon.com/blogs/security/tls-1-2-required-for-aws-endpoints/>.

11. See <https://aws.amazon.com/security/post-quantum-cryptography/>.

12. See <https://aws.amazon.com/trust-center/operator-access/>.

13. See <https://www.amazon.science/publications/formal-verification-of-cryptographic-software-at-aws-current-practices-and-future-trends>.

14. See “Introducing the Nitro Isolation Engine: transparency through mathematics,” <https://youtu.be/hqqKi3E-oG8>.

15. See <https://www.amazon.science/publications/formally-verified-cloud-scale-authorization>; see also “Proving the correctness of AWS authorization,” <https://youtu.be/oshxAJGrwMU>.

16. See <https://www.amazon.science/blog/new-aws-tool-recommends-removal-of-unused-permissions>.

17. See <https://aws.amazon.com/blogs/security/automated-reasoning-and-amazon-s2n>.

18. See <https://aws.amazon.com/blogs/aws/minimize-ai-hallucinations-and-deliver-up-to-99-verification-accuracy-with-automated-reasoning-checks-now-available/>.

19. See <https://aws.amazon.com/blogs/security/how-aws-threat-intelligence-deters-threat-actors>.

sophisticated “honeypot” system. Sonaris,¹⁹ a threat detection and automatic mitigation system, and Mithra,²⁰ 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. AWS security engineers and data scientists have explained in depth this work and the AI systems that power it at recent technical conferences.²¹

- **Security operations, threat hunting, and abuse capabilities.** Our continually improving automated systems are backstopped by the world’s leading human threat intelligence and response experts. Amazon operates global 24/7/365 security operations and response teams to supervise and triage incoming issues using four security 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 advanced threat actor groups, observing their tactics, techniques, and procedures, and when appropriate, taking part in coordinated takedowns of their infrastructure. Several examples of our work in this area can be found on our security blog.²² 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.
- **General third-party audit and compliance.** AWS operates one of world’s most carefully and thoroughly audited IT environments. Independent auditors do regular assessments of our environment and services under a wide variety of compliance regimes. These independent experts then issue reports, attestations, and certifications with supporting evidence that customers and regulators can refer to and rely on when assessing the security of AWS’s offerings.²³ Most of these reports, attestations, and certifications are available directly to customers for searching and downloading on a self-service basis using the AWS Artifact service.²⁴
- **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 developments of frontier models in AWS accounts meet the required security bar for careful security 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 VPC Flow Logs (netflow data) and DNS logs. The AWS GuardDuty intrusion detection service is enabled in all AI development environments, providing automatic monitoring for potential security threats, searching for indicators of compromise, and surfacing high priority alerts as appropriate. Software engineers and data scientists must be members of the correct Critical Permission Groups and authenticate with hardware security tokens from enterprise-managed endpoints to access or operate 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. Datasets used to train AI models are stored in data repositories that implement fine-grained access controls, data encryption, data access flow tracking, version control, and integrity validation.
- **Third party audits and red teaming assessments of AI systems.** We regularly engage outside security experts to test the security posture of our development and operational systems, reviewing evidence of the controls we have in place and conducting simulated attacks to test our assumptions and ensure models are deployed into secure environments. These audits validate that our model weights are not exposed in ways not accounted for in our threat models.

20. See <https://aws.amazon.com/blogs/security/how-aws-uses-active-defense-to-help-protect-customers-from-security-threats>.

21. See <https://aws.amazon.com/blogs/security/how-aws-tracks-the-clouds-biggest-security-threats-and-helps-shut-them-down>.

22. See “How AWS’s global threat intelligence transforms cloud protection,” <https://youtu.be/MsZWR4JJ5IA> (2025); “Discovering emerging threats in cloud security,” <https://youtu.be/38Z9csvyFDg> (2024).

23. See, e.g., <https://aws.amazon.com/blogs/security/>.

24. See <https://aws.amazon.com/compliance/programs/>.

25. See <https://aws.amazon.com/artifact/>.