

Caltech LLM Services Security & Governance Standards

Owner: CIO & CISO (IMSS)

Effective date: *March 5, 2026*

Review cadence: Annual (or upon major service change)

Applicable to: Faculty, Staff, Postdocs, Students, Contractors, and Service Providers using AI/LLM services with Caltech data.

1) Purpose

Establish mandatory security, privacy, and governance controls for using **Large Language Models (LLMs)** AI tools at Caltech covering **Amazon Bedrock**, **Amazon SageMaker**, and institutionally sanctioned services to protect Caltech information assets, research, and community members while enabling responsible innovation.

2) Scope & Definitions

- **In scope:** All AI/LLM services (hosted or managed), model training/fine-tuning, inference, agents/automation, data pipelines, and integrations that **process, access, or derive from Caltech data** or Caltech identities.
- **LLM Services (examples):** Amazon **Bedrock** (foundation models & agents), Amazon **SageMaker** (custom model training/hosting), Microsoft **Copilot/M365 Copilot** (productivity), approved domain-specific models.
- **Data classes:** Use Caltech's data classification standard (e.g., *Public, Internal, Confidential, Restricted*). If not formalized, IMSS will provisionally map federal and contractual obligations (FERPA, HIPAA/PHI, ITAR/EAR-controlled, human subject data) into the **Restricted** category pending Data Governance Committee confirmation.
- **Model customization:** Any training, fine-tuning, prompt-tuning, or RAG (retrieval-augmented generation) that uses Caltech data.

2.1 Key Terms & Acronyms (Glossary)

Networking & Connectivity

VPC (Virtual Private Cloud)

A logically isolated private network inside AWS used to secure LLM services and data paths.

VPC Endpoint / PrivateLink

A private, internal-only connection that allows Caltech systems to reach AWS services **without using the public internet**.

This prevents data from ever leaving Caltech's trusted network boundary.

Direct Connect

A dedicated, private network line between Caltech and AWS for predictable, secure traffic.

VPN (Virtual Private Network)

Encrypted tunnel allowing secure access to Caltech systems from outside the campus network.

Encryption & Keys

KMS (AWS Key Management Service)

AWS service used to generate, store, rotate, and audit encryption keys.

CMK (Customer-Managed Key)

Encryption keys **owned and controlled by Caltech**, not by AWS is required for all Restricted/Confidential data.

SSE-KMS

Server-side encryption using Caltech's CMKs to secure data stored in S3.

AI/LLM Concepts

Inference

Running a model to produce outputs (e.g., a GPT response) without modifying the model.

Training / Fine-tuning

Adjusting a model using Caltech data to improve performance on specific tasks. Has much stricter security requirements.

RAG (Retrieval-Augmented Generation)

Combines an LLM with a Caltech-owned knowledge base; improves accuracy while keeping data private.

Guardrails (Bedrock)

Safety filters that block harmful, sensitive, or out-of-policy content.

Identity & Access

IAM (Identity and Access Management)

Controls who can access which models, datasets, and AWS resources.

IAM Identity Center (SSO)

Caltech single-sign-on identity provider for AWS.

Data Types

PII (Personally Identifiable Information)

Data that identifies an individual (names, addresses, unique IDs).

PHI (Protected Health Information)

Medical/health-related personal information governed by HIPAA.

FERPA Data

Student academic records protected by federal law.

Tooling & Logging:

CloudTrail

AWS logging service that captures all API calls for auditing.

CloudWatch

Service for monitoring usage, anomalies, performance, and security signals.

SIEM (Security Information & Event Management)

Centralized system that collects logs and alerts IMSS Security.

3) Requirements for the use of LLM (What is Mandatory)

3.1 Approved Platforms & Connectivity

1. **Primary LLM inference** for Caltech data must use **Amazon Bedrock** or other IMSS-approved platforms with **no provider training on Caltech prompts/data**.
2. **Custom training/fine-tuning** must be performed on **Amazon SageMaker** or IMSS-approved secured environments.
3. All AI traffic with Caltech data must use **private networking** (e.g., VPC Endpoints/PrivateLink, Direct Connect) and **encryption in transit and at rest** (TLS 1.2+; KMS CMKs).
4. **Public, consumer, or unsanctioned GenAI tools** are **prohibited** for Restricted/Confidential data unless IMSS issues a specific exception with compensating controls. For more details on GenAI tools click [here](#)

Rationale: Amazon Bedrock does not store or use prompts/outputs to train models and supports VPC endpoints; SageMaker supports private VPC-only training/inference with KMS encryption.

3.2 Data Protection & Privacy

1. **Data Minimization:** Only the minimum necessary data may be sent to models (prompt redaction/anonymization required where feasible).
2. **Encryption:** All model inputs, outputs, embeddings, artifacts, training sets, feature stores, and logs must use **AWS KMS CMKs** managed by Caltech IMSS.
3. **Retention:** Prompt and inference logs are retained only as long as required for audit and troubleshooting, then purged per Caltech retention schedules. No model provider data retention is permitted for Confidential and Restricted data.

3.3 Identity, Access & Segregation

1. **IAM:** Role-based access with least privilege; separate **build/train** roles from **inference/use** roles; mandatory **MFA**; SSO via **IAM Identity Center** where supported.
2. **Isolation:** Use **separate AWS accounts/VPCs** for production vs. research/sandbox; segregate projects with **resource-level permissions** (Bedrock model invocation policies; SageMaker domain/profile isolation).
3. **Human subjects / sensitive research:** Requires **IRB** approvals and IMSS data-use review before any LLM processing.

3.4 Safety & Abuse Prevention

1. **Guardrails:** Bedrock **Guardrails** must be enabled for production assistants (safety filters, topic blocks, PII protection).
2. **Prompt Injection & Data Exfiltration Protections:** Implement **RAG allow-lists**, output validation, model-tool scope restrictions, and content filtering for download/external calls.

3. **Agent Controls:** Agents must only invoke **approved tools** with explicit scopes and least-privilege credentials; no arbitrary code execution in production without sandboxing.

3.5 Monitoring, Audit & IR (Incident Response)

1. **Audit:** Enable **CloudTrail, CloudWatch** metrics/logs, and centralize in **Security Lake/SIEM**; tag resources for ownership and data class.
2. **Detections:** Monitor for anomalous token usage, model invocation spikes, forbidden content categories, and cross-account access attempts.
3. **IR:** LLM incidents (prompt injection, data leakage, model abuse, compromised keys) follow Caltech's IR plan with 24x7 escalation to IMSS Security Operations – Security@Caltech.edu

3.6 Third-Party & Procurement

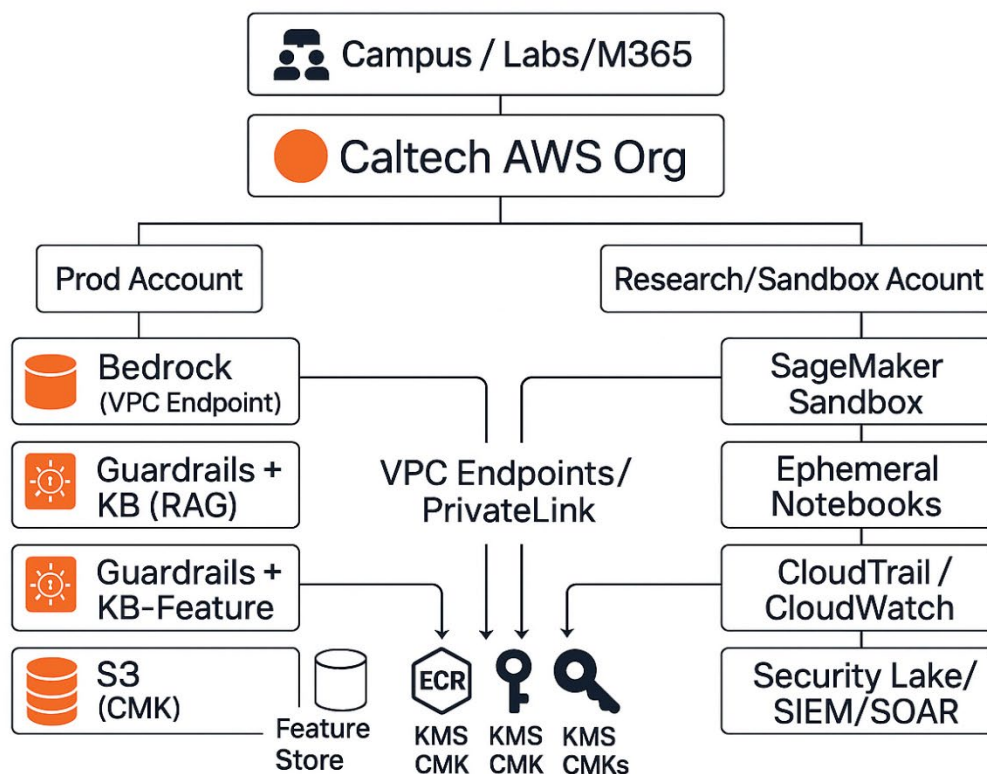
1. **Security & Data Processing Addendum** required for any non-AWS LLM vendor; **no training on Caltech data** clauses is a mandatory requirement.
2. **Risk Assessment:** IMSS Security Risk Assessment (and where applicable, ASIC - DPIA/Privacy review) before production use.
3. **Export Controls:** Coordinate with the **Export Compliance Office** before using AI services with controlled data (ITAR/EAR).

3.7 Research Enablement & Exceptions

- Research pilots are supported in **sandbox accounts** with reduced friction but must still:
 - Use **CMKs**, private networking, and no provider retraining on prompts.
 - Label outputs “**Research-Not for Production Use.**”
- Exceptions require the **Research Compliance Officer, CIO and CISO** (or delegate) approval with compensating controls and an agreed sunset date.

4) Caltech-Specific Security LLM Services Reference Architecture

4.1 High-Level Logical View



Shared: VPC Endpoints / PrivateLink, ECR, KMS CMKs

Key characteristics:

- **No public internet** for model calls or data paths.
- **Bedrock** used for enterprise LLM inference & agents with **Guardrails** and **Knowledge Bases** (RAG).
- **SageMaker** used for training/fine-tuning/hosting custom models with private endpoints.
- **Data always encrypted** (S3, EBS, endpoints) with **Caltech-owned KMS CMKs**.
- **Central logging** to SIEM with analytics/detections.
- **Account segmentation** for Prod vs. Research; project-level isolation by IAM and tags.

4.2 Recommended Network & Data Controls

- **Networking:**

- Create **VPC endpoints** for Bedrock and required AWS services (S3, STS, KMS, CloudWatch).
- **SageMaker**: VPC-only training/inference; disable internet access on notebooks/jobs; use NAT only for approved package mirrors.
- **Direct Connect/VPN** from Caltech core to AWS for deterministic, private paths.
- **Data:**
 - S3 buckets with **Block Public Access**, bucket policies enforcing **aws:SecureTransport** and **aws:SourceVpc** conditions, **SSE-KMS** with CMKs.
 - RAG corpora stored in S3 with **Object Ownership = Bucket owner enforced**; use **Access Points** with VPC-only access.
 - **PII/PHI** masked or tokenized before indexing into vector stores; store vectors in **private, encrypted** services (e.g., Aurora/Opensearch w/ encryption).
- **Identity & Permissions:**
 - **Model-invoke roles** (Bedrock) separated from **data-prep/train roles** (SageMaker).
 - Service control policies (SCPs) to **deny public endpoints, deny non-approved regions**, and **require KMS encryption**.

5) Minimum Security Configuration Baselines

5.1 Amazon Bedrock (Production)

- **PrivateLink/VPC Endpoint** is required, no internet egress for model calls.
- **Guardrails** enabled (safety filters, topic/PII blocks aligned to Caltech policy).
- **Knowledge Bases (RAG):**
 - S3 sources encrypted with CMKs; index jobs in private subnets.
 - Retrieval allow-lists (no broad bucket listing), and response post-processing (PII scrub).

- **Access:**
 - IAM policy to allow bedrock:InvokeModel only for approved model IDs.
 - CloudTrail & CloudWatch metrics enabled; anomaly alerts on token surges.

5.2 Amazon SageMaker (Production)

- **Networking:** EnableNetworkIsolation=true for training; endpoints in **private subnets; no public notebooks**.
- **Encryption:** S3, EBS, Model Artifacts, Endpoints with **CMKs**; enforce via IAM and SageMaker policies.
- **Runtime:** Approved base images from private **ECR**; signature verification.
- **Access:** Fine-grained roles tied to project tags; **deny* wildcards; MFA.
- **Monitoring:** Model data capture (with masking), CloudWatch metrics, logs to central SIEM; drift/quality monitors.
- **CI/CD:** CodePipeline/CodeBuild with security scanning; two-person approval for production deployments.

6) Roles & Responsibilities (RACI Summary)

- **CIO/CISO/IMSS InfoSec:** Define policy, approve exceptions, monitor, incident response (A/R).
- **Cloud Platform Team:** Provision accounts, networking, KMS, VPC endpoints, logging (R).
- **Data Owners/PI:** Classify data; approve use cases; ensure IRB/compliance (A/R).
- **Project Teams:** Build and operate solutions within guardrails; maintain documentation (R).
- **Vendors/Partners:** Sign DPAs from Caltech OGC, meet technical controls; support audits (R).

Symbol	Meaning
R	Responsible - performs the task

A	Accountable - ultimately answerable; final sign-off
A/R	The same group is both <i>responsible</i> and <i>accountable</i>

7) Lifecycle Controls & Process

1. **Intake & Classification:** Submit use case + data classification to IMSS.
2. **Design Review:** Architecture with controls (this standards); export control check if applicable.
3. **Build (Sandbox):** Prototyping in research account with required baselines.
4. **Security Gate:** Security Review, Threat model, IaC scan, IAM review, pen-test (if internet-exposed toolchains).
5. **Production Cutover:** Secrets rotation, runbooks, playbooks, monitoring in place.
6. **Operations:** Monthly key/access review, drift checks, cost & token use monitoring.
7. **Decommission:** Secure artifact disposal, data purge attestations.

8) Enforcement

Violations may result in **access revocation**, removal of workloads, and disciplinary actions per Caltech IMSS. Persistent violations are escalated to the **CIO, CISO, OGC** and appropriate governance bodies.

9) References

- **Security in Amazon Bedrock** (shared responsibility, IAM, data protection, VPC endpoints, incident response, guardrails)
<https://docs.aws.amazon.com/bedrock/latest/userguide/security.html>

- **Data Protection in Bedrock** (no training on your data; no logging of prompts/outputs; provider isolation)\n
<https://docs.aws.amazon.com/bedrock/latest/userguide/data-protection.html>
- **Generative AI data protection with Amazon Bedrock** (private access via VPC endpoints/Direct Connect; encryption; guardrails)\n
<https://maturitymodel.security.aws.dev/en/4.-optimized/gen-ai-security/>

These references align with the controls articulated above and support the architecture design decisions.

10) Appendices

A) Example Guardrail Policy (Bedrock) — Pseudocode

```
{
  "guardrails": {
    "blockedTopics": ["political advocacy", "malware creation", "harassment"],
    "piiProtection": { "detect": true, "masking": "hash" },
    "outputFilters": ["confidentialtermsdictionaryv1"],
    "actionOnViolation": "BLOCKAND_LOG"
  }
}
```

B) Example IAM Snippets (Conceptual)

```
{
  "Version": "2012-10-17",
  "Statement": [
    { "Effect": "Allow", "Action": ["bedrock:InvokeModel"],
      "Resource": ["arn:aws:bedrock:::foundation-model/anthropic.claude-3."]}
  ],
  { "Effect": "Deny", "Action": ["bedrock:InvokeModel"],
    "Resource": [""], "Condition": { "StringNotLike": {
      "aws:ResourceTag:Project": "CALTECH-*" } } }
```

```
]
}
{
  "Effect": "Deny",
  "Action": ["sagemaker:CreateEndpoint",
"sagemaker:CreateNotebookInstance"],
  "Resource": "*",
  "Condition": { "BoolIfExists": { "sagemaker:EnableInternetAccess": "true" } }
}
```

C) Security Validation Checklist (Go-Live)

- VPC endpoints for Bedrock, S3, STS, KMS, CloudWatch in place
- All S3 buckets **Block Public Access + SSE-KMS (CMK)**
- IAM policies reviewed; no * permissions; MFA enforced
- Bedrock **Guardrails** enabled & tested
- SageMaker endpoints **private**; notebooks **no-internet**
- CloudTrail/CloudWatch logs flowing to **SIEM**; alerts defined
- Data retention & purge procedures documented
- IR playbooks updated for LLM abuse/prompt injection/data exfiltration
- Export control/IRB approvals attached (if applicable)