Guide to Using Large Multimodal Models v1.1

Technical Supplement 6 - Security & Governance: Framework for Trustworthy AI Operations

Purpose & Scope
This supplement defines the organizational, technical, and procedural controls required to operate LMMs securely and in compliance with applicable standards. It establishes a layered governance model for trustworthy AI operations.

Audience: Security Officers, Compliance Leads, AI Governance Board, System Architects
Prerequisites: Familiarity with the Core Guide's verification and risk management principles (Sections 1, 4, & 5).
Outcome: A framework of security controls, audit checklists, and an incident management plan to prevent data leakage, ensure compliance, and maintain continuous oversight.

Key Objectives:

- Prevent unauthorized access, leakage, or misuse of sensitive data.
- Ensure model outputs remain compliant with ethics, law, and policy.
- Maintain continuous oversight, logging, and accountability throughout the AI lifecycle.

Developed by Russell Nida

Contents

1. Threat Surface and Risk Domains

Common security and risk domains affecting LMM operation include:

Domain	Example Threat Scenario	Control Strategy
Data Input	Prompt-injection or malicious file upload	Content filtering, sandbox execution, input validation
Model Context	Retrieval poisoning via corrupt RAG index	Versioned vector stores, source signing, periodic re-embedding audits (See Technical Supplement 5 for RAG controls)
User Interface	Insider data exfiltration through model responses	Role-based access, red-team prompt testing, output masking
Infrastructure	API key or token theft	Vault-based secret storage, short-lived tokens, zero-trust networking
Supply Chain	Compromised open-source model or embedding library	Dependency attestation, SBOM tracking, reproducible builds

2. Zero-Trust Architecture for LMM Systems

Adopt 'verify everything' principles across all components. Key practices include:

1. **Identity and Access Management (IAM)** - Enforce least privilege and multifactor authentication.
2. **Segmentation** - Isolate inference servers, RAG indexes, and training pipelines.
3. **Data Tagging & Classification** - Label and encrypt data at rest and in transit.
4. **Continuous Verification** - Every model call includes provenance check (who, what, when, why).
5. **Automated Revocation** - Credentials expire automatically; old model versions are quarantined.

3. Operational Governance Model

Role	Core Responsibility	Key Deliverables
AI Governance Board	Define risk tolerance, approve new deployments	Policy charters, exception logs
Data Steward	Owns RAG sources and fine-tuning datasets	Data inventory, lineage reports
Model Owner	Maintains model registry and validation scores	Model cards, change records
Security Officer	Oversees monitoring, incident response	SIEM dashboards, breach reports
Compliance Lead	Maps practices to NIST 800-171, ISO 27001, EU AI Act	Audit evidence, annual assessment

Governance Cycle: Define → Implement → Monitor → Audit → Improve.

4. Security Controls and Audit Checklist

Control Area	Verification Item	Evidence Type
Access Control	IAM policies reviewed quarterly	Access log, review record
Data Governance	RAG sources signed + versioned	Hash manifest
Model Integrity	Base models checksum verified pre-deployment	Build artifact hash
Logging & Monitoring	Prompts and responses logged with redaction	SIEM extract
Incident Response	Documented plan tested annually	Tabletop exercise report

5. Compliance Mapping

Framework	Relevant Controls in this Supplement
NIST 800-171 / 800-53	Access control (AC-1--AC-7), Audit (AU-2), System Integrity (SI-2)
ISO 27001:2022	A.5.9 Information Security in AI Use
EU AI Act (2024)	Articles 9-15 Risk Management & Data Governance
U.S. EO 14110 (Safe AI)	Sections 4(b) - Testing and Red-Team Protocols

6. Incident Management Framework

1. **Detection:** Automated alerts via log anomaly monitoring.
2. **Containment:** Disable affected API keys / quarantine model instance.
3. **Investigation:** Trace prompt, data source, user ID.
4. **Remediation:** Retrain or rollback model version.
5. **Post-Mortem:** File incident report and update controls.

7. Audit and Continuous Improvement

- Quarterly security audits with independent review.
- Annual red-team exercise simulating prompt injection and data exfiltration.
- Rolling KPIs: incident frequency, mean time to detect, compliance coverage %, audit findings closed %.
- Feed lessons learned into prompt filters, RAG validation logic, and employee training.

8. Key Takeaway

Security and governance are not adjacent to AI operations—they are AI operations. They are the organizational-scale implementation of the verification mindset mandated throughout this guide. Every prompt, dataset, and deployment is a potential policy event. Trustworthy AI depends on transparent design, verifiable behavior, and enforceable accountability.

End of Technical Supplement 6

Cross-Reference: For dataset and fine-tuning controls, see Technical Supplement 5 - Customization & Fine-Tuning.