Agency & Reasoning Threats

Threat	MAESTRO Layer	Likelihood	Impact	Mitigations	Monitoring
Prompt Injection	Data Operations	Medium	High	Enforce strict input validation; sanitize user prompts; limit command execution scope; implement operation allowlists	Log all prompt inputs; anomaly detection on unexpected task execution or command patterns
Hidden Instructions in Documents	Data Operations	Medium	High	Disable execution of embedded instructions; filter document inputs; use static content scanning for malicious patterns	Monitor document ingestion events; flag abnormal metadata or unexpected triggers
Insufficient Logging	Evaluation & Observability	Low	High	Define clear logging policies for sensitive actions; enforce structured logging and audit trails for AI and admin activities	Continuous log integrity checks; alert on missing or inconsistent logs

Memory Based Threats

Threat	MAESTRO Layer	Likelihood	Impact	Mitigations	Monitoring
Cascading Hallucinations	Foundation Model, Data Operations	Medium	High	Implement human-in-the-loop review for critical tasks; apply filtered or verified outputs	Monitor model output drift and hallucination rates; log all corrected outputs for retraining

Tool & Execution-Based Threats

Threat	MAESTRO Layer	Likelihood	Impact	Mitigations	Monitoring
Abusing Agent Tools	Agent Framework	Medium	Medium	Apply least-privilege	Log tool invocations and

				access; restrict high-risk tool calls; enforce output sanitization	correlate with user intents; alert on unexpected tool chains
Privilege Compromise	Deployment & Infrastructure, Security & Compliance	Low	High	Enforce RBAC (role-based access control); isolate credentials; use MFA for admins	Monitor failed authentication attempts; audit configuration changes
Exhausting Compute, Memory, or APIs	Deployment & Infrastructure, Agent Framework	Medium	Low	Implement rate limiting and resource quotas; sandbox heavy computations	Collect and visualize resource usage metrics; alert on spikes or throttled processes
Unexpected RCE or Code Injection Attacks	Deployment & Infrastructure	Low	High	Disable arbitrary code execution; enforce signed and verified scripts; log command invocations	Intrusion detection on runtime processes; monitor for new file or process creation

Authentication & Identity Threats

Threat	**MAESTRO Layer**	**Likelihood**	**Impact**	**Mitigations**	**Monitoring**
Impersonation of Agents, Users, or Services	Security & Compliance	Medium	Medium	Use JWT-based authentication; apply short-lived tokens; auto-logout after inactivity	Log all authentication events; detect concurrent sessions or token reuse

Human-in-the-Loop (HITL) Threats

Threat	**MAESTRO Layer**	**Likelihood**	**Impact**	**Mitigations**	**Monitoring**
Overloading Human Reviewers	Agent Ecosystem	Low	Low	Queue-based request throttling; cap review assignments per user	Track queue depth and review latency metrics
Exploiting User Trust in AI Responses	Agent Framework, Agent	Medium	Medium	Adversarial training to improve robustness;	Monitor error rates and user correction frequency; detect

	Ecosystem			explainability features; clear user feedback loop	overreliance patterns

Multi-Agent System Threats

Threat	**MAESTRO Layer**	**Likelihood**	**Impact**	**Mitigations**	**Monitoring**
Corrupting Inter-Agent Messages	Agent Ecosystem, Agent Framework	Low	High	Encrypt and sign inter-agent communications; apply mutual authentication	Log message integrity failures; detect unauthorized agent IDs
Exploiting Delegation & Workflows	Agent Ecosystem	Medium	High	Implement explicit authorization between agents; enforce transaction boundaries; encrypt communications	Track workflow delegation chains; detect abnormal escalation patterns
Malicious or Compromised Agents	Agent Ecosystem, Agent Framework	Low	Medium	Verify agent identity; use sandboxed environments; monitor agent behavior profiles	Behavior anomaly detection; alert on unauthorized requests or data access
Unsigned Docs	Agent Ecosystem	Medium	High	Apply short-lived tokens; Avoiding data leakage	Detect token reuse