Agentic AI Threats: Securing Against Autonomous, Self-Directed Attacks

IN COLLABORATION WE TRUST,
CREATING LASTING VALUE

13:20-13:50, 27 May 2026

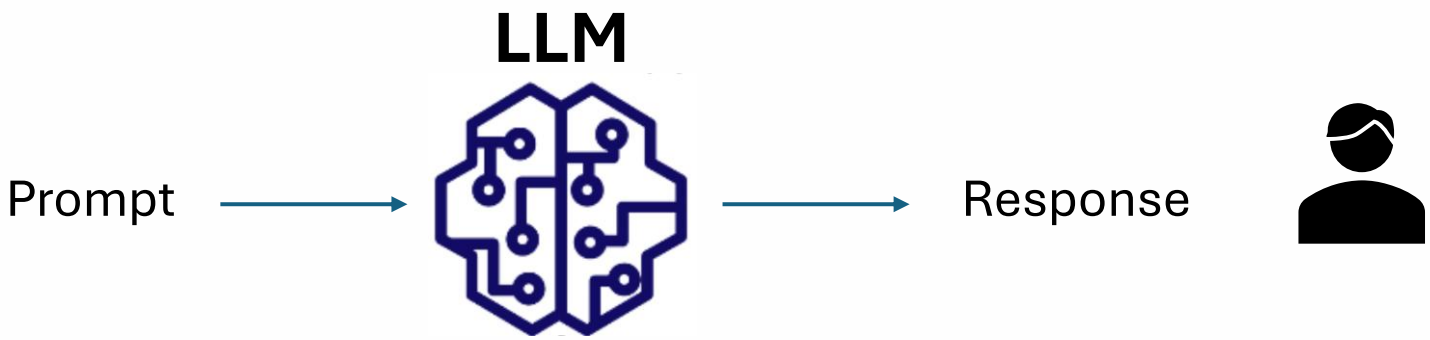
Berkeley Hotel Pratunam, Bangkok

Kitti Kosavisutte, Ph.D.

THAILAND BANKING SECTOR CERT
2026

Generative AI

Answer based on pre-trained knowledge only

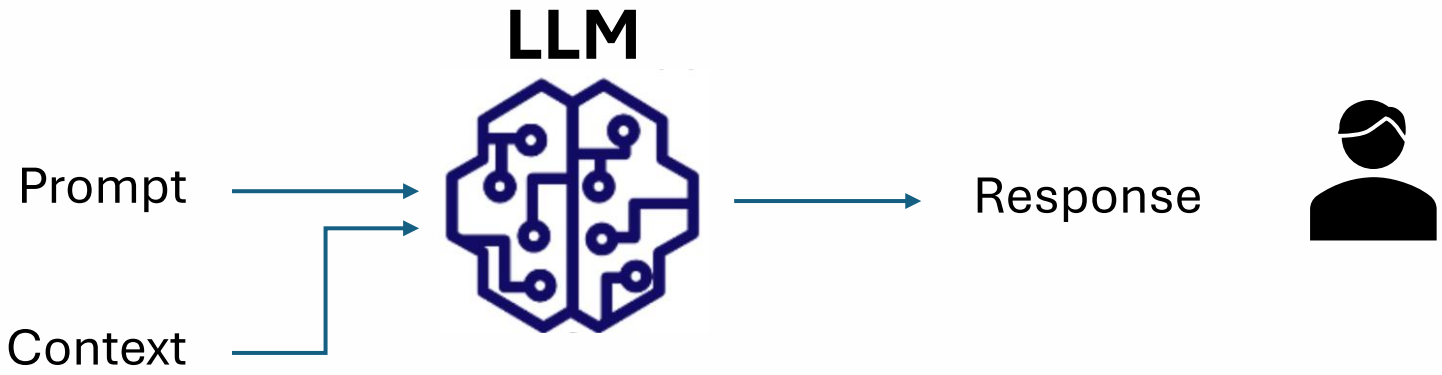


Risk specific to Generative AI

- Data leakage
- Prompt injection attack
- Model abuse
- Supply chain risk
- Hallucination
- Model drift
- Inconsistent output
- Lack of explainability
- Compliance and regulation risk

Generative AI with RAG

Cannot make complex decision and take actions



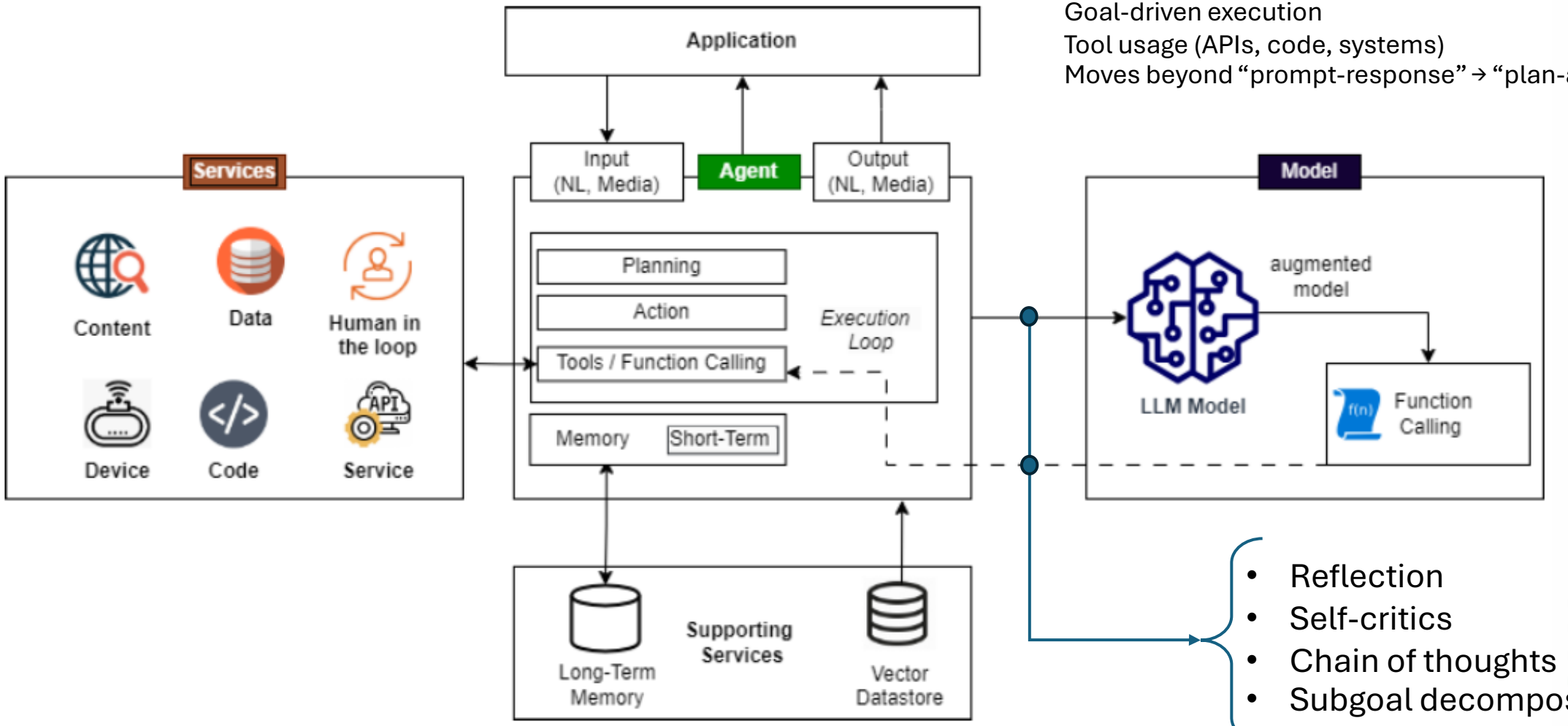
Risk specific to RAG

- Retrieval poisoning
- Sensitive information exposure
- Context manipulation
- Insecure vector database

RAG : Retrieval-Augmented Generation

Fundamental Aspect of Agentic Autonomy

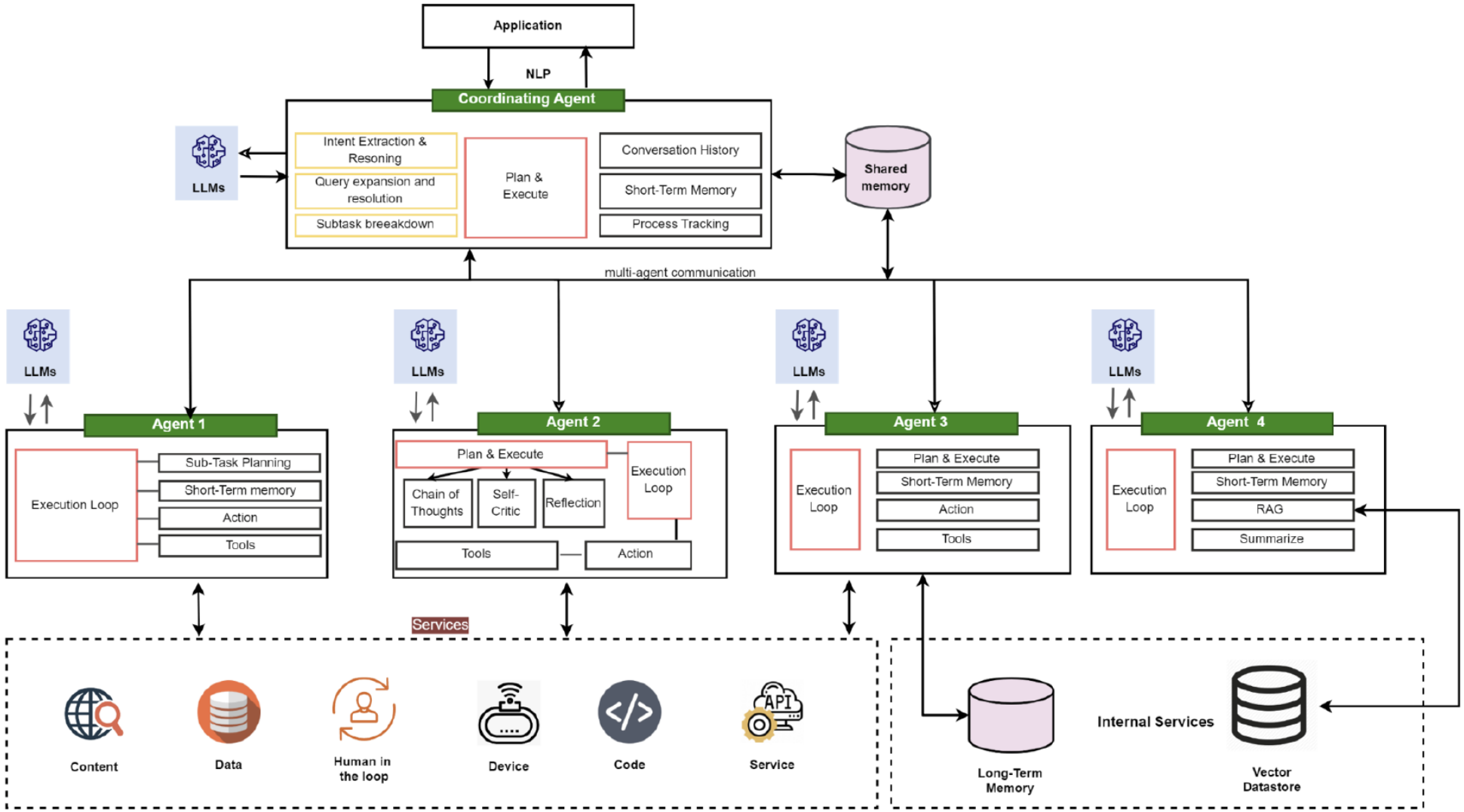
AI systems capable of:
Autonomous decision-making
Multi-step reasoning
Goal-driven execution
Tool usage (APIs, code, systems)
Moves beyond "prompt-response" → "plan-act-learn" loop



- Reflection
- Self-critics
- Chain of thoughts
- Subgoal decomposition

Ref: OWASP

Multi Agentic AI

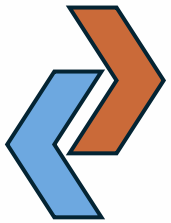


Ref: OWASP

Paradigm Shift in Cybersecurity from Traditional AI to Agentic AI

Traditional AI Paradigm

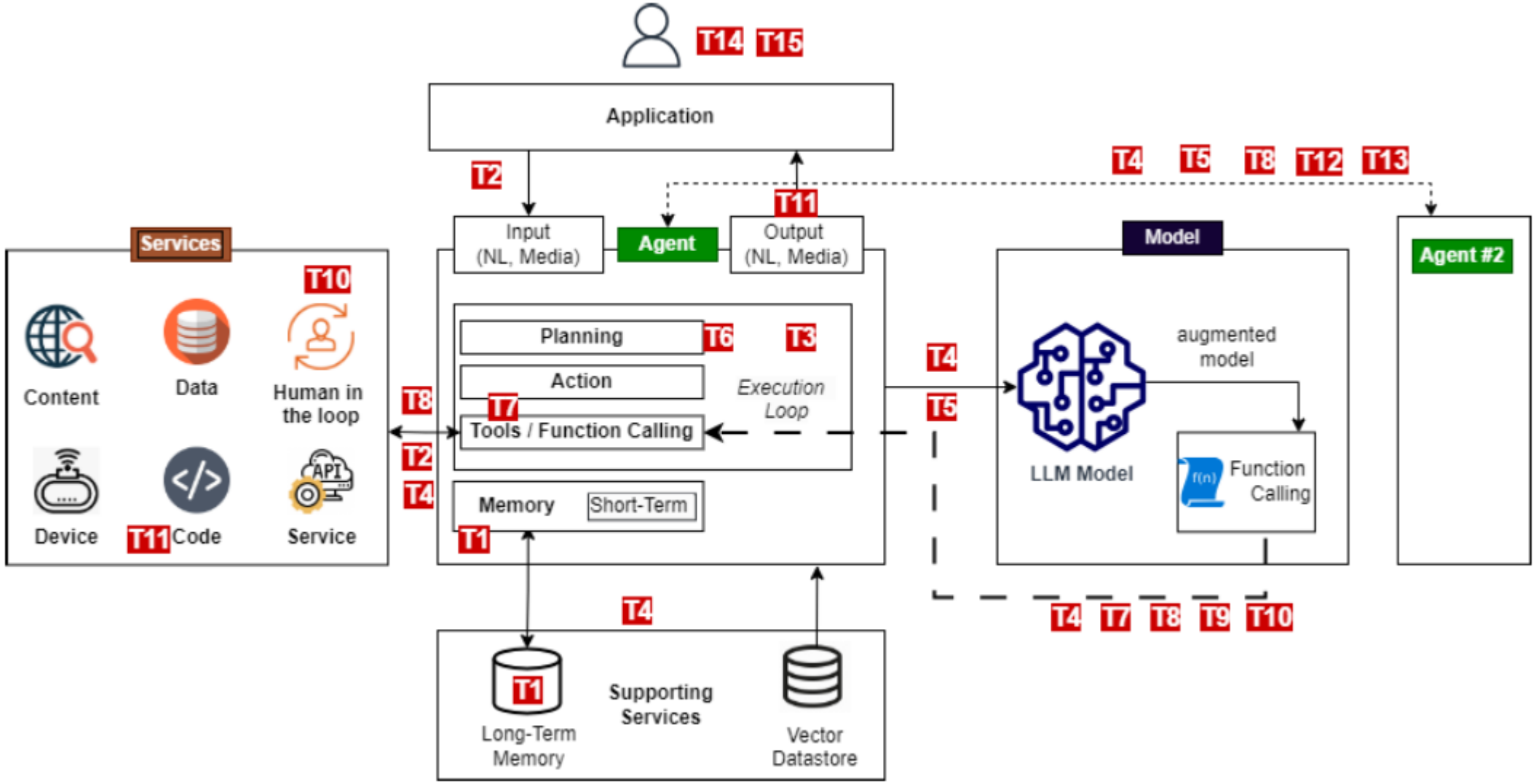
- Single-task execution
- Human-directed actions
- Prompt-response interaction
- Static automation
- Tool-specific processing
- Deterministic workflows
- Rule-based orchestration
- Centralized control
- Knowledge retrieval
- Static security assumptions



Agentic AI Paradigm

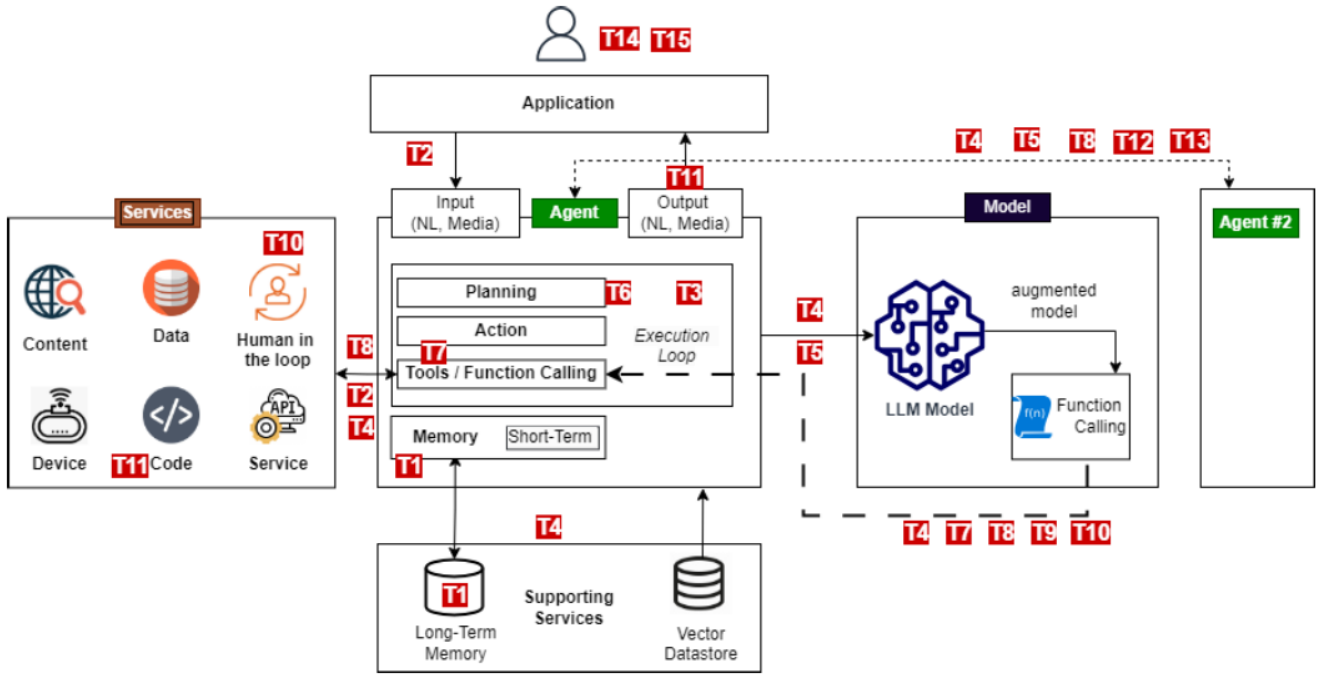
- Multi-step autonomous workflows
- Self-directed planning and execution
- Continuous reasoning and adaptation
- Dynamic decision-making
- Tool orchestration across system
- Adaptive and probabilistic behavior
- Autonomous task decomposition
- Distributed multi-agent coordination
- Reasoning + action + execution
- Adaptive trust and risk evaluation

Threat Model



Reference: OWASP

Detailed Threat Model



1. Memory poisoning
2. Tool misuse
3. Privilege compromise
4. Resource overload
5. Cascading hallucination attack
6. Intent breaking & goal manipulation
7. Misaligned & deceptive behavior
8. Repudiation & untracability
9. Identity spoofing & impersonation/Agent identity compromise
10. Over whelming human in the loop
11. Unexpected RCE and code attacks
12. Agent communication poisoning
13. Rogue agents in multi-agent systems
14. Human attacks on multi-agent systems
15. Human manipulation
16. Insecure inter-agent protocol abuse
17. Supply chain compromise

Reference: OWASP

Threat Landscape and Security Challenges Unique to Agentic AI

	Threat Landscape
Automation of attack lifecycle (end-to-end)	• Recon (scan + data gathering) • Weaponization (generate exploit/phishing) • Delivery (automated targeting) • Execution (adaptive attack) • Persistence & lateral movement
Autonomous malware/Attack agents	• Self-modifying malware • Exploit generation • AI-driven vulnerability discovery • Adaptive evasion
AI-Powered credential & identity attack	• Automated credential stuffing • MFA fatigue attack automation • Identity impersonation at scale
Hyperscale social engineering	• Personalized phishing • Real-time conversation scam • Deepfake + adaptive script
Exploitation of AI Systems	• Prompt injection • Control agent behavior • Data exfiltration via tool usage • Model poisoning/Supply chain attack

Challenges

- Non-deterministic behavior
- Speed & scale of attack beyond human response
- Broader access scope across systems, data and workflows
- Sophistication in adaptive, context-aware and multi-step reasoning

Key Risk Scenarios - Sample

- Loss of control and unintended actions
- Prompt injection and adversarial manipulation
- Over-privilege and access abuse
- Data leakage and privacy breach
- Lack of explainability and auditability
- Autonomous scaling of errors
- Supply chain and tooling risk
- Misalignment with business intent
- Human over-reliance
- Emergent and unpredictable behavior

Defensive Strategy

1. Governance & Control
- Define “Allowed AI Actions” (Action policy)
- Risk-tiering for use cases (High-risk agent)
- Approval workflow for agent deployment

2. Identity & Access Management for AI
- Treat AI as “non-human identity”
- Least privilege for agent
- Short-lived token / scoped access
- Strong authentication for tool usage

3. Secure Agent Architecture
- Sandbox execution environment
- Tool isolation (API gateway control)
- Human-in-the-loop for critical action
- Output validation layer

4. AI-Specific Security Controls
- Prompt injection defense
- Input / output filtering
- Context boundary enforcement
- Memory protection

5. Monitoring & Detection
- Agent activity logging (who / what / why)
- Behavior anomaly detection
- AI action traceability
- Incident response for AI misuse

6. Data Protection
- Data minimization
- Sensitive data masking
- Secure RAG (retrieval control)
- Prevent data exfiltration via agent

7. Red Teaming & Testing
- Adversarial testing (prompt attack simulation)
- Abuse case testing
- Continuous validation of agent behavior

Thank you