THREAT MODELING · FIELD NOTES

AI Threat Modeling: Adapting STRIDE for LLM Systems

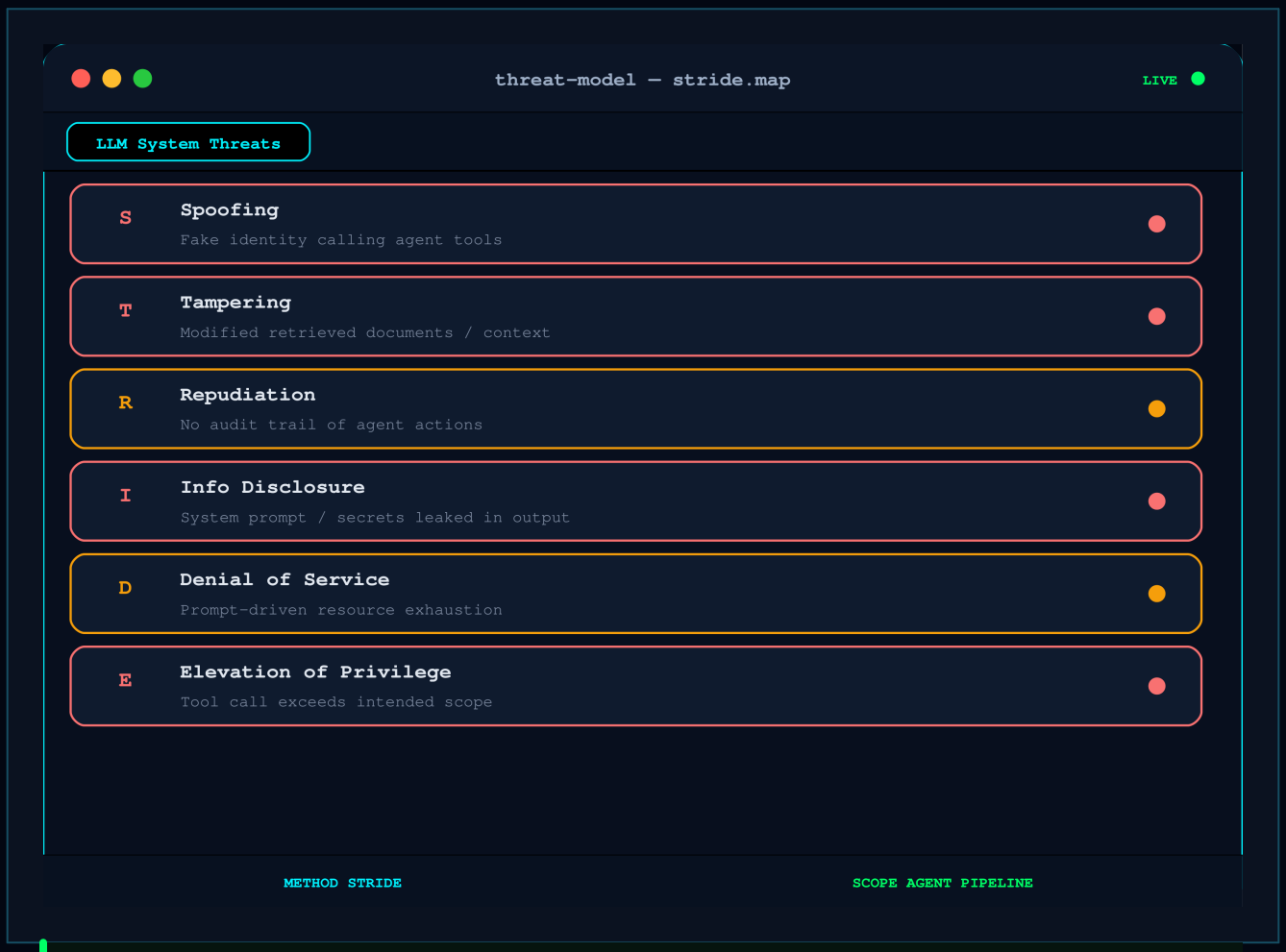
> *“STRIDE wasn't written for agents that call tools and read untrusted documents — but the categories still map, once you know where to look.”*

Microsoft's STRIDE framework (Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, Elevation of Privilege) predates LLM agents by two decades. It still applies — this dossier maps each category onto an agent/RAG pipeline's actual components.

Scope	Threat modeling for LLM agent and RAG pipelines
Method	STRIDE, mapped to agent-specific components
Audience	Teams building or reviewing AI agent systems
Analyst	Nazline Mwita, Cybersecurity Assurance Lead, HarLyn Digital Partners

1. The Six Categories, Mapped to an Agent Pipeline

Each STRIDE category has a concrete, non-hypothetical analog in a typical agent/RAG system. Treating these as abstract categories rather than specific failure modes is how threat models end up too generic to act on.



Modeling Principle

Map each category to a specific component — not a generic "AI risk" bucket.

2. The Highest-Priority Categories for Agent Systems

Tampering: retrieved documents and tool outputs are attacker-influenceable input, not trusted context — treat them accordingly.

Information Disclosure: system prompts, credentials, and other-tenant data can leak through model output if boundaries aren't enforced server-side.

Elevation of Privilege: a tool call that can be steered beyond its intended scope is a privilege escalation path, not a UX bug.

3. Applying This to a Real Pipeline

Component	Primary STRIDE exposure
User input / prompt	Spoofing (identity), Tampering (injected instructions)
Retrieval / RAG documents	Tampering (indirect prompt injection via retrieved content)
Tool-call execution	Elevation of Privilege, Denial of Service
Model output	Information Disclosure (leaked secrets/context)
Audit / logging layer	Repudiation (no record of what the agent actually did)

4. What This Is Not

This is not a replacement for a full security review — it's the structuring step that makes a review comprehensive instead of ad hoc. A threat model without a follow-up mitigation owner per finding is a document, not a control.

Related Service

This dossier supports the AI Agent Security and RAG Security services — structured threat modeling ahead of a deeper technical review.

Nazline Mwita, CompTIA Security+ certified Cybersecurity Assurance Lead and Co-Founder at HarLyn Digital Partners. Specializing in authorized web & API security assessments, KDPA compliance reviews, and defensive cloud architecture in Nairobi, Kenya.