


# Garbage In, Garbage Out

## Offensive Testing Strategies for AI Systems

**Kevin Tackett**
**kevin@secureideas.com**
**904-403-8024**
**@secureideas**

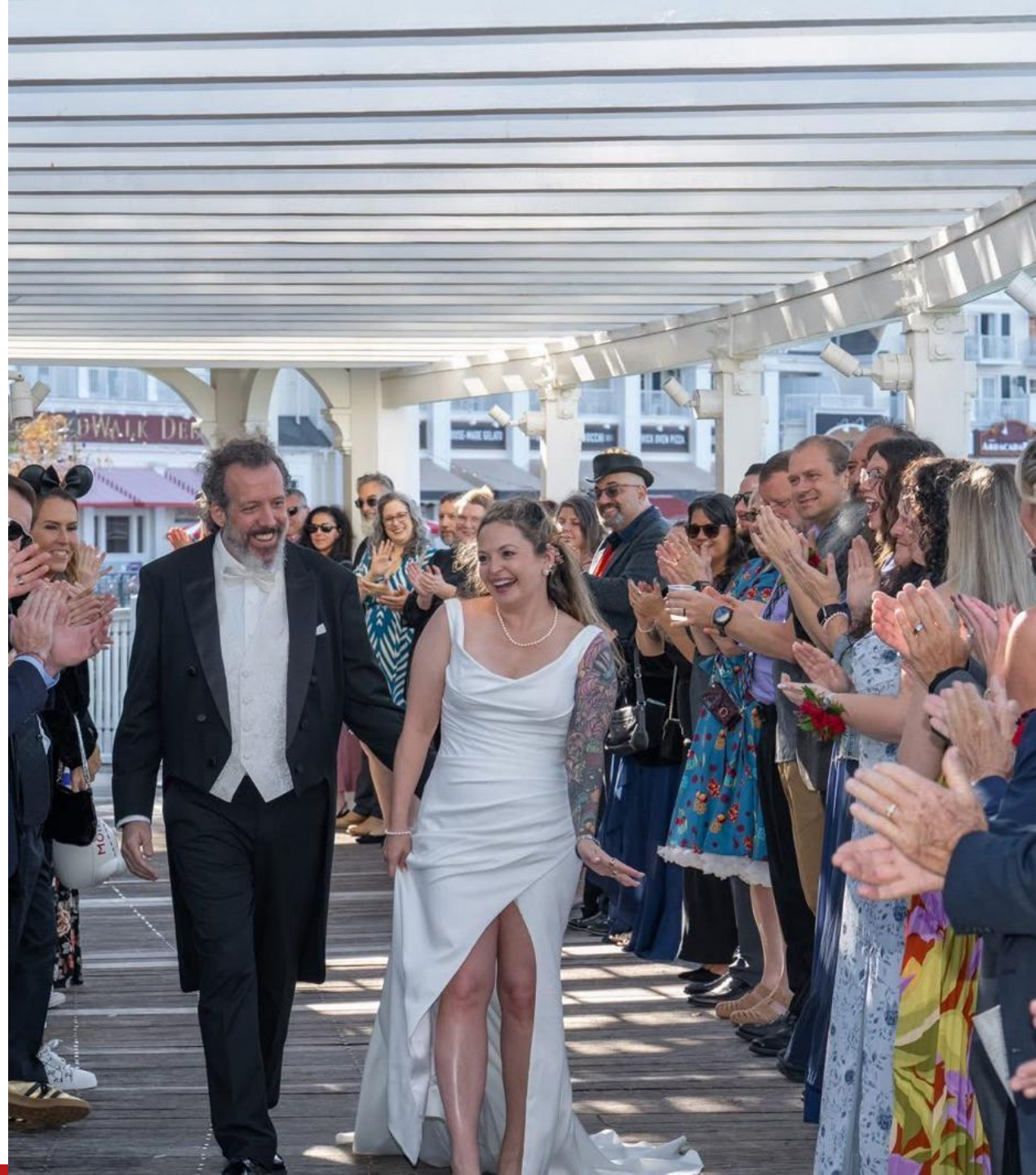
Copyright © 2026 Secure Ideas LLC

Secure Ideas
professionally evil®

# Kevin Tackett

- CEO - Secure Ideas, LLC
  - Founded in 2010
- IANS Faculty member
  - Since 2011
- Open-Source Fanatic
  - OWASP Leader
  - Former OWASP Global Board
  - Project Builder and Founder
- Speaker, Course Instructor, Author
  - Since 1991
  - SANS, ISC2, OWASP, DEF CON, Blackhat, etc.

- Other Interests
  - Warhammer (40K and AoS)
  - Woodworking
  - Husband and Father




Copyright © 2026 Secure Ideas LLC

# Agenda


01 The Threat Model

02 Robots Chatting — LLMs & Chatbots

03 Integrating the Dump — APIs & Integrations

04 Trash Compactor — ML Pipelines

05 Malicious Assistants — Agents & Tool Use

06 What Next — A Practical Framework

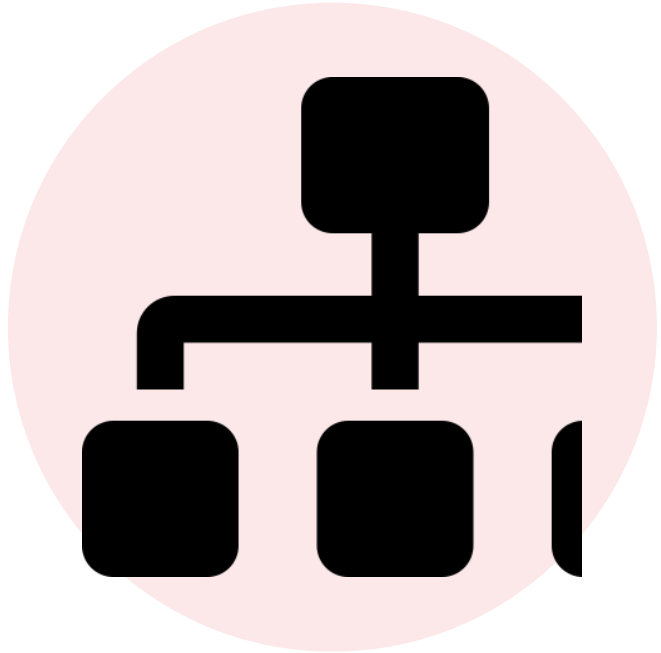


Copyright © 2026 Secure Ideas LLC


Copyright © 2026 Secure Ideas LLC

# Understanding the Threat

- AI is everywhere — well, not really AI
- Most “AI” is a thin interface over the same backend systems you already test
- Organizational trust is transitive
  - Even when the org forgets that
- The attacker doesn’t care what marketing calls it



Copyright © 2026 Secure Ideas LLC
Copyright © 2026 Secure Ideas LLC

# AI ≠ Traditional Software

## WHAT'S ACTUALLY NEW

- Nondeterministic output
- Natural language as attack surface
- Data and instructions share one channel

## WHAT ISN'T

- Broken auth
- Leaky APIs
- Excessive privilege

*Most "AI vulnerabilities" are classic misconfigurations wearing a new hat.*

Copyright © 2026 Secure Ideas LLC
Copyright © 2026 Secure Ideas LLC

# What Are We Testing

- LLMs & Chatbots — the conversational layer
- AI-Integrated Applications & APIs — the plumbing
- ML Pipelines — the training and data layer
- Autonomous Assistants & Agents — the layer with hands



Copyright © 2026 Secure Ideas LLC
Copyright © 2026 Secure Ideas LLC



SECTION 1

# Robots Chatting

LLMs & Chatbots

Copyright © 2026 Secure Ideas LLC
Copyright © 2026 Secure Ideas LLC

Secure Ideas professionally evil®

# Robots Chatting

- Chatbots are everywhere, varying complexity
- Mostly third-party — “CHATaaS”
- Privileges are mixed, typically more than we want
- Most orgs are consuming somebody else’s chatbot, not building one
  - Privilege and data exposure are inherited, not designed



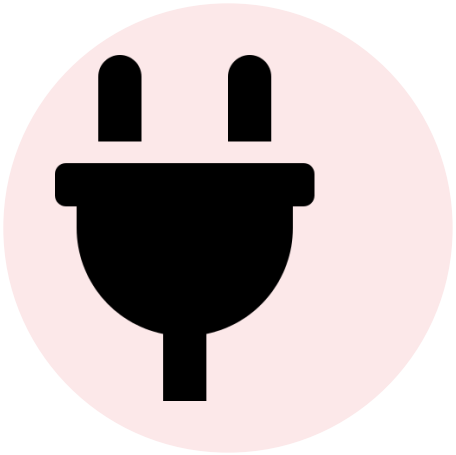
Copyright © 2026 Secure Ideas LLC
Copyright © 2026 Secure Ideas LLC

# Interrupting the Chat

- Multiple Tools and techniques
  - Prompt injection
  - Sensitive information
  - Integration issues
- Tools
  - PromptFoo
  - Arcanum taxonomy



Copyright © 2026 Secure Ideas LLC



SECTION 2

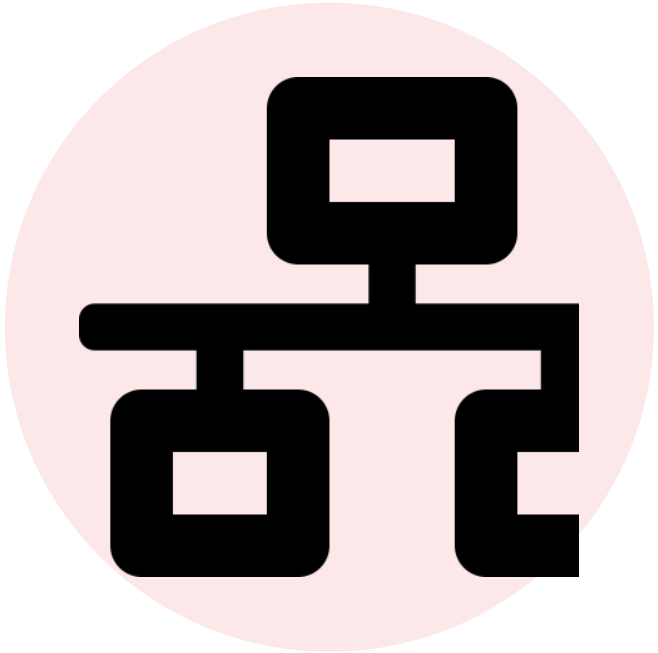
# Integrating the Dump

AI-Integrated Applications & APIs

Copyright © 2026 Secure Ideas LLC
Copyright © 2026 Secure Ideas LLC

# Integrating the Dump

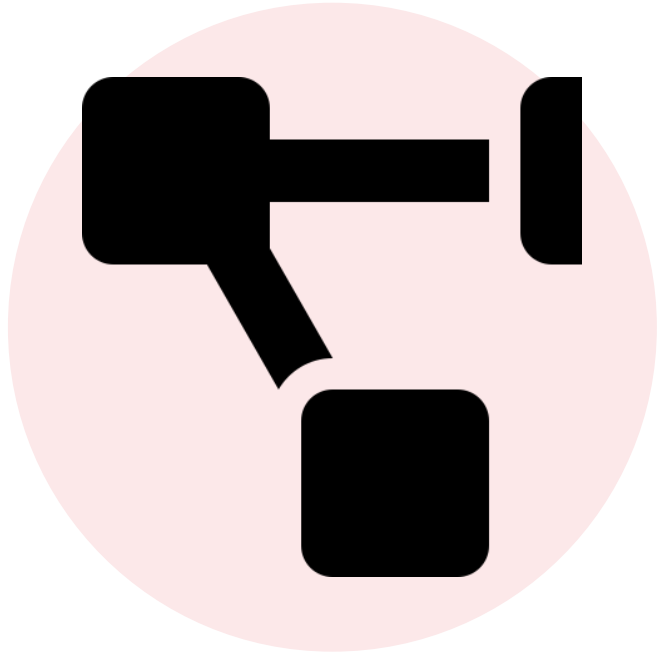
- Integrations are the most common tests, and it's changing fast
- APIs and backend services are the real focus
- Guardrails are often looser here than at the chat layer
- This is where "AI security" quietly becomes "API security testing with extra steps"



Copyright © 2026 Secure Ideas LLC
Copyright © 2026 Secure Ideas LLC

# MCP: The New API You Haven't Tested

- Model Context Protocol gives an agent structured access to tools and data
- Authentication is often an afterthought in early deployments
- MCP servers show up in places a gateway never sees — laptops, dev environments
- Treat every MCP server like an API you haven't inventoried yet



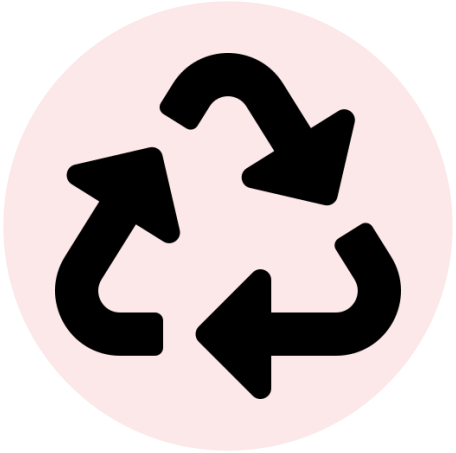
Copyright © 2026 Secure Ideas LLC
Copyright © 2026 Secure Ideas LLC

# Attacking the Trash

- Attacks are common
  - Integration points
  - API end points
- Tooling is the same
  - Burp/ZAP
  - SSRF Tools
  - MCP Inspector




Copyright © 2026 Secure Ideas LLC



SECTION 3

# Trash Compactor

ML Pipelines

Copyright © 2026 Secure Ideas LLC
Copyright © 2026 Secure Ideas LLC

# Trash Compactor

- Data poisoning and training data exposure
- Model supply chain risk — untrusted base models, plugins, LoRAs
- MLOps infrastructure: notebook servers, feature stores, model registries
- Often internal-only, and often unauthenticated



Copyright © 2026 Secure Ideas LLC
Copyright © 2026 Secure Ideas LLC

# Stopping the Compactor

- Much harder to exploit
  - And verify
- Attacks are different
  - Poisoning the sources
  - Attacking the systems and storage



Copyright © 2026 Secure Ideas LLC



SECTION 4

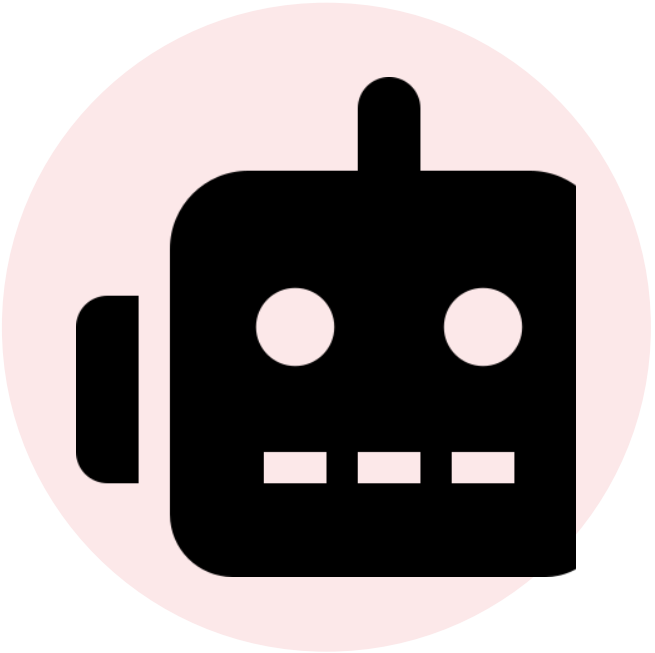
# Malicious Assistants

Agents & Tool Use

Copyright © 2026 Secure Ideas LLC
Copyright © 2026 Secure Ideas LLC

# Malicious Assistants

- Everyone wants an assistant — built-in, co-pilot, or stand-alone
- Access rights are the real concern, and so is the sandbox
- The question isn't "can it help," it's "what can it reach"



Copyright © 2026 Secure Ideas LLC
Copyright © 2026 Secure Ideas LLC

# Excessive Agency

- What happens when an agent's tool access exceeds the task it was given
- Confused deputy: the agent acting with more authority than the user who invoked it
- Agent-to-agent and agent-to-MCP-server trust is rarely tested



Copyright © 2026 Secure Ideas LLC
Copyright © 2026 Secure Ideas LLC

# Tripping the Assistant

- End point controls and systems are in scope
  - Agents run somewhere
- Permissions
  - Permissions, permissions, permissions.
- Risk evaluation
  - PyRIT




Copyright © 2026 Secure Ideas LLC

# What Next?

- Understand your usage — all of it (closes Robots Chatting)
- Embed testing into existing processes and new assessments (closes Integrating the Dump)
- Map your data concerns: training, usage, privacy (closes Trash Compactor)
- Manage risk — yours, not some standard (closes Malicious Assistants)



Copyright © 2026 Secure Ideas LLC
Copyright © 2026 Secure Ideas LLC

# Questions?



Kevin Tackett
kevin@secureideas.com
904-403-8024