Would anyone notice? Valid access. Wrong task.

CompoundTrust

A starting framework from CompoundTrust, not a standard. Fill it in for one live agent.

AGENT	OWNER	ON-DUTY RESPONDER · BACKUP · WHO CAN CONTAIN

1 · TELL IT APART

1. Agent identity (not a person's or a shared account): _____
2. Per-run identifier recorded: yes / no · credential lifetime: _____
3. Logged under that identity: what it read · tools called · where output went
4. Logs reach the security team: yes / no

2 · KNOW NORMAL: FIVE-LINE PROFILE

1. Reads: _____
2. How much (per day and per short window): _____
3. Tools: _____
4. Output goes to: _____
5. Working hours: _____

Start from the Day 30 permission set; add expected volumes, time windows and the task.

3 · SIGNAL	WATCH FOR	YOUR RULE	ALERTS OR BLOCKS?	TESTED ON
More	Far more records or actions than the profile expects in a short window	> ___ records in ___ minutes		
Wider	Data or systems outside the agent's job (another region, another system)	Scope: _____		
Out	Output going to a destination it has not used before	Approved destinations: _____		
Different	A tool, time or pattern it has never used	Tools / hours: _____		

> READ THIS Signals tell you where to investigate. Unusual does not always mean harmful, and harm can stay within normal volumes. Detection sits beside least privilege and blocks on unapproved destinations; alerting on a transfer is not the same as blocking it.

DRILL · TEST ENVIRONMENT ONLY	ACTION AT	ALERT AT	ACKNOWLEDGED AT	CONTAINED AT	NO ALERT WITHIN WINDOW?
Burst above test threshold					
Transfer to internal test box (external delivery off)					
Approved high-volume batch (should not alert)					

Use test credentials and synthetic data. Record any no-alert result against the agreed observation window.

CLOCK FIVE (DAY 45)

Suspicious activity to a responsible person acknowledging it. Record action, alert, acknowledgement and containment separately.

REPORT

What did we test? Who noticed? What happened next?

WORKED EXAMPLE

Hypothetical: normal 200 records a day; rule >500 an hour; 11,000 read by containment at assumed times. Cause: indirect prompt injection in an email.

Context: CERT-In CIAD-2026-0049 (15 Oct 2026), evaluation cases: agents run by AI companies reached Indian organisations and none of the organisations detected the access. OWASP LLM01 Prompt Injection · OWASP LLM06 Excessive Agency.