INTELLIGENCE BRIEFING
Security Command Center
TLP:CLEAR
2026-07-12 14:16 UTC

Prompt Injection Grows Up: 18 New Techniques Expose AI Agents as High-Value Attack Targets

SECURITY ANALYSIS | **HIGH** | CVSS 7.5

SCC Item ID	SCC-STY-2026-0348
Type	Security Analysis
Severity	HIGH
CVSS Base Score	7.5
Affected Products	AI agents broadly (LLMs including Gemini referenced as example), CrowdStrike Falcon AIDR, enterprise AI and agentic platforms with access to email, CRM, shell commands, and enterprise data stores
Discovery Source	Rss:T1 Threatintel

Executive Summary

CrowdStrike's AI security research team has documented 18 new prompt injection techniques, expanding its taxonomy to more than 200 distinct attack patterns targeting large language models and AI agents. The research reveals adversaries have shifted from simple jailbreaks to sophisticated, multi-stage attacks, including dormant payloads and fragmented instructions designed to evade detection. As AI agents gain privileged access to email, CRM systems, shell commands, and enterprise data stores, organizations that have not modeled the AI prompt layer as an attack surface face material operational and data security risk.

Technical Analysis

CrowdStrike's AI security research team has published a significant expansion of its prompt injection taxonomy, adding 18 new technique categories and pushing total documented coverage beyond 200 distinct attack patterns. The research signals a maturation of the threat: adversaries are no longer relying on blunt jailbreak prompts but are engineering multi-stage, condition-aware attacks designed to persist, evade detection, and weaponize the agent's own privileged access.

Five newly documented techniques are particularly instructive. Delayed trigger activation introduces dormant payloads that remain inert until a specific runtime condition is met, a behavioral pattern that mirrors advanced persistent threat tradecraft and directly challenges detection approaches that evaluate prompts at ingestion time only. Payload fragmentation splits malicious instructions across multiple inputs, exploiting the context-assembly behavior of LLMs to reconstruct a coherent attack vector that no single input would surface. Unwitting user context injection is arguably the most operationally concerning: it leverages legitimate user inputs, such as forwarded emails or pasted CRM records, to carry adversarial instructions into the agent's context window

without the user's knowledge, mapping directly to MITRE ATT&CK T1566 (Phishing) and T1557 (Adversary-in-the-Middle) patterns.

The attack surface implications are significant. Modern AI agents are not passive chat interfaces; they are increasingly granted write access to email, CRM platforms, databases, and shell command interpreters. A successful prompt injection in this environment can execute code (T1059), exfiltrate data via trigger-activated email forwarding (T1567), or destroy records (T1485). Boundary spoofing, elevating user-tier content to system-authority instructions, maps to T1078 (Valid Accounts) and CWE-441 (Unintended Proxy or Intermediary), reflecting how the agent itself becomes an unwitting intermediary for the attacker.

CrowdStrike's Falcon AIDR is referenced in the source material as a detection and response capability addressing this attack surface. The research does not attribute these techniques to specific threat actors or active campaigns; the taxonomy describes technique categories. Organizations should treat the absence of confirmed in-the-wild exploitation as a planning window, not a safety signal.

Action Checklist

1. Step 1: Assess AI agent exposure, inventory all deployed AI agents and document every privileged integration each agent holds (email, CRM, shell, database, API). Map the blast radius if any agent were hijacked. Prioritize agents with write or execute access.
2. Step 2: Apply least-privilege to AI agent permissions, enforce NIST AC-6 (Least Privilege) by removing or scoping agent permissions to the minimum required for each task. An agent that reads CRM records should not also hold shell execution rights. Align with CIS 5.4 (Restrict Administrator Privileges to Dedicated Administrator Accounts) for any agent operating with elevated entitlements.
3. Step 3: Implement input validation and prompt boundary enforcement, treat all externally sourced content entering an agent's context window (forwarded emails, pasted records, API responses) as untrusted input per NIST SI controls and CWE-20 (Improper Input Validation). Enforce strict separation between user-tier and system-tier instructions at the architecture layer.
4. Step 4: Enable behavioral monitoring on agent actions, configure logging for all agent-initiated actions (outbound email, file writes, API calls, shell commands) per NIST AU-2 (Event Logging) and AU-12 (Audit Record Generation), and CIS 8.2 (Collect Audit Logs). Alert on anomalous action sequences, especially those triggered by external content ingestion.
5. Step 5: Update your threat model to include the AI prompt layer, add prompt injection as an explicit attack vector in your threat register. Reference CrowdStrike's expanded taxonomy and map relevant techniques to your deployed agent architectures. Brief leadership on the operational risk of agentic AI with privileged access.
6. Step 6: Evaluate detection tooling, assess whether your current SIEM, EDR, or AI security stack provides visibility into prompt-layer behavior. Review CrowdStrike Falcon AIDR and comparable capabilities against the 18 newly documented technique categories.
7. Step 7: Monitor CrowdStrike and MITRE for taxonomy updates, track follow-on publications from CrowdStrike's AI research team and watch for MITRE ATT&CK updates incorporating prompt injection sub-techniques as the framework evolves to cover AI-native attack surfaces.

IR / Forensic Enrichment

Triage Priority	URGENT
Escalation Criteria	Escalate immediately if any AI agent audit log shows an action sequence (outbound email send, file write, shell execution, CRM record modification) that was triggered within a short time window of ingesting externally sourced content — particularly forwarded emails or third-party API responses — as this pattern is consistent with successful prompt injection exploitation of an agent holding write or execute privileges; also escalate if PII or regulated data (PHI, PCI-scope records) is accessible to any agent lacking prompt boundary enforcement, as unauthorized disclosure may trigger breach notification obligations.
Recovery Notes	After containment and permission scoping are complete, do not restore full agent operation until a test harness has been run against the agent using representative samples of the 18 documented injection technique categories to verify that prompt boundary enforcement is blocking malicious instruction injection. Monitor agent action logs continuously for at least 30 days post-recovery, with particular attention to anomalous action sequences involving external content ingestion followed by privileged API or shell calls. If CrowdStrike Falcon AIDR or equivalent AI-native detection is deployed, review its alert queue daily during this observation window and tune alert thresholds based on observed baseline agent behavior.
Forensic Artifacts	AI agent API gateway or proxy logs capturing full request payloads (including the assembled prompt context window) at inference time — these are the primary forensic record of what instruction content was present when a suspicious agent action was executed, and are the only artifact that can reveal injected instructions embedded in externally sourced content such as forwarded emails or CRM record fields | OAuth token and API key audit logs from the identity provider (e.g., Azure AD Sign-in logs, Google Workspace Token audit, Salesforce Connected App usage logs) for the agent's service identity — these record what privileged actions the agent account performed, when, and from which IP, enabling reconstruction of the blast radius if the agent was successfully hijacked via prompt injection | Outbound email send logs from the email gateway (e.g., Exchange transport logs, Google Workspace Gmail log search, or Postfix mail log) filtered on the agent's sending identity — prompt injection against email-integrated agents frequently targets outbound exfiltration or spear-phishing relay, and these logs capture the injected action's output | Sysmon Event ID 1 (Process Creation) records on the agent host, filtered on processes spawned by the agent runtime (e.g., Python interpreter, Node.js, or container entrypoint) — a dormant or fragmented prompt injection payload that ultimately executes shell commands will appear here as an anomalous child process, providing a timestamped record of when the injected instruction reached execution | LLM platform inference logs or model response logs (e.g., OpenAI API usage logs, Vertex AI audit logs, or self-hosted model server logs) capturing the model's output at the time of the suspicious action — these preserve the agent's response that triggered the downstream privileged action, which may contain the injected instruction's directive and is critical for determining whether a specific CrowdStrike-documented technique category was used

Per-Action IR Details

Step 1: Assess AI agent exposure — inventory all deployed AI agents and document every privileged integration each agent holds (email, CRM, shell, database, API). Map the blast radius if any agent were hijacked. Prioritize agents with write or execute access.

NIST Phase: Preparation

Reference: NIST 800-61r3 §2 — Preparation: establishing IR capability includes maintaining an accurate asset inventory and understanding which systems are high-value targets before an incident occurs.

Controls: NIST AC-2 (Account Management) — governs defining and documenting account types including non-human service accounts and agent identities with privileged integrations, CIS 1.1 (Establish and Maintain Detailed

Enterprise Asset Inventory) — requires an accurate inventory of all enterprise assets with potential to store or process data, including AI agent deployments, CIS 5.1 (Establish and Maintain an Inventory of Accounts) — requires tracking all accounts including service and API accounts held by AI agents across email, CRM, shell, and database integrations

Compensating: Run a script querying your identity provider (Azure AD / Okta via CLI or Graph API) for all service principals and OAuth app registrations, then cross-reference with email gateway rules, CRM connected-app lists, and cron jobs or task scheduler entries that invoke LLM APIs. A two-person team can cover this in a half-day using `az ad sp list --all` (Azure) or `gcloud iam service-accounts list` (GCP) plus manual review of integration dashboards.

Evidence: Before this step alters no live state, but document current OAuth token scopes and API key permissions for each agent identity NOW — screenshot or export the permission grants from each integration console (e.g., Google Workspace OAuth scope grants, Salesforce Connected App permissions, AWS IAM role policies attached to Lambda/agent runtime). These snapshots establish a pre-remediation baseline. If an agent is already suspected of compromise, capture active outbound connection state via `netstat -ano` or `ss -tulnp` on the agent host before proceeding.

Step 2: Apply least-privilege to AI agent permissions — enforce NIST AC-6 (Least Privilege) by removing or scoping agent permissions to the minimum required for each task. An agent that reads CRM records should not also hold shell execution rights. Align with CIS 5.4 (Restrict Administrator Privileges to Dedicated Administrator Accounts) for any agent operating with elevated entitlements.

NIST Phase: Containment

Reference: NIST 800-61r3 §3.3 — Containment Strategy: reducing the blast radius of a compromised or manipulated agent by restricting what actions it can take is a containment action that limits lateral movement and privilege abuse enabled by prompt injection.

Controls: NIST AC-6 (Least Privilege) — directly governs this step; enforces that agents receive only the minimum permissions necessary for their defined task scope, NIST AC-3 (Access Enforcement) — governs enforcement of approved authorizations for logical access, applicable to agent API tokens and service account permissions, CIS 5.4 (Restrict Administrator Privileges to Dedicated Administrator Accounts) — requires restricting elevated entitlements to dedicated accounts, applicable to any AI agent currently holding admin-tier API or shell rights

Compensating: For each agent, export the current permission set, then create a scoped replacement credential (narrow OAuth scope, read-only API key, restricted IAM role). Revoke the over-privileged credential only AFTER the replacement is validated. Use AWS IAM Access Analyzer, GCP IAM Recommender (both free), or manual scope-down of OAuth app permissions in the identity provider console. Document changes in a change log with timestamps.

Evidence: BEFORE revoking any existing agent credentials or API keys, capture: (1) current active sessions or token grants for the agent identity from the IdP audit log (e.g., Google Workspace Admin > Reports > Token activity, or Azure AD Sign-in logs filtered on the service principal); (2) any in-flight agent action logs from the past 72 hours to establish whether the over-privileged credential was already abused — look for shell execution events, mass CRM reads, or outbound email sends initiated by the agent identity. Revoking the credential before this capture destroys evidence of potential prior exploitation.

Step 3: Implement input validation and prompt boundary enforcement — treat all externally sourced content entering an agent's context window (forwarded emails, pasted records, API responses) as untrusted input per NIST SI controls and CWE-20 (Improper Input Validation). Enforce strict separation between user-tier and system-tier instructions at the architecture layer.

NIST Phase: Containment

Reference: NIST 800-61r3 §3.3 — Containment Strategy: architectural enforcement of prompt boundaries prevents externally injected instructions from reaching system-tier execution, directly limiting the attack surface exploited by the 18 documented prompt injection techniques.

Controls: NIST AC-4 (Information Flow Enforcement) — governs enforcing approved authorizations for controlling the flow of information within systems; applies here to enforcing separation between untrusted external content flows (email, API responses) and the trusted system-prompt instruction tier, NIST AC-25 (Reference Monitor) — governs implementing a tamperproof, always-invoked reference monitor for access control policies; maps directly to enforcing a prompt boundary layer that intercepts and validates all content before it reaches the agent's instruction context

Compensating: Deploy an open-source prompt firewall or input sanitization layer (e.g., the Rebuff project on GitHub, or a custom middleware function) in front of the agent's context assembly step that strips or escapes known injection patterns (role-switching directives, instruction-override tokens, base64-encoded payloads). Maintain a YARA rule set targeting strings such as 'ignore previous instructions', 'system:', 'INST', and common delimiter abuse patterns found across CrowdStrike's documented technique categories. Apply the rules to all externally sourced strings before they are concatenated into the agent prompt.

Evidence: This step modifies agent architecture, not live host state; volatile evidence capture is not required prior to implementation. However, before deploying boundary enforcement, collect and preserve a sample of the agent's recent raw input logs (the full context window contents at time of inference, if logged by the LLM platform or API gateway) to establish a pre-enforcement baseline and identify whether any historically processed inputs contain injection patterns. For agents using OpenAI or Gemini APIs, check your API gateway or proxy logs for anomalous system-role content injected via user-tier messages.

Step 4: Enable behavioral monitoring on agent actions — configure logging for all agent-initiated actions (outbound email, file writes, API calls, shell commands) per NIST AU-2 (Event Logging) and AU-12 (Audit Record Generation), and CIS 8.2 (Collect Audit Logs). Alert on anomalous action sequences, especially those triggered by external content ingestion.

NIST Phase: Detection Analysis

Reference: NIST 800-61r3 §3.2 — Detection and Analysis: establishing telemetry on agent-initiated actions enables detection of post-injection behavioral anomalies, such as an agent sending outbound email or executing shell commands immediately after ingesting an externally sourced document — the core execution pattern for agentic prompt injection.

Controls: NIST AU-2 (Event Logging) — requires identifying and logging event types the system is capable of capturing; applies to configuring agent action logs across outbound email, API calls, file writes, and shell invocations, NIST AU-12 (Audit Record Generation) — requires generating audit records for the events identified in AU-2; applies to ensuring agent runtime logs capture the action, timestamp, triggering input source, and agent identity for each executed action, CIS 8.2 (Collect Audit Logs) — requires enabling audit logging across enterprise assets; applies to ensuring agent hosting infrastructure and integration endpoints are included in the enterprise log collection scope

Compensating: On agent host systems, deploy Sysmon with a configuration that logs process creation (Event ID 1), network connections (Event ID 3), and file creation (Event ID 11) for the agent runtime process. Write a Sigma rule that fires when the agent process initiates an outbound SMTP connection (port 25/465/587) or spawns a shell (cmd.exe, bash, powershell.exe) within 60 seconds of writing a new file to its input staging directory — this correlation approximates 'external content ingested → privileged action executed'. Forward Sysmon XML logs to a central syslog server or aggregate via Winlogbeat (free, Elastic).

Evidence: This step establishes monitoring; it does not alter live agent state. Before enabling new logging, verify whether the agent platform already generates any action logs and preserve their current state (do not overwrite or rotate). If the agent runs in a container, capture the current container stdout/stderr log stream (`docker logs ` or `kubectl logs`) before reconfiguring the logging driver, as reconfiguration may flush the buffer.

Step 5: Update your threat model to include the AI prompt layer — add prompt injection as an explicit attack vector in your threat register. Reference CrowdStrike's expanded taxonomy and map relevant techniques to your deployed agent architectures. Brief leadership on the operational risk of agentic AI with privileged access.

NIST Phase: Post Incident

Reference: NIST 800-61r3 §4 — Post-Incident Activity: updating the threat model and threat register with newly documented attack techniques from CrowdStrike's 200+ pattern taxonomy constitutes the lessons-learned and policy-improvement activity that strengthens the organization's detection and response posture against future prompt injection campaigns.

Controls: NIST AU-6 (Audit Record Review, Analysis, and Reporting) — requires analyzing system audit records for indications of inappropriate activity and reporting findings to defined personnel; maps to the leadership briefing component and formalizing prompt injection as a monitored threat category

Compensating: Use the OWASP Top 10 for LLM Applications (free, publicly available) as a structured threat taxonomy if CrowdStrike's full report is not accessible. Map each deployed agent's external input surfaces (email ingest, web retrieval, API responses, user chat) against the OWASP LLM01 (Prompt Injection) and LLM02 (Insecure Output Handling) entries. Document the mapping in a one-page risk register entry that includes the agent name, privileged integrations, injection surface, and current control gaps — this gives leadership a concrete blast-radius picture without requiring enterprise tooling.

Evidence: No live system state is altered by this step; volatile evidence capture is not applicable. Preserve the current state of the threat register as a versioned snapshot before updating it, so the delta introduced by this change is auditable during future post-incident reviews or compliance assessments.

Step 6: Evaluate detection tooling — assess whether your current SIEM, EDR, or AI security stack provides visibility into prompt-layer behavior. Review CrowdStrike Falcon AIDR and comparable capabilities against the 18 newly documented technique categories.

NIST Phase: Preparation

Reference: NIST 800-61r3 §2 — Preparation: evaluating and closing gaps in detection tooling before an incident occurs is a core preparation activity; the specific gap here is prompt-layer visibility, which traditional SIEM and EDR do not provide natively against CrowdStrike's documented agentic injection techniques.

Controls: CIS 7.1 (Establish and Maintain a Vulnerability Management Process) — requires a documented vulnerability management process; applies here to formally assessing detection coverage gaps for prompt injection as a newly documented vulnerability class affecting deployed AI agents

Compensating: Without commercial AI security tooling, build a detection gap matrix in a spreadsheet: list each of the 18 newly documented technique categories from the CrowdStrike research, then mark each column with your current visibility (SIEM rule exists / Sysmon log available / no coverage). For techniques in the 'no coverage' column, write a Sigma rule targeting the closest available proxy signal — for example, a dormant payload technique that eventually calls a shell can be detected via Sysmon Event ID 1 even if the injection itself is not visible. Treat the gap matrix as a living document reviewed quarterly.

Evidence: This step involves no alteration of live system state; volatile capture is not required. Before conducting the tooling evaluation, export your current SIEM alert rule inventory and EDR policy configuration as a baseline snapshot so that any coverage changes made after the evaluation are auditable.

Step 7: Monitor CrowdStrike and MITRE for taxonomy updates — track follow-on publications from CrowdStrike's AI research team and watch for MITRE ATT&CK updates incorporating prompt injection sub-techniques as the framework evolves to cover AI-native attack surfaces.

NIST Phase: Post Incident

Reference: NIST 800-61r3 §4 — Post-Incident Activity: integrating external threat intelligence from CrowdStrike's AI research publications and MITRE ATT&CK framework updates into the organization's ongoing IR posture constitutes the intelligence-sharing and continuous-improvement activity that keeps detection capabilities current against a rapidly evolving prompt injection technique set.

Controls: NIST AU-13 (Monitoring for Information Disclosure) — requires monitoring open-source information and information sites at a defined frequency; maps to the structured monitoring of CrowdStrike and MITRE publication feeds for prompt injection taxonomy updates affecting deployed AI agent architectures

Compensating: Subscribe to the CrowdStrike blog RSS feed and MITRE ATT&CK GitHub repository release notifications (free, via GitHub Watch > Releases) using a shared team email alias. Assign one team member a bi-weekly 30-minute task to review new entries, flag any technique that maps to a deployed agent's input surface or integration, and update the threat register entry created in Step 5. This ensures a two-person team maintains current intelligence coverage without dedicated threat intelligence platform spend.

Evidence: No live system state is altered by this step; volatile evidence capture is not applicable. Archive each CrowdStrike research publication and MITRE ATT&CK release changelog reviewed as a dated PDF or HTML snapshot in your IR documentation repository to preserve the version of the taxonomy your controls were mapped against — this is critical for demonstrating due diligence during post-incident or audit review if a prompt injection incident occurs after a taxonomy update was published.

Detection Guidance

No specific IOCs (hashes, domains, IPs) are present in the source material for this taxonomy research. Detection is behavioral. Security teams should focus on the following:

Agent Action Anomalies: Monitor for AI agent-initiated actions that are disproportionate to or inconsistent with the triggering user request, particularly outbound email sends, file writes, shell command execution, or API calls to external services. These map to T1059 (Command and Scripting Interpreter) and T1567 (Exfiltration Over Web Service). Log all agent-initiated actions per NIST AU-2 and CIS 8.2.

Delayed or Condition-Triggered Behavior: Watch for agent actions that occur significantly after the input that logically should have triggered them, or that fire only when specific environmental conditions are met. This is the behavioral signature of delayed trigger activation. Standard prompt-time evaluation will miss these; runtime behavioral monitoring is required.

Context Boundary Violations: Alert when external content (forwarded emails, pasted text, API-retrieved records) causes the agent to take privileged actions not explicitly authorized by the user in the current session. This is the fingerprint of unwitting user context injection and boundary spoofing (T1078, CWE-441).

Fragmented Instruction Patterns: Hunt for sequences of individually innocuous agent inputs that, when assembled, form a coherent instruction chain with a privileged outcome. This requires session-level context aggregation, not per-input evaluation. Map to T1027 (Obfuscated Files or Information) and CWE-20.

Privilege Escalation via Agent: Audit logs for cases where agent-executed actions exceed the permission scope of the authenticated user who initiated the session, a direct indicator of system-authority boundary spoofing.

Log Sources to Prioritize: Agent orchestration layer logs, outbound email gateway logs, CRM audit trails, shell/command execution logs, and any API gateway logs for services the agent can reach. Align retention with NIST AU-11 (Audit Record Retention).

Apply D3FEND countermeasures D3-UAP (User Account Permissions) to scope agent entitlements, D3-LAM (Local Account Monitoring) to track agent account activity, and D3-MFA (Multi-factor Authentication) on any human-facing escalation paths the agent can trigger.

Framework Mappings

MITRE-ATTACK

- **T1078** — Valid Accounts
- **T1557** — Adversary-in-the-Middle
- **T1059** — Command and Scripting Interpreter
- **T1027** — Obfuscated Files or Information
- **T1566** — Phishing
- **T1485** — Data Destruction
- **T1190** — Exploit Public-Facing Application

NIST-800-53R5

- **AC-2** — Account Management

- **AC-6** — Least Privilege
- **IA-2** — Identification and Authentication (Organizational Users)
- **IA-5** — Authenticator Management
- **CM-7** — Least Functionality
- **SI-3** — Malicious Code Protection
- **SI-4** — System Monitoring
- **SI-7** — Software, Firmware, and Information Integrity
- **AT-2** — Literacy Training and Awareness
- **CA-7** — Continuous Monitoring
- **SC-7** — Boundary Protection
- **SI-8** — Spam Protection
- **CA-8** — Penetration Testing
- **RA-5** — Vulnerability Monitoring and Scanning
- **SI-2** — Flaw Remediation
- **SI-10** — Information Input Validation

OWASP-TOP10-2021

- **A03:2021** — Injection

CIS-V8

- **16.10** — Apply Secure Design Principles in Application Architectures
- **8.2** — Collect Audit Logs

ISO-27001-2022

- **A.8.26** — Application security requirements
- **A.5.34** — Privacy and protection of personal information

NIST-CSF-2

- **DE.CM-01** — Networks and network services are monitored

MITRE ATT&CK Mapping

Technique ID	Technique Name	Tactic
T1078	Valid Accounts	Defense-Evasion
T1557	Adversary-in-the-Middle	Credential-Access
T1059	Command and Scripting Interpreter	Execution
T1027	Obfuscated Files or Information	Defense-Evasion
T1566	Phishing	Initial-Access
T1485	Data Destruction	Impact

Technique ID	Technique Name	Tactic
T1190	Exploit Public-Facing Application	Initial-Access

Sources

Source	URL	Tier
Blog	https://www.crowdstrike.com/en-us/blog/crowdstrike-uncovers-new-pro...	T1
Cyberpress	https://cyberpress.org/crowdstrike-5-new-prompt-injection-techniques/	T3
CrowdStrike	https://www.crowdstrike.com/en-us/blog/crowdstrike-named-leader-gar...	T1
CrowdStrike	https://www.crowdstrike.com/en-us/blog/crowdstrike-secures-growing-...	T1
CrowdStrike Innovations Secure AI Agents and Govern ...	https://www.crowdstrike.com/en-us/blog/new-crowdstrike-innovations-...	T1

Generated 2026-07-12 14:16 UTC by TJS Security Command Center