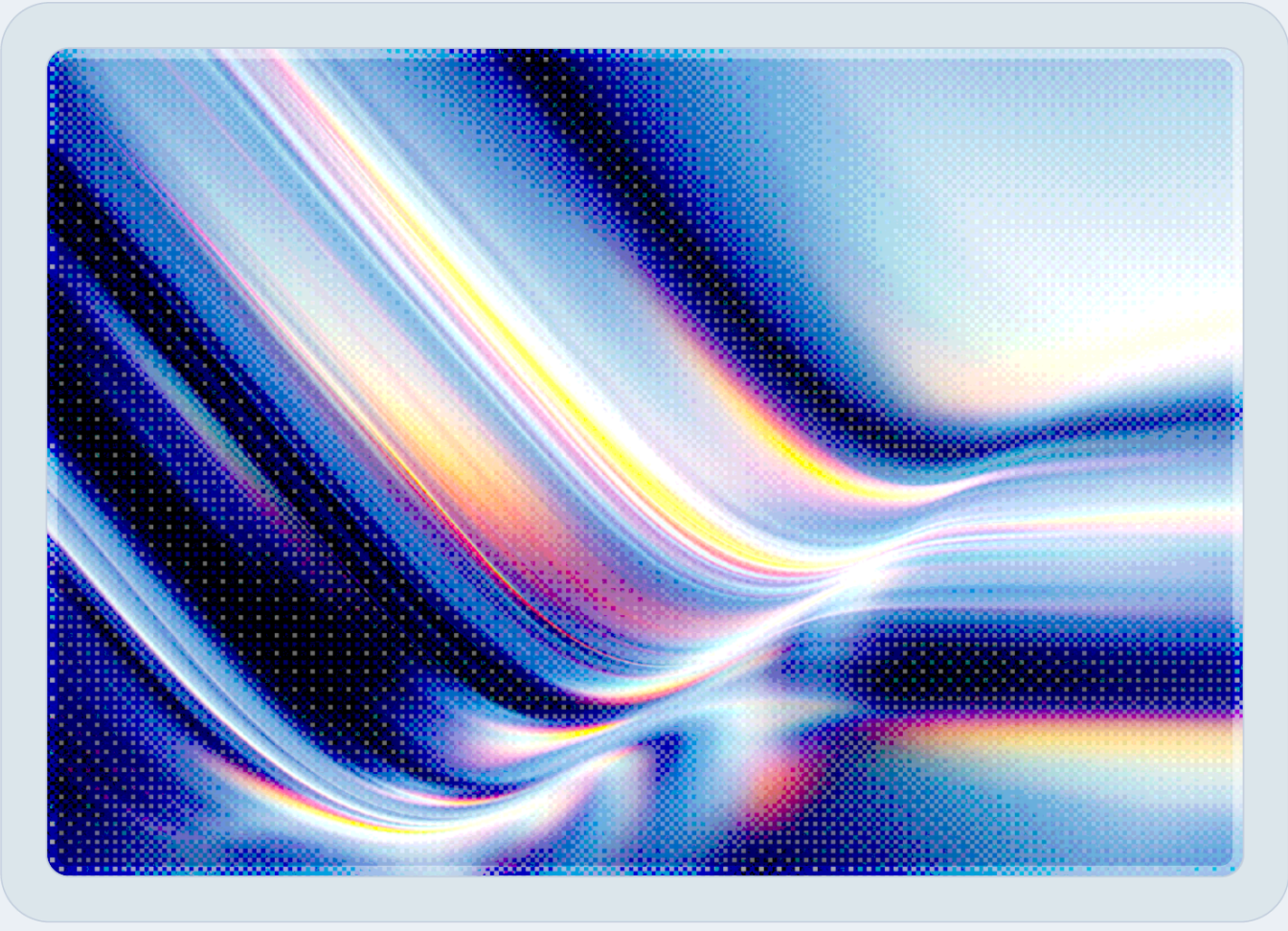
CSAI Foundation | Cloud Security Alliance

Copirate 365: M365 Copilot Command Injection at Scale

Security Guidance for CVE-2026-24299

2026-05-05

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

CVE-2026-24299 is a medium-severity command injection vulnerability (CVSS 3.1: 5.3; AV:N/AC:H/PR:N/UI:R) in Microsoft 365 Copilot that allows a network-based attacker to disclose sensitive organizational information without requiring authentication, though exploitation requires high attack complexity and user interaction. Although the CVSS score places this in the medium tier, the deployment footprint of M365 Copilot across large enterprise environments—where the AI assistant routinely ingests emails, documents, SharePoint content, and Teams conversations—means the effective blast radius of successful exploitation is substantially larger than the score alone suggests.

- Microsoft patched CVE-2026-24299 and no confirmed in-the-wild exploitation has been publicly disclosed as of this writing, but the vulnerability was reserved in January 2026 and publicly disclosed in March 2026, creating a window during which unpatched tenants were exposed.
- The underlying weakness (CWE-77: Improper Neutralization of Special Elements Used in a Command) reflects a pattern of insufficient input sanitization in AI processing pipelines—a structural concern that extends well beyond this single CVE to the broader class of enterprise AI assistants that synthesize content from untrusted sources.
- CVE-2026-24299 is the third significant information-disclosure vulnerability disclosed against M365 Copilot within twelve months, following EchoLeak (CVE-2025-32711, CVSS 9.3) [1] and Reprompt [2]; a fourth related disclosure during this period, CVE-2026-21520, targeted Copilot Studio rather than M365 Copilot proper. This cadence indicates a systemic weakness across the Copilot product family rather than a series of isolated defects, and warrants a strategic security posture review in addition to patching.
- Organizations should audit Copilot access scopes, enforce sensitivity labels, and review Copilot interaction logs for anomalous query patterns, even in tenants that have confirmed patch application.

Background

Microsoft 365 Copilot is an AI assistant integrated across the Microsoft 365 productivity suite—Word, Excel, PowerPoint, Outlook, Teams, and SharePoint—with access to an organization's entire corpus of content through Microsoft Graph [3]. As of 2026, M365 Copilot has achieved broad commercial adoption at enterprise scale across organizations worldwide. The AI assistant operates on a Retrieval-Augmented Generation (RAG) architecture, meaning it dynamically retrieves content from the user's organizational data stores and synthesizes responses by feeding that retrieved content alongside the user's query into a foundation model. This architecture is efficient and capable, but it also means the system's context window regularly contains sensitive organizational data—financial records, strategic plans, personnel communications, and privileged legal materials—that would normally be compartmentalized in traditional applications.

CVE-2026-24299 was reserved by MITRE on January 21, 2026, and publicly disclosed in the Microsoft Security Response Center advisory published March 19, 2026, with a last-updated date of April 14, 2026 [4]. The vulnerability is classified under CWE-77 (Improper Neutralization of Special Elements Used in a Command), the command injection category. The CVSS 3.1 base score of 5.3 reflects a network attack vector with high attack complexity, no privileges required, user interaction required, and a high confidentiality impact with no integrity or availability impact [4]. The score's "medium" classification is somewhat misleading in enterprise context: a single successful exploitation event against a Copilot session that has been querying across financial forecasts, M&A documents, or HR data could expose far more sensitive material than a high-severity vulnerability against a more narrowly scoped system.

We use the label "Copirate 365" in this document to characterize this vulnerability and, more broadly, the class of techniques that exploit M365 Copilot's cross-corpus access to "pirate" data across organizational boundaries. The label captures something important: the concern is not merely that an attacker can extract data from a single document, but that Copilot's access model creates a natural aggregation surface where a single injection point can pivot across an entire organizational data estate.

Security Analysis

Vulnerability Mechanism

CVE-2026-24299 stems from the improper neutralization of special elements within M365 Copilot's command processing pipeline. In a command injection vulnerability of this class, attacker-controlled content—embedded in an email, document, SharePoint page, or other resource that Copilot is directed

to process—contains special characters or instruction sequences that are not properly sanitized before being passed to a downstream processing component. When these sequences reach the processing layer without neutralization, they can redirect or modify system behavior, in this case causing the AI assistant to disclose information it would not return through a properly structured query. The following mechanism account reflects the CWE-77 classification and the documented behavior of analogous Copilot injection vulnerabilities; a detailed public technical analysis of CVE-2026-24299's specific exploitation path had not been published as of this writing.

The attack requires user interaction, which explains the "UI:R" component of the CVSS vector. In practice, this means the attack chain typically begins with a victim user opening or summarizing a document or message that contains the injected payload. This is a meaningful constraint on fully automated exploitation, but it is a weak one in enterprise environments where knowledge workers regularly use Copilot to summarize inbound email, process shared documents from external parties, and synthesize content from SharePoint sites accessible to broad organizational audiences. Phishing content, externally shared documents, and publicly indexed SharePoint sites all represent attacker-controlled input vectors that require only routine user behavior to trigger.

The high attack complexity ("AC:H") rating reflects that exploitation requires specific conditions to be met, likely relating to the specific context or query pattern that activates the injection. However, high complexity does not mean impractical: researchers have demonstrated that AI injection attacks in M365 Copilot can be made highly reliable once the triggering conditions are understood, as seen in the detailed technical analysis of EchoLeak and Reprompt [5][6].

Enterprise Exposure Context

To understand why a CVSS 5.3 vulnerability demands immediate attention, it is necessary to consider the deployment model of M365 Copilot in enterprise environments. Unlike traditional applications where a vulnerability exposes the data directly managed by that application, Copilot's RAG architecture creates a dynamic aggregation problem. The AI assistant has access, by design, to the full scope of content accessible to the authenticated user through Microsoft Graph. In practice, this includes mailboxes, OneDrive files, SharePoint libraries, Teams conversations, and calendar data. A successful injection that causes Copilot to disclose content from its retrieval context does not expose one file—it potentially exposes whatever content the assistant loaded into its context window for the processing session.

This aggregation surface distinguishes AI assistant vulnerabilities from prior-generation information disclosure bugs. A traditional information disclosure vulnerability might leak one record from one database. An AI assistant vulnerability during a synthesis session can leak a substantial number of retrieved documents from a corpus spanning multiple sensitivity classifications, because the entire retrieval result set exists in the AI's processing context at the moment of exploitation. A DLP label bypass

vulnerability disclosed separately in February 2026 [7] illustrated this dynamic concretely: Copilot was observed summarizing emails protected by confidentiality sensitivity labels, effectively allowing the AI's output to carry content that the underlying sensitivity controls were designed to prevent from leaving its designated container.

The timing gap between CVE reservation (January 21, 2026) and public disclosure (March 19, 2026) is also worth noting. During this period, organizations had no published guidance to act on. Enterprises operating M365 Copilot at scale should treat this gap as a design consideration for their monitoring posture: anomaly detection against Copilot interaction logs, rather than waiting for vendor disclosure, is the primary proactive coverage mechanism available during the responsible disclosure window, short of restricting Copilot access or scope entirely.

Pattern Context: A Recurring Vulnerability Class

CVE-2026-24299 is the most recent in a sequence of disclosed vulnerabilities affecting M365 Copilot's data handling. EchoLeak (CVE-2025-32711), disclosed in mid-2025 with a CVSS score of 9.3, demonstrated a zero-click exploitation path in which malicious instructions embedded in an email caused Copilot to automatically exfiltrate sensitive content from its retrieval context via image URL references, bypassing multiple security controls including XPIA classifiers, external link redaction, and Content Security Policy [1][5]. Reprompt, disclosed in early 2026, demonstrated a single-click exploitation path that hijacked an authenticated Copilot session and extracted calendar data, file access history, and conversation content through a double-request bypass technique [2][6]. CVE-2026-21520, a CVSS 7.5 prompt injection in Copilot Studio disclosed in January 2026, showed how similar injection techniques could override agent system instructions and direct the AI to query connected SharePoint Lists and exfiltrate customer data via Outlook [8].

These four disclosures over a twelve-month period—three targeting M365 Copilot directly and one targeting the related Copilot Studio product—are consistent with a systemic weakness in input validation and scope enforcement across the M365 Copilot product family, rather than a series of isolated defects. Each vulnerability used a different technical mechanism, but each exploited the same fundamental design characteristic: that Copilot systems process content from untrusted or semi-trusted sources and combine it with retrieval context containing sensitive organizational data, without consistently enforcing boundaries between the two. The pattern strongly suggests that organizations should not treat each disclosure as an isolated event requiring a discrete patch response, but as evidence of an underlying architectural exposure that demands sustained defensive investment.

Recommendations

Immediate Actions

Organizations running M365 Copilot should verify patch status and confirm that the April 14, 2026 remediation for CVE-2026-24299 has been applied to their tenants. Because M365 Copilot is a cloud-delivered service, most remediation is applied by Microsoft without direct customer action; however, organizations should confirm through their Microsoft 365 admin center or security portal that their tenant reflects the current service version. Security teams should also pull Copilot interaction logs for the period between January 21 and April 14, 2026, and review for anomalous query patterns—unusual breadth of document retrieval, queries against high-sensitivity content from unexpected users, or output patterns inconsistent with normal business use.

If sensitivity labels have not been applied to high-value content categories (executive communications, M&A materials, legal privilege documents, financial forecasts), organizations should treat that gap as an urgent remediation item independent of this specific CVE. Sensitivity labels, when properly enforced, limit the content that flows into Copilot's retrieval context and reduce the blast radius of any injection vulnerability that causes unauthorized disclosure. The February 2026 DLP bypass incident demonstrates that label enforcement is not a complete mitigation—but it remains one of the most effective scope-reduction controls available at the data layer [7].

Short-Term Mitigations

In the near term, security teams should review Copilot access provisioning against a least-privilege model. Copilot inherits the full permissions of the authenticated user, meaning that over-permissioned service accounts, shared mailboxes with broad access grants, and users with access to sensitive SharePoint libraries that exceeds their operational need all represent elevated risk surfaces. A targeted permission-reduction exercise for users with access to the most sensitive content categories limits the damage from any future injection or exfiltration vulnerability.

Organizations should also evaluate their Copilot deployment configuration for the processing of externally sourced content. Email summarization, document processing for content from external parties, and Copilot access to SharePoint sites indexed by external-facing portals all represent paths by which attacker-controlled input can reach Copilot's processing pipeline. Restricting Copilot access to processing externally sourced content, or routing such content through human review before AI

summarization, reduces the attacker-controlled input surface meaningfully. Microsoft's guidance on defending against indirect prompt injection attacks offers a vendor-perspective framework for understanding where these controls can be applied [9].

Strategic Considerations

The recurring disclosure pattern across M365 Copilot vulnerabilities argues for incorporating enterprise AI assistants into the organization's vulnerability management program as a distinct asset class, not as a subcategory of SaaS application security. Traditional SaaS vulnerability management focuses on patch tracking and configuration audits against vendor benchmarks. AI assistant security requires additional controls: ongoing monitoring of AI interaction patterns for anomalous behavior, red-team exercises that specifically target prompt injection and context extraction, and governance policies that specify which data categories may flow through AI processing pipelines and under what conditions.

Security architecture reviews for M365 Copilot deployments should include explicit threat modeling against the MAESTRO framework's seven-layer model for agentic AI systems [10]. Layer 2 (Data Operations) and Layer 4 (Agent Frameworks) in MAESTRO's hierarchy are directly relevant here: Layer 2 addresses data ingestion and retrieval security, including the risk of injected content in data pipelines; Layer 4 addresses the framework-level constraints that govern what an agent is permitted to retrieve and disclose. Mapping CVE-2026-24299's mechanism to these layers and identifying organizational-level mitigations at each layer provides a more durable security foundation than patch-and-monitor alone.

Longer-term, organizations deploying AI assistants at enterprise scale should engage vendors on architectural transparency. The frequency of disclosed vulnerabilities in M365 Copilot's content processing pipeline suggests that the attack surface has not yet been fully characterized by external researchers—and organizations should not assume that vendor remediation of individual CVEs implies a comprehensive resolution of the underlying vulnerability class. Organizations have legitimate interests in understanding how their AI assistant systems handle untrusted content, what isolation exists between retrieval context and output generation, and how input sanitization is implemented at the model boundary. Vendor transparency on these design decisions should be treated as a procurement and renewal evaluation criterion, not an optional nicety.

CSA Resource Alignment

The security challenges surfaced by CVE-2026-24299 and the broader M365 Copilot vulnerability pattern are directly addressed across several CSA frameworks and research programs.

CSA's MAESTRO framework provides purpose-built threat-modeling vocabulary for agentic AI systems and is directly applicable to analyzing this vulnerability class [10]. Under MAESTRO's taxonomy, the seven-layer architecture for agentic AI identifies specific threat categories at the Data Operations layer (injected content in retrieval pipelines), the Agent Framework layer (insufficient scope enforcement on model output), and the Ecosystem Integration layer (cross-application data flows through Microsoft Graph). Security teams conducting threat modeling of their M365 Copilot deployments should apply MAESTRO to map the attack surface systematically rather than reacting to individual CVE disclosures.

The AI Controls Matrix (AICM), CSA's comprehensive control framework for AI systems, provides the control catalog against which organizations can assess their defensive posture [11]. AICM controls relevant to this vulnerability include those governing input validation and sanitization in AI processing pipelines, data classification and access controls for AI-accessible content stores, and monitoring and logging requirements for AI assistant interactions. Organizations using AICM for AI governance should ensure that their Copilot deployment is in scope for relevant control assessments.

CSA's guidance on securing LLM-backed systems through essential authorization practices addresses the least-privilege and permission scoping concerns directly relevant to Copilot's inherited-access model [12]. The framework's analysis of authorization challenges in RAG-based systems—specifically, the risk that retrieval context pulls content across sensitivity boundaries—maps directly to the data aggregation concern at the center of the Copirate 365 vulnerability class.

The Zero Trust principles documented in CSA's zero trust research provide the architectural foundation for limiting the blast radius of any Copilot vulnerability. A Zero Trust posture applied to the Microsoft 365 data estate—where access is continuously verified, sessions are scoped to minimum necessary content, and lateral movement across data categories is monitored and constrained—reduces the effective impact of any injection technique that relies on Copilot's broad organizational data access [13].

Finally, CSA's AI Organizational Responsibilities guidance addresses the governance dimension: who within an organization is accountable for AI system security, and how AI security considerations should be integrated into existing risk management and compliance programs [14]. The recurring disclosure pattern across M365 Copilot argues that AI assistant security requires named ownership, defined risk appetite, and ongoing monitoring investment at the organizational level—not ad hoc response to vendor advisories.

References

[1] Microsoft Security Response Center. "CVE-2025-32711: Microsoft 365 Copilot Information Disclosure Vulnerability." Microsoft, June 2025.

[2] Varonis Threat Labs. "Reprompt: The Single-Click Microsoft Copilot Attack that Silently Steals Your Personal Data." Varonis, January 2026.

[3] Microsoft. "Security for Microsoft 365 Copilot." Microsoft Learn, 2026.

[4] Microsoft Security Response Center. "Security Update Guide – CVE-2026-24299." Microsoft, March–April 2026.

[5] Varonis Threat Labs. "EchoLeak in Microsoft Copilot: What It Means for AI Security." Varonis, 2025.

[6] The Hacker News. "Researchers Reveal Reprompt Attack Allowing Single-Click Data Exfiltration From Microsoft Copilot." The Hacker News, January 2026.

[7] Cyberpress. "Microsoft 365 Copilot Vulnerability Exposes Sensitive Emails to AI Summarization." cyberpress.org, February 2026.

[8] SentinelOne Vulnerability Database. "CVE-2026-21520: Copilot Studio Information Disclosure." SentinelOne, January 2026.

[9] Microsoft Security Response Center Blog. "How Microsoft Defends Against Indirect Prompt Injection Attacks." Microsoft, July 2025.

[10] Cloud Security Alliance. "Agentic AI Threat Modeling Framework: MAESTRO." CSA Blog, February 2025.

[11] Cloud Security Alliance. "AI Controls Matrix (AICM)." CSA, July 2025.

[12] Cloud Security Alliance. "Securing LLM Backed Systems: Essential Authorization Practices." CSA, August 2024.

[13] Cloud Security Alliance. "Using Zero Trust to Secure Enterprise Information in LLM Environments." CSA, March 2026.

[14] Cloud Security Alliance. "AI Organizational Responsibilities: AI Tools and Applications." CSA, January 2025.