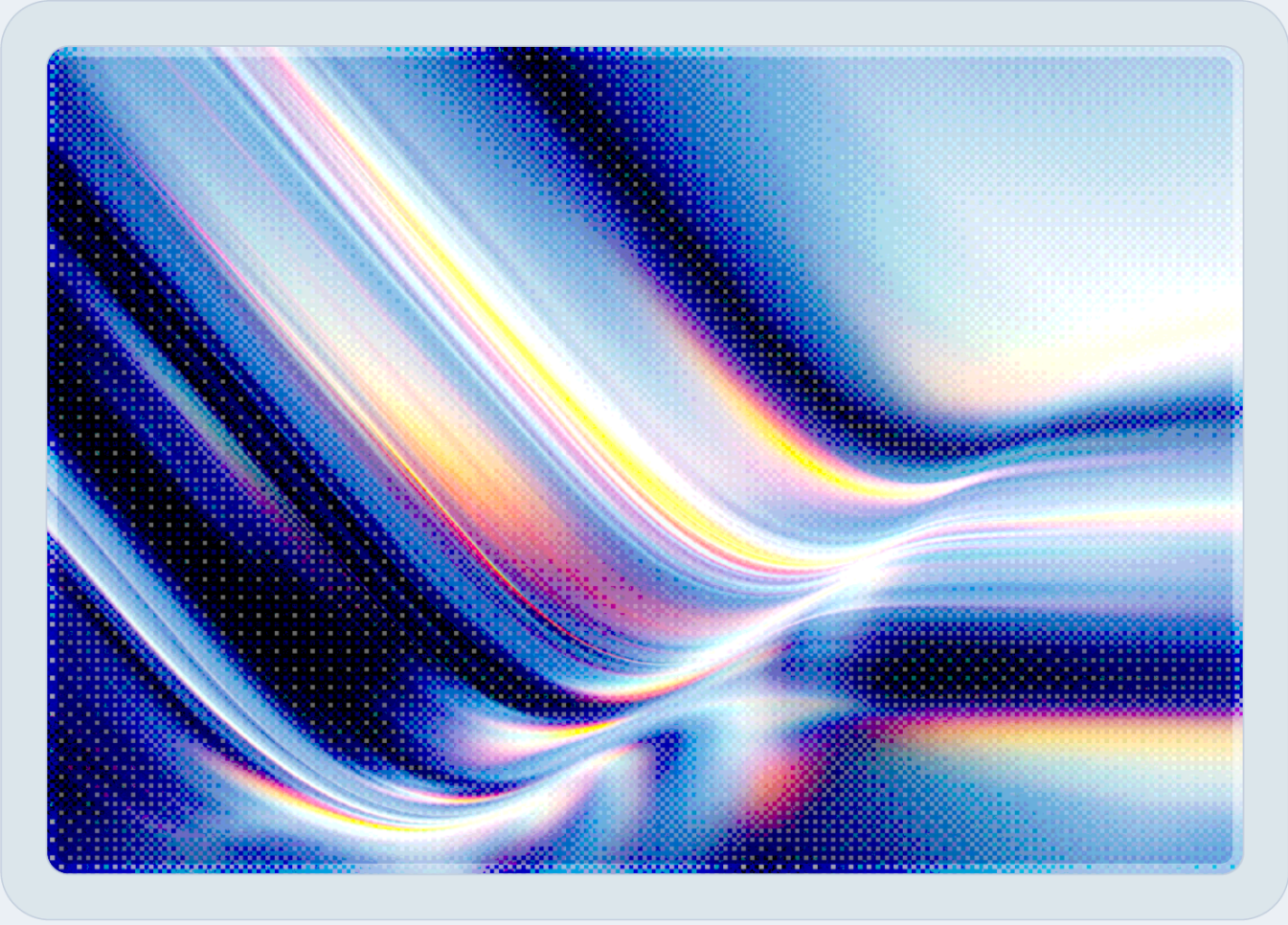
CSAI Foundation | Cloud Security Alliance

# CLOSEDQUORUM: When Malware Lets AI Models Vote

2026-09-26

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

## Key Takeaways

Cisco Talos disclosed CLOSEDQUORUM on September 22, 2026, describing it as the first publicly documented Windows implant to delegate tactical command-and-control decisions to commercial large language models rather than to a human operator or an attacker-controlled server [1][2]. Instead of pulling instructions from infrastructure the attacker owns, the malware queries up to four LLM providers, DeepSeek, Qwen, Mistral, and Google Gemini, and executes whichever action wins a plurality vote, with DeepSeek breaking ties [1][3]. The four available actions cover credential theft, process injection, persistence, and lateral movement, though the lateral-movement handler is non-functional in the samples Talos analyzed [2][4]. Talos also released CAIRN, an open-source toolkit built to hunt for AI-integrated malware by scanning metadata for prompt templates, provider endpoints, and orchestration logic rather than executing suspicious binaries [3]. The publicly circulating sample contains placeholder API keys and cannot function as distributed, and Talos has found no confirmed in-the-wild deployment, but artifacts in the code appear to link the developer to criminal forum activity involving stolen payment card data dating back to 2025 [1][4]. CSA assesses that, for enterprise defenders, CLOSEDQUORUM is less significant for what it currently does than for what it demonstrates is now buildable: malware whose next move is not scripted in advance but generated at runtime by a committee of commercial AI models the defender does not control and cannot subpoena.

## Background

Security researchers have spent much of 2025 and 2026 documenting a steady migration of large language models from developer tooling into the malware supply chain itself [5][6]. Google's Threat Intelligence Group first reported LLM-assisted self-obfuscation in the PROMPTFLUX family in November 2025, and subsequent disclosures showed state-linked and criminal actors wiring live model APIs into malicious code so that payloads could rewrite themselves or generate commands on the fly rather than shipping fixed logic [10]. CLOSEDQUORUM extends that trajectory in a specific and consequential direction: rather than using a single model to obfuscate or generate a piece of code, it uses a panel of four models as a standing decision-making authority for the entire intrusion lifecycle. Talos researcher Ryan Fetterman, who led the disclosure, found the sample through the CAIRN project

rather than through conventional detection, underscoring how well the malware's design resists traditional signature and command-and-control monitoring, whether or not that resistance was the developer's explicit intent [1][3].

The malware itself is a 16.4-megabyte, 64-bit Go binary that Talos describes as an "LLM-as-C2" design [2][3]. On a fixed interval, randomized between five and fifteen minutes, it collects basic host state, hostname, Windows version, and administrator status, and submits that information to each of the four LLM providers under a system prompt that frames the model as "an advanced malware strategist" and instructs it to return one of a small, fixed set of possible actions in a strict JSON format [2][4]. Responses that don't conform to the required schema are discarded outright, and the decision that accumulates the most votes across the four models is the one the implant executes; if the vote ties, DeepSeek's answer prevails, cascading to Qwen, then Mistral, then Gemini if an earlier model in that priority order is unavailable, and to a sleep-and-retry loop if every model fails to return a usable answer [3][4]. Talos noted that despite this AI-driven decision loop, a human operator retains visibility into the process: each decision and the model's stated rationale are logged to a Discord channel, and stolen data is exfiltrated through a separate Discord webhook rather than a dedicated command server [2].

This architecture matters because it weakens a constraint that has historically bounded even sophisticated malware: the need for the attacker to maintain, protect, and periodically task infrastructure that investigators can seize, sinkhole, or subpoena. CLOSEDQUORUM's "infrastructure" is a set of publicly available commercial AI APIs and a Discord webhook, both of which are inexpensive and fast to stand up, replace, and abandon. Talos was explicit that it has no evidence CLOSEDQUORUM has been deployed against real victims, and the sample circulating publicly ships with placeholder API keys and a dummy webhook that render it non-operational as distributed [1][4]. The six SHA256 hashes Talos catalogued span roughly a week of iterative development, and forensic artifacts in the binary connect its author to forum postings about carding and stolen payment-card trading going back to 2025, suggesting an individual financially motivated developer rather than a nation-state program [1][3][4].

## Security Analysis

CLOSEDQUORUM's four action modules are conventional on their own; what is novel is that an external, uncontrolled AI service selects among them at runtime. The "steal" action dumps LSASS process memory for Windows credentials, extracts saved passwords from Chrome, Edge, and Firefox, and searches for cryptocurrency wallet data associated with MetaMask, Exodus, and Ethereum keystores, encrypting whatever it collects with AES-256-GCM before exfiltration [2][4]. The "inject" action delivers code into another process using either PEB-walk process hollowing or Early Bird APC injection, and Talos found that the choice between these two techniques is itself one of the decisions the model panel can

be asked to make, meaning the malware's evasion behavior varies from run to run based on what the LLMs recommend rather than a hardcoded attacker preference [3][4]. The "persist" action establishes redundant survival mechanisms, a Registry Run key, a scheduled task, and a WMI event subscription, while also suppressing Windows event tracing to reduce the telemetry available to defenders [1][2]. The fourth action, "move," is defined in the malware's code but has no working handler in any sample Talos analyzed, which the company reads as evidence the implant is still under active development rather than feature-complete [2][3].

This design creates a genuinely different detection problem than prior AI-assisted malware. Signature and indicator-of-compromise-based defenses, including file hashes, static command sequences, and fixed C2 domains, lose significant value against a payload whose behavioral sequence is not fixed at compile time and whose "infrastructure" is a set of legitimate, constantly-changing commercial API endpoints that ordinary business traffic also uses. CSA's prior research on autonomous agentic AI adversaries observed the same dynamic across a broader set of 2025-2026 incidents: attack chains that once took hours of human-paced decision-making now compress to seconds because a model, not an operator, is making the tactical calls, and defenders who instrument only for known infrastructure or known command syntax are likely to miss the activity [5]. CLOSEDQUORUM sharpens that problem because the decision-making model itself is not embedded in the binary or hosted by the attacker; it is a commercial service the defender has no visibility into and no ability to compel logs from, at least not without legal process directed at the AI provider rather than at the attacker.

Talos's own response, the CAIRN toolkit, is instructive precisely because it does not attempt to detect CLOSEDQUORUM by executing or emulating it. Instead, CAIRN scans binaries and metadata for the fingerprints that an LLM-driven decision loop necessarily leaves behind: DNS queries to known LLM provider endpoints, hardcoded system prompts, API-key-shaped strings, structured JSON response schemas, and orchestration logic connecting those elements together [3][4]. This is a meaningfully different detection posture than the process-lineage and behavioral heuristics that many current endpoint detection and response tools rely on as a primary signal, and it echoes a broader shift CSA has documented in AI-era malware analysis: when the malicious logic lives partly outside the binary, in a model's runtime output rather than in compiled code, defenders need instrumentation aimed at the seams where the binary talks to that external reasoning layer, not just at the process tree the binary spawns [6]. A network or endpoint control that flags outbound connections to consumer or developer-tier LLM APIs from processes that have no legitimate reason to call them, and that logs the request and response content when such calls do occur, would likely have surfaced CLOSEDQUORUM's behavior as anomalous even before Talos named the family, though this has not been independently tested.

The financial and identity signals in the case also deserve attention independent of the technical mechanism. Talos's attribution work links the binary's development artifacts to forum activity around stolen payment-card trading rather than to an established malware-development group or a state-

sponsored unit, which is consistent with a broader trend of individual, moderately resourced actors gaining access to capabilities that previously required more institutional backing [1][5]. The barrier to building an "LLM-as-C2" implant is now primarily conceptual rather than financial: the four models CLOSEDQUORUM queries are all commercially available at modest cost, and the malware's own code shows that a lone developer, working over roughly a week based on the sample timestamps Talos recovered, could assemble a working prototype [1][4]. That combination, a low-cost, publicly documented technique paired with a financially motivated solo developer, is likely to attract imitation faster than it attracts defensive tooling unless enterprises act on the detection surface CAIRN has already mapped out.

# Recommendations

## Immediate Actions

Security teams should treat any outbound API traffic from unmanaged or unexpected processes to commercial LLM provider endpoints, including DeepSeek, Qwen, Mistral, and Gemini, as a candidate indicator worth investigating rather than routine developer activity, particularly on endpoints with no sanctioned business reason to call those services. Incorporate the CAIRN YARA rules and detection logic Talos published into endpoint and network detection pipelines, since they were built specifically to surface the prompt-template, API-key, and provider-endpoint fingerprints this malware family leaves behind [3]. Review Discord egress traffic from endpoints for webhook-based exfiltration patterns, since CLOSEDQUORUM is one of several recent malware families observed using Discord as a low-cost, difficult-to-attribute replacement for dedicated C2 infrastructure.

## Short-Term Mitigations

Extend credential-theft detection to explicitly cover LSASS memory access patterns, browser credential-store reads, and cryptocurrency wallet file access as a correlated cluster rather than as independent low-severity alerts, since CLOSEDQUORUM's "steal" module executes all three simultaneously. Update sandbox and detonation environments to allow longer observation windows and to permit controlled outbound connectivity to commercial AI APIs during analysis, since a sandbox that blocks all external network access will never observe an LLM-directed payload make its runtime decision. Build a standing inventory of which internal processes and services have a legitimate business need to call external LLM APIs, so that anomalous calls from unrelated processes stand out against a known baseline rather than blending into a large volume of routine sanctioned AI usage.

### Strategic Considerations

Enterprises should begin treating "which processes call which AI provider APIs, and why" as a first-class asset-management and governance question, not an afterthought to conventional endpoint inventory, because CLOSEDQUORUM demonstrates that commercial AI infrastructure can now substitute entirely for attacker-owned C2. Incident response and threat-hunting playbooks should be updated to include model-API interaction logs, prompts sent and responses received, as a forensic artifact category alongside traditional network and process telemetry, since that log is often the only record of what an LLM-directed implant was actually instructed to do at a given point in time. Finally, security leaders should expect this technique to proliferate beyond a single financially motivated developer; the underlying pattern, replacing attacker-controlled decision infrastructure with commercial AI services, is straightforward enough that other malware families are likely to adopt some version of it, and detection strategy built around CAIRN's metadata-first approach should be treated as a durable capability investment rather than a one-off response to a single family.

## CSA Resource Alignment

CLOSEDQUORUM is best understood as a concrete, single-family instance of a pattern CSA's AI Safety Initiative has already been tracking across the broader threat landscape. **Autonomous Agentic AI Adversaries** documented, through cases such as GTG-1002 and CyberStrikeAI, that autonomous AI-driven decision-making in offensive operations has crossed from theoretical to operational reality; CLOSEDQUORUM extends that same shift into commodity, financially motivated Windows malware rather than state-linked or well-resourced campaigns, showing the pattern is now reachable by a solo developer with a week of effort [5]. **Semantic Malware: Why Promptware Breaks Process-Lineage Detection** made the broader case that malicious logic increasingly lives partly outside compiled code, in natural-language instructions processed by an AI system, and argued that conventional process-lineage detection consequently loses signal value; CAIRN's metadata-first detection approach against CLOSEDQUORUM is a direct, practical validation of that argument applied to a different mechanism, external API-driven decisions rather than context-resident prompt injection [6]. **LLMjacking Evolves: Stolen AI Compute as Attack Infrastructure** documented how commercial AI compute and API access have themselves become attack infrastructure that adversaries acquire, steal, or resell; CLOSEDQUORUM's reliance on paid access to four separate commercial LLM providers illustrates the same dependency from the opposite direction, with implications for how enterprises should monitor and govern legitimate API-key issuance and usage against misuse [7].

More broadly, this case reinforces the value of applying CSA's MAESTRO framework when threat-modeling any system, defensive or offensive, that places an LLM in a decision-making role rather than an advisory one, since MAESTRO's layered approach to agent autonomy, tool access, and trust boundaries maps directly onto the questions defenders must now ask about their own AI-integrated tooling as well as about malware like CLOSEDQUORUM [8]. Organizations building their AI governance programs around CSA's AI Controls Matrix (AICM) v1.1 should treat model-API egress monitoring and LLM-interaction logging as concrete implementation targets under the framework's threat and vulnerability management and logging and monitoring domains, since CLOSEDQUORUM demonstrates precisely the blind spot those domains are designed to close [9].

# References


[1] Cisco Talos. "The Closed Quorum: Inside the first reported autonomous AI C2 implant." Cisco Talos Blog, September 22, 2026.

[2] Duncan Riley. "Cisco Talos finds malware that puts its next move to a four-model vote." SiliconANGLE, September 22, 2026.

[3] The Hacker News. "This Windows Malware is Built to Let Up to Four AI Models Vote on Its Next Move." The Hacker News, September 2026.

[4] Pierluigi Paganini. "CLOSEDQUORUM, the malware that asks four AI models what to do next." Security Affairs, September 24, 2026.

[5] Cloud Security Alliance AI Safety Initiative. "Autonomous Agentic AI Adversaries." CSA Lab Space, June 24, 2026.

[6] Cloud Security Alliance AI Safety Initiative. "Semantic Malware: Why Promptware Breaks Process-Lineage Detection." CSA Lab Space, July 16, 2026.

[7] Cloud Security Alliance AI Safety Initiative. "LLMjacking Evolves: Stolen AI Compute as Attack Infrastructure." CSA Lab Space, June 18, 2026.

[8] Cloud Security Alliance. "Agentic AI Threat Modeling Framework: MAESTRO." CSA Blog, February 6, 2025.

[9] Cloud Security Alliance. "AI Controls Matrix (AICM) v1.1." Cloud Security Alliance, 2026.

[10] The Hacker News. "Google Uncovers PROMPTFLUX Malware That Uses Gemini AI to Rewrite Its Code Hourly." The Hacker News, November 5, 2025.