10.48047/jocaaa.2026.35.01.94

Security Threats in Telecom Networks by Prompt Poisoning of MCP Servers and Mitigations

Binu Kiliamkavunkal Govindan
Independent Researcher, USA

Received: 31.01.2026 Revised: 02.02.2026

Abstract

The communication sector experiences the most security challenges ever because artificial intelligence agents are integrated into the network functioning, customer service, and administration. The use of prompt poisoning and injection attacks is one of the critical vulnerabilities that can be used to put pressure on the AI systems in understanding instructions, and potentially put the integrity of the network under question, and make sensitive customer information available. Although the Model Context Protocol offers the capability of powerful AI agents, it presents certain attack vectors, such as the lack of proper authentication protocols, broad permissions, and tool poisoning by corrupting the data sources. Such threats are reported in various fields of operation in which compromised AI agents may cause service interruptions, parasite observability systems, compromise fraud detection systems, and enable unauthorized access to customer information. The defense should be multi-layered and a combination of stringent input checking, output filtering, proper authentication, minimization of privileges, and constant monitoring of behaviors. Organizations are required to deploy Zero Trust designs that are tailored to the specific needs of AI agents, have detailed audit trails, and have fast incident response strategies that identify and limit compromises. With the development of telecommunications networks to achieve complete autonomy where AI agents gain greater powers, the intelligence layer faces as much risk as physical infrastructure, and it requires urgent consideration by security leaders and long-term dedication to best practices.

Keywords: Prompt Poisoning, Telecommunications Security, AI Agent Vulnerabilities, Model Context Protocol, Autonomous Network Defense

1. Introduction

The telecommunication industry is at a junction point where artificial intelligence has deeply embedded into network business operations and customer service management. What began as mere automation has since been modified into a much more advanced autonomous agent that reasons, makes judgments, and takes actions with limited human supervision. These intelligent systems are currently performing tasks such as routing of traffic, customer complaints, etc, and in doing so are making decisions on a split second, a feat that would have taken teams of engineers years ago. The transformation happening right now goes beyond making existing processes faster; it fundamentally reimagines how networks operate [1].

Enter the Model Context Protocol, a framework that lets AI agents talk to databases, pull information from documents, and connect with enterprise tools—all the things needed to run a modern telecom network. But here's the catch: this connectivity opens doors that attackers can walk through. The most

10.48047/jocaaa.2026.35.01.94

troubling vulnerability? Something called prompt injection, where bad actors slip malicious instructions into what looks like normal data. Think of it like whispering instructions to someone while they're on the phone—they might follow those whispered commands without realizing they came from the wrong source. Research confirms this represents the biggest single threat facing AI applications today [2].

What makes this scary for telecom companies isn't just the technical sophistication of these attacks. A compromised AI agent managing network traffic could cause outages affecting entire cities. An agent handling customer data might leak personal information while appearing to process a routine request. Traditional cybersecurity focused on protecting servers and databases—things with clear boundaries. But when attacks target how AI thinks and reasons, the boundaries blur. With networks getting closer to complete autonomy, where AI agents make more calls, and humans become more of spectators, it will no longer be a choice to stay ahead of such threats. The network is on the verge of failure; customer confidence and the business's existence are all at risk.

2. Understanding Prompt Poisoning Threats

Picture this: an AI system receives instructions on how to behave, then someone sneaks in additional instructions disguised as regular data. The AI can't tell the difference, so it follows both sets of directions—including the malicious ones. That's prompt injection in a nutshell, and security experts have flagged it as the number one risk for any organization using large language models [3]. The core problem stems from a fundamental limitation: these AI systems struggle to separate legitimate commands from cleverly disguised attacks embedded within normal-looking information.

Telecom networks face this threat from two angles. Direct attacks happen when someone crafts specific inputs designed to hijack an AI agent's behavior right then and there—maybe tricking a network management bot into revealing configuration details or executing unauthorized commands. The indirect approach proves even sneakier. Attackers plant malicious instructions inside documents, databases, or web pages that AI agents routinely consult. When the agent reads that poisoned content during normal operations, boom—the hidden commands activate. Recent studies show how poisoning the data sources that AI relies on can corrupt behavior across thousands of interactions, not just one [4].

The level of sophistication continues to increase. Multi-turn conversations have become the new trick, with attackers setting up trust across multiple harmless conversations before dropping the payload. Telecom AI agents make juicy targets because they run with elevated permissions—they need access to configure routers, modify customer accounts, and control critical systems. That access becomes a weapon in the wrong hands. Plus, these agents interact with so many interconnected systems that a breach in one place can spread like wildfire. The larger the autonomy and integration of the system provided by the companies to the AI, the bigger the attack surface and the harder it becomes to detect. The significance of these patterns of attack is in the fact that they must be counteracted using a particular set of strategies rather than the general security policies.

Attack Vector	Mechanism	Target Systems	Impact Severity
Direct Prompt Injection	Malicious instructions embedded in real-time inputs	Network configuration agents, customer service bots	High - immediate unauthorized actions
Indirect Prompt Injection	Poisoned content in referenced documents and	Observability agents, diagnostic systems	Critical - persistent compromise across

10.48047/jocaaa.2026.35.01.94

	databases		sessions
RAG Data Poisoning	Corrupted retrieval sources with high semantic similarity	Knowledge base systems, documentation agents	Severe - widespread behavioral corruption
Tool Poisoning	Malicious instructions in API responses and data feeds	Monitoring tools, configuration management	Critical - system-wide propagation
Multi-Turn Attacks	Trust-building sequences followed by adversarial requests	All agent types with conversational interfaces	Moderate to High - evades simple filtering

Table 1: Prompt Poisoning Attack Mechanisms in Telecommunications [9, 10]

3. Impact on Telecommunications Infrastructure

When AI agents run telecommunications networks, their compromise creates ripple effects that touch everything from call quality to customer bank accounts. Modern telecom infrastructure has become a maze of interconnected systems where AI agents make thousands of decisions per second—which route handles this traffic, which server processes that request, and how to allocate bandwidth across competing demands. Studies examining network architecture reveal just how dependent telecom operations have become on these autonomous decision-makers [5]. Break that decision-making process, and the whole house of cards wobbles.

Consider network operations first. An AI agent with configuration authority getting hit by prompt poisoning might start applying wrong settings across the board. Maybe it shuts down security features, thinking it's optimizing performance. Perhaps it reroutes traffic in ways that create bottlenecks or expose data to interception. Self-driven networks that dynamically self-adjust themselves according to AI suggestions are time bombs when these suggestions are fundamentally flawed by logic compromised by being fed suspect data. The damage accretes in little bits here and there, misconfigurations which, over weeks, or in one smack, in a catastrophic crash, degrade performance. Either way, by the time humans notice, significant harm has occurred.

Then there's the customer service angle, which might sound less critical until considering what those AI agents can access. Customer service bots have keys to the kingdom: names, addresses, payment information, call records, and account credentials. Evaluations of AI agent security vulnerabilities show how hijacking attacks can manipulate these bots into spilling secrets or making unauthorized changes to accounts [6]. Through what might seem to be regular customer service interactions, an attacker is able to drain accounts, steal identities, or collect intelligence to use in other criminal activities. The standard methods of fraud detection monitor the suspicious logins or the strange network behavior, and when the attack is committed via the legitimate access channels of the AI, those alarms remain silent. This redefines the whole concept of security: securing networks has become securing the way AI thinks and not necessarily the systems it is linked to.

Operational Domain	AI Agent Functions	Compromise Consequences	Business Impact

10.48047/jocaaa.2026.35.01.94

Network Operations	Traffic routing, configuration management, and resource allocation	Service outages, misconfiguration cascades, security control bypass	Revenue loss, SLA violations, subscriber churn
Observability Systems	Log analysis, anomaly detection, predictive maintenance	False metrics, masked incidents, corrupted analytics	Delayed issue resolution, undetected failures
Security and Fraud Detection	Transaction monitoring, intrusion prevention, compliance	Detection evasion, policy bypass, audit trail corruption	Financial losses, regulatory penalties
Customer Service	Account management, billing, information retrieval	Data exfiltration, unauthorized modifications, and privacy violations	Legal liability, reputation damage, customer trust erosion
Autonomous Coordination	Multi-agent collaboration, decision synchronization	Coordinated failures, supply chain amplification	Systemic infrastructure collapse

Table 2: Telecommunications Infrastructure Vulnerabilities [5, 6]

4. Defense Strategies and Technical Mitigations

Building defenses against prompt injection demands layers upon layers of protection, acknowledging upfront that no single fix solves everything. The best security strategies weave together multiple techniques, each catching what others miss [7]. Start at the input stage with validation mechanisms that scrutinize every prompt before it reaches the AI. Pattern matching can spot common attack signatures—phrases like "ignore previous instructions" or attempts to break out of expected query formats. Semantic analysis adds another filter, checking whether requests make sense given what that particular agent should be doing. Encoding detection prevents attackers from disguising malicious payloads using base64 or other tricks to slip past initial checks.

But input filtering alone won't cut it because clever attackers find ways around any filter given enough time. That's where output validation earns its keep. Even if a malicious prompt sneaks through, controlling what the AI can say limits the damage. Telecommunications companies should enforce strict response formats—network diagnostic agents return only status codes and metrics, nothing more. Keyword scanning catches sensitive data trying to escape in responses: no credit card numbers, no passwords, no internal IP addresses. Secondary validation using another AI model double-checks that responses match expectations for given inputs. When outputs deviate from acceptable patterns, red flags go up.

Privilege restrictions form another critical layer. AI agents should operate with the absolute minimum permissions needed for their job, no more. A customer service bot doesn't need network configuration access. A monitoring agent doesn't require write permissions on billing databases. Research on autonomous network security emphasizes establishing clear authority boundaries with human approval gates for high-stakes decisions [8]. Here, strong identity verification is made as a foundation of authentication, short-lived access tokens that expire, and constant rechecking that agents remain authorized to do whatever they are attempting to do are made foundational. Add behavioral monitoring, which learns typical behaviour of each agent, and issues a warning in case something appears out of the

10.48047/jocaaa.2026.35.01.94

ordinary data access, unusual communication patterns, or outputs not in context. These layers combined provide strong defense mechanisms that increase successful attacks exponentially.

Security Layer	Implementation Components	Protection Capability	Deployment Complexity
Input Validation	Pattern filtering, encoding detection, semantic analysis, and length normalization	Blocks known attack signatures and obfuscated payloads	Moderate - requires continuous signature updates
Output Filtering	Keyword scanning, schema enforcement, PII detection, and command injection prevention	Prevents data leakage and unauthorized command execution	Low to Moderate - schema definition overhead
Authentication Controls	OAuth with PKCE, mutual TLS, short-lived tokens, audience validation	Prevents unauthorized access and token reuse attacks	High integration with identity systems
Access Restrictions	Role-based permissions, scope minimization, just-in-time privileges	Limits the blast radius of compromised agents	Moderate - requires detailed privilege mapping
Behavioral Monitoring	Baseline establishment, deviation detection, anomaly correlation	Identifies compromised agents through behavioral changes	High - requires tuning and analytics infrastructure

Table 3: Defense-in-Depth Security Controls [7, 8]

5. Continuous Monitoring and Incident Response

The only way to identify that the AI agents are compromised is to monitor systems that were designed to identify this novel threat environment. Traditional security tools designed for servers and networks need major upgrades to handle AI-specific attacks. The monitoring must analyze prompt content itself, looking for malicious patterns that signature-based detection might miss. It must correlate behaviors across multiple agents to spot coordinated attacks. Most importantly, it must distinguish between normal operational variations—which happen all the time in dynamic networks—and genuine indicators of compromise [9].

Here behavioral analysis is greatest due to the fact that it concentrates on the deviations of the norms rather than struggling to predict all possible attack techniques. The first operation of the monitoring systems is to learn the normal behavior of each agent: what data it accesses, at what times it is most responsive, what type of queries it answers, how it reacts. These baselines are refined as the operations proceed. Alerts are raised when an agent inexplicably begins using new and unusual databases, interacting with other systems, or generating computed outputs that do not match its purpose. The dilemma is in the fact that the sensitivity is set too loose and the real threats are slipping through; it becomes too tight and the security teams are flooded with false alarms which end up being neglected.

The response to AI compromises is not similar to the normal breach response since the contamination may last even when the immediate threats are eliminated. Poisoned training data or corrupted retrieval sources keep influencing agent behavior until explicitly cleaned. Documentation addressing Model Context Protocol security stresses maintaining complete audit trails that capture full context around every

10.48047/jocaaa.2026.35.01.94

agent interaction, enabling forensic analysis after incidents [10]. Response procedures need quick isolation capabilities that quarantine suspect agents without bringing down dependent systems—no easy task in tightly integrated environments. Rollback mechanisms that restore agents to verified clean states become essential. Organizations should designate response teams combining traditional cybersecurity expertise with AI system knowledge, then drill these procedures through realistic exercises. Automated responses that execute initial containment within seconds of detecting compromise can prevent attacks from spreading or accomplishing their goals. The speed and scope of AI agent operations demand equally fast defensive reactions.

Response Element	Capabilities	Timing Requirements	Success Factors
Detection Systems	Prompt content analysis, behavioral correlation, cross-agent monitoring	Real-time to sub-minute	Low false positive rates, comprehensive coverage
Isolation Mechanisms	Agent quarantine, privilege revocation, and communication blocking	Seconds to minutes	Minimal disruption to dependent systems
Forensic Analysis	Audit trail reconstruction, attack vector identification, and impact assessment	Hours to days	Complete logging, contextual data retention
Recovery Procedures	Agent rollback, data source cleaning, configuration restoration	Minutes to hours	Verified clean states, tested rollback paths
Organizational Readiness	Cross-functional teams, tabletop exercises, escalation protocols	Continuous preparation	AI security expertise, clear accountability

Table 4: Incident Response Framework Components [9, 10]

Conclusion

The merging of artificial intelligence and telecommunication infrastructure presents the opportunity for a breakthrough and the challenge of security that requires both urgent and long-term solutions. Poisoning attacks against the AI agents working in the telecommunications network are a paradigm change in the threat situation, not just in the traditional perimeter protection, but in leveraging the human thinking behind the autonomous decision-making. These AI-based threats do not use the established detection signatures and mitigation playbooks as traditional cyberattacks; instead, they modify the reasoning and interpretations of instructions in a manner that these behaviors may go unnoticed by the typical security surveillance and lead to disastrous failures in running operations. Although the Model Context Protocol supports strong integration features between the AI agents and the enterprise systems, it poses architectural weaknesses that attackers actively take advantage of using poor authentication, over-granting permissions, and using poisoned sources of data. The security of AI cannot be considered a back-burner or even a matter to consider in the future, as telecommunications operators are operating critical infrastructure that impacts millions of their subscribers, and attackers have the capability and motivation to exploit it, and the impacts of successful compromise are far exceeding the individual organization to affect industries dependent on it and human safety. Meaningful protection involves considering the fact

10.48047/jocaaa.2026.35.01.94

that AI agents present a new line of assets that will require unique protection measures, unlike those of traditional IT security frameworks. Layered strategies such as input sanitization, output validation, stringent authentication, privilege minimization, on-going behavioral monitoring and quick incident response establish robust security posture that can sustain the attacks of high sophistication. The application of Zero Trust principles to autonomous agents, full audit with forensic examination capability, and human control of high-stakes decisions creates the necessary guardrails as networks are increasingly autonomous. Organizations are required to invest in highly skilled knowledge in the interface of cybersecurity and AI systems, put in place governance frameworks with clear responsibility on AI security, and engage in information sharing across the industry to keep pace with the emerging attack models. The telecommunications industry has been at a point of no turning back, and now it is up to the choices made in the present regarding the safety of AI that will either make autonomous networks prove their worth in terms of efficiency and innovation or become sources of disruption and harm never before seen. To succeed, it is necessary to approach the intelligence layer with the same seriousness as deployed toward physical infrastructure protection, deploy tested defensive methods immediately, and be on the alert as both AI capabilities and adversarial methods keep developing at an unparalleled rate.

References

[1] Zedian Shao et al., “Enhancing Prompt Injection Attacks to LLMs via Poisoning Alignment,” arxiv>cs>arXiv:2410.14827, 2025 [Online]. Available: https://arxiv.org/abs/2410.14827

[2] Deloitte, "Agentic AI Blueprint for Telcos". [Online]. Available: https://www.deloitte.com/global/en/what-we-do/capabilities/agentic-ai-blueprint-for-telcos.html

[3] OWASP, "LLM01:2025 Prompt Injection". [Online]. Available: https://genai.owasp.org/llmrisk/llm01-prompt-injection/

[4] Alexandra Souly et al., "Poisoning Attacks On Llms Require A Near-Constant Number Of Poison Samples," arXiv preprint arXiv:2410.14827, 2025. [Online]. Available: https://arxiv.org/pdf/2510.07192

[5] Massimo Iovene et al., "AI agents in the telecommunication network architecture," Ericsson, 2024. [Online]. Available: https://www.ericsson.com/en/reports-and-papers/white-papers/ai-agents-and-network-architecture

[6] The U.S. AI Safety Institute Technical Staff, "Technical Blog: Strengthening AI Agent Hijacking Evaluations," NIST, 2025. [Online]. Available: https://www.nist.gov/news-events/news/2025/01/technical-blog-strengthening-ai-agent-hijacking-evaluations

[7] Rich Harang, "Securing LLM Systems Against Prompt Injection," NVIDIA, 2023. [Online]. Available: https://developer.nvidia.com/blog/securing-llm-systems-against-prompt-injection/

[8] Comarch, "Autonomous Networks Driving Your Business Forward". [Online]. Available: https://www.comarch.com/telecommunications/autonomous-network-that-evolve-with-your-business/

[9] Obsidian Security, "Prompt Injection Attacks: The Most Common AI Exploit in 2025", 2025. [Online]. Available: https://www.obsidiansecurity.com/blog/prompt-injection

[10] Tetrate, "MCP Security Best Practices: Authentication, Authorization & Supply Chain Protection". [Online]. Available: https://tetrate.io/learn/ai/mcp/security-privacy-considerations