

AI Infrastructure LiteLLM Supply Chain Poisoning Alert

 nsfocusglobal.com/ai-infrastructure-litellm-supply-chain-poisoning-alert

NSFOCUS

March 27, 2026

- Check the current version `pip show litellm #`
- If it is a malicious version, immediately downgrade to the safe version `pip install litellm==1.82.6 #`
- Or fix the version in requirements.txt `echo "litellm==1.82.6" >> requirements.txt #`
- Check for malicious .pth files `find $(python -c "import site; print(site.getsitepackages()[0])") -name "litellm_init.pth" -delete #`
- Verify the installation of `python -c "import litellm; print(litellm.__version__)"`

Overview

Recently, NSFOCUS Technology CERT detected that the GitHub community disclosed that there was a credential stealing program in the new version of LiteLLM. Analysis confirmed that it had suffered supply chain poisoning by the TeamPCP group on PyPI. It stole the publishing permission credentials by hacking into the security scanning tool Trivy used in the LiteLLM release process and released the malicious version at around 18:00 (GMT+8) on March 24; After the user triggers, it will collect host information, steal SSH keys, cloud service certificates, K8s keys, SSL certificate private keys and cryptocurrency wallets and other sensitive data. If a K8s cluster is detected, it will move horizontally to deploy privileged Pods on each node and implant a persistent systemd backdoor. The impact is wide-ranging, and relevant users are requested to take measures for investigation and protection as soon as possible.

In today's technological evolution driven by artificial intelligence, AI model gateways have become a key infrastructure for enterprise technology architecture. LiteLLM is a mainstream open source AI model gateway that can simplify the calling and management of large language models from hundreds of different suppliers through a unified interface, with functions such as cost tracking, security protection, load balancing and logging.

Reference links:

<https://github.com/BerriAI/litellm/issues/24518>

<https://github.com/pypa/advisory-database/blob/main/vulns/litellm/PYSEC-2026-2.yaml>

Attack Group

TeamPCP is a cybercrime group active from 2025 to early 2026, also known as DeadCatx3, PCPcat and ShellForce. It mainly launches large-scale automated attacks against global cloud native infrastructure and involves advanced threat activities such as supply chain poisoning.

TeamPCP builds a large-scale distributed proxy and scanning infrastructure to invade servers through stolen credentials to steal data, deploy ransomware, conduct ransomware activities and mine cryptocurrencies. It has mature tool sets such as proxy.sh (for environmental identification and load deployment), scanner.py (vulnerability scanning), kube.py (Kubernetes related operations), etc.

Major recent attacks

- 2026-03-19: Trivy vulnerability scanner that hacked into Aqua Security, released malicious version v0.69.4
- 2026-03-23: Attack Checkmarx KICS GitHub Action, by forcing push tags to malicious commits
- 2026-03-24: Supply chain poisoning of LiteLLM, release malicious versions 1.82.7 and 1.82.8
- 2026-03-25: Deployment of Kubernetes erasers against Iran, attack intent escalates from stealing to destructive

Scope of Impact

Affected versions

- LiteLLM = 1.82.7
- LiteLLM = 1.82.8

Applications including but not limited to the following that reference LiteLLM are also affected:

- DSPy
- MLflow
- CrewAI
- OpenHands
- Prodigy

Note: LiteLLM is widely used, with about 95 million downloads per month and 3.4 million downloads per day, affecting more than two thousand dependent packages. The malicious version was released at around 18:00 (GMT+8) on March 24 and survived for about 3 hours. It is known that the attackers may have stolen about 300GB of data, including up to 500,000 credentials.

Unaffected version

LiteLLM <= 1.82.6

Event Analysis

Event timeline (GMT+8):

- March 19 – TeamPCP hacks into Trivy security scanning tool
- March 23 – TeamPCP breaches Checkmarx KICS
- March 24 10:39 – UTC Malicious version 1.82.7 released to PyPI
- March 24 10:52 – UTC Malicious version 1.82.8 released to PyPI
- March 24 13:52 – UTC PyPI removed malicious version

TeamPCP's supply chain poisoning against LiteLLM did not directly break into its code base, but rather exploited the configuration flaw of the `pull_request_target` trigger in GitHub Actions to submit malicious pull requests to the dependent security scanning tool Trivy, inducing the workflow to execute the attacker's implanted code in a high-privilege basic repository. Thereby stealing the `PYPI_PUBLISH` token with the highest release authority; The attacker used this token to bypass all code review and CI automation testing processes, and directly pushed a malicious version (1.82.7/1.82.8) with no corresponding Git commit record to the PyPI repository, demonstrating a high-level supply chain attack mode of "using security tools to create insecurity", exposing the structural vulnerabilities caused by excessive trust in third-party integration and improper permission configuration in modern software delivery chains.

Malicious version behavior

In version 1.82.7, malicious code was implanted in the `"litellm/proxy/proxy_server.py"` file, which may have been injected during or after the wheel build process. The code is designed to be executed automatically when a module is imported, and any process that imports `"litellm.proxy.proxy_server"` will trigger the attack payload without user interaction.

Version 1.82.8 takes a more aggressive approach-adding a malicious `"litellm_init.pth"` file to the wheel root directory, causing the logic to automatically execute when all Python processes in the environment are started, not just when `litellm` is imported. And the `.pth` launcher will generate a child Python process through `subprocess.Popen`, so that the attack payload can run in the background.

After the attack payload is decoded, it will release credential stealing tools and persistent backdoors. When a Kubernetes service account token is present, the credential stealing tool enumerates all nodes in the cluster and deploys privileged Pods on each node. These Pods then enter the host file system through `chroot` and install a persistent backdoor as a `systemd` user service on each node.

The systemd service is configured to launch a Python script (~/.config/sysmon/sysmon.py)-the same name used in the Trivy hacking incident-that accesses “checkmarx[.]zone/raw” every 50 minutes to get the next stage payload URL. The script will terminate if the URL contains youtube[.]com, which is a common kill switch pattern in observed events so far.

Risk Investigation

Relevant users can conduct troubleshooting according to the following steps:

- Check the current version `pip show litellm #`
- If it is a malicious version, immediately downgrade to the safe version `pip install litellm==1.82.6 #`
- Or fix the version in requirements.txt `echo "litellm==1.82.6" >> requirements.txt #`
- Check for malicious .pth files `find $(python -c "import site; print(site.getsitepackages()[0])") -name "litellm_init.pth" -delete #`
- Verify the installation of python `-c "import litellm; print(litellm.__version__)"`

If it is confirmed that there are malicious packages in the environment, the infected assets should be taken offline immediately and the hosts should be physically or logically isolated.

Find and mitigate the suspicious process named /tmp/pglog, delete the /tmp/pglog and /tmp/.pg_state files, and forcibly remove the persistent component; An attacker may have left a malicious Wheel file in the local pip cache, which can be used to perform a pip cache purge or manually delete directories such as ~/.cache/uv to prevent re-poisoning due to the cache mechanism after rollback.

Cloud environment users can check whether there are malicious Pods in the Kubernetes cluster; Execute in the full namespace: `kubectl get pods -A | grep -E "node-setup|host-provisioner"`; If found, it needs to be deleted immediately and reviewed node by node:

```
ls -la /root/.config/sysmon/2>/dev/null
```

Summary and Recommendations

This poisoning incident is different from a conventional cyber attack, but an industry warning to supply chain security; in the era of rapid development of AI ecology, software supply chain risks are no longer limited to a single technology stack or platform, but have gradually evolved into a global challenge for the entire IT ecosystem. Relevant units must build a sound supply chain security governance system to deal with such new potential threats.

It is recommended that affected users take the following measures to deal with it:

1. Revoke and rotate all SSH keys, cloud account keys, API Tokens, K8s tokens, database passwords and other credentials;
2. Block and investigate malicious IoC;

3. Implement the principle of least privilege, limit the access scope of CI/CD tools, and rotate secrets regularly;
4. Build an internal private PyPI image, scan and verify the signature for security before synchronizing;
5. Establish a normalized open source component security audit mechanism and increase multiple approval processes before release;
6. Subscribe to PyPI security alerts, GitHub Security Advisories, etc. for supply chain monitoring.

IOCs

Malicious files

File name: litellm_init.pth

SHA256: 71e35aef03099cd1f2d6446734273025a163597de93912df321ef118bf135238

File name: proxy_server.py

SHA256: a0d229be8efcb2f9135e2ad55ba275b76ddcfcb55fa4370e0a522a5bdee0120b

File name: litellm-1.82.7-py3-none-any.whl

SHA256: 8395c3268d5c5dbae1c7c6d4bb3c318c752ba4608cfcd90eb97ffb94a910eac2

File name: litellm-1.82.8-py3-none-any.whl

SHA256: d2a0d5f564628773b6af7b9c11f6b86531a875bd2d186d7081ab62748a800ebb

Malicious domain name

checkmarx.zone

models.litellm.cloud

Statement

This advisory is only used to describe a potential risk. NSFOCUS does not provide any commitment or promise on this advisory. NSFOCUS and the author will not bear any liability for any direct and/or indirect consequences and losses caused by transmitting and/or using this advisory. NSFOCUS reserves all the rights to modify and interpret this advisory. Please include this statement paragraph when reproducing or transferring this advisory. Do not modify this advisory, add/delete any information to/from it, or use this advisory for commercial purposes without permission from NSFOCUS.

NSFOCUS, a pioneering leader in cybersecurity, is dedicated to safeguarding telecommunications, Internet service providers, hosting providers, and enterprises from sophisticated cyberattacks.

Founded in 2000, NSFOCUS operates globally with over 4000 employees at two headquarters in Beijing, China, and Santa Clara, CA, USA, and over 50 offices worldwide. It has a proven track record of protecting over 25% of the Fortune Global 500 companies, including four of the five largest banks and six of the world's top ten telecommunications companies.

Leveraging technical prowess and innovation, NSFOCUS delivers a comprehensive suite of security solutions, including the Intelligent Security Operations Platform (ISOP) for modern SOC, DDoS Protection, Continuous Threat Exposure Management (CTEM) Service and Web Application and API Protection (WAAP). All the solutions and services are augmented by the Security Large Language Model (SecLLM), ML, patented algorithms and other cutting-edge research achievements developed by NSFOCUS.