# CVE Detail – CVE-2024-34359

llama-cpp-python is the Python bindings for llama.cpp. `llama-cpp-python` depends on class `Llama` in `llama.py` to load `.gguf` llama.cpp or Latency Machine Learning Models. The `\_\_init\_\_` constructor built in the `Llama` takes several parameters to configure the loading and running of the model. Other than `NUMA, LoRa settings`, `loading tokenizers,` and `hardware settings`, `\_\_init\_\_` also loads the `chat template` from targeted `.gguf` 's Metadata and furtherly parses it to `llama\_chat\_format.Jinja2ChatFormatter.to\_chat\_handler()` to construct the `self.chat\_handler` for this model. Nevertheless, `Jinja2ChatFormatter` parse the `chat template` within the Metadate with sandbox-less `jinja2.Environment`, which is furthermore rendered in `\_\_call\_\_` to construct the `prompt` of interaction. This allows `jinja2` Server Side Template Injection which leads to remote code execution by a carefully constructed payload.

## Threat-Mapped Scoring

Score: 1.8

Priority: P4 - Informational (Low)

## EPSS

EPSS Score: N/A

Percentile: 0.97825

## CVSS Scoring

CVSS v3.1 Score: 9.6

Severity: CRITICAL

## Mapped CWE(s)

* CWE-76: Improper Neutralization of Equivalent Special Elements