Primeiro malware com GPT-4 integrado descoberto: chega o MalTerminal

Data: 2025-09-22 13:52:47

Autor: Inteligência Against Invaders

Redazione RHC**:22 Setembro 2025 15:52**

SentinelLABS pesquisadores descobriram o que descrevem como o **primeiro exemplo conhecido de malware com funcionalidade LLM integrada** Apelidado *Terminal Maligno* . A descoberta foi apresentada em **LABScon 2025** , onde uma grande variedade de artefatos foi exibida: um binário do Windows, vários scripts Python e ferramentas auxiliares demonstrando como **GPT-4 foi explorado para gerar código malicioso dinamicamente** , como ransomware ou shells reversos.

A amostra analisada continha um endpoint de API referente ao antigo **Conclusões de bate-papo OpenAI** serviço, que foi desativado em novembro de 2023. Isso sugere que *Terminal Maligno* foi desenvolvido antes dessa data, tornando-se um **Exemplo inicial de malware** com LLM incorporado. Ao contrário do malware tradicional, parte de sua lógica não é pré-compilada, mas é **criado em tempo de execução** via consultas GPT-4: o operador pode escolher entre os modos "criptografador" ou "shell reverso", e o modelo gera o código correspondente em tempo real.

Dentro do kit, os pesquisadores também encontraram scripts que replicavam o comportamento do binário, bem como um **Scanner de segurança baseado em LLM** , capaz de avaliar arquivos Python suspeitos e produzir relatórios: um exemplo claro do **Uso duplo** de modelos generativos, aplicáveis tanto para fins ofensivos quanto defensivos.

Os autores também demonstrou uma nova metodologia para detectar malware LLM, com base em **Artefatos de integração inevitáveis** : chaves de API incorporadas e prompts codificados. Ao analisar prefixos de chave (por exemplo, *sk-ant-api03*) e fragmentos reconhecíveis relacionados à OpenAI, eles desenvolveram regras eficazes para retrocesso em larga escala. Uma análise de um ano sobre o VirusTotal revelou **milhares de arquivos contendo chaves** , variando de vazamentos acidentais de desenvolvedores a amostras maliciosas. Paralelamente, eles testaram uma técnica de pesquisa baseada em prompt: extrair strings de texto de arquivos binários e avaliar sua intenção usando **classificação LLM leve** , que se mostrou altamente eficaz na detecção de ferramentas anteriormente invisíveis.

O estudo destaca um paradoxo crucial: o uso de um modelo externo oferece flexibilidade e adaptabilidade aos invasores, mas também introduz **Vulnerabilidades** . Sem chaves de API válidas ou prompts armazenados, o malware perde muito de sua eficácia. Isso abre novos caminhos defensivos, como a busca por "prompts como código" e chaves incorporadas, especialmente nos estágios iniciais da evolução dessas ameaças.

Até o momento, não há evidências de *Terminal Maligno* amplamente implantado: pode ser um **prova de conceito** ou uma ferramenta de equipe vermelha. No entanto, a técnica em si representa uma

mudança de paradigma, impactando **assinaturas, análise de tráfego e atribuição de ataque** .

O SentinelLABS recomenda prestar maior atenção à análise de aplicativos e repositórios: além de bytecodes e strings, agora é essencial procurar **Rastreamentos textuais, estruturas de mensagens e artefatos relacionados a modelos de nuvem** , onde os mecanismos de malware de próxima geração podem ser ocultados.

Os autores concluem enfatizando que a integração de geradores de comandos e lógica de tempo de execução **enfraquece os detectores tradicionais** e complica significativamente a atribuição de ataques, abrindo um novo capítulo na luta entre a defesa cibernética e o crime cibernético.

Redação
A equipe editorial da Red Hot Cyber é composta por um grupo de indivíduos e fontes anônimas que colaboram ativamente para fornecer informações e notícias antecipadas sobre segurança cibernética e computação em geral.

Lista degli articoli