Malterminal: New GPT-4 Malware que escreve seu próprio ransomware - A

Data: 2025-09-22 08:14:03

Autor: Inteligência Against Invaders

Uma descoberta inovadora na pesquisa de segurança cibernética revelou o surgimento de malware de ‘Malterminal’, potencialmente o exemplo mais antigo de malware habilitado para Modelo de Linguagem (LLM) que utiliza a API do GPT-4 do OpenAI para gerar dinamicamente o código de ransomware e os conchos reversos no tempo de execução.

Essa descoberta representa uma evolução significativa na sofisticação de malware, apresentando desafios sem precedentes para os métodos tradicionais de detecção.

Os pesquisadores do Sentinellabs têm identificado Uma nova categoria de malware que muda fundamentalmente o cenário de ameaças descarregando funcionalidade maliciosa para sistemas de inteligência artificial.

Diferentemente do malware tradicional com código malicioso incorporado, essas ameaças habilitadas para o que geram lógica maliciosa exclusiva durante a execução, tornando quase impossível a detecção de assinatura estática.

A metodologia de pesquisa empregada pelos sentinellabs focou na correspondência de padrões contra chaves de API incorporadas e estruturas rápidas específicas para identificar essas ameaças sofisticadas.

Através de uma extensa análise de mais de 7.000 amostras contendo mais de 6.000 teclas de API exclusivas descobertas através de retrohunting VIRUSTOTALos pesquisadores desenvolveram técnicas inovadoras de detecção projetadas especificamente para esta nova categoria de malware.

Malterminal: um mergulho profundo técnico

O malware do Malterminal se destaca como executável baseado em Apython que consulta dinamicamente o GPT-4 para gerar código de ransomware ou estabelecer conchas reversas sob demanda.

O que torna essa descoberta particularmente significativa é a presença de um endpoint de conclusão de chat do OpenAI depreciado a partir de novembro de 2023, sugerindo que as amostras de malware do Malterminal antecedem previamente documentados LLM.

O pacote de malware inclui vários componentes: os principais scripts do MalTerminal.exe, vários scripts de prova de conceito testapi.py e até uma ferramenta defensiva chamada ‘FalConshield’ projetada para analisar arquivos Python suspeitos.

Este abrangente kit de ferramentas demonstrou o desenvolvimento de implementações simples de prova de conceito.

Os pesquisadores descobriram que os usuários do Malterminal solicitam que os usuários escolham entre as opções de ‘ransomware’ ou ‘shell reversa’ e, em seguida, aproveita os recursos de geração de código do GPT-4 para produzir a funcionalidade maliciosa solicitada.

Essa abordagem permite que o malware tenha seu comportamento com base em condições ambientais e requisitos operacionais, tornando -o significativamente mais versátil do que o malware estático tradicional.

O surgimento de malware habilitado para LLM apresenta desafios e oportunidades para os profissionais de segurança cibernética. Os métodos de detecção tradicionais enfrentam obstáculos significativos, como a lógica, é gerada em tempo de execução, em vez de incorporado no código.

As assinaturas estáticas se tornam ineficazes quando o código exclusivo é gerado para cada execução, e a análise de tráfego de rede se torna complicada à medida que as chamadas de API maliciosas se misturam com o uso legítimo de LLM.

No entanto, essas ameaças avançadas também introduzem novas vulnerabilidades. Malware habilitado para LLM deve incorporar Chaves da API e solicitações em seu código, criando artefatos detectáveis ??que os pesquisadores de segurança podem caçar.

A dependência dos serviços de IA externa também torna essas ameaças quebradiças – se as chaves da API forem revogadas ou os serviços ficarem indisponíveis, o malware perde sua funcionalidade principal.

Estrutura de segurança cibernética orientada pela IA para desenvolvimento seguro de software ilustrando revisão de literatura, pesquisas especializadas, redes neurais, modelagem estrutural e processos de mitigação

Metodologias de detecção

A Sentinellabs desenvolveu duas estratégias de caça primárias para identificar malware habilitado para LLM. A API de todo o API DetectionAproach usa regras YARA para identificar as teclas da API dos principais fornecedores de LLM com base em seus padrões estruturais exclusivos.

Por exemplo, as teclas antrópicas são prefixadas com “SK-ANT-API03”, enquanto as teclas OpenAI contêm a substring codificada por Base64 “T3blbkfj” representando “Openai. ”

O HuntingMethodology processa a HuntingModology por estruturas rápidas e prontas e formatos de mensagens em binários e scripts.

Os pesquisadores emparelharam essa técnica com classificadores leves da LLM para obter instruções para obter intenções maliciosas, permitindo uma identificação eficiente de ameaças de grandes conjuntos de amostras.

Interface do terminal do OpenAI Codex com um prompt de usuário para implementar o Modo Dark, ilustrando a geração de código AI-aprimorada

Essas metodologias se mostraram altamente eficazes, descobrindo não apenas o Malterminal, mas também várias aplicações ofensivas de LLM, incluindo agentes de pesquisa de pessoas, utilitários de benchmarking da equipe vermelha e ferramentas de injeção de vulnerabilidade.

A pesquisa revelou aplicativos criativos, como navegação do navegador, com assistência LLM para desvio antibot, controle de tela móvel por meio de análise visual e assistentes de pentesting para ambientes Kali Linux.

Antes do Malterminal, os pesquisadores de segurança documentaram outros notáveis ??malware de LLM. Ai movido O ransomware da ESET foi mais tarde revelado como pesquisa de prova de conceito da universidade.

Escrito em Golang com versões para Windows, Linux X64 e Arm ARM Architecturas, o PromptLock incorporou técnicas sofisticadas de solicitação para contornar os controles de segurança LLM, enquadrando solicitações em contextos de especialistas em segurança cibernética.

O LameHug (PromptSeal) da Apt28 representou outra etapa evolutiva, utilizando LLMs para gerar comandos de shell do sistema para coleta de informações.

Este malware incorporou 284 teclas de API Huggingface exclusivas para redundância e longevidade, demonstrando como os atores de ameaças se adaptam à API Key Blackisting e Interrupções de Serviço.

Paisagem de ameaças futuras

As implicações do malware habilitado para LLM se estendem muito além dos recursos atuais. À medida que os sistemas de inteligência artificial se tornam mais sofisticados e acessíveis, os atores de ameaças provavelmente se desenvolverão mais autônomos e adaptáveis ??de desdém da tomada de decisão em tempo real e adaptação ambiental.

O potencial de geração de malware autônoma em larga escala, embora atualmente limitado pelas alucinações LLM e instabilidade de código, representa um cenário futuro sobre o futuro.

Os profissionais de segurança devem se preparar para ameaças que canmodificam seu comportamento com base em ambientes de destino, gerar conteúdo convincente de engenharia social e adaptar suas táticas em resposta a medidas defensivas.

O jogo tradicional de gato e rato entre atacantes e defensores está evoluindo para uma dinâmica mais complexa, onde a inteligência artificial atende a papéis ofensivos e defensivos.

A descoberta de ameaças mal terminais e semelhantes marca o início de uma nova era na segurança cibernética, onde a inteligência artificial se torna uma ferramenta poderosa para os invasores e um componente crítico das estratégias de defesa.

As organizações devem adaptar suas posturas de segurança para abordar essas ameaças emergentes, desenvolvendo novas metodologias de detecção que explicam a natureza dinâmica e adaptativa dos malware habilitado para LLM.

Como esta pesquisa demonstra, enquanto o malware habilitado para LLM introduz desafios significativos para abordagens de segurança tradicionais, as dependências e artefatos que essas

ameaças exigem também criam novas oportunidades de detecção e mitigação.

A chave para a defesa eficaz reside na compreensão desses novos vetores de ataque e no desenvolvimento de técnicas inovadoras de caça que podem identificar ameaças antes que elas causem danos significativos.

Encontre esta história interessante! Siga -nos LinkedIn**e**X**Para obter mais atualizações instantâneas**.