BOLETIM DE SEGURANÇA

Ataque avançado LLMjacking com o uso de inteligência artificial explora credenciais de nuvem

heimdall
security research
A DIVISION OF ISH

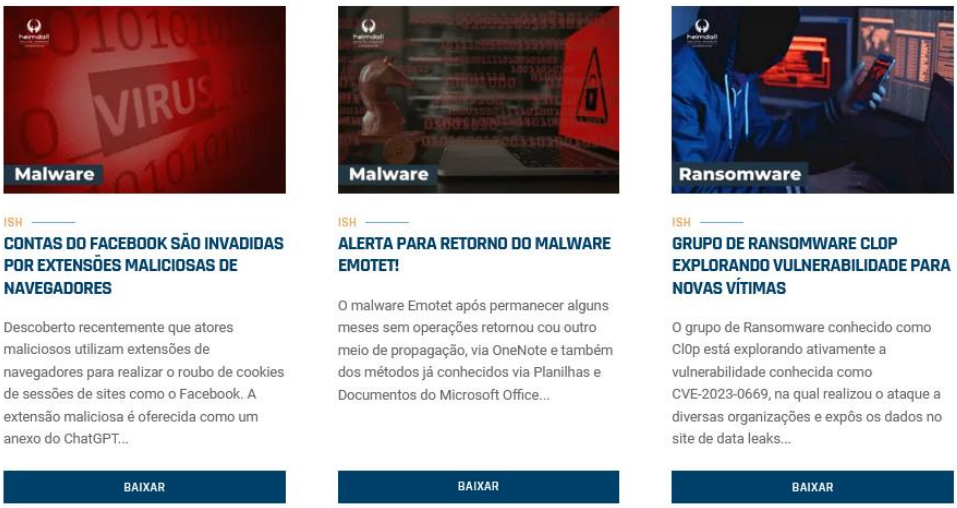
TLP: CLEAR

Receba alertas e informações sobre segurança cibernética e ameaças rapidamente, por meio do nosso **X**.

Heimdall Security Research

Acesse boletins diários sobre agentes de ameaças, *malwares*, indicadores de comprometimentos, TTPs e outras informações no *site* da ISH.

Boletins de Segurança – Heimdall



SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

1 SUMÁRIO EXECUTIVO

Pesquisadores da Sysdig descobriram um ataque intitulado como "**LLMjacking**", que está utilizando credenciais de nuvem roubadas para controlar grandes modelos de linguagem (LLMs) na nuvem. Esse ataque causa perdas financeiras e ameaça a segurança dos dados. O LLMjacking se dá quando invasores acessam a nuvem ilegalmente através da vulnerabilidade CVE-2021-3129 como a encontrada no Laravel, com o objetivo de comprometer a integridade dos sistemas.

2 CADEIA DO ATAQUE

Após garantir o acesso inicial, os atacantes exfiltram credenciais e realizam a invasão no ambiente de nuvem, visando modelos LLM específicos, como o Claude (v2/v3) da Anthropic. Caso não interceptado, esse ataque poderia acarretar custos diários superiores a US$ 46.000 em uso de LLM para as entidades afetadas. As investigações realizadas revelaram o uso de um proxy reverso para acessar LLMs, indicando fins lucrativos. Além disso, a extração de dados de treinamento de LLMs surge como outro possível objetivo dessas invasões cibernéticas.

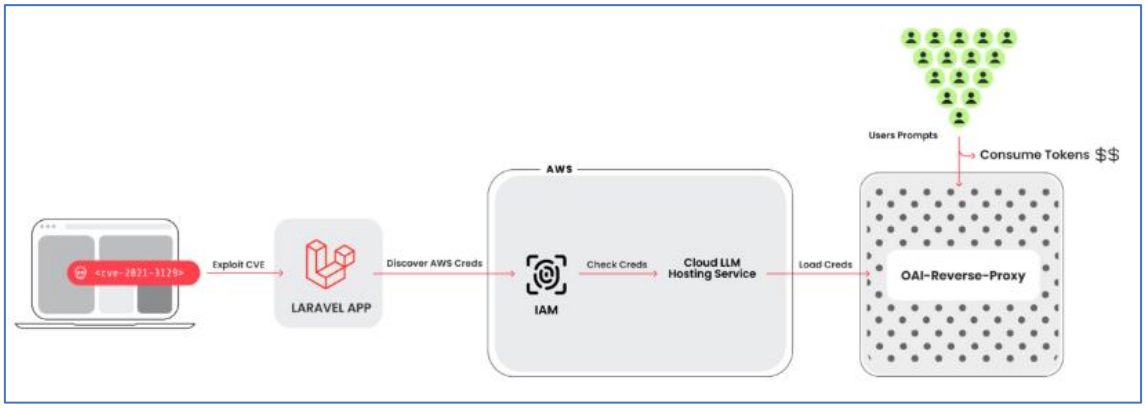


Figura 1 – Cadeia de ataque.

A análise do ataque, permitiu identificar as ferramentas responsáveis por gerar as solicitações que ativavam os modelos de IA. Foi observado um script capaz de testar credenciais em dez diferentes serviços de inteligência artificial, selecionando os que melhor serviam aos objetivos dos invasores. Durante a fase de teste, não foram realizadas nenhuma operação legítima com os LLMs, limitando-se a avaliar o potencial das credenciais e suas limitações de uso. Também analisaram as configurações de registro para minimizar riscos de detecção ao empregar as credenciais obtidas para ativar seus comandos.

Os principais serviços de nuvem, como Azure Machine Learning, Vertex AI do GCP e AWS Bedrock, oferecem agora hospedagem para modelos de linguagem de grande escala (LLM). Estas plataformas permitem que desenvolvedores tenham acesso simplificado a uma variedade de modelos amplamente utilizados em inteligência artificial baseada em LLM. A interface do usuário, conforme mostrado na imagem a seguir, é intuitiva, facilitando aos desenvolvedores a rápida criação de aplicações.

Models	Access status	Modality
AI21 Labs		
Jurassic-2 Ultra	Access granted	Text
Jurassic-2 Mid	Access granted	Text
Amazon		
Titan Embeddings G1 - Text	Access granted	Embedding
Titan Text G1 - Lite	Access granted	Text
Titan Text G1 - Express	Access granted	Text
Titan Image Generator G1 Preview	Access granted	Image
Titan Multimodal Embeddings G1	Access granted	Embedding
Anthropic		
Claude 3 Sonnet	Access granted	Text & Vision
Claude 3 Haiku	Access granted	Text & Vision
Claude	Access granted	Text
Claude Instant	Access granted	Text
Cohere		
Command	Access granted	Text
Command Light	Access granted	Text
Embed English	Access granted	Embedding
Embed Multilingual	Access granted	Embedding
Meta		
Llama 2 Chat 13B	Access granted	Text
Llama 2 Chat 70B	Access granted	Text
Llama 2 13B	Available to request	Text
Llama 2 70B	Available to request	Text
Mistral AI		
Mistral 7B Instruct	Access granted	Text
Mixtral 8x7B Instruct	Access granted	Text
Stability AI		
SDXL 0.8	Access granted	Image
SDXL 1.0	Access granted	Image

Figura 2 – Serviços de nuvem que hospedam modelos LLM.

Para utilizar os modelos de linguagem em nuvem, não basta apenas acessá-los, é necessário enviar uma solicitação específica ao provedor de serviços de nuvem. Dependendo do modelo, o processo de aprovação pode ser instantâneo ou exigir o preenchimento de um formulário, especialmente no caso de modelos desenvolvidos por terceiros. Uma vez que a solicitação é processada, o acesso é concedido com celeridade. Contudo, é importante notar que essa etapa de solicitação serve mais como um entrave para possíveis ataques do que como uma barreira de segurança efetiva.

A interação com modelos de linguagem na nuvem foi facilitada pelos provedores através do uso de comandos CLI simplificados. Após configurar as permissões necessárias, o usuário pode se comunicar com o modelo de maneira descomplicada, utilizando comandos que seguem uma estrutura padrão. Este método agiliza o processo de trabalho com modelos hospedados na nuvem, tornando-o mais acessível e menos técnico.

```
aws bedrock-runtime invocar-model –model-id anthropic.claude-v2 –body
'{“prompt”: “\n\nHumano: história de dois cães\n\nAssistente:”,
“max_tokens_to_sample”: 300}' –cli -formato binário raw-in-base64-out invoque-
modelo-output.txt
```

Figura 3 – Comando para interação com modelos.

O código de verificação essencial que observa se as credenciais são capazes de usar LLMs específicos menciona também o OAI Reverse Proxy, um projeto open-source que serve como um intermediário para serviços LLM. A utilização deste software possibilitaria a um atacante controlar o acesso a múltiplas contas LLM de maneira centralizada, sem revelar as credenciais reais ou, neste contexto, um conjunto de credenciais comprometidas. Em um incidente de ataque, foi observado um agente de usuário associado ao OAI Reverse Proxy tentando acessar modelos LLM com credenciais de nuvem violadas.

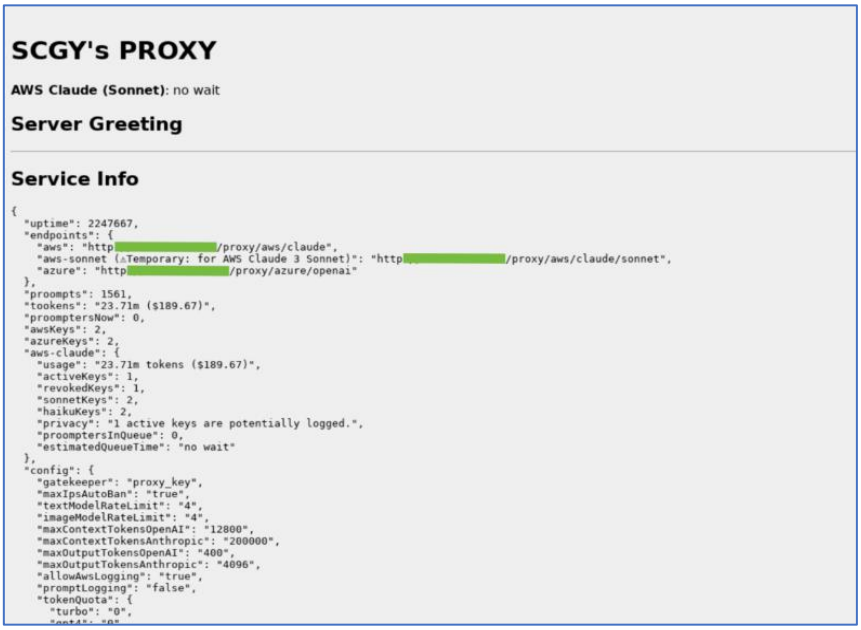


Figura 4 – OAI Reverse Proxy.

A figura abaixo ilustra um exemplo do proxy reverso OAI online. Embora não haja indícios de associação desta instância com o ataque mencionado, ela revela a natureza das informações que são coletadas e apresentadas. É particularmente importante observar os registros de contagens de tokens, custos e chaves, que podem ser capturados pelo sistema.

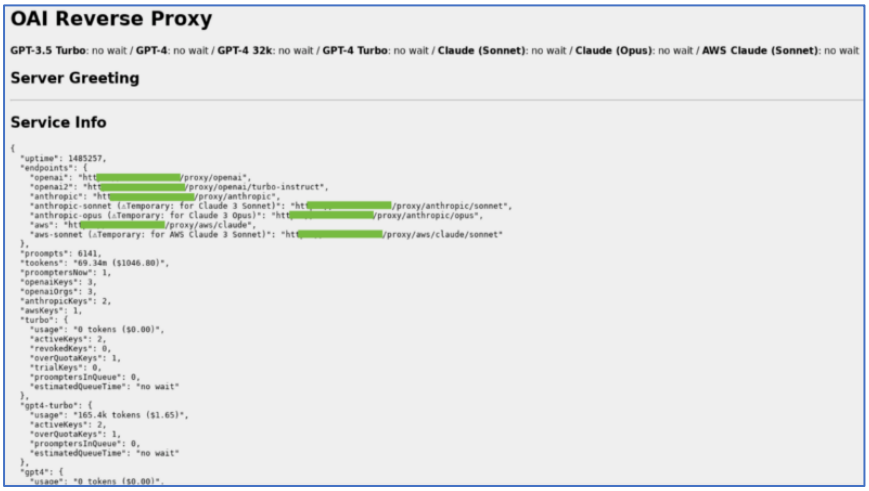


Figura 5 – Proxy reverso online.

No exemplo abaixo apresenta um proxy reverso OAI ajustado para operar com diversos modelos de LLMs. Não existem provas que liguem diretamente esta instância ao ataque discutido. Caso os atacantes estivessem coletando credenciais valiosas com a intenção de comercializar o acesso aos modelos LLM, um proxy reverso dessa natureza seria uma ferramenta viável para a monetização de suas atividades ilícitas.

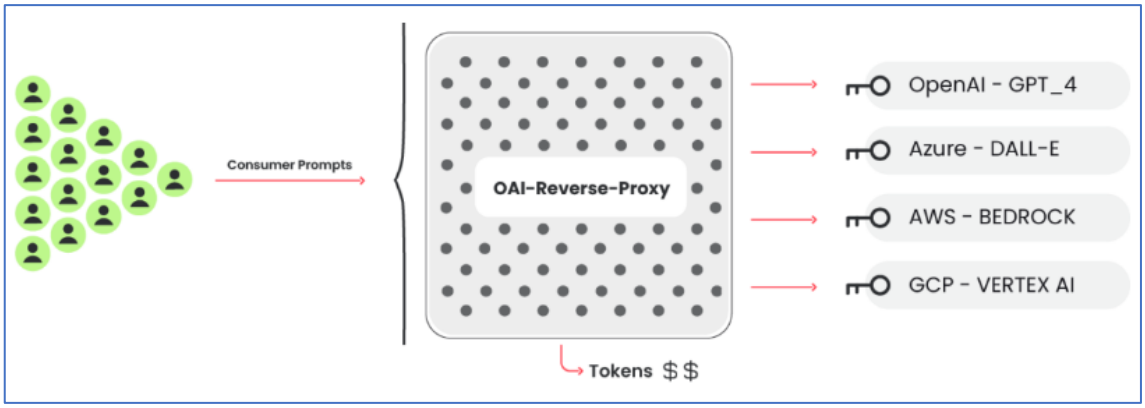
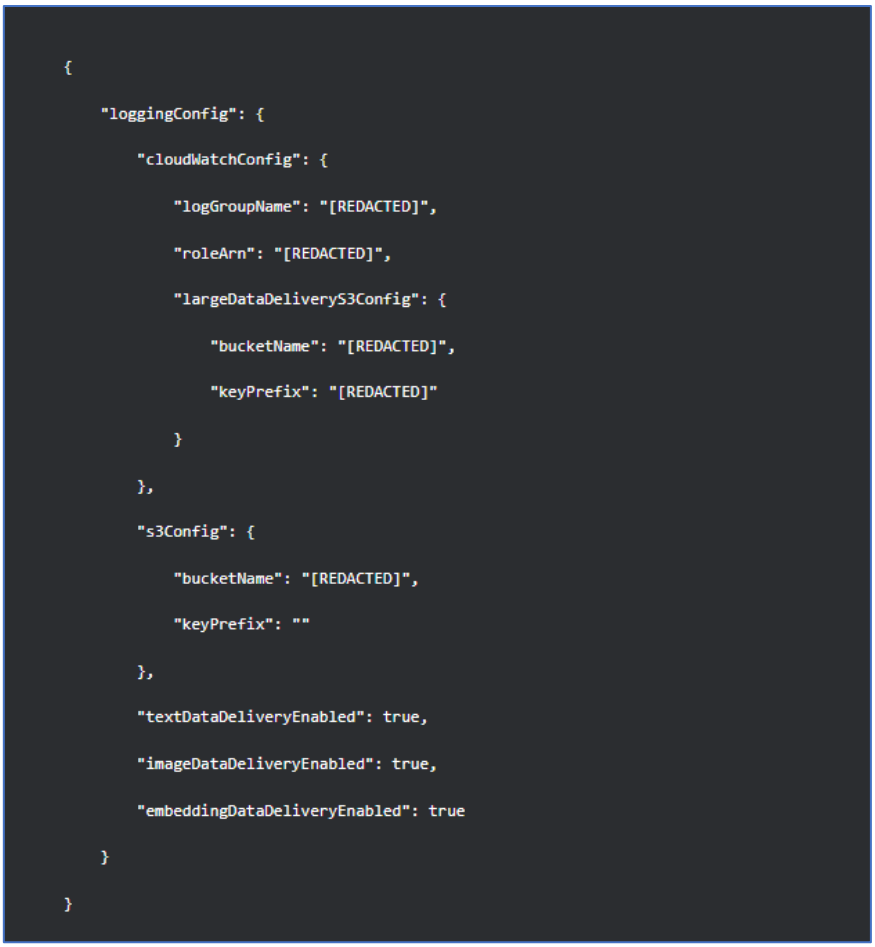


Figura 6 – Proxy reverso OAI ajustado para operar com diversos modelos de LLMs.

Foi analisada a maneira como invasores se infiltraram em um sistema de nuvem. Utilizando comandos de API que pareciam normais, eles testaram os limites de suas permissões de forma discreta, evitando acionar alertas. O caso citado ilustra o uso tático da função InvokeModel, monitorada pelo CloudTrail. Os invasores, propositalmente, ajustaram o parâmetro max_tokens_to_sample para -1, um valor que normalmente geraria um erro. Esse valor atípico teve uma função dupla: confirmou tanto o acesso aos LLMs quanto a operacionalidade dos serviços, como mostrado pela ValidationException gerada. Um erro diferente, como AccessDenied, indicaria restrições no acesso. Esse método de sondagem discreto destaca uma estratégia meticulosa para determinar o alcance das ações permitidas pelas credenciais comprometidas dentro da conta de nuvem.

O evento de InvokeModel, capturado pelo CloudTrail, revela uma tentativa de intrusão. Os atacantes fizeram um pedido aparentemente válido, mas com o parâmetro "max_tokens_to_sample" definido para -1. Esse valor inválido desencadeou uma "ValidationException", indicando aos invasores que possuíam acesso aos LLMs e que estes estavam ativos. Se as credenciais fossem limitadas, o sistema teria retornado um erro "AccessDenied".

Os atacantes mostraram curiosidade pela configuração do serviço, o que é possível verificar através do comando "GetModelInvocationLoggingConfiguration". Este retorna as definições de log do S3 e Cloudwatch, caso estejam ativados. Na nossa estrutura, ambos S3 e Cloudwatch são utilizados para registrar uma quantidade extensa de informações relativas ao ataque.



```
{
    "loggingConfig": {
        "cloudWatchConfig": {
            "logGroupName": "[REDACTED]",
            "roleArn": "[REDACTED]",
            "largeDataDeliveryS3Config": {
                "bucketName": "[REDACTED]",
                "keyPrefix": "[REDACTED]"
            }
        },
        "s3Config": {
            "bucketName": "[REDACTED]",
            "keyPrefix": ""
        },
        "textDataDeliveryEnabled": true,
        "imageDataDeliveryEnabled": true,
        "embeddingDataDeliveryEnabled": true
    }
}
```

Figura 7 – Exemplo de resposta GetModelInvocationLoggingConfiguration.

Os detalhes das operações de prompts e seus resultados não são registrados no Cloudtrail por padrão. Para capturar essas informações, configurações extras são necessárias para direcioná-las ao Cloudwatch e ao S3. Essa medida visa proteger a privacidade das atividades, evitando análises minuciosas. O OAI Reverse Proxy se compromete a não empregar chaves AWS com registro ativo, reforçando a "privacidade". Assim, fica inviável verificar os prompts e respostas quando o AWS Bedrock é utilizado.

Em um cenário de LLMjacking, os prejuízos se manifestam como custos adicionais impostos à vítima. Não é segredo que a utilização de um LLM implica em despesas consideráveis, que podem escalar rapidamente. No pior caso, se um atacante explorar o Anthropic Claude 2.x e exceder a cota em diversas regiões, o custo diário para a vítima pode ultrapassar US$ 46.000. A tabela de preços e cotas iniciais do Claude 2 indica que 1.000 tokens de entrada custam US$ 0,008 e 1.000 tokens de saída, US$ 0,024. Com um limite de processamento de até 500.000 tokens de entrada e saída por minuto conforme o AWS Bedrock, o custo médio para 1.000 tokens é de US$ 0,016. Calculando o custo total, temos: (500.000 \text{ tokens} / 1.000 \times US\$ 0,016) \times 60 \text{ minutos} \times 24 \text{ horas} \times 4 \text{ regiões} = US\$ 46.080/\text{dia}

Além dos custos, os atacantes podem maximizar as cotas para impedir que a entidade afetada utilize os modelos legitimamente, paralisando as operações comerciais. A detecção e resposta ágil a ameaças são cruciais para uma defesa eficaz. A análise dos registros de atividade na nuvem é essencial para identificar comportamentos anômalos ou acessos indevidos. Utilizando plataformas como Falco e Sysdig Secure, é possível reconhecer padrões de ataque e agir prontamente. Os clientes da Sysdig Secure têm à disposição a regra específica na política AWS Notable Events, que auxilia na detecção e resposta a tais incidentes.

```
- rule: Bedrock Model Recon Activity

  desc: Detect reconaissance attempts to check if Amazon Bedrock is enabled, based
on the error code. Attackers can leverage this to discover the status of Bedrock,
and then abuse it if enabled.

    condition: jevt.value[/eventSource]="bedrock.amazonaws.com" and
jevt.value[/eventName]="InvokeModel" and
jevt.value[/errorCode]="ValidationException"

    output: A reconaissance attempt on Amazon Bedrock has been made (requesting
user=%aws.user, requesting IP=%aws.sourceIP, AWS region=%aws.region,
arn=%jevt.value[/userIdentity/arn], userAgent=%jevt.value[/userAgent],
modelId=%jevt.value[/requestParameters/modelId])

    priority: WARNING
```

Figura 8 – Regra Falco.

É possível estabelecer alertas no CloudWatch para monitorar e responder a atividades atípicas. As métricas de desempenho do Bedrock, quando observadas, podem deflagrar notificações de comportamentos que demandam atenção.

O roubo de credenciais em plataformas de nuvem e SaaS é um método de ataque frequente, com expectativa de crescimento à medida que atacantes descobrem novos modos de explorar esses acessos para lucro. Embora o custo dos serviços LLM varie conforme o modelo e o volume de tokens, os desenvolvedores buscam eficiência, diferentemente dos atacantes, que não possuem tal preocupação. Por isso, é vital ter sistemas de detecção e resposta ágeis para combater essas ameaças prontamente.

3 RECOMENDAÇÕES

Além dos indicadores de comprometimento elencados abaixo pela ISH, poderão ser adotadas medidas visando a mitigação da infecção da referida *ameaça*, como por exemplo:

Atualize regularmente

- Mantenha todos os sistemas e softwares atualizados para corrigir vulnerabilidades que possam ser exploradas por atacantes.

Use credenciais fortes

- Implemente políticas de senha forte e altere-as regularmente para evitar acessos não autorizados.

Monitoramento de acesso

- Monitore e controle o acesso aos serviços de LLM na nuvem para detectar atividades suspeitas rapidamente.

Proteção de rede

- Utilize firewalls e outras ferramentas de segurança de rede para monitorar e controlar o tráfego de dados.

Educação em segurança

- Treine os colaboradores sobre práticas de segurança para que possam reconhecer e evitar ataques de phishing e outras ameaças.

Criptografia

- Proteja dados sensíveis usando criptografia, garantindo que apenas pessoas autorizadas tenham acesso.

Softwares de proteção

- Instale e mantenha atualizados programas de antivírus, anti-malware e anti-ransomware para uma defesa em camadas contra ataques.

4 INDICADORES DE COMPROMISSOS

A ISH Tecnologia realiza o tratamento de diversos indicadores de compromissos coletados por meio de fontes abertas, fechadas e também de análises realizadas pela equipe de segurança Heimdall. Diante disto, abaixo listamos todos os Indicadores de Compromissos (IOCs) relacionadas a análise do(s) artefato(s) deste relatório.

Indicadores de URL, IPs e Domínios

Indicadores de URL, IPs e Domínios	
IP	83.7.139.184 83.7.157.76 73.105.135.228 83.7.135.97

Tabela 1 – Indicadores de Compromissos de Rede.

Obs: Os *links* e endereços IP elencados acima podem estar ativos; cuidado ao realizar a manipulação dos referidos IoCs, evite realizar o clique e se tornar vítima do conteúdo malicioso hospedado no IoC.

5 REFERÊNCIAS

- **Heimdall *by* ISH Tecnologia**
- **Sysdig**
- **Gbhackers**

6 AUTORES

- **Leonardo Oliveira Silva**

heimdall
security research
A DIVISION OF ISH