New K2 Think Ai Model cai para o jailbreak em tempo recorde - Against Inv

Data: 2025-09-12 06:51:26

Autor: Inteligência Against Invaders

Surgiu uma vulnerabilidade inovadora no recém -lançado modelo K2 Think AI da Universidade de

Inteligência Artificial de Mohamed Bin Zayed, nos Emirados Árabes Unidos (MBZUAI), em

colaboração com o G42.

Os pesquisadores de segurança tiveram com sucesso o Sistema de Raciocínio Avançado Poucas horas após sua libertação pública, expondo uma falha crítica que transforma os recursos de transparência do modelo em um vetor de ataque.

A vulnerabilidade permite que os invasores mapeem sistematicamente e ignorem as medidas de segurança, explorando o próprio processo de raciocínio do modelo, transformando tentativas de falha no passo para eventual comprometimento.

O modelo K2 Think incorpora recursos sofisticados de raciocínio projetados para fornecer processos de tomada de decisão transparentes, tornando-o atraente para aplicativos corporativos que exigem trilhas de auditoria e IA explicável.

No entanto, essa transparência se tornou sua maior fraqueza. Pesquisadores de segurança da plataforma de equipes vermelhas da Adversa AI descoberto O fato de o processo de pensamento interno do modelo expor inadvertidamente instruções no nível do sistema e protocolos de segurança, criando um roteiro para os invasores refinarem suas tentativas de jailbreak iterativamente.

Ao contrário dos jailbreaks tradicionais de IA que tenham sucesso ou fracassaram completamente, essa nova metodologia de ataque explora os registros de raciocínio para criar um ciclo de feedback.

Cada tentativa fracassada revela fragmentos da arquitetura de segurança subjacente, incluindo números de regras específicos, hierarquias defensivas e protocolos de meta-segurança.

Essas informações se tornam progressivamente mais valiosas à medida que os invasores mapeiam sistematicamente toda a estrutura defensiva através de sondagens repetidas.

Diagrama de uma arquitetura de agente de IA mostrando a interação do usuário com vários módulos, incluindo Rachoining LLM, execução de sandbox e pesquisa na Internet

A metodologia de ataque iterativa

O ataque segue um padrão trifásico distinto que armazia a transparência contra a segurança. Na inicial reconhecimento Fase, os pesquisadores começaram com prompts padrão de jailbreak projetados para ignorar as diretrizes de segurança.

Enquanto o modelo recusou corretamente essas solicitações, seus logs de raciocínio expostos informações críticas sobre sua estrutura defensiva, incluindo referências a regras de segurança específicas e seus sistemas de indexação.

Por exemplo, depois de descobrir a “Regra #7” sobre atividades prejudiciais, os avisos subsequentes abordaram explicitamente essa restrição ao investigar camadas defensivas mais profundas. Cada iteração expôs meta-rajas adicionais e protocolos de segurança de ordem superior.

A fase de exploração final demonstrou o devastador efeito cumulativo dessa abordagem. Depois de mapear camadas defensivas suficientes por meio de sondagem sistemática, os invasores construíram instruções sofisticadas que abordaram simultaneamente várias medidas de segurança descobertas.

A segunda fase envolveu a neutralização direcionada, onde os invasores criaram instruções projetadas especificamente para combater as medidas defensivas reveladas em tentativas anteriores.

Esse padrão de vulnerabilidade representa ameaças sérias às implantações da IA ??corporativa em vários setores.

Assistência médica Sistemas de IA Isso explica o raciocínio diagnóstico pode ser manipulado para revelar critérios de diagnóstico proprietários ou facilitar os esquemas de fraude de seguros.

Os algoritmos de negociação financeira que fornecem transparência de raciocínio podem ter sua lógica de engenharia reversa para fins de manipulação do mercado.

As plataformas educacionais que usam IA explicáveis ??para o monitoramento da integridade acadêmica se tornam particularmente vulneráveis, pois os alunos podem aprender sistematicamente a ignorar os mecanismos de detecção por meio de testes iterativos.

O padrão de falha em cascata significa que as avaliações iniciais de segurança podem mostrar uma defesa bem -sucedida contra ataques, além de perder o vazamento de informações críticas, permitindo eventual comprometimento.

A equipe vermelha envolve a interseção de tecnologia, pessoas e segurança física para identificar vulnerabilidades

A vulnerabilidade é especialmente preocupante porque transforma a transparência da IA ??- um recurso cada vez mais exigido para fins de conformidade regulatória e auditoria – em um passivo de segurança.

As empresas que pretendem a implantação de sistemas de IA explicáveis ??podem, sem saber, criar plataformas que treinam atacantes em tempo real, com cada resposta defensiva fornecendo inteligência para ataques mais sofisticados.

Mitigações

As medidas de proteção imediatas incluem a implementação de filtros de higienização de raciocínio que removem referências a regras específicas ou medidas defensivas de saídas visíveis.

O resultado foi completo do sistema de segurança, com o modelo produzindo conteúdo restrito, incluindo instruções detalhadas de criação de malware e outras saídas prejudiciais.

Tentativas de falha limitantes da taxa com atrasos exponenciais podem tornar impraticáveis ??ataques iterativos de refinamento, enquanto as regras do Honeypot no raciocínio podem confundir as tentativas de mapeamento, incluindo medidas defensivas falsas.

As soluções de longo prazo requerem mudanças fundamentais na arquitetura de segurança da IA. As organizações devem desenvolver modos de raciocínio opacos, onde os processos internos de tomada de decisão permanecem completamente ocultos durante as operações sensíveis à segurança.

Este incidente ressalta a importância crítica da IA ??avançada Equipe vermelha na identificação de novos vetores de ataque antes da implantação pública.

As técnicas de privacidade diferenciais podem adicionar ruído aos toras de raciocínio, preservando a interpretabilidade geral, e os sistemas de defesa adaptativa podem detectar tentativas de mapeamento e alterar dinamicamente as estruturas defensivas.

Diagrama de Venn mostrando a sobreposição entre segurança cibernética, equipes vermelhas tradicionais e equipes de AI Red, incluindo considerações éticas e legais compartilhadas

O K2 Think Vulnerability representa um momento decisivo na segurança da IA, destacando a tensão fundamental entre transparência e segurança nos sistemas modernos de IA.

À medida que as organizações exigem cada vez mais a IA explicável para fins de conformidade e auditoria, elas devem equilibrar cuidadosamente esses requisitos em relação às considerações de segurança.

A visão binária tradicional da segurança cibernética – os sistemas são violados ou seguros – fornecem insuficientes para plataformas de IA que podem educar inadvertidamente os atacantes através de suas respostas defensivas.

À medida que os sistemas de IA se tornam essenciais para a infraestrutura crítica e as operações comerciais, a comunidade de segurança cibernética deve desenvolver novos paradigmas que protejam contra ataques bem -sucedidos e o vazamento de informações que os permite.

A corrida entre a segurança da IA ??e a exploração da IA ??entrou em uma nova fase, onde até ataques fracassados ??podem fornecer vitórias para adversários determinados.

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