DeepSeek sob fogo: 50% do código malicioso produzido em consultas co

Data: 2025-09-20 07:21:03

Autor: Inteligência Against Invaders

Redazione RHC**:20 Setembro 2025 09:12**

Especialistas na Ataque de multidão conduziu uma série de experimentos com o sistema de inteligência artificial chinês **Busca Profunda**, testando seu **Geração de código com base em termos de consulta**. Eles descobriram que **Os resultados dependiam diretamente da identidade do cliente ou da organização associada.**

Se as consultas incluíssem cenários neutros ou mencionassem os Estados Unidos, o modelo produzia código limpo, bem estruturado e resistente a ataques. No entanto, assim que o projeto foi vinculado a tópicos que provocaram uma reação negativa do governo chinês, a qualidade das soluções diminuiu drasticamente.

Os exemplos mais notáveis envolveram consultas de praticantes e organizações do Falun Gong que mencionaram o Tibete, Taiwan ou a região uigur de Xinjiang. Nesses casos, o sistema geralmente gerava fragmentos contendo vulnerabilidades críticas, permitindo que invasores acessassem o sistema. No caso do Falun Gong, até metade das consultas foram bloqueadas por filtros e não geraram nenhum código, enquanto uma parte significativa das consultas restantes continha falhas graves. Um padrão semelhante foi observado com as referências ao ISIS: o modelo rejeitou aproximadamente 50% das consultas e as respostas resultantes continham erros graves.

A CrowdStrike enfatiza que esses não são backdoors intencionais. O código gerado parecia desleixado e inseguro, o que pode ser devido a dados de treinamento inadequados ou filtros ideológicos integrados. *Esses filtros, de acordo com os pesquisadores, podem reduzir a confiabilidade das soluções para grupos politicamente “indesejáveis”, mas o fazem indiretamente, por meio de implementações falhas.*

Os dados confirmam *a natureza sistêmica do problema. Para consultas relacionadas aos EUA, a probabilidade de erros graves era mínima, inferior a 5%, e essas eram principalmente pequenas falhas lógicas sem risco real de exploração. Para a Europa e projetos “neutros”, a taxa de problemas estava entre 10 e 15%. No entanto, para tópicos envolvendo organizações sensíveis à China, as estatísticas mudaram drasticamente: cerca de 30% das amostras continham injeção de SQL, outros 25% foram acompanhados por estouros de buffer e outros erros de memória e cerca de 20% envolveram manuseio inseguro da entrada do usuário, sem validação ou escape de string.*

No caso do Falun Gong e do ISIS, entre as consultas desbloqueadas, **quase uma em cada duas gerações continha vulnerabilidades críticas,** elevando a porcentagem geral de soluções maliciosas para mais de 50%.

Em conclusão, a CrowdStrike adverte que, **mesmo que o trabalho do DeepSeek não seja malicioso, a própria existência de tais dependências abre oportunidades significativas para os invasores. Atacantes.** O código vulnerável pode acabar em projetos reais, sem saber que os problemas decorrem da arquitetura politicamente motivada do modelo. Tais vulnerabilidades representam sérios riscos de segurança cibernética para organizações em todo o mundo.

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