Vencedores do AI Cyber Challenge revelados no relatório de segurança ci

Data: 2025-08-09 11:45:00

Autor: Inteligência Against Invaders

Inteligência Against Invaders

2025-08-09 08:45

Após dois anos de competição, os vencedores do AI Cybersecurity Challenge (AIxCC) foram revelados no evento de hackers DEFCON 33 em 9 de agosto.

A equipe de Atlanta foi revelada como a equipe vencedora. O grupo é uma colaboração poderosa de especialistas do Instituto de Tecnologia da Geórgia (Georgia Tech), Samsung Research, Instituto Avançado de Ciência e Tecnologia da Coreia e da Universidade de Ciência e Tecnologia de Pohang. Eles ganharam um prêmio de US $ 4 milhões.

A Trail of Bits, uma empresa de segurança cibernética com sede em Nova York especializada em pesquisa de segurança de ponta, ficou em segundo lugar, garantindo um prêmio de US$ 3 milhões no AI Cyber Challenge de alto risco.

A terceira equipe com melhor desempenho foi a Theori, um grupo de pesquisadores de IA e profissionais de segurança dos EUA e da Coreia do Sul, completando o pódio na vitrine competitiva da Agência de Projetos de Pesquisa Avançada de Defesa (DARPA), com um prêmio de US$ 1,5 milhão.

Os três sistemas de raciocínio cibernético desenvolvidos pelo trio fazem parte de um conjunto de quatro modelos que foram abertos e já estão disponíveis para uso de todos.

“Os outros três modelos serão disponibilizados nas próximas semanas”, disse o diretor da DARPA, Stephen Winchell, durante a sessão de anúncio no DEFCON 33.

AIxCC: dois anos em construção

Anunciado na Black Hat 2023 por Perri Adams, gerente de programa da DARPA, o AIxCC foi uma competição para cientistas da computação, especialistas em IA, desenvolvedores de software e outros especialistas em segurança cibernética para criar uma nova geração de ferramentas de segurança cibernética baseadas em IA para proteger a infraestrutura crítica dos EUA e os serviços governamentais.

Especificamente, a DARPA e a Agência de Projetos de Pesquisa Avançada para a Saúde (ARPA-H), outra agência do governo dos EUA, financiaram este projeto para explorar se a IA pode ajudar a encontrar e corrigir vulnerabilidades de software de forma mais eficaz e inaugurar um futuro em que os ataques podem ser interrompidos tão rápido quanto são detectados.

Os sete finalistas (Team Atlanta, Trail of Bits, Theori, All You Need IS A Fuzzing Brain, Shellphish, 42-b3yond-6ug e Lacrosse) foram anunciado no DEFCON 32 em agosto de 2024 e receberam US$ 2 milhões cada.

Os gigantes da tecnologia Google, Microsoft, Anthropic e OpenAI apoiaram coletivamente a competição com mais de US$ 1 milhão cada em créditos de modelo de IA, garantindo que as equipes tivessem o poder de fogo computacional necessário para enfrentar os desafios críticos de segurança da infraestrutura.

Falando antes do anúncio dos vencedores, Jim O'Neill, vice-secretário do Departamento de Saúde e Serviços Humanos dos EUA (HHS), disse que a DARPA e a ARPA-H injetarão US$ 1,4 milhão adicionais além dos US$ 29,5 milhões planejados para prêmios em dinheiro.

Durante uma coletiva de imprensa pós-anúncio, Andrew Carney, gerente de programa do AIxCC, revelou que o financiamento adicional apoiará os finalistas no refinamento de suas ferramentas para implantação no mundo real.

A distribuição desses fundos adicionais ocorrerá em incrementos faseados, sujeito às equipes vencedoras demonstrarem a adoção mensurável de suas ferramentas pelas principais organizações de infraestrutura.

Abordagens alimentadas por IA corrigem falhas mais rapidamente por US$ 152 por correção

Durante a fase final do AIxCC, realizada no ano passado, as equipes participantes foram obrigadas a implantar seus sistemas em um ambiente controlado e simulado, deliberadamente semeado com falhas introduzidas pelos organizadores da competição.

As sete equipes finalistas descobriram 54 das 70 vulnerabilidades sintéticas intencionalmente incorporadas ao desafio, representando uma taxa de detecção de 77%.

Esta é uma melhoria significativa em comparação com a rodada semifinal do ano passado, durante a qual as equipes descobriram apenas 37% das vulnerabilidades conhecidas.

Eles foram capazes de corrigir 43 desses 54.

As sete equipes finalistas também detectaram 18 falhas do mundo real anteriormente desconhecidas que não foram plantadas pelos organizadores e corrigiram 11 delas.

Essas descobertas de dia zero destacam a capacidade dos modelos de identificar pontos fracos críticos além dos ambientes de teste controlados.

“Estamos agora no processo de divulgação [these real-world zero-day vulnerabilities] aos mantenedores”, disse Carney no palco.

Velocidade e eficiência eram pontos fortes definidores. Em média, os sistemas de IA corrigiram vulnerabilidades em apenas 45 minutos, uma melhoria dramática em relação aos processos manuais tradicionais.

Jennifer Roberts, diretora de sistemas resilientes da ARPA-H, disse à imprensa que essas capacidades são particularmente importante no setor da saúde, em que são necessários, em média, 491 dias para corrigir uma vulnerabilidade, em comparação com 60 a 90 dias noutros setores.

Além disso, o custo unitário para a conclusão da tarefa no A concorrência foi quantificada em US$ 152, demonstrando uma vantagem de custo acentuada sobre os gastos tradicionais com força de trabalho humana.

“Este é o novo piso – vai melhorar rapidamente. “Para nos tornarmos mais seguros, precisamos tornar todos mais seguros. Este é o caminho”, disse Carney.

Winchell acrescentou: “Estamos vivendo em um mundo agora que tem andaimes digitais antigos que estão segurando tudo. Muitas das bases de código, muitas linguagens, muitas das maneiras como fazemos negócios e tudo o que construímos em cima disso incorreram em enormes dívidas técnicas ao longo dos anos.”

Prêmio em dinheiro alimenta futuras pesquisas de segurança de IA para as principais equipes

A equipe vencedora, Team Atlanta, alcançou sucesso em várias competições de hackers e conferências acadêmicas. Para AIxCC, eles usaram principalmente métodos tradicionais de descoberta de vulnerabilidades (por exemplo, análise dinâmica, análise estática, fuzzing) com os modelos de linguagem grande (LLMs) da OpenAI, como o4-mini, GPT-4o e o3.

Eles lideraram todas as categorias, exceto uma, e descobriram as vulnerabilidades mais reais das sete equipes.

Questionado sobre o que sua equipe faria com o dinheiro, Taesoo Kim, o principal líder da equipe e professor da Georgia Tech, disse que concordou em oferecer uma grande parte do prêmio em dinheiro ao instituto para ajudar a apoiar futuros desenvolvimentos na pesquisa de vulnerabilidades alimentada por IA.

A vencedora da medalha de prata, Trail of Bits, é uma pequena empresa composta por 10 engenheiros com profunda experiência no desenvolvimento de novas ferramentas de segurança de

software, incluindo seu próprio sistema de raciocínio cibernético, Buttercup.

Um de seus parceiros mais notáveis é o Reino Unido Instituto de Segurança de IA.

Para AIxCC, Trail of Bits combinou Buttercup e métodos tradicionais de descoberta de vulnerabilidades com LLMs como Claude Sonnet 4, GPT-4.1 e GPT-4.1 mini da Anthropic. Suas conquistas incluem o maior número de categorias de vulnerabilidade exclusivas, também conhecidas como Fraquezas Comuns e Categorias de Enumeração (CWEs).

O terceiro vencedor, Theori, tem uma longa história de vitórias em competições de segurança, incluindo oito vitórias nas finais de captura da bandeira da DEFCON.