ESTUDO DA UNIVERSIDADE DE COIMBRA, APOIADO PELA INDRA GROUP, REVELA VULNERABILIDADES EM MODELOS AVANÇADOS DE IA

- **Investigação propõe enquadramento inovador para avaliar a segurança dos modelos de IA e alerta para vulnerabilidades que desafiam a nova geração de sistemas inteligentes**
- **Estudo mostra que mais de 80% dos modelos testados geraram código inseguro quando expostos a ataques de manipulação dissimulados**

Lisboa, 26 de novembro de 2025. – A Universidade de Coimbra, em colaboração com a Indra Group, acaba de apresentar o whitepaper "IA e Cibersegurança: O Desafio da Confiança Digital", que analisa os riscos, vulnerabilidades e dilemas éticos da Inteligência Artificial (IA) e propõe um enquadramento inovador para testar a segurança dos grandes modelos de linguagem (LLMs) — a tecnologia que sustenta a atual geração de assistentes de IA generativa.

Desenvolvido por João Donato, sob a orientação de João Campos, investigadores do Centro de Informática e Sistemas da Universidade de Coimbra (CISUC) e do Laboratório de Sistemas Inteligentes (LASI), este estudo faz um retrato real da maturidade da IA, uma tecnologia poderosa e transformadora, mas ainda longe de ser invulnerável.

Entre as conclusões mais relevantes, destaque para:

- Mais de 80% dos modelos testados geraram código inseguro quando expostos a ataques de manipulação dissimulados;
- As técnicas multi-turno e de role-play (como Crescendo ou Mr. Robot) continuam capazes de contornar mecanismos de segurança considerados robustos;
- Os modelos mais recentes, como o Llama 3.1:70b, mostram avanços na distinção entre risco real e aparente, mas mantêm fragilidades contextuais que exigem vigilância constante.

Segundo os investigadores da Universidade de Coimbra, o verdadeiro desafio da próxima geração de IA será o de encontrar o equilíbrio entre a utilidade e o risco, construindo sistemas "seguros por design", onde a inovação e a segurança evoluem em conjunto.

"A segurança da IA precisa de ser mensurável, comparável e contínua. Só assim será possível criar confiança digital real e sustentável", referem João Donato e João Campos, investigadores da Universidade de Coimbra e autores do estudo.

O Enquadramento proposto pelos investigadores permite precisamente avaliar e comparar a robustez dos modelos face a diferentes tipos de ataques, combinando métricas objetivas, cenários realistas e um "júri automatizado" de modelos independentes. O objetivo é o de transformar a investigação científica em valor prático, contribuindo para uma IA mais ética, transparente e segura.

O papel da tecnologia é decisivo na deteção precoce destas vulnerabilidades. Ferramentas avançadas de monitorização, algoritmos de análise comportamental e sistemas automatizados de auditoria são essenciais para identificar riscos antes que possam comprometer a integridade dos modelos de IA, garantindo maior segurança e confiança digital. Desde a otimização de processos até à criação de soluções inovadoras, a tecnologia é o motor que impulsiona competitividade, sustentabilidade e segurança.

Para a Indra Group, que apoiou o desenvolvimento e publicação do estudo, esta colaboração reflete um compromisso claro com o avanço do conhecimento e com a criação de conteúdo científico relevante, produzido em Portugal, na área da tecnologia e da Inteligência Artificial, que pode ser transformado em valor real para as empresas.

"Acreditamos que a confiança digital tem de ser o novo pilar da transformação tecnológica. Ao apoiar o desenvolvimento deste género de investigações, a Indra Group está a reforçar o seu compromisso em tornar

a cibersegurança um motor de valor e confiança, antecipando riscos e promovendo uma IA ética e responsável", afirma António Ribeiro, responsável de Cibersegurança da Minsait (Indra Group) em Portugal.

O whitepaper está também disponível em formato e-book, para consulta online, permitindo uma perspetiva inovadora sobre como Portugal pode posicionar-se na vanguarda da confiança digital e da inovação responsável por parte das empresas que estão a adotar estas tecnologias

Sobre a Indra Group

A Indra Group é uma holding que promove o progresso tecnológico, que inclui a Indra, empresa global em defesa, tráfego aéreo e espaço; e a Minsait, líder em novos ambientes digitais e tecnologias disruptivas. A Indra Group impulsiona um futuro mais seguro e conectado através de soluções inovadoras, relações de confiança e o melhor talento. A sustentabilidade faz parte da sua estratégia e cultura, com o objetivo de responder aos desafios sociais e ambientais presentes e futuros. No final de 2024, a Indra Group tinha um volume de negócios de 4.843 milhões de euros, presença local em 46 países e operações comerciais em mais de 140 países.

Em Portugal desde 1997, a Indra, com escritórios em Lisboa, Porto e Amarante, conta com uma sólida equipa de profissionais com elevada especialização para o desenvolvimento e implementação das suas soluções e serviços. A empresa integra alguns dos projetos mais inovadores que são chave para o desenvolvimento económico e tecnológico do país nos sectores de Defesa, Aeroespaço e Mobility e através da sua filial Minsait, nas Tecnologias de Informação.

Contactos de Comunicação

Corpcom - Cátia Gil
catia.gil@corpcom.pt

Corpcom – Rodrigo Almeida Fernandes
rodrigo.fernandes@corpcom.pt