Trend Micro warnt: Unkontrollierter Einsatz von KI birgt versteckte Geschäftsrisiken

ein Kommentar von
Robert McArdle
Director of Cybersecurity Research
Trend Micro
www.trendmicro.com

Trend Micro, einer der weltweit führenden Anbieter von Cybersicherheitslösungen, veröffentlicht neue Forschungsergebnisse, die davor warnen, dass Unternehmen beim Einsatz von Large Language Models (LLMs) ohne angemessene Regularien, Verifikation und Aufsicht erhebliche rechtliche, finanzielle und reputationsbezogene Risiken eingehen.

Die Studie Risks of Unmanaged AI Reliance: Evaluating Regional Biases, Geofencing, Data Sovereignty, and Censorship in LLM Models zeigt, dass KI-Systeme je nach geografischem Standort, Sprache, Modelldesign und integrierten Kontrollmechanismen teils deutlich unterschiedliche Ergebnisse lieferten. In kundenorientierten oder entscheidungsunterstützenden Einsatzszenarien könnten solche Inkonsistenzen das Vertrauen untergraben, im Widerspruch zu lokalen regulatorischen oder kulturellen Rahmenbedingungen stehen und kostspielige geschäftliche Folgen nach sich ziehen.

Mehr als 100 KI-Modelle getestet

Trend-Micro-Forscher testeten mehr als 100 KI-Modelle mit über 800 gezielt entwickelten Prompts, um Verzerrungen, politisches und kulturelles Kontextverständnis, Geofencing-Verhalten, Signale zur Datensouveränität sowie kontextuelle Einschränkungen zu analysieren. Tausende wiederholte Experimente dienten dazu, Veränderungen der Ausgaben über Zeit und Standorte hinweg zu messen. Insgesamt werteten die Forscher mehr als 60 Millionen Eingabetokens und über 500 Millionen Ausgabetokens aus.

Identische Eingaben - unterschiedliche Antworten

Die Ergebnisse zeigen, dass identische Eingaben je nach Region und Modell unterschiedliche Antworten erzeugen und selbst bei wiederholten Interaktionen mit demselben System variieren. In politisch sensiblen Szenarien, etwa bei der Darstellung umstrittener Gebiete oder nationaler Identitäten, weisen die Modelle klare regionale Ausrichtungsunterschiede auf. In weiteren Tests lieferten die Systeme inkonsistente oder veraltete Ergebnisse in Bereichen, die hohe Präzision erfordern, darunter Finanzberechnungen und zeitkritische Informationen.

KI und klassische Software

„Viele Unternehmen gehen davon aus, dass KI wie klassische Software funktioniert, bei der dieselbe Eingabe zuverlässig dieselbe Ausgabe erzeugt“, erklärt Robert McArdle, Director of Cybersecurity Research bei Trend Micro. „Unsere Forschung zeigt, dass diese Annahme nicht zutrifft. LLMs verändern ihre Antworten abhängig von Region, Sprache und Schutzmechanismen und diese Antworten können sich von einer Interaktion zur nächsten unterscheiden. Werden KI-Ausgaben direkt in Customer Journeys oder Geschäftsentscheidungen eingebunden, riskieren Unternehmen den Verlust der Kontrolle über Markenkommunikation, Compliance-Positionierung und kulturelle Anschlussfähigkeit.“

Verstärktes Risiko

Der Bericht hebt hervor, dass sich diese Risiken insbesondere für globale Unternehmen verstärken, die KI grenzüberschreitend einsetzen. Dort müsse ein KI-gestützter Service häufig gleichzeitig unterschiedlichen rechtlichen Rahmenbedingungen, politischen Sensibilitäten und gesellschaftlichen Erwartungen gerecht werden. Zudem verweist die Studie auf besondere Herausforderungen für den öffentlichen Sektor: KI-generierte Inhalte könnten dort als offizielle Orientierung wahrgenommen werden, während der Einsatz nicht lokalisierter Modelle Risiken für Souveränität und Zugänglichkeit berge.

Fazit

„KI sollte nicht als Plug-and-Play-Produktivitätswerkzeug betrachtet werden“, ergänzt Robert McArdle. „Unternehmen müssen sie als Abhängigkeit mit hohem Risiko behandeln, die klare Regularien, eindeutige Verantwortlichkeiten und die menschliche Verifikation aller nutzerseitigen Ausgaben erfordert. Dazu gehört auch, von KI-Anbietern Transparenz darüber einzufordern, wie sich Modelle verhalten, auf welchen Daten sie basieren und wo Schutzmechanismen greifen. Wenn KI mit einem klaren Verständnis ihrer Grenzen eingesetzt wird und mit Kontrollmechanismen, die dem tatsächlichen Verhalten dieser Systeme in realen Umgebungen Rechnung tragen, kann sie Innovation und Effizienz erheblich vorantreiben.“

Vollständige Studie

Die vollständige Studie Risks of Unmanaged AI Reliance: Evaluating Regional Biases, Geofencing, Data Sovereignty, and Censorship in LLM Models finden Sie hier:
https://www.trendmicro.com/vinfo/de/security/news/cybercrime-and-digital-threats/how-unmanaged-ai-adoption-puts-enterprises-at-Risk ◄