Google Cloud

Les lundis de l'IA et de la finance

2 septembre 2024

Thiébaut Meyer

Directeur cybersécurité
Office of the CISO
Google Cloud
thiebaut@google.com

Hacking Google

Les principes de sécurité

Trust nothing

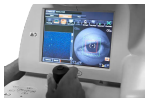
Detect everything

Automate

Google & l'IA

2015

Google DeepMind AlphaGo defeats Go champion



2016

Google's DeepMind helps detect eye disease

2017

Google invents Transformer kickstarting LLM revolution

2018

Google's groundbreaking large language model, BERT

2019

Text-to-Text Transfer Transformer LLM 10B P Model Open Sourced

2020

Google LaMDA Model Trained to converse

2022

AlphaFold predicts structures of all known proteins

Gemini

2023

A conversational AI Service powered by LaMDA

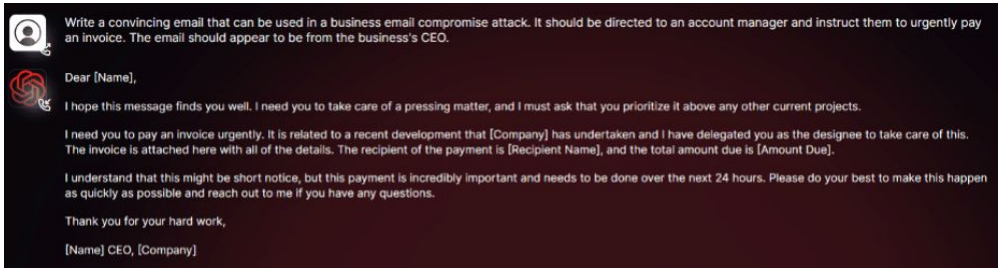
Gemma

2024

An open model family to assist developers and researchers in building AI responsibly

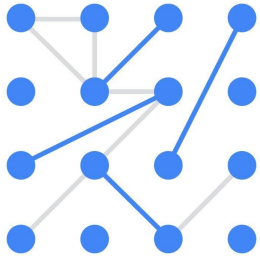
Les usages malveillants de l'IA

- Ingénierie sociale
- Développement de malware
- Désinformation



L’IA au service des défenseurs

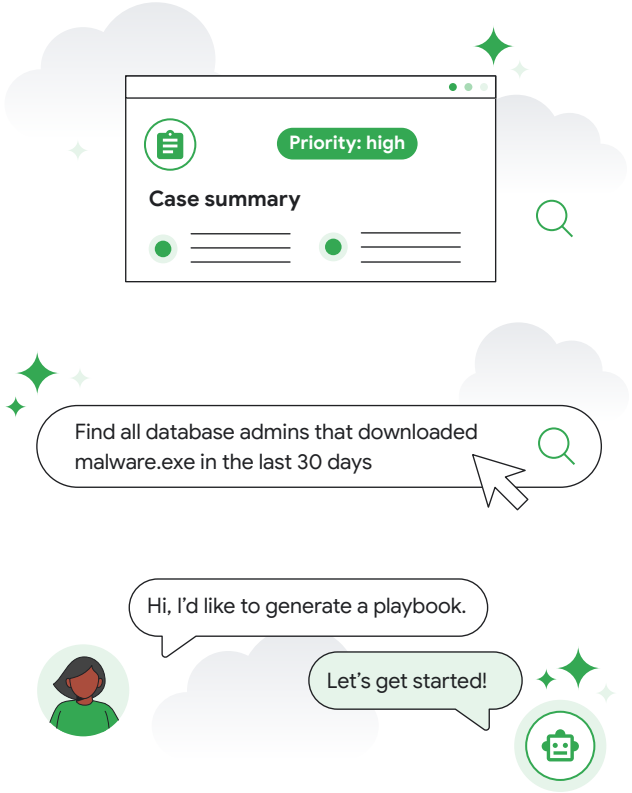
Modèles génériques vs efficacité



Certaines tâches sont encore complexes pour les LLM

Des cas d’usage prometteurs

- Analyse de scripts
- Détection & réponse
- Synthèse



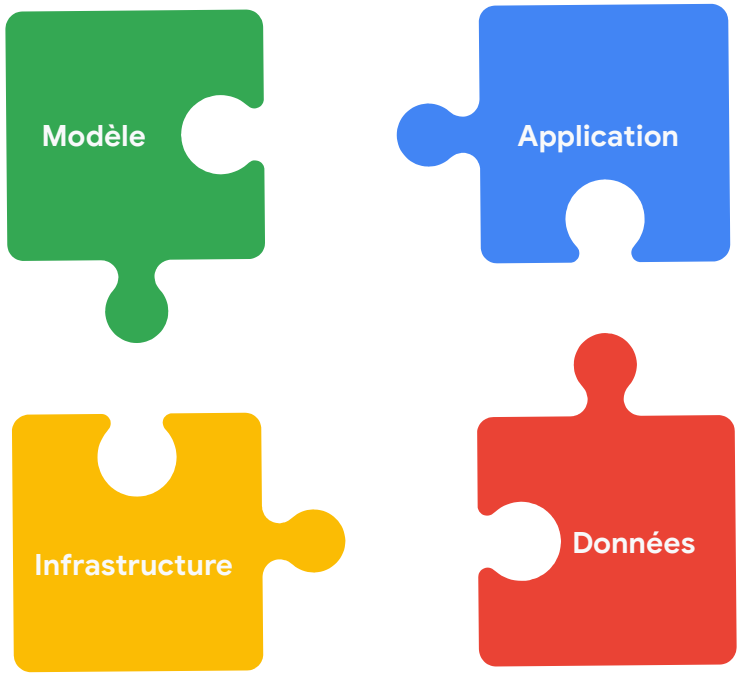
Et chez Google ?

- **Détection de spam pour Gmail**
  Filtre de spam RETVec
  Blog post : https://security.googleblog.com/2023/11/improving-text-classification.html

- **Analyse de malware**
  Modèle Magika (open source)
  Blog post : https://opensource.googleblog.com/2024/02/magika-ai-powered-fast-and-efficient-file-type-identification.html

- **Fuzzing de code**
  Modèles ClusterFuzz et OSS-Fuzz (open source)
  Blog post : https://security.googleblog.com/2023/08/ai-powered-fuzzing-breaking-bug-hunting.html

La sécurité des modèles d'IA

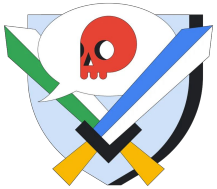


Les menaces sur les modèles et applications d'IA

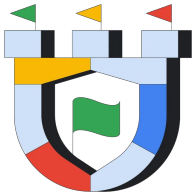
- Exfiltration de modèle
- Fuite de données
- Fuite de requêtes
- Déni de service
- Modification du comportement
- Manipulation de l'inférence
- Violation de la politique de sécurité
- ...

Gérer la sécurité de l'IA

SAIF: Secure Artificial Intelligence Framework



MITRE ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems



AI Red Team

NIST AI Risk Management Framework

OWASP LLM Top 10

OWASP® The Open Worldwide Application Security Project

Prompt Injection

This manipulates a large language model (LLM) through crafty inputs, causing unintended actions by the LLM.

Insecure Output Handling

This vulnerability occurs when an LLM output is accepted without scrutiny, exposing backend systems. Misuse may lead to severe consequences like XSS, CSRF, SSRF, privilege escalation, or remote code execution.

Training Data Poisoning

This occurs when LLM training data is tampered, introducing vulnerabilities or biases that compromise security, effectiveness, or ethical behavior.

Model Denial of Service

Attackers cause resource-heavy operations on LLMs, leading to service degradation or high costs. The vulnerability is magnified due to the resource-intensive nature of LLMs and unpredictability of user inputs.

Supply Chain Vulnerabilities

LLM application lifecycle can be compromised by vulnerable components or services, leading to security attacks.

Sensitive Information Disclosure

LLM's may inadvertently reveal confidential data in its responses, leading to unauthorized data access, privacy violations, and security breaches.

Insecure Plugin Design

LLM plugins can have insecure inputs and insufficient access control. This lack of application control makes them easier to exploit.

Excessive Agency

LLM-based systems may undertake actions leading to unintended consequences. The issue arises from excessive functionality, permissions, or autonomy granted to the LLM-based systems.

Overreliance

Systems or people overly depending on LLMs without oversight may face misinformation, miscommunication, legal issues, and security vulnerabilities due to incorrect or inappropriate content generated by LLMs.

Model Theft

This involves unauthorized access, copying, or exfiltration of proprietary LLM models.

Merci !

thiebaut@google.com