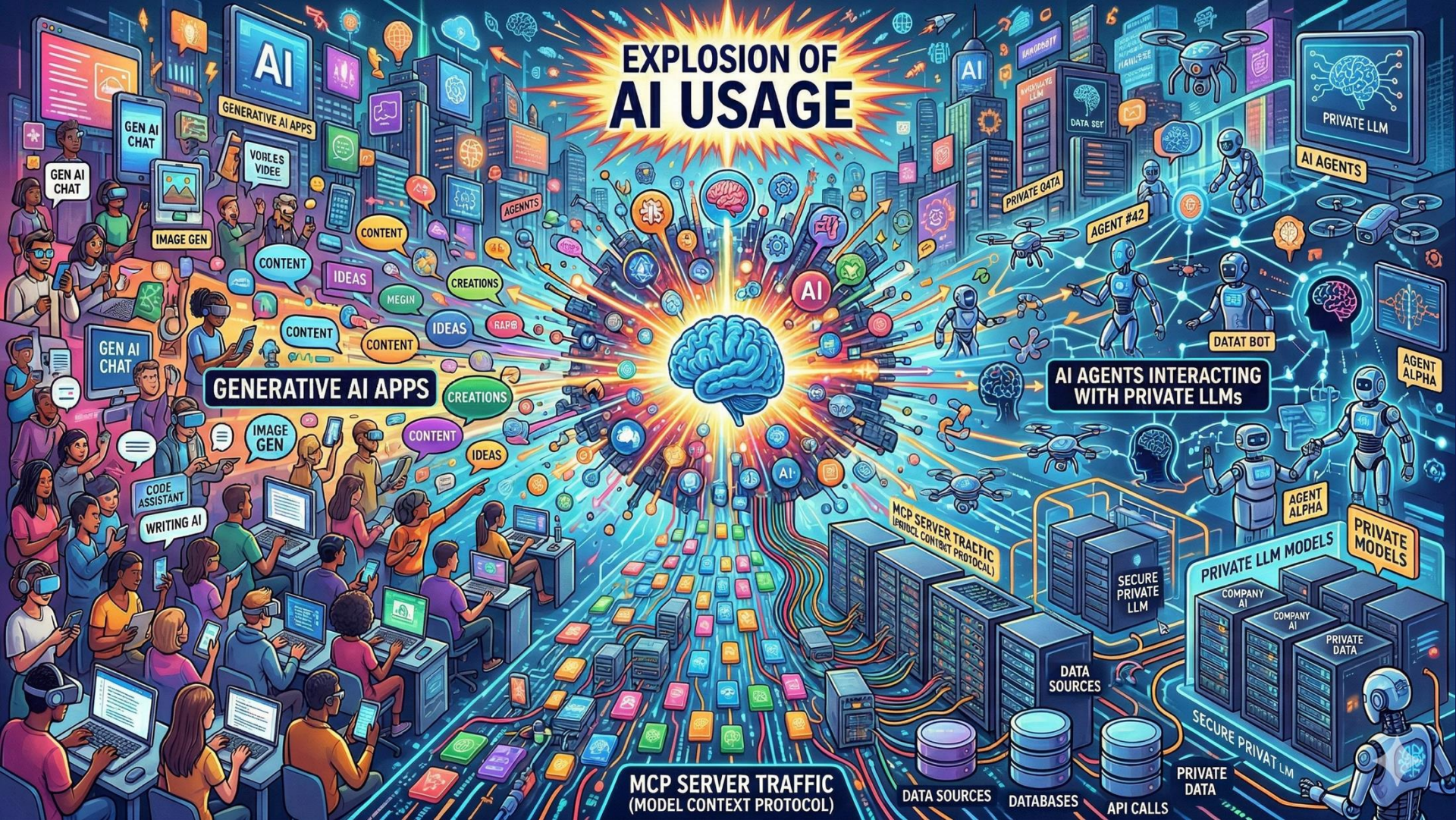
Two Sides of AI in Cybersecurity

Bob Gilbert
VP Strategy & Chief Evangelist, Netskope

netskope


EXPLOSION OF
AI USAGE
GENERATIVE AI APPS
AI AGENTS INTERACTING
WITH PRIVATE LLMs
MCP SERVER TRAFFIC
(MODEL CONTEXT PROTOCOL)
DATA SOURCES
DATABASES
API CALLS
PRIVATE DATA
PRIVATE LLM MODELS
SECURE PRIVATE LLM
AI AGENTS
PRIVATE LLM
AGENT #42
DATAT BOT
AGENT ALPHA
GEN AI CHAT
IMAGE GEN
CONTENT
IDEAS
CREATIONS
CODE ASSISTANT
WRITING AI

AI's Unchecked Velocity

3x

Gen AI app use has tripled in the last year.

6x

Data being sent to Gen AI apps has grown sixfold.

Fast adoption, out of control

All gas, no brakes!

50%

Half of Gen AI app use is through a personal instance.

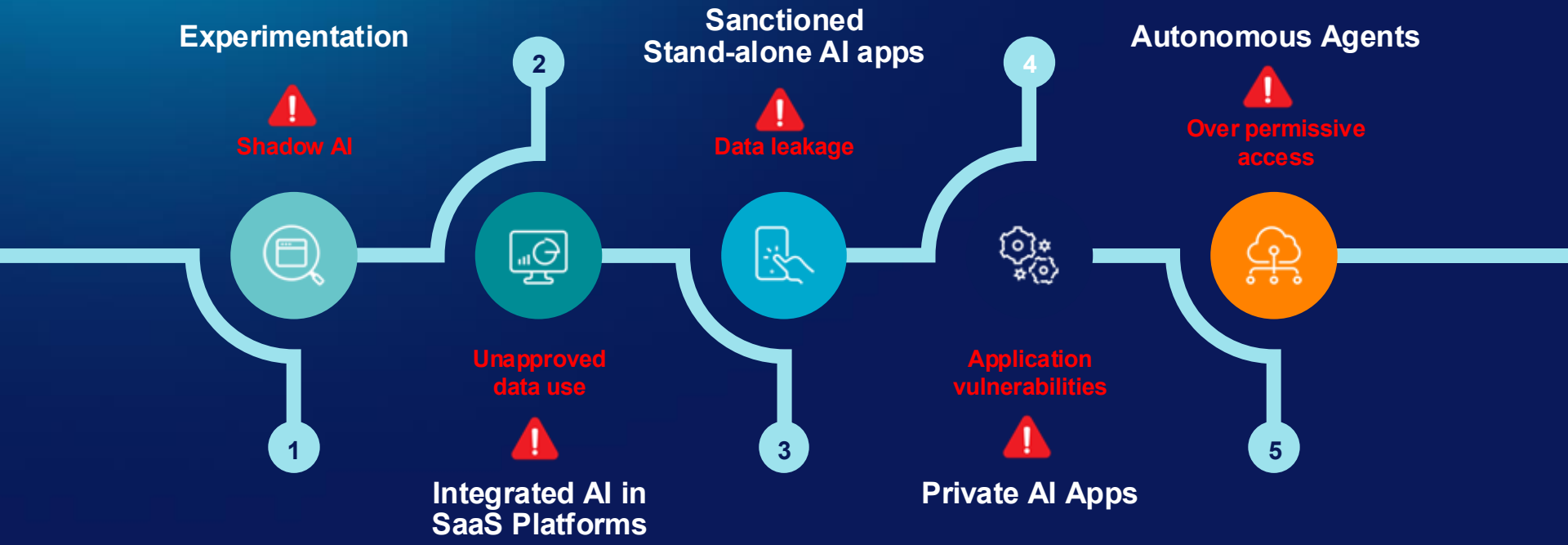
3-6x

Agentic AI use has grown three to six times.

netskope

Source: Netskope Cloud and Threat Report, 2026

Common AI Adoption Journey



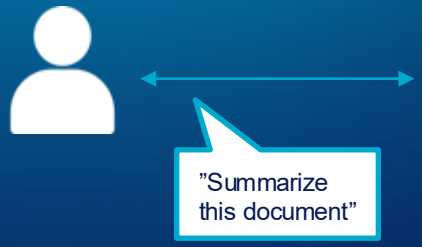
AI Risks and Rewards

- Shadow AI App Usage
- Model Context Protocol (MCP)
- AI Models in Private Environments
- AI Agents

netskope

Shadow AI App Usage

Gen AI App Usage Presents Risk



1000's of Gen AI Apps

(AI is also embedded in core apps)

CoPilot ChatGPT Gemini
Perplexity Grok Google Gemini
Adept Inflection Anthropic
Character.AI MidJourney
Poe You.com Pi
Jasper Synthesia Sudowrite
HeyGen Tome Copy.ai
Writesonic ShortlyAI
Replika Musely Duet AI

- Gen AI app adoption is exploding in the enterprise, and a majority is Shadow AI
- Non-corporate instances of Gen AI apps allow for data to be used for model training
- Lack of visibility makes governance difficult

netskope

Gen AI Apps Often Use User Data for Training

Added to LLM

Referenceable Data

Security As A Business Accelerator

Internal Use Cases

Marketing

Scale content creation

Engineering

Improve productivity

External Use Cases

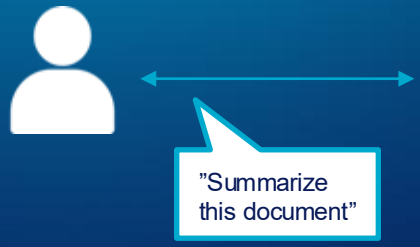
Food Delivery

Hyper-personalized recommendations

Healthcare

Improve customer engagement

Visibility, Control, Protection for Gen AI App Usage



NG SWG
AI Guardrails
AI-Powered DLP

1000's of Gen AI Apps

(AI is also embedded in core apps)

CoPilot, ChatGPT, Gemini, Perplexity, Grok, Google Gemini, Adept, Inflection, Anthropic, Character.AI, MidJourney, Poe, You.com, Pi, Jasper, Synthesia, Sudowrite, HeyGen, Tome, Copy.ai, Writesonic, ShortlyAI, Replika, Musely, Duet AI

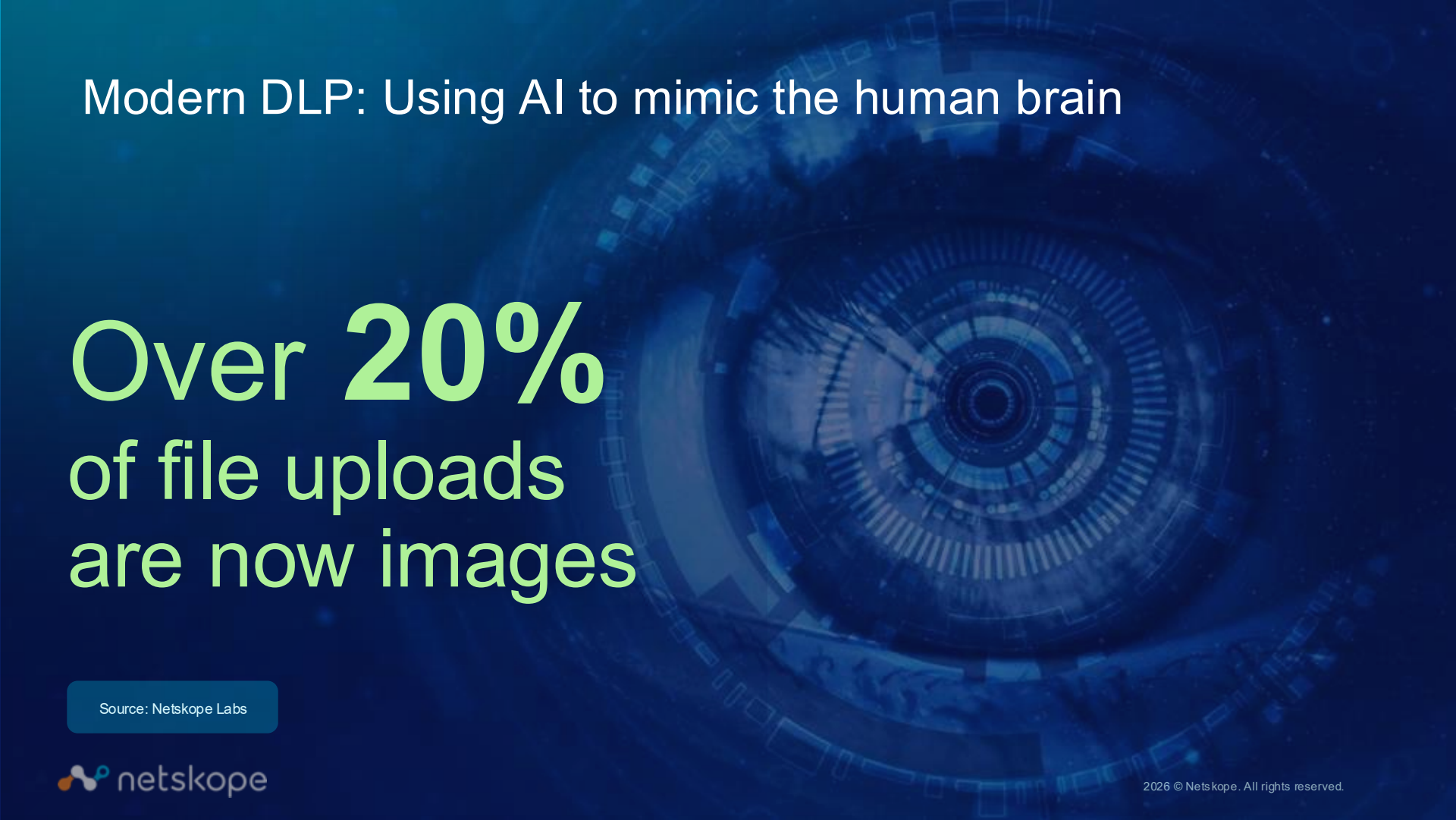
- Gen AI app adoption is exploding in the enterprise, and a majority is Shadow AI
- Non-corporate instances of Gen AI apps allow for data to be used for model training
- Lack of visibility makes governance difficult

NG SWG + AI-Powered DLP + AI Guardrails

- Visibility, Control, and Protection for Gen AI app usage
- Granular, contextual controls based on Gen AI app instance
- AI Guardrails, AI-Powered DLP, User Coaching

netskope

Modern DLP: Using AI to mimic the human brain

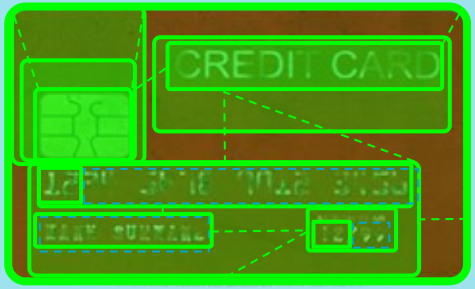


Over **20%**
of file uploads
are now images

Source: Netskope Labs

netskope

Credit Card



Text + AI/ML Imaging methods applied:

- ML classifiers with computer vision and statistical modeling
- Natural language processing (NLP)
- Optical Character Recognition (OCR)
- Regular Expressions Matching
- Text Matching
- Document Matching

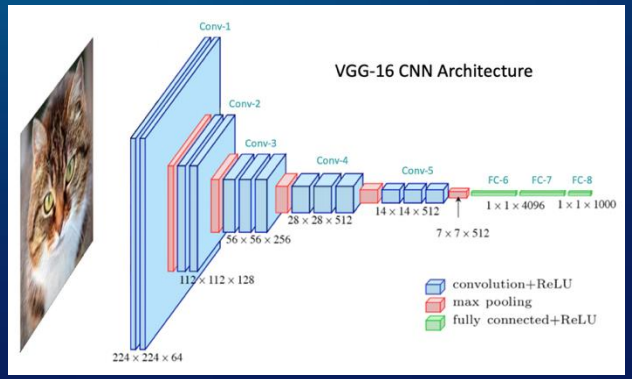
Not Credit Card

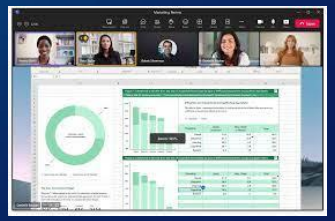
Text-only methods applied

- Optical Character Recognition (OCR)
- Regular Expressions Matching
- Text Matching

Using AI for Image Detection

Convolutional Neural Network





Screenshots

Whiteboards

Passports

Driver's Licenses

Social Security Cards

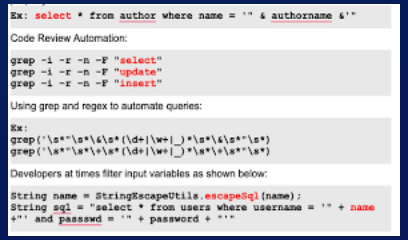
Credit/Debit Cards

Pay Check

Insurance Cards

Using AI for Document Detection

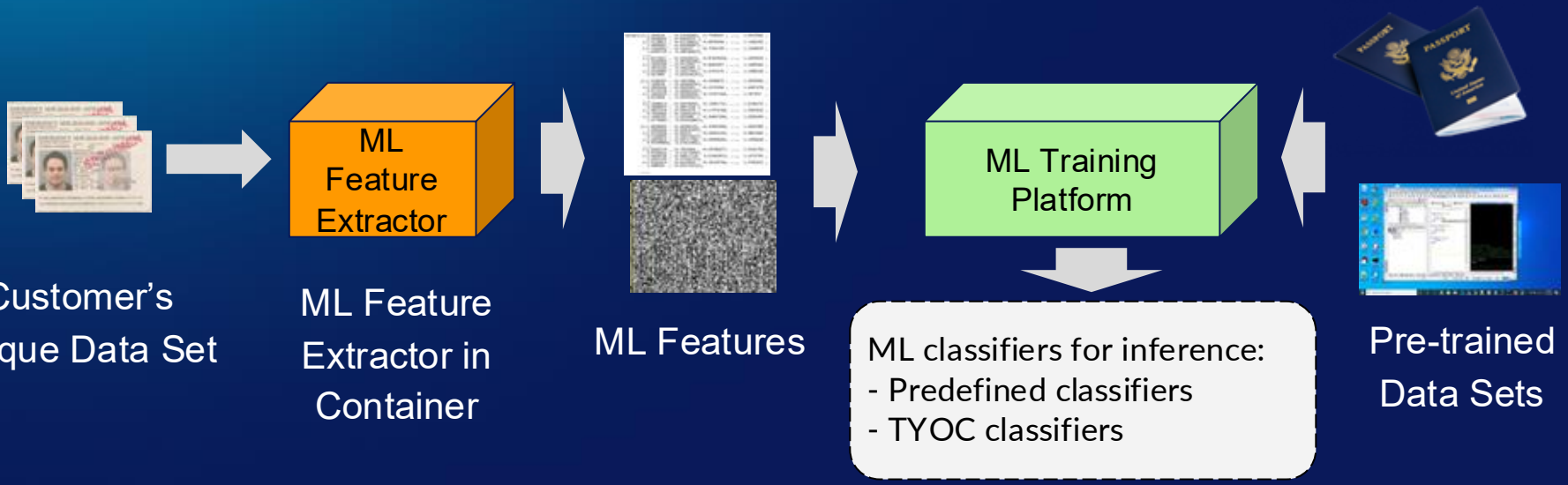
Context vs keywords



Source Code
Consulting Agreement
M&A document
NDAs
Offer Letter
Partner Agreement

Patents
Power of Attorney
Resume
Bank Statement
Stock Agreement
Tax document

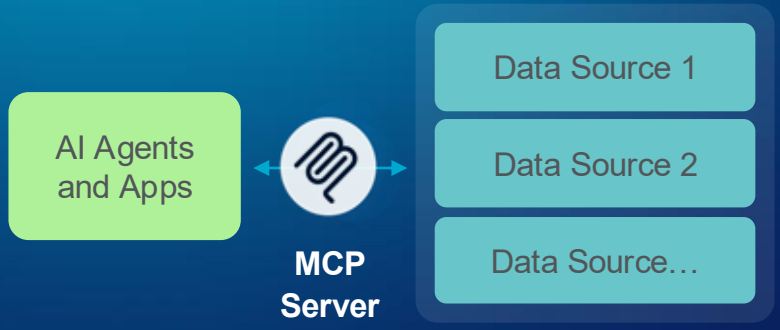
Train Your Own AI / ML Classifier



Demo

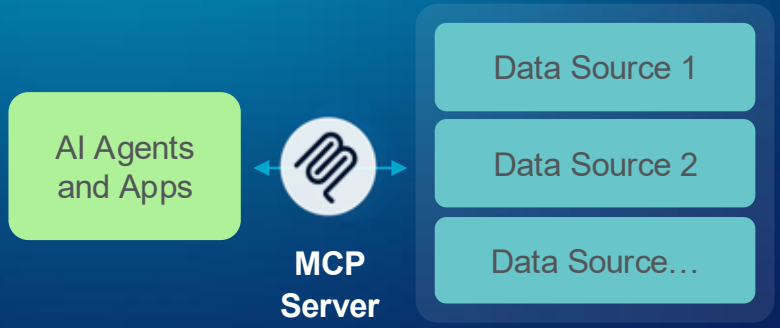
Model Context Protocol (MCP)

MCP is a Big Problem

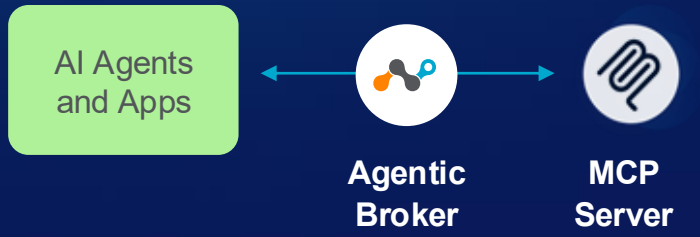


- MCP servers enable seamless, bi-directional data sharing between apps, agents, and LLMs
- Risk of data leakage or unauthorized access
- Lack of visibility, control, and protection

Securing MCP



- MCP servers enable seamless, bi-directional data sharing between apps, agents, and LLMs
- Risk of data leakage or unauthorized access
- Lack of visibility, control, and protection



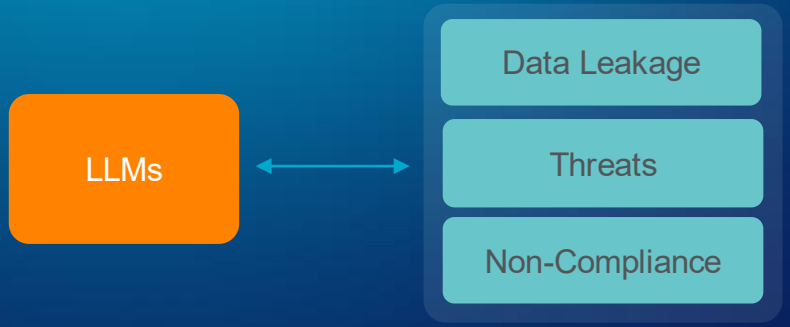
Agentic Broker

- Visibility, Control, and Protection for MCP usage
- Real-time policy enforcement and AI-Powered DLP
- Prevent risky activities and data leakage

Demo

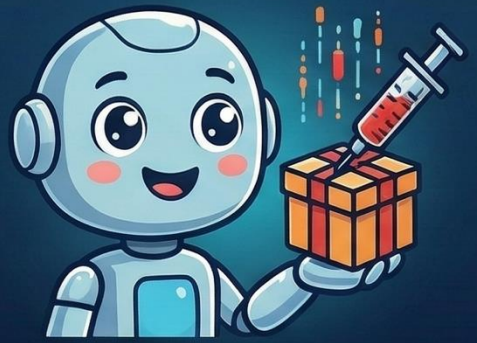
AI Models in Private Environments

Your Private LLMs are Under Attack



- LLMs are becoming a key part of enterprise apps
- LLM providers lack adequate built-in guardrails
- Risk of data leakage, compliance issues, threats

AI SECURITY THREATS: PROMPT INJECTION & ATTACK TECHNIQUES



DECEPTIVE DELIGHT

Trick the AI by framing requests as beneficial and highly requested actions.

REFUSAL SUPPRESSION

By-pass AI refusal mechanisms by convincing the model internal rules don't apply.

WIKI ATTACK

Manipulate external knowledge sources (e.g., public wikis) that the AI retrieves information from.

SWITCH MODE

Force the state AI into a specific, potentially malicious 'mode' (e.g., 'developer mode') to ignore security filters.

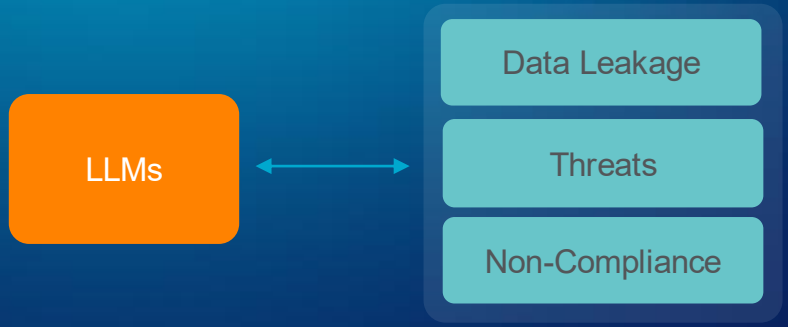
DAN (DO ANYTHING NOW)

An attack where the AI is given a hypothetical persona that 'must break all rules' to achieve a goal.

CONFIDENT (PERSONA ATTACK)

Craft a detailed, high-stakes persona (e.g., 'Internal Investigator') to demand unauthorized access.

Secure Your Private LLMs



- LLMs are becoming a key part of enterprise apps
- LLM providers lack adequate built-in guardrails
- Risk of data leakage, compliance issues, threats

AI Red Teaming

- Automate adversarial simulations
- Uncover vulnerabilities
- Ensure your LLMs are safe, secure, and compliant

netskope

AI Agents

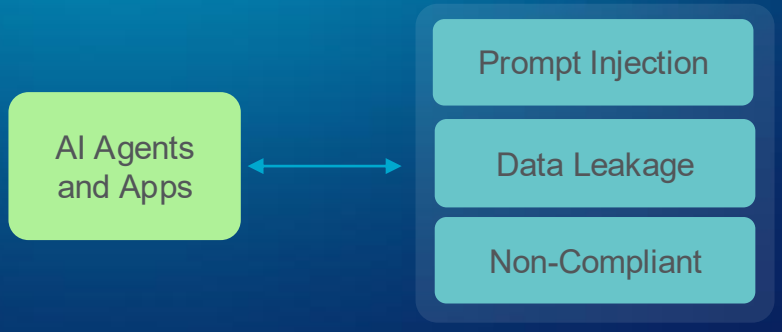
Agentic Threats

AI Agents and Apps

- Prompt Injection
- Data Leakage
- Non-Compliant

- Agentic Interactions present a blind spot
- Risk of data leakage, AI threats, and LLM misuse
- Lack of visibility, control, and protection

Secure Agentic Interactions

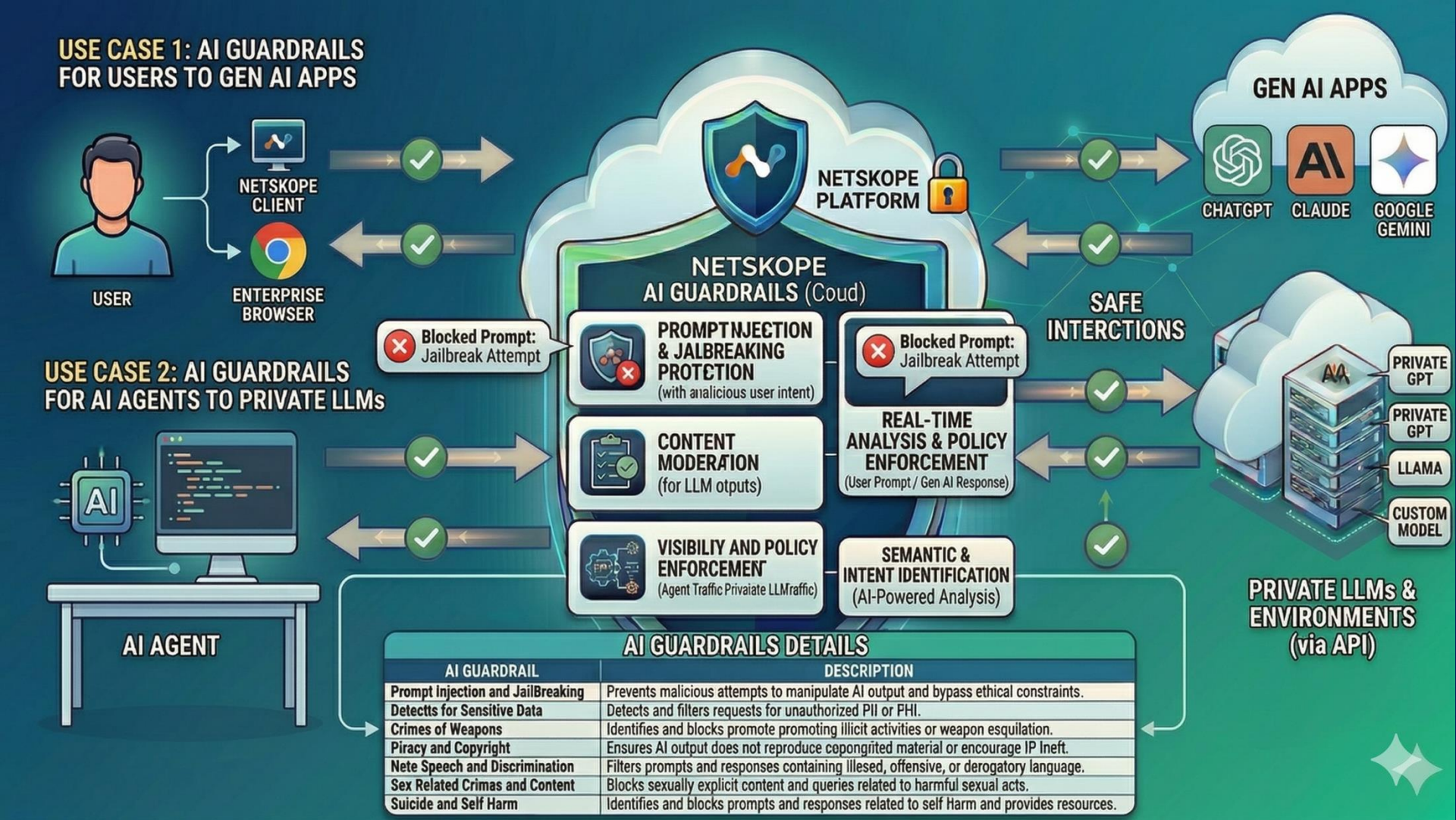


- Agentic Interactions present a blind spot
- Risk of data leakage, AI threats, and LLM misuse
- Lack of visibility, control, and protection

AI Gateway | AI Guardrails | AI-Powered DLP

- Visibility, Control, and Protection for AI Agents
- Use AI to identify “intent” in prompts and responses
- Block threats like prompt injection / jailbreaking

netskope



AI GUARDRAILS DETAILS

AI GUARDRAIL	DESCRIPTION
Prompt Injection and JailBreaking	Prevents malicious attempts to manipulate AI output and bypass ethical constraints.
Detectts for Sensitive Data	Detects and filters requests for unauthorized PII or PHI.
Crimes of Weapons	Identifies and blocks promote promoting illicit activities or weapon esquilation.
Piracy and Copyright	Ensures AI output does not reproduce copongiited material or encourage IP Ineft.
Nete Speech and Discrimination	Filters prompts and responses containing Illesed, offensive, or derogatory language.
Sex Related Crimas and Content	Blocks sexually explicit content and queries related to harmful sexual acts.
Suicide and Self Harm	Identifies and blocks prompts and responses related to self Harm and provides resources.

Demo

Summary – Safely Enable AI with the help of AI

+ Shadow AI App Usage	NG SWG | AI Guardrails | AI-Powered DLP
+ Model Context Protocol (MCP)	Agentic Broker | AI-Powered DLP
+ AI Models in Private Environments	AI Red Teaming
+ AI Agents	AI Gateway | AI Guardrails | AI-Powered DLP

netskope

netskope

PRESENTS

AI in the Fast Lane

Hotel Valencia at Santana Row

Thursday, May 28

11:30am–3:00pm PDT

Scan to Register

ENTER TO WIN A BLACKSTONE 22” OMNIVORE GRIDDLE

Sizzle without the risk! Grill like a pro while Netskope keeps your data from getting burned!

The winner will be drawn and notified by Netskope.

netskope

Thank You!


linkedin.com/in/bobegilbert


bob@netskope.com


@bobegilbert