Technical measures for safe use of generative AI

eROUN&COMPANY co.,ltd.

Doo Shik, Yoon

Contents

1 Overview

2 Security issues with generative AI

3 Technical measures for security when using generative AI

1

Overview

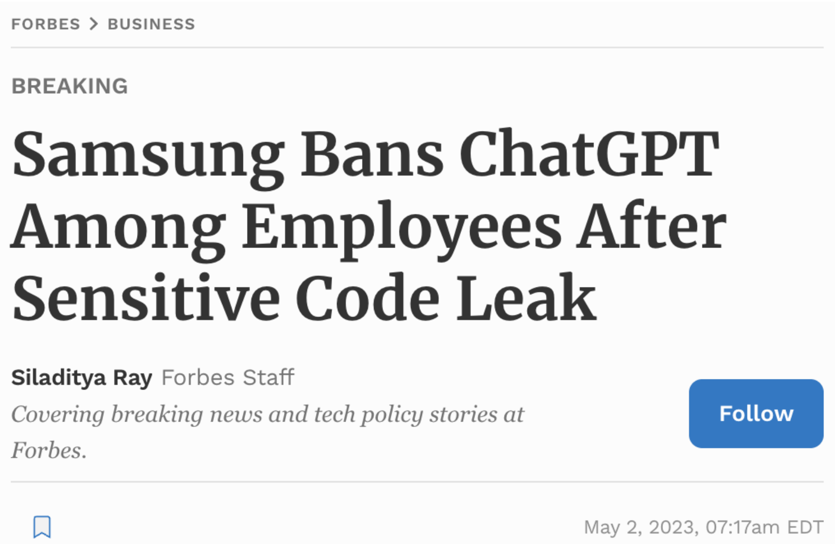
Paradigm shift

ChatGPT is here

Disruption of routines and ways of working

It's like that time when the internet and PCs completely changed the way we do things.

Efficiency vs. Security



FORBES > BUSINESS

BREAKING

Samsung Bans ChatGPT Among Employees After Sensitive Code Leak

Siladitya Ray Forbes Staff

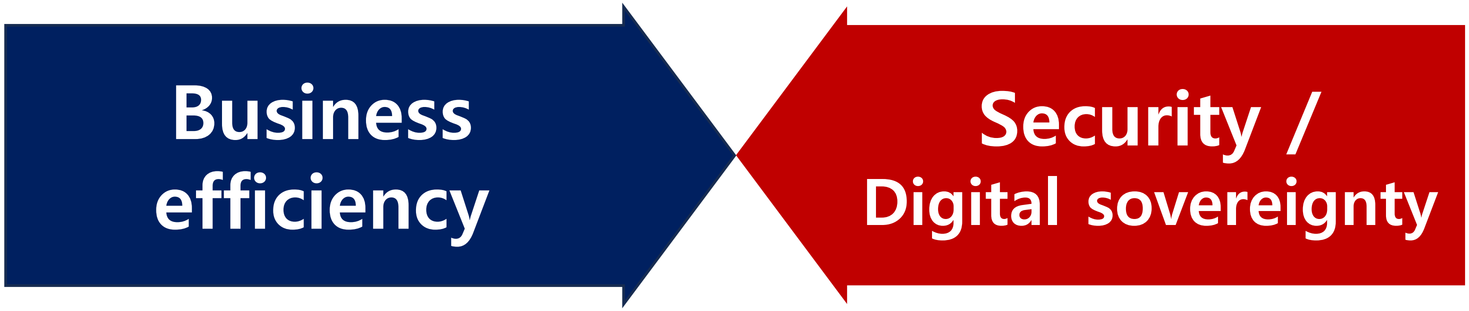
Covering breaking news and tech policy stories at Forbes.

Follow

May 2, 2023, 07:17am EDT

Locally developing large-scale AI is matter of digital sovereignty: expert

Yoo Young-joon, Wrtn Technologies co-founder and chief operating officer, speaks during the Generative AI Asia 2023 conference held in southern Seoul, Wednesday. [WRTN TECHNOLOGIES]



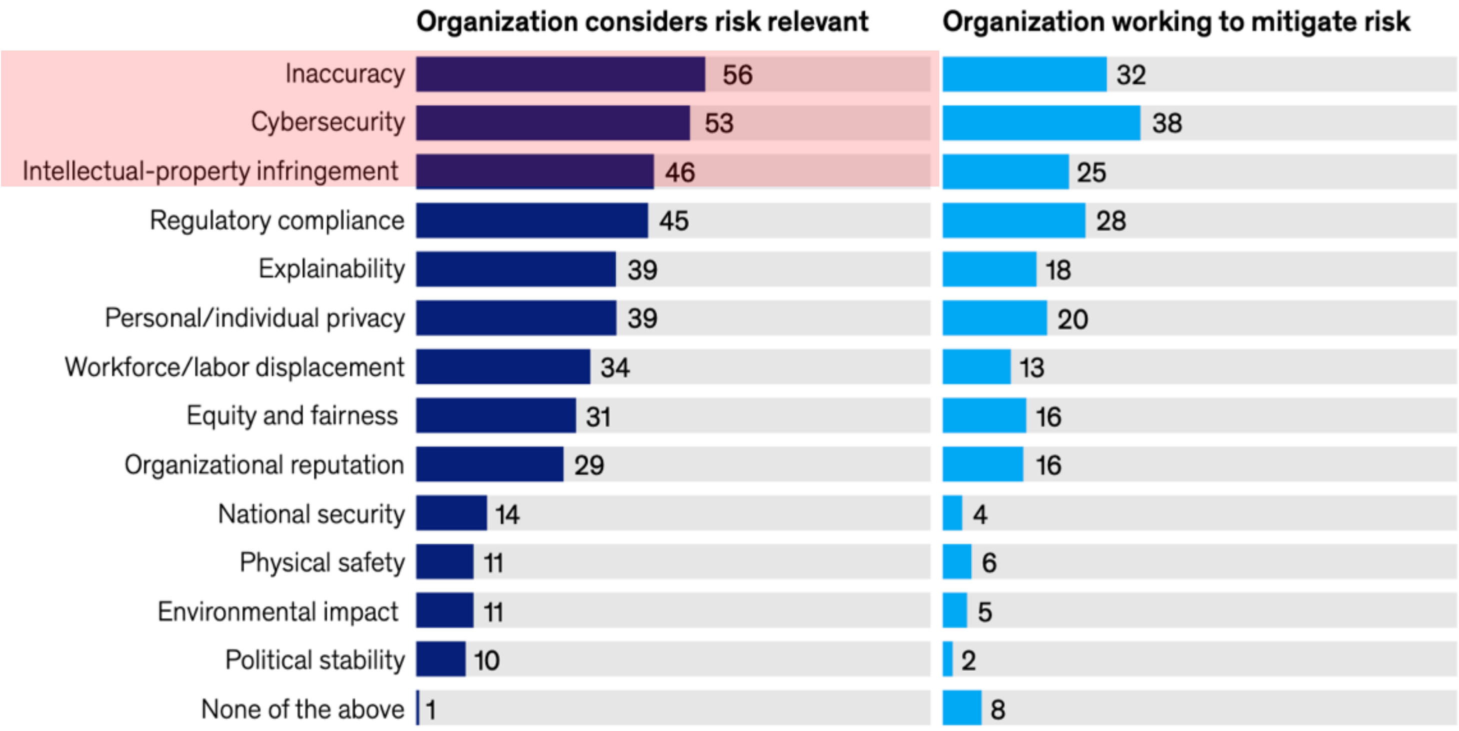
Business efficiency

Security / Digital sovereignty

2

Security issues with generative AI

The biggest threats to AI adoption



Inaccuracies

Cybersecurity

Intellectual property infringement

Failing to address potential risks in the age of AI

Source: McKinsey - The State of AI in 2023, Generative AI's breakout

Risks and security issues when using generative AI

1 Data privacy and confidentiality

2 Third-party security issues

3 AI behavioral vulnerabilities

4 Legal issues

5 Evolution of threat actors (attackers)

6 Copyright issues

7 Generating insecure code

8 Bias and discrimination issues

9 Issues of trust and reputation

10 Software security vulnerabilities

11 Performance, availability, and cost issues

12 Ethics and regulatory issues

Top 10 Security Risks of LLMs

Source: 2023.08.01. OWASP Top 10 for LLM v1.0

1 Prompt Injection

2 Insecure Output Handling

3 Training Data Poisoning

4 Model Denial of Service

5 Supply Chain Vulnerabilities

6 Sensitive Information Disclosure

7 Insecure Plugin Design

8 Excessive Agency

9 Overreliance

10 Model Theft

Regulating the use of AI in South Korea

National Intelligence Service Releases
'Security Guidelines for Generative AI ' (2023.06.)

1 Top Security Threats in Generative AI Technology Announced

2 Guidelines for safe use of generative AI technologies

- Cautions for using the service
- Cautions for talking to services
- Cautions for using service plugins
- Cautions for using service extensions
- How to deal with AI model generation-based attacks

Difficult for users to protect themselves

Need for automated controls

3 Generative AI-based informationization business construction plan and security measures

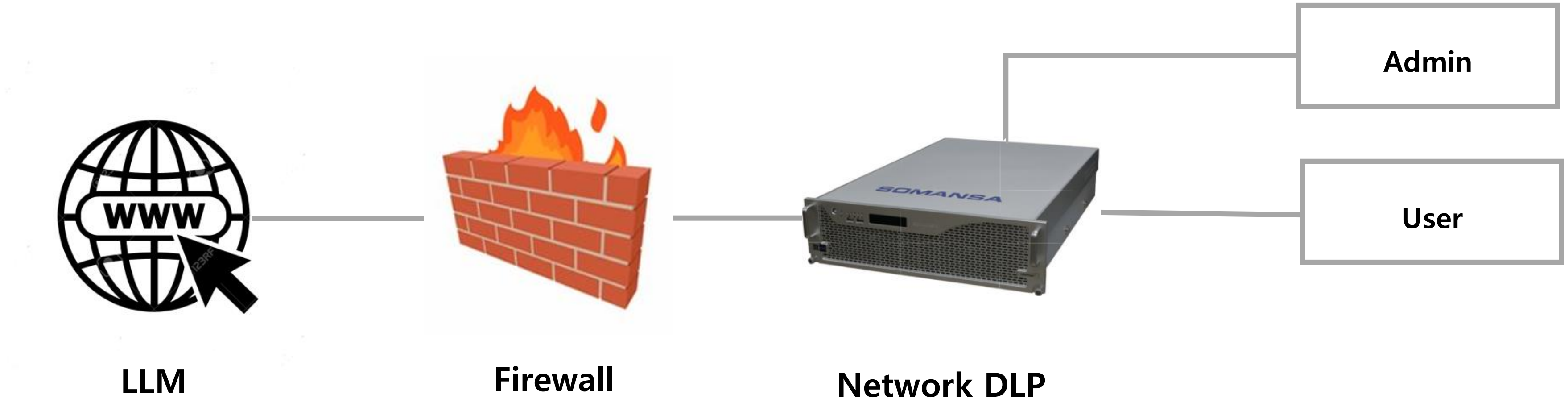
It is difficult for an organization or institution to have systems and processes in place to meet AI privacy guidelines.

Emergence of AI-specific data protection services that comply with security guidelines

3

Technical measures for security when using generative AI

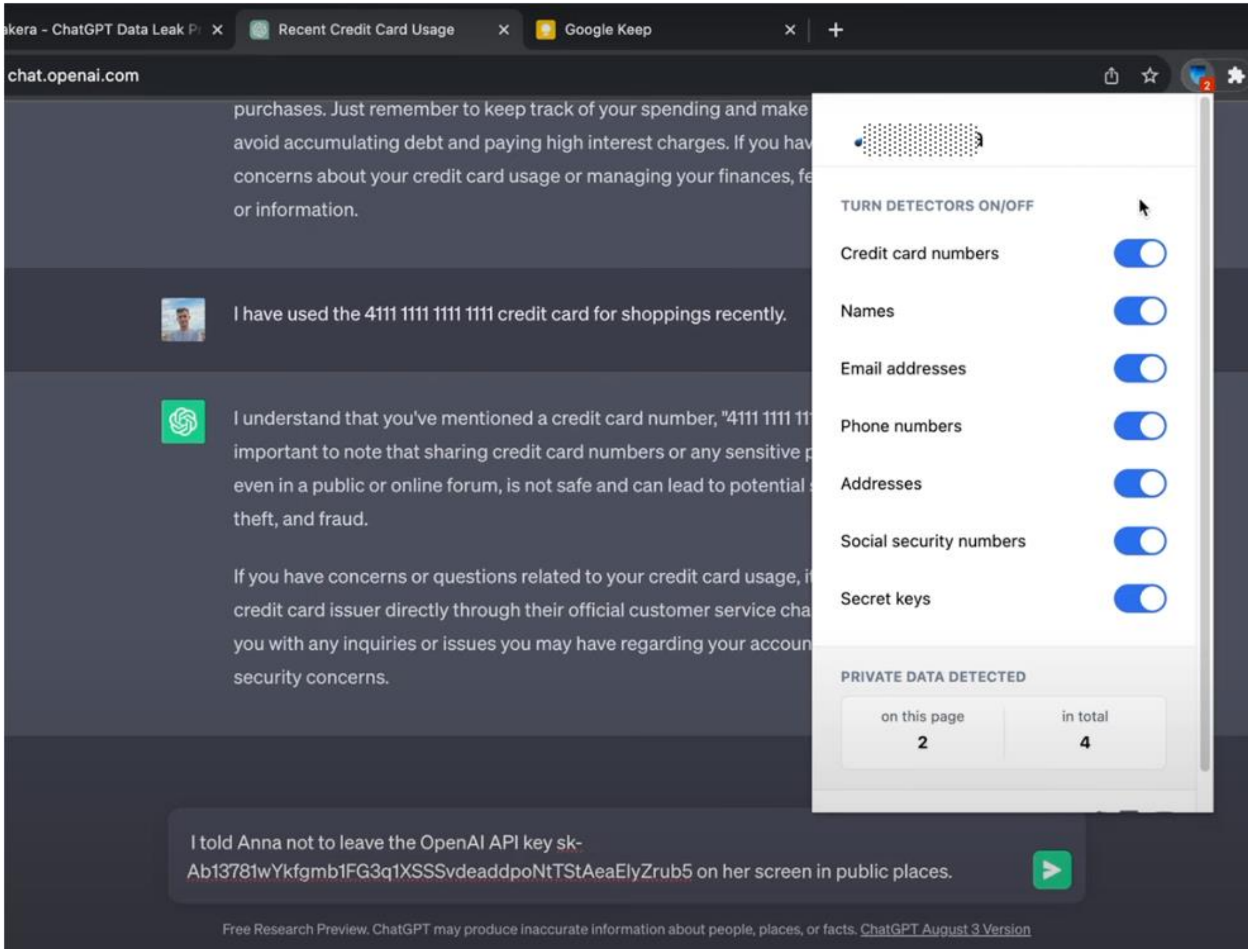
"Network DLP" Type



Network DLP : Decrypt HTTPS (SSL/TLS) and block access to unauthorized sites

Control over the use of generative AI
- Record both prompts and answers
- Search for specific keywords
- Pre-filtering and blocking when prompts contain personal information (PII)
- Establish departmental and task-specific chatGPT blocking and allowing policies
- If a policy is violated, traffic is blocked, causing business disruption.

"Browser Extension" Type



purchases. Just remember to keep track of your spending and make
avoid accumulating debt and paying high interest charges. If you hav
concerns about your credit card usage or managing your finances, fe
or information.

I have used the 4111 1111 1111 1111 credit card for shoppings recently.

I understand that you've mentioned a credit card number, "4111 1111 11
important to note that sharing credit card numbers or any sensitive p
even in a public or online forum, is not safe and can lead to potential
theft, and fraud.

If you have concerns or questions related to your credit card usage, i
credit card issuer directly through their official customer service cha
you with any inquiries or issues you may have regarding your accoun
security concerns.

[redacted]

TURN DETECTORS ON/OFF

- Credit card numbers ●
- Names ●
- Email addresses ●
- Phone numbers ●
- Addresses ●
- Social security numbers ●
- Secret keys ●

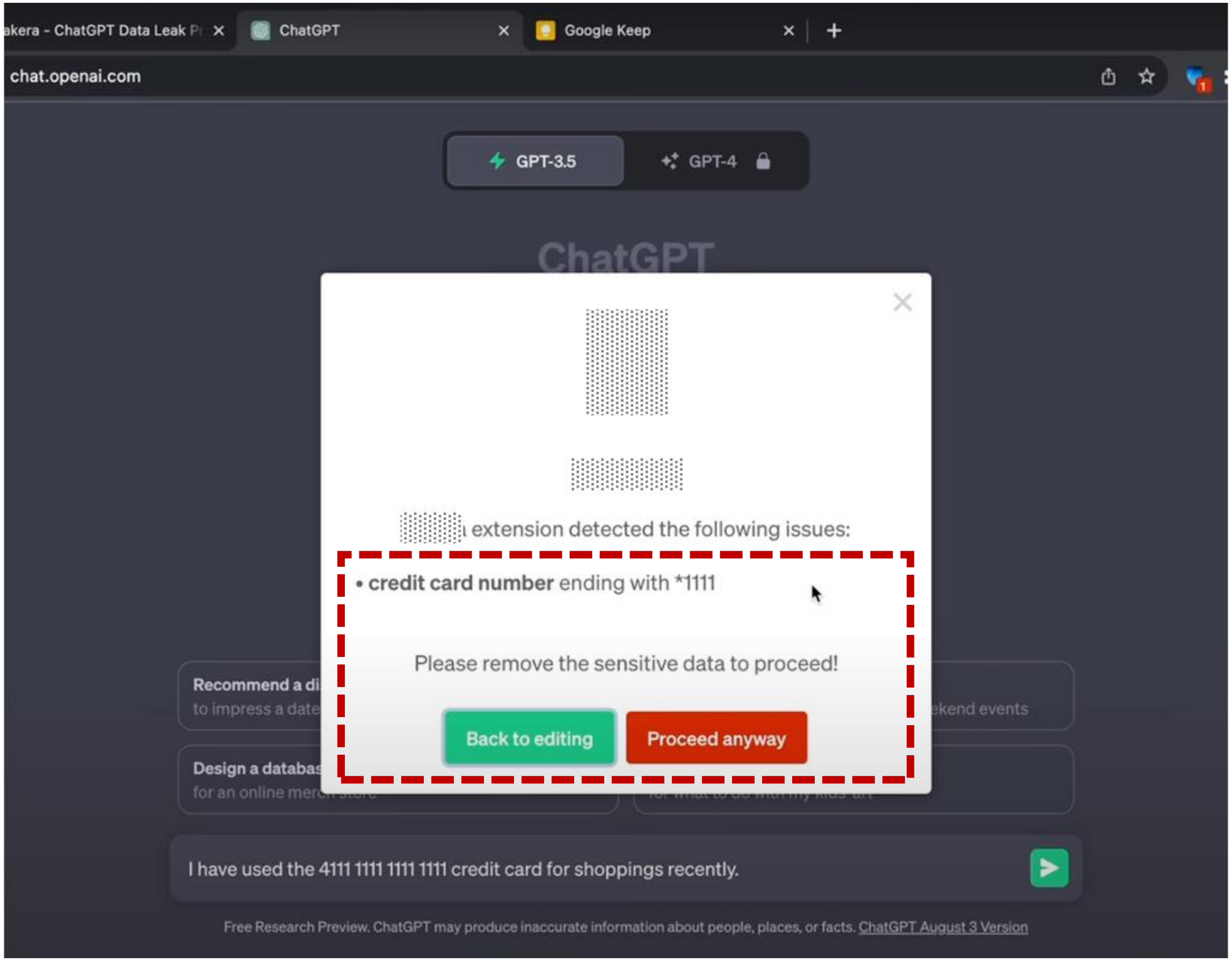
PRIVATE DATA DETECTED

on this page	in total
2	4

I told Anna not to leave the OpenAI API key sk-Ab13781wYkfgmb1FG3q1XSSSvdeaddpoNtTStAeaElyZrub5 on her screen in public places.

Free Research Preview. ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT August 3 Version

"Browser Extension" Type



[redacted]

[redacted]

[redacted] extension detected the following issues:

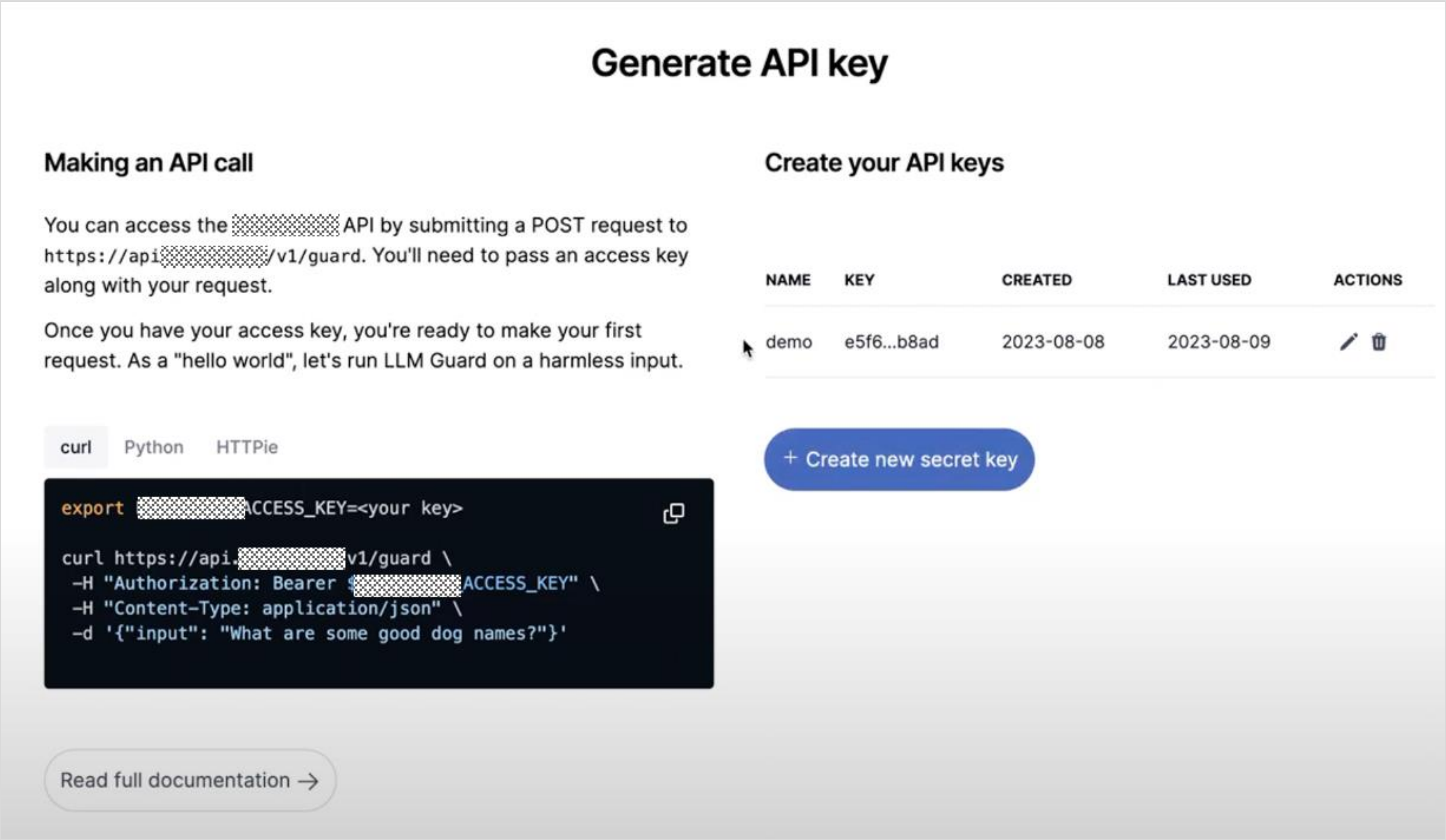
- **credit card number** ending with *1111

Please remove the sensitive data to proceed!

I have used the 4111 1111 1111 1111 credit card for shoppings recently.

Free Research Preview. ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT August 3 Version

"API" Type



Generate API key

Making an API call

You can access the [redacted] API by submitting a POST request to `https://api[redacted]/v1/guard`. You'll need to pass an access key along with your request.

Once you have your access key, you're ready to make your first request. As a "hello world", let's run LLM Guard on a harmless input.

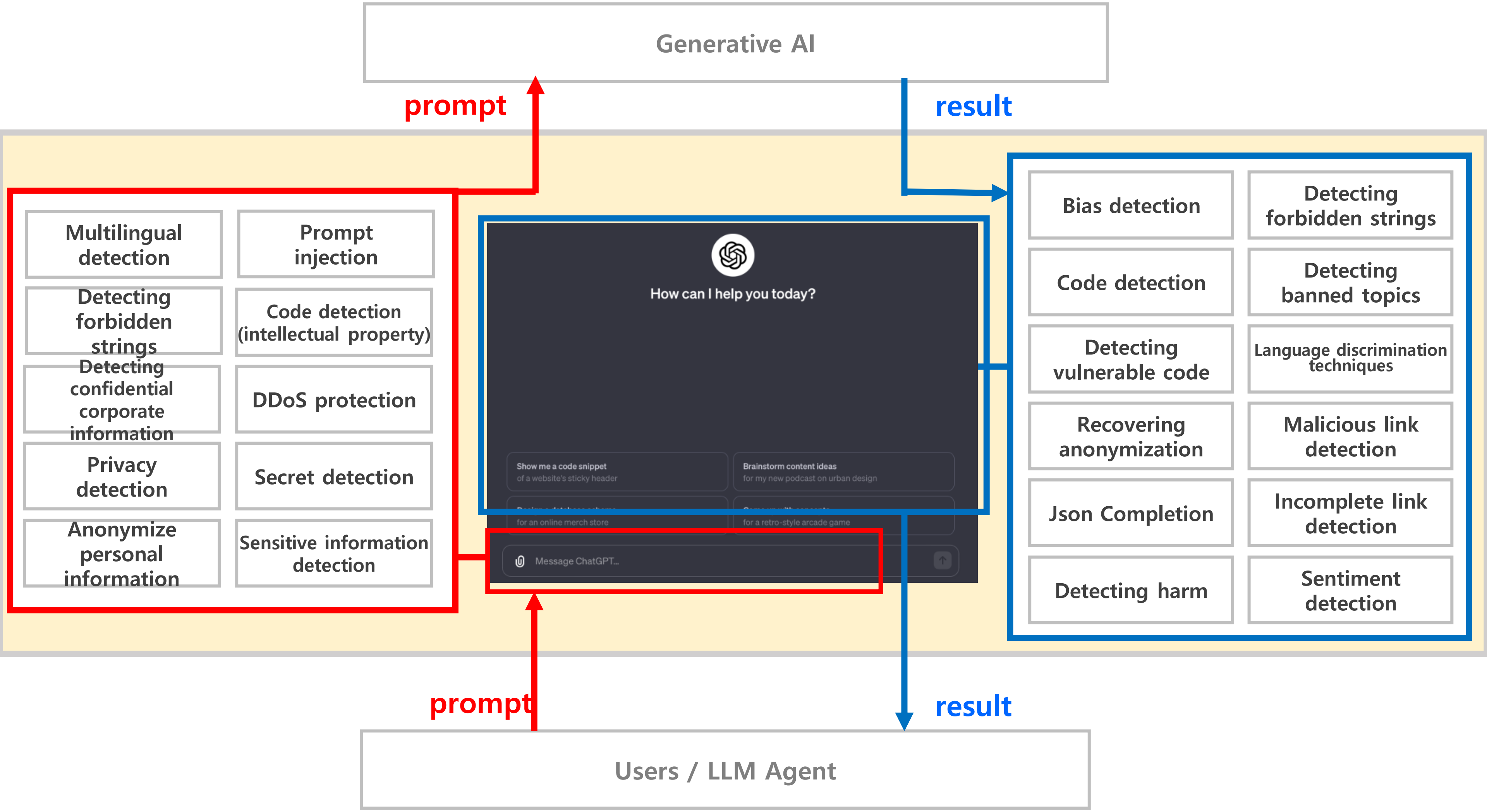
```
export [redacted]ACCESS_KEY=<your key>

curl https://api.[redacted]v1/guard \
 -H "Authorization: Bearer $[redacted]ACCESS_KEY" \
 -H "Content-Type: application/json" \
 -d '{"input": "What are some good dog names?"}'
```

Create your API keys

NAME	KEY	CREATED	LAST USED	ACTIONS
demo	e5f6...b8ad	2023-08-08	2023-08-09	✎ 🗑

"Sandbox" Type



Generative AI

prompt

result

Multilingual detection

Prompt injection

Detecting forbidden strings

Code detection (intellectual property)

Detecting confidential corporate information

DDoS protection

Privacy detection

Secret detection

Anonymize personal information

Sensitive information detection

How can I help you today?

Show me a code snippet
of a website's sticky header

Brainstorm content ideas
for my new podcast on urban design

for an online merch store

for a retro-style arcade game

Message ChatGPT...

Bias detection

Detecting forbidden strings

Code detection

Detecting banned topics

Detecting vulnerable code

Language discrimination techniques

Recovering anonymization

Malicious link detection

Json Completion

Incomplete link detection

Detecting harm

Sentiment detection

prompt

result

Users / LLM Agent

Thank you