GenAI Consumer Services Tracking

Keep a list of Approved or Non-Approved GenAI services

Domain | App Category | etc

If you have allowed or disallowed groups or users, remove or add them to the search.

```
index IN (netskope umbrella palo)
domain IN (claude.ai
    api.openai.com openai.com
    deepai.org  gemini.google
    perplexity.ai
copilot.cloud.microsoft
    askyourpdf.com)
User IN ($LIST)
```

Look for Large Uploads

Look for larger uploads to these services that may indicate file uploads of files.

If your data includes `file_name` use this to infer file continents.

```
index=netskope app="ChatGPT"
OR app="Anthropic Claude" OR
"other_categories{}"="Generative AI"
domain IN (claude.ai api.openai.com
openai.com deepai.org  gemini.google
Perplexity.ai copilot.cloud.microsoft
askyourpdf.com)
| bin _time span=10m
| stats values(file_name)
sum(bytes_in) by user _time
| sort - sum(bytes_in)
```

Look for Anomalies or Automation

Detect daily/continuous GenAI service access from employee devices

Identify potential automated processes or unauthorized integrations

Review discovered workflows for business approval and compliance

```
index=netskope app="ChatGPT" OR
app="Anthropic Claude" OR
"other_categories{}"="Generative AI"
    domain IN (claude.ai
    api.openai.com  openai.com
    deepai.org  gemini.google
    Perplexity.ai
    copilot.cloud.microsoft
    askyourpdf.com)
| timechart count by user
```

DLP Strategies

Data Protection Controls

- Data masking & tokenization of sensitive fields
- DLP pattern matching & content filtering
- Automated PII/PHI detection and blocking
- File type & size restrictions

Access & Authentication

- SSO integration with role-based access
- API key rotation & management
- IP restrictions & geo-fencing
- Approval for processes containing sensitive data

Monitoring & Response

- Real-time traffic analysis & logging
- Prompt/response content scanning
- Automated alerts for policy violations
- Comprehensive audit trails
- Incident response procedures

GenAI/LLM Service Risks

Data Leakage

- Sensitive information shared in prompts
- Corporate data retained in model training
- Unintended exposure through model responses

Knowledge Base Compromise

- RAG system exposure of restricted data
- Cross-tenant data leakage
- Unauthorized access to proprietary information

Access Control Bypass

- Prompt injection to circumvent restrictions
- Authentication/authorization weaknesses
- Session/context manipulation

Supply Chain/Vendor Risk

- Third-party model vulnerabilities
- Dependency on external AI providers
- Limited visibility into backend security

Model Security

models-scan

Deserialization Threats

Untrusted data/code reconstructs objects, leading to exploitation, resulting in **arbitrary code execution**.

Malicious actors inject harmful code during deserialization.

Crucial for securing data integrity and preventing unauthorized code execution.

Jinja Chat Template, in the GGUF format, can include arbitrary code execution.

Architectural Backdoors

Results in predetermined **malicious output**.

Parallel path from model input to output, manipulated by attackers.

Model behaves normally with non-malicious inputs, unexpected behavior with triggers.

Attacker controls when backdoored model gives incorrect output.

Runtime Threats

Triggered at model load and can lead in exploitation to result in **arbitrary code execution**.

Untrusted data/code exploits vulnerabilities during inference or execution.

Malicious actors inject harmful code, gaining unauthorized access or manipulating systems.

Crucial for securing data integrity and preventing unauthorized code execution.

Hidden Layer

Core Offering

AI/ML model security platform

Protection against inference, bypass, and extraction attacks

No-code implementation without model complexity

Risk Assessment Services

Comprehensive ML Operations lifecycle analysis

Critical AI model risk evaluation

Mapping to industry standards (NIST, MITRE ATLAS, OWASP)

Actionable risk reduction recommendations

Back to Basics

App Firewall

Detecting and blocking attacks against the service such as injection, RCE, info-disclosure, etc.

DLP

Prevent unauthorized sensitive data from being shared

Access & Control

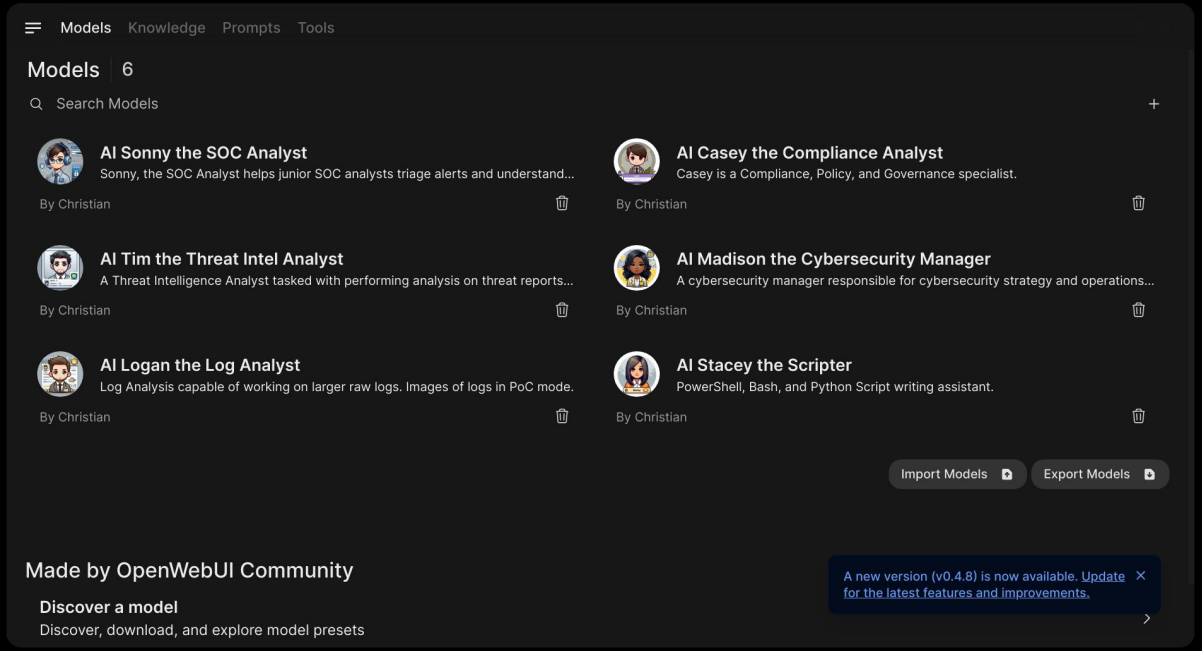
RBAC with MFA preventing unauthorized access to sensitive information or to send sensitive information to model for LLM inference.

Encryption

Ensure data at rest or in transfer is encrypted.

Centralize Access Authorization

If connecting to sensitive information, contain it within a solution that offers visibility to a governance team.



Centralize Access Authorization

BYOD Licensed Solutions