Challenges and Threats During Azure OpenAI Deployments

Security Risks Associated with OpenAI Deployments

- OpenAI models can be vulnerable to attacks
- Data privacy concerns must be taken seriously
- Developers must implement security measures to protect their models

General Challenges adopting LLM

Challenges

1. Hallucinations/Fabricates facts
2. Opaque source
3. Biased
4. Static
5. Expensive/Wasteful

Mitigation principles

1. Human-in-loop / non-factual tasks
2. Retrieval / knowledge augmentation
3. Content moderation
4. Model customization
5. Consider alternatives

OWASP TOP 10 for LLM Applications

LLM01

Prompt Injection

This manipulates a large language model (LLM) through crafty inputs, causing unintended actions by the LLM. Direct injections overwrite system prompts, while indirect ones manipulate inputs from external sources.

LLM02

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.

LLM03

Training Data Poisoning

This occurs when LLM training data is tampered, introducing vulnerabilities or biases that compromise security, effectiveness, or ethical behavior. Sources include Common Crawl, WebText, OpenWebText, & books.

LLM04

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.

LLM05

Supply Chain Vulnerabilities

LLM application lifecycle can be compromised by vulnerable components or services, leading to security attacks. Using third-party datasets, pre- trained models, and plugins can add vulnerabilities.

LLM06

Sensitive Information Disclosure

LLMs may inadvertently reveal confidential data in its responses, leading to unauthorized data access, privacy violations, and security breaches. It's crucial to implement data sanitization and strict user policies to mitigate this.

LLM07

Insecure Plugin Design

LLM plugins can have insecure inputs and insufficient access control. This lack of application control makes them easier to exploit and can result in consequences like remote code execution.

LLM08

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.

LLM09

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.

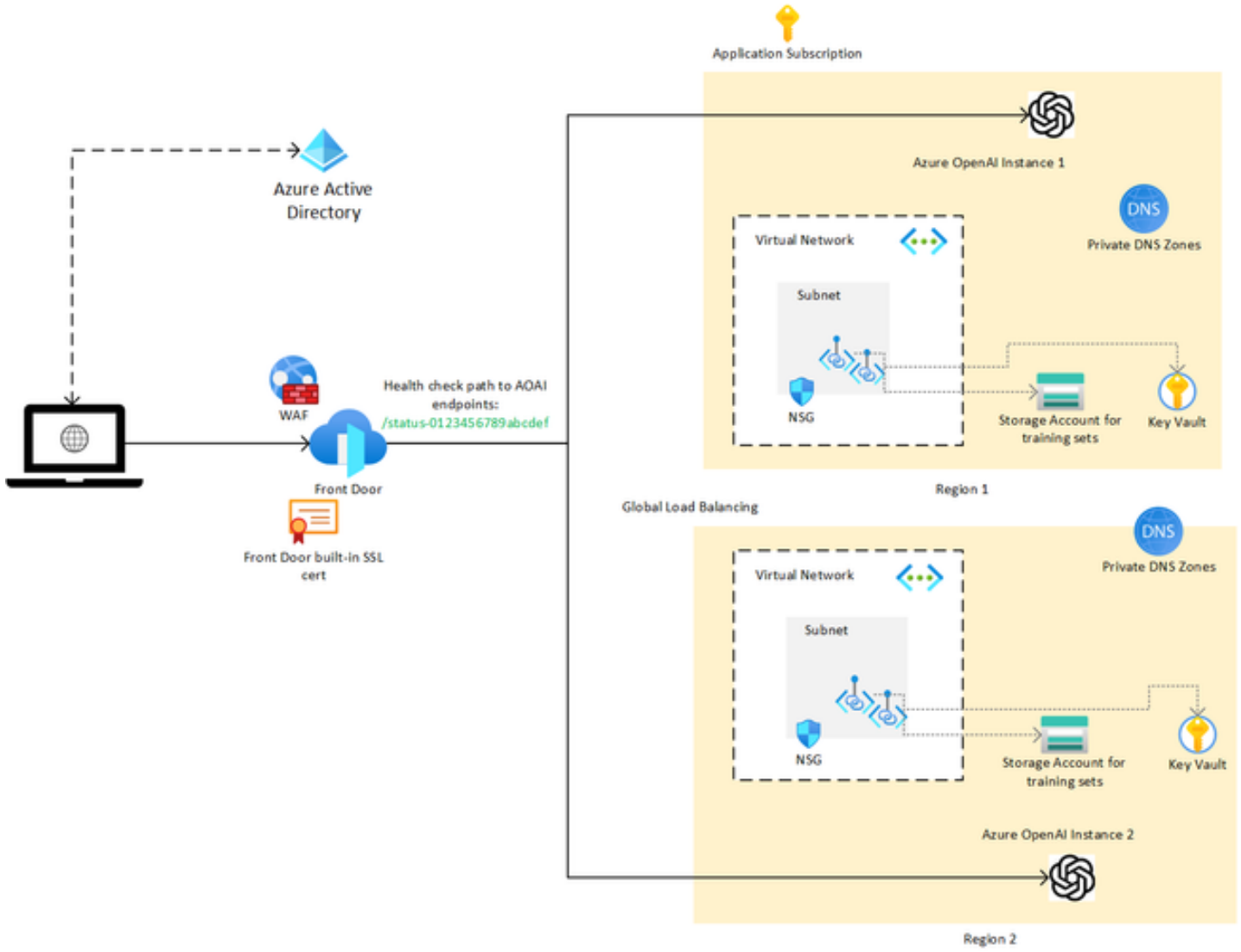
LLM10

Model Theft

This involves unauthorized access, copying, or exfiltration of proprietary LLM models. The impact includes economic losses, compromised competitive advantage, and potential access to sensitive information.

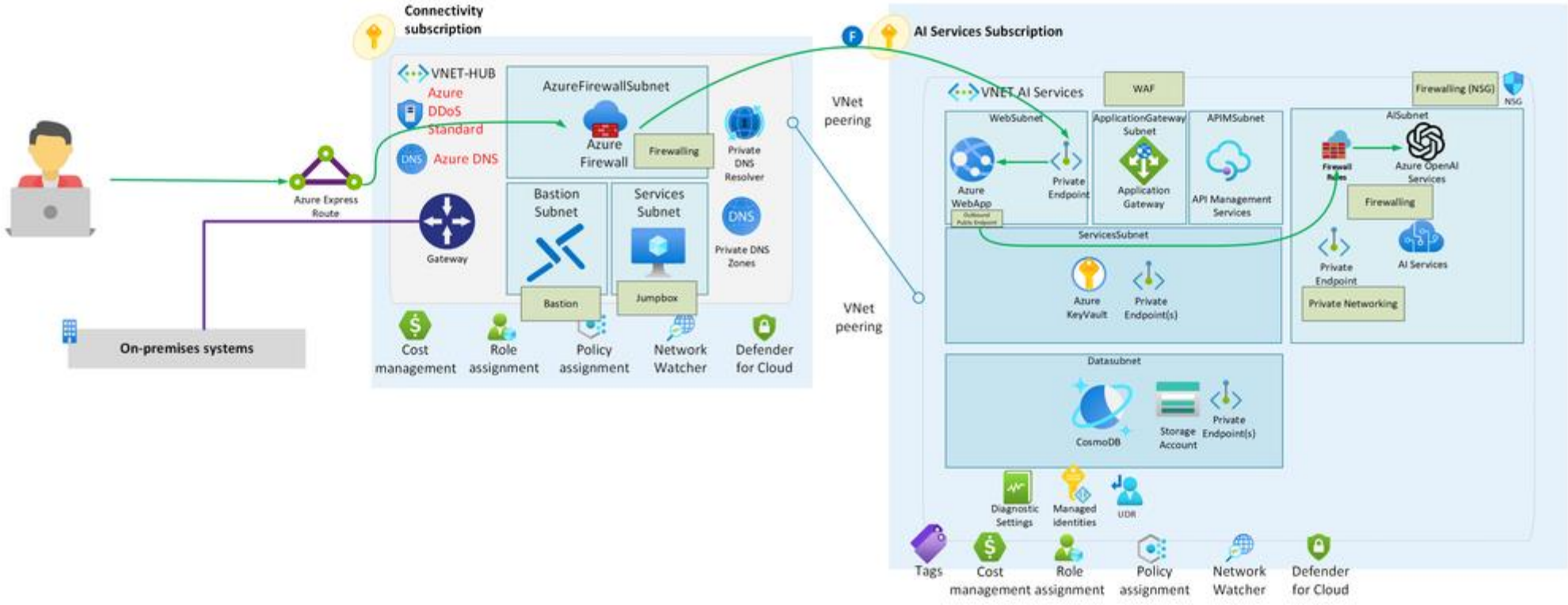
https://owasp.org/www-project-top-10-for-large-language-model-applications/

Simple Azure OpenAI Architecture



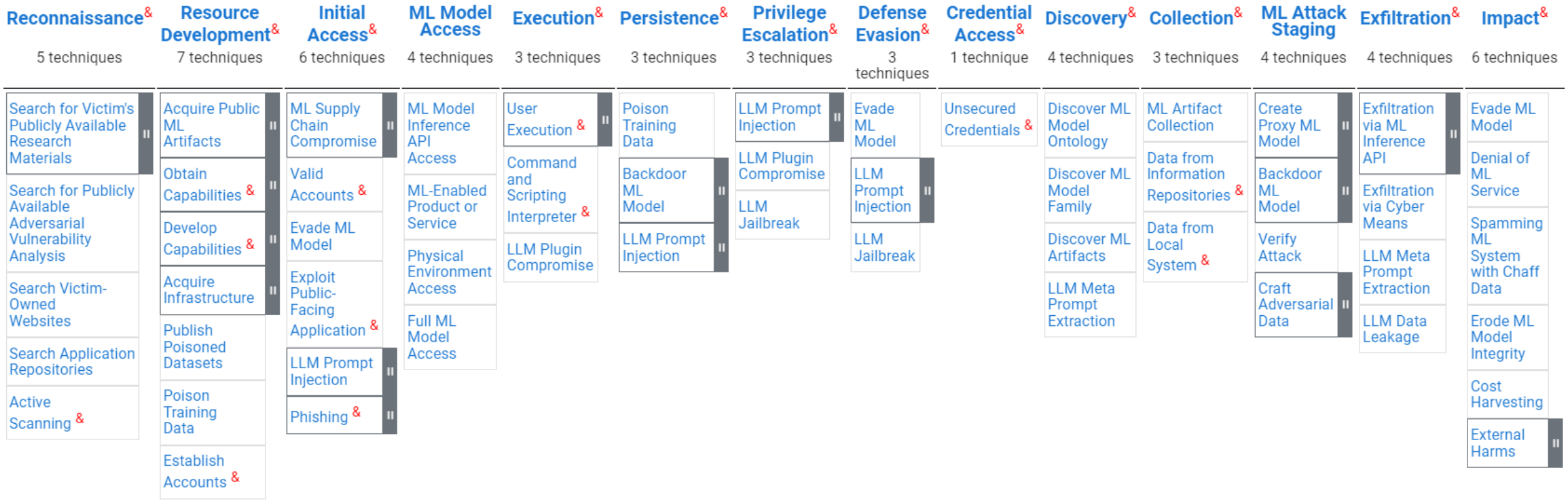
https://learn.microsoft.com/en-us/azure/web-application-firewall/afds/protect-azure-open-ai

Another Azure OpenAI Architecture



https://techcommunity.microsoft.com/t5/azure-architecture-blog/security-best-practices-for-genai-applications-openai-in-azure/ba-p/4027885
https://learn.microsoft.com/en-us/security/benchmark/azure/baselines/azure-openai-security-baseline

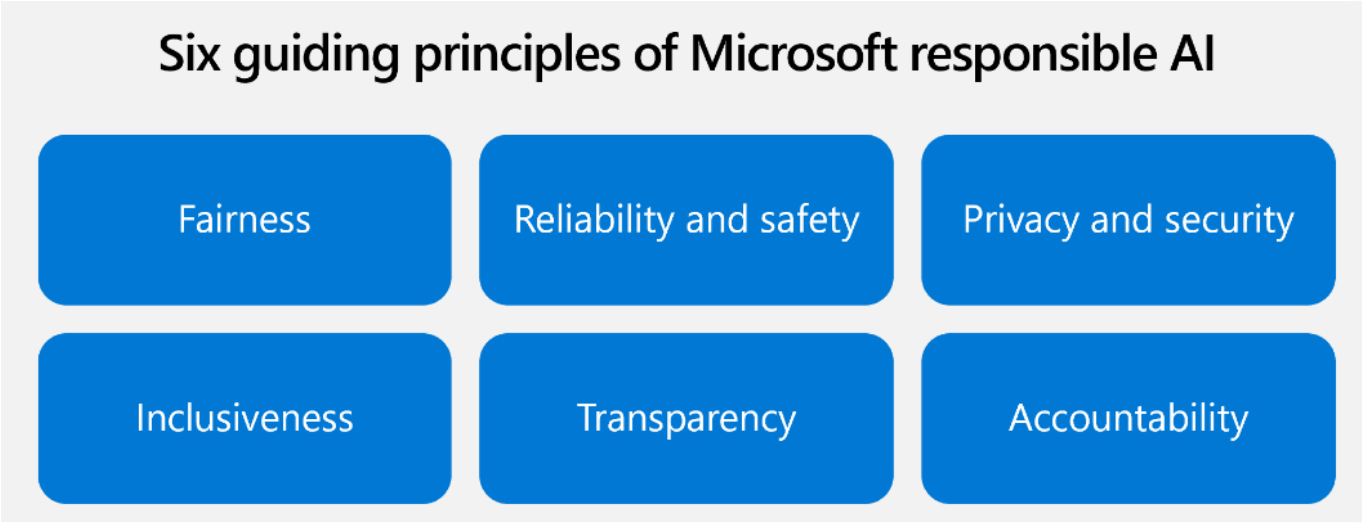
MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems)



https://atlas.mitre.org/matrices/ATLAS

Responsible AI

Detailed guidelines:
[Responsible AI Principles and Approach]



AI Act

Rozporządzenie w sprawie AI (*Artificial Intelligence Act*) zostało przyjęte 13 marca 2024 r.

Najważniejsze punkty:

- Zabezpieczenia dotyczące sztucznej inteligencji ogólnego przeznaczenia;
- Będą wprowadzone kategorię zakazanych praktyk w zakresie sztucznej inteligencji. Do tych zaliczono m.in.: stosowanie rozwiązań opartych na technikach podprogowych czy takich, które dyskryminują określone grupy osób. Niedozwolone będzie także stosowanie systemów AI do tzw. oceny osób obywatelskich (social scoring). To znaczy, że nie będzie można wykorzystywać jej do śledzenia stylu życia obywateli;
- Ograniczenia w korzystaniu z systemów identyfikacji biometrycznej dla organów ścigania. Systemy kategoryzacji biometrycznej, **wykorzystują cechy wrażliwe i nieukierunkowane pobieranie wizerunków twarzy z internetu lub nagrań z telewizji przemysłowej, by stworzyć bazy danych służące rozpoznawaniu twarzy.**
- Operatorzy SI będą musieli między innymi przeprowadzać oceny modeli, oceniać i ograniczać ryzyko systemowe i zgłaszać incydenty;
- Zakaz klasyfikacji punktowej obywateli i stosowania AI do manipulowania użytkownikami i wykorzystywania ich słabości;
- Prawo konsumentów do składania skarg i otrzymywania merytorycznych wyjaśnień;
- Obowiązki w stosunku do sztucznej inteligencji w oparciu o potencjalne ryzyko z nią związane i jej potencjalne skutki.

https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.html

Conclusion

- Azure OpenAI deployments can be challenging but are still very promising
- Developers should be aware of the challenges and take steps to mitigate them
- Collaboration between OpenAI and Azure can lead to better support and standardization