👮 Cyber Security for Sprint 3

In Sprint 3, we continue the development of our Q&A platform utilizing the Furhat robot. Cyber security remains a top priority as we integrate and refine components established in previous sprints. This sprint involved completing LLM training, enhancing the RAG framework, and initiating testing phases. Notable changes included deploying functionalities directly on the Furhat robot and implementing image recognition features.

To address cyber security considerations comprehensively, we adopt a threat modeling approach, assessing each consideration in terms of impacts, likelihood, and consequences of failing to comply with established security protocols.

Development

1. We **incorporated security from the design phase**, including threat modeling and security reviews at each development stage.
2. We utilized **isolated development environments** segmented from production networks.

Deployment

1. We have concluded an **Incident Response Plan** for the project to handle the unexpected breach of our system.
2. We will have an **information exchange with our clients** regarding cybersecurity issues during the final delivery and deployment of the project.

Incident Response Plan

Detection:

1. We advise our clients to hire a cybersecurity specialist to regularly monitor their systems.
2. We use role-based access controls and logging to monitor access and changes to data.

Threat 1: Phishing Attack

The potential phishing attack might be conducted by injecting code into our RAG framework and modifying the prompts, thereby tricking our clients into entering their sensitive data.

Response:

Disable the system to prevent further attacks, identify the vulnerability in our code, and rewrite it to eliminate the weakness preventing code injection in the future.

Threat 2: System Breaches

Interfacing with the GPT-4 API poses risks that could potentially expose our systems to breaches.

Response:

Revoke any compromised API keys and generate new ones.

Threat 3: Denial of Service (DoS) Attack

Disrupt service availability by overwhelming the system with traffic.

Response:

Identify and block the source of the attack using firewall rules and network traffic analysis.

Threat 4: Others Unidentified

Response:

Shut down the system to stop any ongoing attacks, identify the vulnerabilities in our code, and rewrite the code to eliminate the weaknesses, ensuring protection against those unidentified attacks.