Datenschutzkonzept Ayunis

Ayunis ist ein KI-Dienst, der unterschiedliche Funktionen zur Vereinfachung der Geschäftsabläufe seinen Nutzern anbietet. Im vorliegenden Dokument werden die technischen und organisatorischen Maßnahmen beschrieben, um ein Sicherheitsniveau nach Stand der Technik zu gewährleisten.

Nutzerdaten werden ausschließlich in den Datenbanken innerhalb der EU gespeichert. Daten, die zur Verarbeitung an die Sprachmodelle der unterschiedlichen Anbieter weitergeleitet werden, werden nicht für Trainingszwecke verwendet. Dazu wurden entsprechende Verträge mit den Anbietern abgeschlossen.

Im Sinne des AI Acts wird die Nutzung der Ayunis Dienste als “begrenztes Risiko” (Stufe 2) eingestuft.

Ayunis Core

Nutzer interagieren per Chatbot mit den LLMs unterschiedlicher Anbieter. Die Administratoren eines Kundenkontos können festlegen, welche Anbieter (Mistral, Anthropic, OpenAI, MS Azure, Ollama etc.) von den Nutzern verwendet werden. Nutzer können neben Textnachrichten auch Dokumente und Websites in Konversationen mit der KI einbeziehen. Der Dienst beinhaltet die Möglichkeit, Inhalte der Dokumente zu extrahieren und mit KI-Unterstützung zu bearbeiten. Des Weiteren können über Schnittstellen weitere Tools aus Drittsystemen eingebunden werden und die Nutzer können eigene Prompt-Bibliothek verwalten. Der Zweck dieses Dienstes umfasst nicht die Verarbeitung von personenbezogenen Daten von Personen im Beschäftigungsverhältnis oder Bewerberdaten.

Risikobewertung:

- Halluzination des LLM
- Prompt Injection / Jailbreaking über Konversationen oder Prompt Library
- Prompt Injection über selbst konfigurierte Tools

Maßnahmen zur Risikominimierung:

- Die Nutzer werden darauf hingewiesen, dass KI-generierte Inhalte Fehler beinhalten können.
- Der Webdienst zur Interaktion mit dem Nutzer beinhaltet Sicherheitsfunktionen, die das Ausführen von schädlichen Aktionen blockiert.

Zur Risikobewertung und der daraus folgenden Risikobehandlung wurden der Standard „OWASP Top 10 für LLMs“ hinzugezogen. Nachfolgend werden die unterschiedlichen Top 10 der bekannten Risiken behandelt.

Prompt Injection

- Durch Tool Prompts: Interne Tools sind klar definiert mit klaren Input und Output Mustern. Es ist nicht möglich, Tool-Responses zu hijacken und bösartige Prompts zu platzieren

- Durch Tools konfiguriert vom Nutzer: Der Nutzer hat die Verantwortung, von ihm genutzte Dritt-Systeme darauf zu prüfen, ob deren Inhalte, insbesondere deren Antworten, ggf. Gefahren für ihn bergen.
- Durch Jailbreaking Versuche des Nutzers: Die von uns angebotenen Modelle haben bereits alle umfangreichen Schutz gegen Jailbreaks.

Sensitive Information Disclosure

- LLM hat ausschließlich Zugriff auf vom Nutzer bereitgestellte Daten. Der Nutzer steht daher in der Verantwortung, dafür zu sorgen, dass keine Daten eingegeben werden, die nicht eingegeben werden sollen

Supply Chain

Trifft auf Ayunis nicht zu, da nur Basismodelle verwendet werden.

Data & Model Poisoning

Bzgl. Embeddings die durch User Input entstehen: Siehe Prompt Injection

Improper Output Handling

Der Output des Modells erfolgt in vordefinierten Strukturen und wird von Parsen überprüft und abgelehnt, wenn der Output nicht der geforderten Struktur entspricht. Textuelle Inhalte haben keinen Einfluss auf Datenflüsse oder Aktionen innerhalb des Systems und stellen daher in diesem Sinne keine Bedrohung dar.

Excessive Agency

Modelle haben ausschließlich Zugriff auf Tools, die unbedingt erforderlich sind, um ihre Aufgabe zu erfüllen. Dabei haben Modelle nie direkten Zugriff auf Datenbanken oder Kommandozeilen, sondern immer nur abstrahierten Zugriff über feste Pfade.

System Prompt Leakage

Der Zweck der zu verarbeitenden Daten beinhaltet keine sensiblen Informationen

Vector and Embedding Weakness

Wir nutzen ausschließlich etablierte Anbieter für Embeddings, die vertrauenswürdig sind und deren Dienste bereits praxiserprobt sind.

Misinformation

- Halluzinationen können nicht ausgeschlossen werden. Der User wird informiert, dass Modelle Fehler machen können.

Unbounded consumption

- Es gibt eine Obergrenze für Agent Loops (aktuell 20).
- Jeder kostenrelevante Endpunkt ist rate limited.

Technisch-organisatorische Maßnahmen im Sinne des Art. 32 DSGVO

- E2E-Verschlüsselung für Datenübertragung:
  Die internen Komponenten unseres Systems kommunizieren in einem isolierten System und werden nicht über das Internet exponiert. Datenübertragungen zum Frontend bzw. externen Services erfolgen gängigen Web Standards (HTTPS, via TLS 1.3).

- Datenbank-Verschlüsselung (AES-256):
  Eine Verschlüsselung der gespeicherten Daten auf Datenbankebene (Encryption at Rest, z. B. AES-256) ist aktuell nicht aktiviert, kann jedoch bei Bedarf implementiert werden.

- API-Keys sicher verwalten:
  API-Keys und Zugangstoken werden sicher verwaltet und vor unbefugtem Zugriff geschützt. Die Speicherung erfolgt ausschließlich in geschützten Secret-Management-Systemen oder sicheren Konfigurationsspeichern und nicht im Quellcode. Der Zugriff auf API-Keys ist auf autorisierte Systemkomponenten und Administratoren beschränkt. API-Schlüssel werden regelmäßig überprüft und können bei Bedarf jederzeit rotiert oder widerrufen werden. Darüber hinaus werden API-Keys nicht in Log-Dateien oder Fehlermeldungen gespeichert.

- Verschlüsselung der API-Schnittstellen:
  Alle extern erreichbaren API-Schnittstellen werden über HTTPS abgesichert. Dabei wird mindestens TLS 1.3 eingesetzt, um die Vertraulichkeit und Integrität der übertragenen Daten zu gewährleisten. Die TLS-Zertifikate werden von Ayunis selbst verwaltet und regelmäßig erneuert. Unsichere Protokolle und Cipher Suites sind deaktiviert.

- Rollenbasierter Zugang:
  Ist gewährleistet. Zugriff erfolgt über ein Berechtigungsmodell, das an Benutzerrollen geknüpft ist.

- MFA:
  Derzeit noch nicht vorhanden, jedoch geplant für Q3 2026.

- Hashing sensibler Daten (SHA-256):
  Passwörter werden nach gängiger Methode gehasht und verschlüsselt in der Datenbank abgelegt. Wir nutzen dabei die bcryp library, ich glaube es ist genau dieser SHA-256 Hash – zumindest einer der großen Standards. Die Library hat wöchentlich knapp 2,5 Millionen Downloads.
  - Hier der Link zur Library: https://github.com/kelektiv/node.bcrypt.js
  - Ansonsten keine Daten so sensibel, dass sie verschlüsselt werden müssten.

- Verschlüsselung der Anwendungsdaten:
  Daten, die innerhalb der Ayunis-Plattform gespeichert werden, werden verschlüsselt gespeichert. Die Verschlüsselung erfolgt mit etablierten kryptographischen Verfahren nach Stand der Technik.

- Datenminimierung durch Strukturierung der Eingabe (Durch vordefinierte Felder in Formularen): Ist gegeben

- Isolation von KI-Anfragen (Prompt-Isolation):
  Eingaben eines Nutzers werden ausschließlich innerhalb der jeweiligen Sitzung verarbeitet. Prompts werden nicht mit Eingaben anderer Nutzer oder Organisationen vermischt. Eine mandantenübergreifende Verarbeitung oder Speicherung findet nicht statt.

- Datenfilterung - Anonymisierung bzw. Pseudonymisierung:
  Vor der Übermittlung von Eingaben an externe Sprachmodelle wird ein Anonymisierungsmodul auf unserer Infrastruktur eingesetzt. Dieses erkennt personenbezogene Daten (z. B. Namen, E-Mail-Adressen, Telefonnummern, Identifikationsnummern) und ersetzt diese durch neutrale Platzhalter. Die Verarbeitung durch das Sprachmodell erfolgt somit grundsätzlich ohne direkte personenbezogene Identifikatoren.

- Mandantentrennung:
  Daten unterschiedlicher Organisationen werden logisch voneinander getrennt gespeichert. Ein Zugriff zwischen Mandanten ist technisch ausgeschlossen.

- Temporäre Speicherung (dynamisches Löschen):
  Aktuell werden in Ayunis Core Daten sofort hart gelöscht, wenn sie gelöscht werden. Alle Daten bleiben solange vorhanden, wie der zugehörige Nutzer bzw. Organisationsaccount existiert. Wird ein Nutzer-Account gelöscht, werden alle zugehörigen Daten hart gelöscht. Wird ein Organisationsaccount gelöscht, werden alle Nutzer und zugehörigen Daten gelöscht. Daten werden nicht zeitbasiert gelöscht.

- Logging: Alle Abläufe im System werden geloggt. Ausgeschlossen davon sind sensible Daten wie z.B. Chat-Inhalte oder Passwörter. Personenbezogene Daten werden teilweise geloggt, beispielsweise teilweise die E-Mail-Adresse eines Nutzers.