Anmerkung für alle Fragen: Wir können keine rechtliche Beratung erteilen, die Antworten geben unsere persönliche Ansicht wieder.

Sprechen sie uns gerne unverbindlich an:
Giscard Venn, SVA Solution Sales Big Data & Analytics
Mail: Giscard.Venn@sva.de
Mobil: +49 151 26427874

****

F: Gibt es schon Rahmenverträge für Bundesbehörden, um LLMs on premises zu betreiben?

A: Ja, Cloudera ist im IBM Software Rahmenvertrag enthalten. Sprechen sie bitte SVA an.

F: Was müsste beim Urheberrecht beachtet werden, wenn ich die Original-Daten nicht kenne?

A: Wenn sie nicht selbst trainieren und nur das Modell nutzen, dann dürfte das eine rechtliche Grauzone sein.

A-FR: Die verwendeten Open Source LLM verweisen auf die verwendeten Daten.

F: Können Sie bitte welche kritischen Punkte von ChatGPT aus Sicht der Informationssicherheit kurz aufzählen?

A: Neben den Grauzonen sind vor allem die Übermittlung/Ablage der Trainingsdaten in die Cloud zu beachten: Bei OpenAI ist das Azure in den USA, bei MS Bing kann eine deutsche Azure Region gewählt werden.

A-FR: Bereits die gestellte Frage kann kritische Informationen enthalten und sollte das Haus nicht verlassen.

F: Ist bei derzeitigen LLMs eine Beachtung von unterschiedlichen Vertraulichkeitsstati von Informationen beim anlernen des Modells behördenspezifisch wirtschaftlich überhaupt möglich?

A-FR: Die Sprachmodelle werden nicht neu erstellt, sondern lediglich ergänzt um den Behörden-Kontext. Daraus ergeben sich sowohl die Wirtschaftlichkeit des Betriebs der Sprachmodelle wie auch Datensicherheit im Betrieb.

JB: Ein Betrieb in private cloud clustern on premises für Cluster mit unterschiedlichen Sicherheitseinstufungen ist in der Regel die Lösung für das von Ihnen geschilderte Problem.

F: wie gehen Sie ,mit Mitarbeiterdatenschutz um und in welcher Form erfolgt ein Missbrauchstracking?

A: Diese Frage muss OpenAI oder Microsoft gestellt werden. Bei einer lokalen Lösung wie von SVA und Cloudera vorgestellt, können hier Maßnahmen ergriffen werden.

F: Was wäre die Rechtsgrundlage, dass LLM / DM mit Behörden-Daten trainiert wird? Die Daten werden regelmäßig personenbezogene Daten von Kund:innen enthalten.

A: Neben den geltenden Regelungen wir DSGVO fehlen einige Regeln, welche derzeit von dem EU AI Act adressiert werden.

A-FR: Die Sprachmodelle werden nicht neu trainiert, sondern lediglich ergänzt um den Behörden-Kontext. In Ihrer Fragestellung um personenbezogene Daten.

JB: Ein Betrieb in private cloud clustern on premises für Cluster mit unterschiedlichen Sicherheitseinstufungen ist in der Regel die Lösung für den datenschutzkonformen Betrieb der Lösung.

F: Welches Open Source LLM wird von Cloudera genutzt?

A-FR: Der vorgestellte Prototyp verwendet h2oai/h2ogpt-oig-oasst1-512-6_9b verfügbar auf Huggingface.com. Grundsätzlich können eine Vielzahl von LLMs aus der Open Source Community oder Closed Source LLMs genutzt werden.

F: Ist OpenSource nicht eigentlich anfälliger was CyperAngriffe angeht?

A: Keinesfalls. Die Cloudera Data Platform ist entwickelt mit den Prinzipien “security by design” und “governance by design”. Sie ist zudem ausgestattet mit high-end security Komponenten.

F: Ist eine Verwendung von LLM auch in Bezug auf Verschlusssachen (alles unterhalb Einstufung "Geheim") in einem behördeninternen "Intranet" grds. denkbar?

A: ja, aber nur mit einem lokalen LLM.

F: Wie sehen Sie hier das Thema Datenschutzfolgeabschätzung, wenn sie im großen Stil Datenquellen kombinieren…

A: Diese Frage muss OpenAI oder Microsoft gestellt werden. Bei einer lokalen Lösung kann dies erfolgen.

JB: Ein Betrieb in private cloud clustern on premises für Cluster mit unterschiedlichen Sicherheitseinstufungen ist in der Regel die Lösung für den datenschutzkonformen Betrieb der Lösung.

F: Ist als Datenquelle der Wissensdatenbank auch eine behördliche Internetseite möglich?

A: Ja, nahezu jede Datenquelle ist möglich.

F: Es wäre doch auch eine TIA mit Open AI nötig, damit wir chatGPT überhaupt nutzen dürfen? Gibt es die von Open AI?

JB: Die Dienste ChatGPT, DALL:E und andere Dienste, die über Webinterfaces zur Verfügung gestellt werden, fallen in die Kategorie „Non-API-Content“. OpenAI bietet diese Consumer-Dienste direkt den Endkunden an und ist datenschutzrechtlich für die Verarbeitung der Daten verantwortlich.

Adressat von Maßnahmen der Aufsichtsbehörden ist daher in erster Linie auch OpenAI direkt.

https://www.datenschutzkanzlei.de/chatgpt-und-openai-api-in-unternehmen/

F: Wie sieht denn die Lizenzierung aus? Nach welchen Modell werden denn die Lizenzkosten gestaltet? Per User, per Datenmenge oder auf andere Art und Weise?

A-FR: Die Nutzung der Open Source LLM ist i.d.R. ohne Lizenzkosten möglich. Lizenzkosten fallen für die Cloudera Data Platform an, nicht jedoch für die Nutzung des darin eingebetteten LLMs.

https://www.cloudera.com/products/pricing.html

Die Lizenzmetrik erläutern wir Ihnen gerne bei Bedarf.

F: Ich verstehe es so, dass ChatGPT nicht genutzt wird. Es werden Modelle eingesetzt, die gleiche Funktionalität wie ChatGPT (was ja auch nur ein Tool ist) mitbringen, die jedoch in ein "sicheres" Umfeld implementiert werden.

A: Ja, bei der lokalen Lösung.

JB: Ein Betrieb des LLM in private cloud clustern on premises für Cluster mit unterschiedlichen Sicherheitseinstufungen vermeidet die data privacy Probleme, die bei der Nutzung von SaaS basierten Lösungen in der Cloud auftreten.

F: Ist die Anwendung auch in der Bundescloud erhältlich?

A: Nein, noch nicht.

JB: ITZBund betreibt die Bundescloud Lösung und Cloudera Data Platform. Sie sind gegenwärtig noch nicht harmonisiert. Fragen dazu richten Sie bitte an: **kisc@itzbund.de**

F: Können Sie Beispielkosten für ein intern gespeistes Pilotprojekt mit einem ersten use case skizzieren?

A: Das können wir gerne individuell machen, sprechen sie uns an.

F: Wie ersetze ich veraltete oder fehlerhafte Daten in der Wissensdatenbank?

A: Gar nicht, das komplette Modell muss neu trainiert werden.

JB: In der vorgestellten Cloudera Data Platform Lösung entfernen Sie veraltete oder fehlerhafte Daten aus der Wissendatenbank und fügen aktuelle oder fehlerfreie Daten in der Wissensdatenbank hinzu.

Die Vektoren oder embeddings, die das Auffinden der Daten ermöglichen werden im Prozess des Entfernens oder Hinzufügens von Daten in der Wissensdatenbank automatisch neu generiert.

A-FR: Die Änderungen bzw. Löschungen sollten in der Wissensdatenbank erfolgen. Alternativ wäre eine Filterung veralteter Dokumente bei der Erstellung der Vector DB.

F: muss nicht allein schon über die Anfrage an die VektorDB schon eine KI laufen, um zu verstehen, welches Wissen aus der WissensDB gesucht, gefunden und an das LLM für die Generierung der Antwort geliefert werden muss?

A-FR: Das ist eben nicht notwendig, die Fähigkeit der Vector DB besteht darin, die relevanten Dokumente für die Anfrage schnell und zuverlässig zu finden. Die Vektor DB ist eine Funktion der Cloudera Data Platform und nicht des LLMs.

F: Wie kann man sicherstellen, dass die benötigten Informationen auch nur berechtigte Empfänger zur Verfügung gestellt werden (Asyl - Führerschein - Kfz-Zulassung, usw.)?

A-FR: Der Zugriff auf die Dokumente des Behörden-Kontext unterliegen auch bei der Nutzung des Sprachmodells einem in Cloudera Data Platform bereitgestellten granularen Berechtigungskonzept.

Der Zugriff auf Daten - auch über das Sprachmodell - wird verweigert, wenn der Nutzer nicht auch das Recht hat, sie zu verwenden.

Dies ist eine Standard Funktion der Cloudera Data Plaform.

F: Wie sieht es mit Rahmenverträgen in Hessen aus?

A: Sprechen sie uns oder die HZD an.