Merkblatt für den Einsatz von KI in vhb-Lehrangeboten

Die Virtuelle Hochschule Bayern (vhb) stellt für ihre Nutzenden KI-Systeme bereit. Um die Sicherheit und Vertrauenswürdigkeit von KI-Systemen zu erreichen, wird deren Einsatz in der Europäischen Union insbesondere durch die KI-Verordnung ein Rahmen gegeben.

Mit diesem Merkblatt möchten wir die KI-Kompetenz stärken.

Eine Missachtung rechtlicher Regelungen bei der Nutzung von KI-Systemen kann neben der zivilrechtlichen und strafrechtlichen Verantwortung zu prüfungs-, dienst- oder arbeitsrechtlichen Maßregelungen führen. Für einen verantwortungsvollen Einsatz müssen zudem Nutzungsregeln sowie Vorgaben der Anbieter der KI-Systeme beachtet werden.

Freigegebene Dienste

Open AI

OpenAI Ireland Ltd., The Liffey Trust Centre, 117-126 Sheriff Street Upper, Dublin 1, D01 YC43, Ireland. Diese ist Teil der Unternehmensgruppe OpenAI, deren Mutterkonzern seinen Sitz in den USA hat: OpenAI, L.L.C., 1455 3rd Street, San Francisco, CA 94158, USA

Allgemeine Hinweise zu KI-Systemen der Virtuellen Hochschule Bayern (vhb)

Die vhb ermöglicht die Nutzung von KI-Systemen, welche auf KI-Modellen für allgemeine Zwecke aufsetzen, die als hochgradig leistungsfähige Modelle für eine Vielzahl von Aufgaben eingesetzt werden können.

Folgende Funktionen können die KI-Systeme abbilden:

1. Dokumente: Texte verfassen, Tabellen und Diagramme erstellen, Präsentationen entwerfen
2. Informationssuche: Fragen beantworten, Internet durchsuchen, Nachrichten bereitstellen
3. Kreativität: Projektideen, Textunterstützung, Designhilfe
4. Technik: Fehlerbehebung, Anleitungen, technische Fragen
5. Übersetzung: Texte übersetzen, mehrsprachige Kommunikation
6. Datenanalyse: Daten analysieren, Diagramme und Berichte erstellen

Supporthinweis

Für die Nutzung der Funktionen der KI-Systeme kann kein Support gegeben werden, da die Technologie ständig weiterentwickelt wird und individuelle Anfragen nicht abgedeckt werden können. Die von der vhb genutzten KI-Systeme haben eine fortschrittliche Selbsthilfefunktion.

Wir bieten folgende Supportangebote:

Bei grundsätzlichen Fragen zu von uns eingesetzten KI-Systemen, die sich nicht unmittelbar auf deren Nutzung beziehen, können Sie uns unter open@vhb.org kontaktieren.

Allgemeine Hinweise zur Klassifikation von Informationen

TLP-STUFE[1]	HOCH-SCHUL-INTERN	BESCHREIBUNG	WEITERGABE
TLP: RED	Verschluss-sachen	Nur für bekannte Empfänger	Informationen dürfen nur an die direkt anwesenden Personen weitergegeben werden. Keine Weitergabe an Dritte.
TLP: AMBER +STRICT	Streng ver-traulich	Eingeschränkte organisations-interne Verteilung	Informationen dürfen nur innerhalb der vhb-Geschäftsstelle und der vhb-Gre-mien und auf einer „Need-to-know“-Ba-sis weitergegeben werden.
TLP: AMBER	Vertraulich	Eingeschränkte organisations-interne Verteilung	Informationen dürfen innerhalb der vhb-Geschäftsstelle und der vhb-Gremien weitergegeben werden, jedoch nicht an Dritte.
TLP: GREEN	Intern	Organisations-übergreifende Weitergabe	Informationen dürfen innerhalb des vhb-Verbundes weitergegeben, jedoch nicht veröffentlicht werden.
TLP: CLEAR	Öffentlich	Uneingeschränkte Weitergabe	Informationen dürfen uneingeschränkt an jeden weitergegeben werden (ehe-mals TLP:WHITE).

Risiken für Nutzende und Maßnahmen

Der Einsatz von KI-Systemen ist mit Risiken verbunden.

Überblick über zentrale Risiken und notwendige Maßnahmen:

Szenario	Risiken	Maßnahmen zur Risi-kobehandlung
Ordnungs-gemäße Nutzung	Unerwünschte Ausgaben und Bias: KI-Modelle kön-nen unerwünschte oder voreingenommene Inhalte generieren, die auf den Trainingsdaten basieren.	Information durch vor-liegendes Merkblatt
	Fehlende Qualität und Faktizität: Die generierten In-halte können fehlerhaft oder erfunden sein, was als „Halluzinieren“ bezeichnet wird.	Information durch vor-liegendes Merkblatt
	Fehlende Aktualität: Modelle ohne Echtzeitzugriff kön-nen keine aktuellen Informationen liefern.	Allgemeine Sicher-heitsmaßnahmen des Anbieters
	Fehlende Reproduzierbarkeit und Erklärbarkeit: Die Ausgaben sind oft nicht reproduzierbar und schwer nachvollziehbar.	Allgemeine Sicher-heitsmaßnahmen des Anbieters
	Fehlende Sicherheit von generiertem Code: Generier-ter Code kann Sicherheitslücken enthalten.	Allgemeine Sicher-heitsmaßnahmen des Anbieters
	Fehlerhafte Reaktion auf spezifische Eingaben: Kleine Änderungen in den Eingaben können zu gro-ßen Unterschieden in den Ausgaben führen.	Information durch vor-liegendes Merkblatt
	Automation Bias: Nutzende könnten den generierten Inhalten zu viel Vertrauen schenken.	Information durch vor-liegendes Merkblatt

[1] Das Traffic Light Protocol (TLP) ist eine standardisierte Vereinbarung zum Austausch schutzwürdiger, aber nicht formell eingestufter Infor-mationen. Alle Dokumente werden in TLP-Stufen eingeteilt, die die Bedingungen für ihre Weitergabe regeln.
https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/TLP/merkblatt-tlp.pdf

	Anfälligkeit für die Interpretation von Text als Anweisung: KI-Modelle können Texte als Anweisungen interpretieren, was zu unerwünschten Aktionen führen kann.	Information durch vorliegendes Merkblatt
	Fehlende Vertraulichkeit der eingegebenen Daten: Daten könnten während der Übertragung abgegriffen werden.	Information durch vorliegendes Merkblatt
	Selbstverstärkende Effekte und Model Collapse: Eine wiederholte Nutzung von KI-generierten Daten kann zu Verzerrungen führen.	Information durch vorliegendes Merkblatt
	Die Ausgaben erhalten keinen Schutz bei Rechten am geistigen Eigentum.	Information durch vorliegendes Merkblatt
	Die Trainingsdaten können rechtswidrig verarbeitete personenbezogene Daten oder Rechte am geistigen Eigentum Dritter verletzen.	Information durch vorliegendes Merkblatt, vertragliche Regelungen mit dem Dienstanbieter
	Die Ausgaben können rechtswidrig personenbezogene Daten verarbeiteten oder Rechte am geistigen Eigentum Dritter verletzen.	Information durch vorliegendes Merkblatt
Missbräuchliche Nutzung	Die Ausgaben können ein strafbarer Inhalt sein.	Information durch vorliegendes Merkblatt
	Falschmeldungen: KI-Modelle können zur Generierung von Falschinformationen missbraucht werden.	Information durch vorliegendes Merkblatt
	Social Engineering: Kriminelle können KI-Modelle nutzen, um überzeugende Phishing-E-Mails zu erstellen.	Information durch vorliegendes Merkblatt
	Re-Identifizierung von Personen: Anonymisierte oder pseudonymisierte Daten können durch KI-Modelle re-identifiziert werden.	Information durch vorliegendes Merkblatt
	Wissenssammlung für Cyberangriffe: Angreifer können KI-Modelle nutzen, um Informationen für Cyber-Angriffe zu sammeln.	Information durch vorliegendes Merkblatt
	Generierung und Verbesserung von Malware: KI-Modelle können zur Erstellung von Schadcode (Malware) verwendet werden.	Information durch vorliegendes Merkblatt
	Platzierung von Malware: Angreifer können schadhaften Code (Malware) in öffentlichen Bibliotheken platzieren.	Information durch vorliegendes Merkblatt, allgemeine Sicherheitsmaßnahmen des Anbieters
	RCE-Angriffe: „Remote Code Execution“-Angriffe können durch generierten Code ermöglicht werden.	Information durch vorliegendes Merkblatt, allgemeine Sicherheitsmaßnahmen des Anbieters
	Privacy Attacks: Angriffe, die darauf abzielen, Trainingsdaten oder Modellinformationen zu rekonstruieren.	Information durch vorliegendes Merkblatt, allgemeine Sicherheitsmaßnahmen des Anbieters
Angriffe	Evasion Attacks: Angriffe, die darauf abzielen, die Eingaben so zu verändern, dass das Modell Fehlverhalten zeigt.	Information durch vorliegendes Merkblatt, allgemeine Sicherheitsmaßnahmen des Anbieters

	Poisoning Attacks: Angriffe, die darauf abzielen, das Modell durch manipulierte Trainingsdaten zu vergiften.	Information durch vorliegendes Merkblatt, allgemeine Sicherheitsmaßnahmen des Anbieters

Nutzungsregeln

1. Grundlagen:
   - Beachten Sie die spezifischen Regeln für die Nutzung, etwa mit Blick auf das Prüfungsrecht, Arbeitsrecht oder Dienstrecht.
   - Es gelten die gleichen Anforderungen wie an eine dienstliche Recherche im offenen Internet, wenn KI-Systeme zur Recherche im Internet eingesetzt werden.
   - Die Nutzung des KI-Systems muss bedacht, sorgsam und kritisch erfolgen.
   - Die Nutzung des KI-Systems darf ausschließlich als (Lern-)Unterstützung in Zusammenhang mit dem vhb-Angebot erfolgen, im Rahmen dessen das KI-System angeboten wird. Die Nutzung des KI-Systems zu anderweitigen Zwecken ist unzulässig.
2. Nutzende sind für die Eingabe der Prompts verantwortlich:
   - Prompts dürfen nicht gegen rechtliche Regelungen verstoßen.
   - Prompts dürfen keine personenbezogenen Daten (z. B. Namen, Adressen, E-Mail-Adressen, Telefonnummern) beinhalten.
   - Prompts dürfen keine vertraulichen Inhalte (z. B. interne Dokumente, Geschäftsgeheimnisse, nur gemäß TLP:Clear oder TLP:Green klassifizierte Informationen oder Informationen mit direktem oder indirektem Bezug zu derart klassifizierten Informationen) offenlegen.
   - Prompts dürfen keine urheberrechtlich geschützten Werke Dritter (z. B. urheberrechtlich geschützte Textwerke) beinhalten.
   - Prompts dürfen nicht dazu verwendet werden, das KI-System zu manipulieren.
   - Prompts sollen keine Schutzmechanismen des KI-Systems umgehen.
3. Nutzende sind für die Nutzung der Inhalte (= Ausgaben/Antworten des KI-Systems) verantwortlich:
   - Inhalte können frei erfunden sein und sollten daher immer überprüft werden.
   - Inhalte können diskriminierend sein und sollten mit Vorsicht behandelt werden.
   - Inhalte können strafbar sein und gegen geltende Gesetze verstoßen.
   - Inhalte können Rechte Dritter verletzen, wie z. B. Urheberrechte oder Persönlichkeitsrechte.
4. Nutzende müssen transparent sein:
   - Der Einsatz von KI-generierten Inhalten ist offenzulegen, z. B. durch einen Hinweis im Dokument oder der Präsentation.
   - Verwaltungshandeln und personenbezogene Bewertungen dürfen grundsätzlich nicht auf KI-generierten Inhalten basieren.

Dienste

OpenAI

Beschreibung:

OpenAI bietet Zugriff auf verschiedene generative KI-Modelle. Aus diesen Modellen bedient sich beispielsweise auch der bekannte KI-Dienst ChatGPT von OpenAI. Die generativen KI-Modelle von OpenAI erstellen auf Basis einfacher Texteingaben (sog. Prompts) des Nutzenden Textausgaben/-antworten und kreative Inhalte wie Bilder.

Die vhb nutzt die generativen KI-Modelle von OpenAI in einzelnen Kursen auf der OPEN vhb-Plattform (https://open.vhb.org/). Die Nutzung erfolgt in der Art, dass das Moodle-Plugin „OpenAI Chat Block“ (https://moodle.org/plugins/block_openai_chat) in diese Kurse integriert ist. Dieses Plugin bedient sich der generativen KI-Modelle von OpenAI, indem über die OpenAI-API auf diese Modelle zugegriffen wird. So haben die Teilnehmenden in den Kursen die Möglichkeit, durch dieses Plugin eine unmittelbare, KI-basierte Unterstützung beim Lernen bzw. beim Absolvieren des Kurses zu erhalten.

Zugang und Nutzung:

Mit OPEN vhb-Account nach Einschreibung in solche OPEN vhb-Kurse, in die das Moodle-Plugin „OpenAI Chat Block“ integriert ist.

Dokumentation unter https://platform.openai.com/docs/api-reference/introduction

Klassifikation: TLP:CLEAR – Der Dienst darf nur mit öffentlichen Informationen genutzt werden.