Herausgeber: Zentren für Kommunikation und Informationsverarbeitung in Lehre und Forschung e.V.
Titel: Merkblatt für KI-Nutzende = Anlage zu KI-Verordnung/AI Act: KI-Kompetenz ab 02.02.2025 Pflicht
Autoren: Johannes Nehlsen, Karola Möhring, Stephan Rehfeld
Datum: 31.01.2025
Kontakt: geschaeftsstelle@zki.de

Anlage zu:
KI-Verordnung / Artificial Intelligence Act:
KI-Kompetenz ab 2. Februar 2025 Pflicht

Anwenderhinweis: Das Merkblatt ist an die individuelle Situation der Hochschule anzupassen, insbesondere die mit ***markierten Passagen.

KI-Merkblatt für Nutzende

Die Hochschule stellt für ihre Mitglieder und Angehörigen 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.

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

***Freigegebene Dienste

- ***Namen/Bezeichnungen eintragen

Allgemeine Hinweise zu KI-Systemen der Hochschule

Die Hochschule ermöglicht die Nutzung von KI-Systemen, die 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 KI-Funktionen kann kein Support gegeben werden, da die Technologie ständig weiterentwickelt wird und individuelle Anfragen nicht abgedeckt werden können. Unsere KI-Systeme haben eine fortschrittliche Selbsthilfefunktion.

Wir bieten folgende Supportangebote:

- ***Betreffende Supportangebote eintragen

***Allgemeine Hinweise zur Klassifikation von Informationen

TLP-STUFE[1]	HOCHSCHUL 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 vertraulich	Eingeschränkte organisationsinterne Verteilung	Informationen dürfen nur innerhalb der Hochschule und auf einer „Need-to-know“-Basis weitergegeben werden.
TLP: AMBER	Vertraulich	Eingeschränkte organisationsinterne Verteilung	Informationen dürfen innerhalb der Hochschule und an Partner weitergegeben werden, jedoch nicht an Dritte.
TLP: GREEN	Intern	Organisations-übergreifende Weitergabe	Informationen dürfen innerhalb der Hochschulgemeinschaft weitergegeben werden, jedoch nicht veröffentlicht werden.
TLP: CLEAR	Öffentlich	Uneingeschränkte Weitergabe	Informationen dürfen uneingeschränkt an jeden weitergegeben werden (ehemals TLP:WHITE).

[1] Das Traffic Light Protocol (TLP) ist eine standardisierte Vereinbarung zum Austausch schutzwürdiger, aber nicht formell eingestufter Informationen. 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?__blob=publicationFile&v=13

***Risiken für Nutzende und Maßnahmen

Der Einsatz von KI-Systemen ist mit Risiken verbunden.
Über zentrale Risiken und notwendige Maßnahmen geben wir Ihnen nun einen Überblick.

Szenario	Risiken	Maßnahmen zur Risikobehandlung
Ordnungs-gemäße Nutzung	Unerwünschte Ausgaben und Bias: KI-Modelle können unerwünschte oder voreingenommene Inhalte generieren, die auf den Trainingsdaten basieren.	Nutzungsregeln, Warnhinweis bei Ausgaben
	Fehlende Qualität und Faktizität: Die generierten Inhalte können fehlerhaft oder erfunden sein, was als „Halluzinieren“ bezeichnet wird.	Nutzungsregeln, Warnhinweis bei Ausgaben
	Fehlende Aktualität: Modelle ohne Echtzeitzugriff können keine aktuellen Informationen liefern.	Allgemeine Sicherheitsmaßnahmen des Anbieters
	Fehlende Reproduzierbarkeit und Erklärbarkeit: Die Ausgaben sind oft nicht reproduzierbar und schwer nachvollziehbar.	Allgemeine Sicherheitsmaßnahmen des Anbieters
	Fehlende Sicherheit von generiertem Code: Generierter Code kann Sicherheitslücken enthalten.	Allgemeine Sicherheitsmaßnahmen des Anbieters und der Hochschule
	Fehlerhafte Reaktion auf spezifische Eingaben: Kleine Änderungen in den Eingaben können zu großen Unterschieden in den Ausgaben führen.	Information, Nutzungsregeln
	Automation Bias: Nutzende könnten den generierten Inhalten zu viel Vertrauen schenken.	Information, Nutzungsregeln
	Anfälligkeit für die Interpretation von Text als Anweisung: KI-Modelle können Texte als Anweisungen interpretieren, was zu unerwünschten Aktionen führen kann.	Information, Nutzungsregeln
	Fehlende Vertraulichkeit der eingegebenen Daten: Daten könnten während der Übertragung oder durch die betreibende Hochschule abgegriffen werden.	Nutzungsregeln, Klassifikationsregelung, Cloud-Nutzung

	Selbstverstärkende Effekte und Model Collapse: Eine wiederholte Nutzung von KI-generierten Daten kann zu Verzerrungen führen.	Information, Nutzungsregeln
	Abhängigkeit vom entwickelnden/betreibenden Hochschule: Nutzende sind stark von den Betreibern abhängig.	Information, Angebot von verschiedenen KI-Systemen
	Die Ausgaben erhalten keinen Schutz wie durch die Hochschule selbst erstellen Inhalte bei Rechten am geistigen Eigentum.	Information
	Die Trainingsdaten können rechtswidrig verarbeitete personenbezogene Daten oder Rechte am geistigen Eigentum Dritter verletzen.	Nutzungsregeln, vertragliche Regelungen mit dem Diensteanbieter
Miss-bräuchliche Nutzung	Die Ausgaben können rechtswidrig personenbezogene Daten verarbeiteten oder Rechte am geistigen Eigentum Dritter verletzen.	Nutzungsregeln, Information
	Die Ausgaben können ein strafbarer Inhalt sein.	Nutzungsregeln, Information
	Falschmeldungen: KI-Modelle können zur Generierung von Falschinformationen missbraucht werden.	Nutzungsregeln, Information
	Social Engineering: Kriminelle können KI-Modelle nutzen, um überzeugende Phishing-E-Mails zu erstellen.	Nutzungsregeln, Information
	Re-Identifizierung von Personen: Anonymisierte oder pseudonymisierte Daten können durch KI-Modelle re-identifiziert werden.	Nutzungsregeln, Information
	Wissenssammlung für Cyberangriffe: Angreifer können KI-Modelle nutzen, um Informationen für Angriffe zu sammeln.	Nutzungsregeln, Information
	Generierung und Verbesserung von Malware: KI-Modelle können zur Erstellung von Schadcode verwendet werden.	Nutzungsregeln, Information
	Platzierung von Malware: Angreifer können schadhaften Code in öffentlichen Bibliotheken platzieren.	Information, allgemeine Sicherheitsmaßnahmen des Anbieters und der Hochschule

	RCE-Angriffe: „Remote Code Execution“-Angriffe können durch generierten Code ermöglicht werden.	Information, allgemeine Sicherheitsmaßnahmen des Anbieters und der Hochschule
Angriffe	Privacy Attacks: Angriffe, die darauf abzielen, Trainingsdaten oder Modellinformationen zu rekonstruieren.	Nutzungsregeln, allgemeine Sicherheitsmaßnahmen des Anbieters
	Evasion Attacks: Angriffe, die darauf abzielen, die Eingaben so zu verändern, dass das Modell Fehlverhalten zeigt.	Nutzungsregeln, allgemeine Sicherheitsmaßnahmen des Anbieters
	Poisoning Attacks: Angriffe, die darauf abzielen, das Modell durch manipulierte Trainingsdaten zu vergiften.	Nutzungsregeln, 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 muss bedacht, sorgsam und kritisch erfolgen.
2. Nutzende sind für die Eingabe der Prompts verantwortlich:
   - Prompts sollen keine unveröffentlichten personenbezogenen Daten (z.B. Namen, 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 nicht dazu verwendet werden, die KI-Anwendung zu manipulieren.
- Prompts sollen keine Schutzmechanismen der KI-Anwendung umgehen.

3. Nutzende sind für die Nutzung der Inhalte 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 dürfen nicht 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

***KI-gestützte mehrsprachige Dienste der EU für Hochschulen

Beschreibung:

Die EU-Kommission bietet eine Reihe von KI-gestützten Sprachdiensten an, die darauf abzielen, die mehrsprachige Kommunikation innerhalb der EU effizienter und zugänglicher zu gestalten. Diese Dienste umfassen maschinelle Übersetzung, Dokumentenzusammenfassungen, KI-unterstützte Antworterstellung und vieles mehr.

Zugang über:
https://language-tools.ec.europa.eu/

Nutzung:
Für die Nutzung des Dienstes ist ein EU-Login-Konto erforderlich.
***Sie dürfen hierfür die E-Mail-Adresse der Hochschule verwenden.

Dokumentation unter:
https://commission.europa.eu/resources-partners/etranslation_de#translateonline

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

GWDG Chat AI

***Beschreibung:**
Der GWDG Chat AI ist ein Chatbot-Dienst der Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG), der seit dem 22. Februar 2024 verfügbar ist. Dieser Dienst ermöglicht es Nutzern mit einer generativen KI zu interagieren und stellt verschiedene Modelle bereit.

***Zugang und Nutzung:**
***Login über die AcademicCloud oder direkt über diesen Link einfach per SSO einloggen.

***Dokumentation unter:**
https://www.uni-goettingen.de/de/686446.html

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

***Besonderheiten:**
Die GWDG weist darauf hin, dass aus ihrer Sicht eine Nutzung zu privaten Zwecken nicht ausgeschlossen sei.

Fiktive Einsatzbeispiele:

- **Fiktives Ersatzbeispiel „Fenster-Assistent“**

Beschreibung:
„Fenster-Assistent ist eine Assistentenfunktion mit Künstlicher Intelligenz für ABC-Anwendungen und -Dienste, Betriebssystem und Suchmaschine, mit dem Ziel, bei Aufgaben zu helfen, diese schneller fertigzustellen und die Produktivität zu steigern.“

Zugang:

Nutzen Sie Fenster-Assistent immer nur angemeldet mit einem autorisierten Account. Nur mit dieser Anmeldung ist die dienstliche und studentische Nutzung freigegeben.

Nutzung:
Sie können über Fenster-Assistent auch im Web mittels Suchmaschine recherchieren. Sollte im Einzelfall die Verarbeitung einer internen Information notwendig sein, tätigen Sie zuvor die Eingabe „Recherchiere nicht im Web“. Nach dieser Eingabe erfolgt in dem aktiven Chat mit Fenster-Assistent keine Recherche mit Suchmaschine.

Dokumentation:
Weitere Informationen finden Sie in der offiziellen Dokumentation.

Klassifikation:

- TLP:CLEAR bei Recherche mit Suchmaschine: Der Dienst darf nur mit öffentlichen Informationen genutzt werden.
- TLP:AMBER ohne Webrecherche: Der Dienst darf mit vertraulichen Informationen genutzt werden // TLP:GREEN ohne Webrecherche: Der Dienst darf mit internen Informationen genutzt werden.
    - **Fiktives Ersatzbeispiel „Glühwürmchen-Dienst“**

Beschreibung:
„Glühwürmchen-Dienst ist eine generative KI-Plattform, die kreative Inhalte wie Bilder und Texteffekte durch einfache Texteingaben erstellt. Sie bietet Funktionen zur Erstellung von hochwertigen, detaillierten und farblich verbesserten Bildern.“

Zugang und Nutzung:
Mit Hochschul-Account über die autorisierte Plattform.

Dokumentation:
Weitere Informationen finden Sie in der offiziellen Dokumentation.

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