Zentraler API-Zugriff auf LLMs am KIT über die KI-Toolbox

Version 1; 19.11.2025

Nutzen Sie moderne KI-Sprachmodelle einfach und sicher über eine zentrale Schnittstelle (API). Diese Lösung richtet sich an alle Mitarbeitenden und erleichtert die Integration von Large Language Models (LLMs) in eigene Anwendungen und Arbeitsabläufe.

Dieser Service wird vom Scientific Computing Center (SCC) im Rahmen des Strategieprojekts „GenAI@KIT“ bereitgestellt.

Ihre Vorteile auf einen Blick:

- **Kein eigener Account bei Anbietern nötig:** Melden Sie sich mit Ihrem KIT-Account an und nutzen Sie einen selbst generierten API-Schlüssel. Separate Registrierungen z.B. bei OpenAI oder Microsoft sind nicht erforderlich.
- **Zentrale Token-Beschaffung:** Das KIT kauft die nötigen Nutzungskontingente ein und stellt diese allen zur Verfügung. So profitieren Sie von besseren Konditionen.
- **Verbesserte Datenschutz- und Sicherheitsstandards:** Dank Rahmenverträgen, etwa mit Microsoft für die Azure Cloud, gelten beim KIT strengere Vorgaben für den Umgang mit Ihren Daten als bei einer individuellen Nutzung.

Abrechnung und Nutzungshinweise (Fair-Use-Prinzip):
Aktuell entstehen Ihnen keine persönlichen Kosten für die Nutzung des Dienstes. Die Abrechnung erfolgt zentral über das KIT. Der Dienst ist dafür gedacht, Ihre tägliche Arbeit und Forschung unkompliziert zu unterstützen. Gerne können Sie ihn in üblichem Umfang nutzen.

Falls Sie jedoch größere, automatisierte Projekte planen – z.B. das Verarbeiten kompletter Forschungsdatensätze mit vielen tausend API-Aufrufen – setzen Sie sich bitte vorher mit dem zuständigen Team über die Service Adresse ki-toolbox@scc.kit.edu in Verbindung. So können wir gemeinsam die Ressourcen planen und einen reibungslosen Ablauf für alle Nutzende sicherstellen.

Sonderfall: Entwicklung von Diensten und zukünftige Service Accounts
Wenn Sie planen, einen automatisierten Dienst, Bot oder eine Anwendung zu entwickeln, die die API dauerhaft nutzt, muss hierfür ein dedizierter Service-Account eingerichtet werden. Lassen Sie diesen Service-Account bitte durch Ihren ITB (IT-Beauftragten) pro Dienst erstellen. Der Service-Account muss danach per E-Mail-Anfrage (ki-toolbox@scc.kit.edu) an das KI-Toolbox-Team für die Nutzung freigeschaltet werden.

Diese Vorgehensweise stellt eine klare Trennung zwischen persönlicher Nutzung und dienstlicher Nutzung sicher.

Die momentanen Ressourcen sind für den individuellen, dienstlichen Gebrauch gedacht. Daher ist eine Absprache zwingend erforderlich. Ebenfalls ist eigenverantwortlich die Konformität mit dem EU-Ai-Act und anderen gesetzlichen Vorgaben zu klären, z.B. Pflichten aus der Anbieter und Betreiber Rolle.

Systemüberblick: Wie funktioniert der Dienst?

Die gesamte KI-Toolbox, sowohl die Weboberfläche als auch die API, ist zunächst nur innerhalb des KIT-Netzes erreichbar. Für den Zugriff von außerhalb ist eine Verbindung über das KIT-VPN zwingend erforderlich.

Das System besteht aus zwei Hauptkomponenten:

1. Die Web-Oberfläche:
   - Erreichbar unter: https://ki-toolbox.scc.kit.edu/
   - Bietet eine interaktive Chat-Umgebung, die mit der von ChatGPT vergleichbar ist.
   - Hier können Sie auch Ihren persönlichen API-Schlüssel generieren.

2. Die API:
   - Bietet einen OpenAI-kompatiblen Endpunkt, sodass Sie bestehende Anwendungen und Skripte, die für die OpenAI-API entwickelt wurden, mit minimalen Anpassungen weiterverwenden können.
   - Der API-Endpunkt lautet: https://ki-toolbox.scc.kit.edu/api

Über diesen Dienst haben Sie Zugriff auf eine Auswahl verschiedener Modelle:

- **Lokale Modelle**: Diese werden vollständig auf der Infrastruktur des SCC am KIT betrieben.
- **Cloud-Modelle**: Modelle von externen Anbietern wie OpenAI, z.B. bereitgestellt über die Microsoft Azure Cloud an einem deutschen Standort.
- Bitte **verwenden** Sie die beiden Modelle standard-external und standard-local nicht über die API. Die beiden Modelle haben einen angepassten Systemprompt, der den eigenen Systemprompt bei der Verwendung über die API verändern kann und so zu ungewollten Ergebnissen führt.

Informationssicherheit, Datenschutz und Urheberrecht

Der verantwortungsvolle Umgang mit Daten ist bei der Nutzung von KI-Modellen von größter Bedeutung.

Ihre Verantwortung: Urheberrecht und Lizenzen: Wenn Sie Dokumente in das System hochladen (z. B. für RAG-Anwendungen), liegt es in Ihrer persönlichen Verantwortung sicherzustellen, dass Sie über die dafür notwendigen Nutzungsrechte verfügen. Prüfen Sie

die Lizenz des jeweiligen Dokuments und stellen Sie sicher, dass eine Verarbeitung in diesem Kontext erlaubt ist.

Der Datenfluss: Wo sind meine Daten wann?
Das System verfügt über eine lokal am SCC gehostete Vektordatenbank für Retrieval-Augmented Generation (RAG).

- **Datenspeicherung (RAG):** Wenn Sie Dokumente hochladen, werden diese verarbeitet und in der lokalen Vektordatenbank gespeichert. In diesem Zustand verlassen Ihre Daten das KIT nicht.
- **Datenverarbeitung (Anfrage):** Wenn Sie eine Anfrage stellen, werden Ihr Prompt sowie die vom RAG-System gefundenen relevanten Datenausschnitte an das ausgewählte Modell gesendet.
- **WICHTIG**: Wenn Sie ein Cloud-Modell (z. B. *azure.gpt-4.1-mini*) auswählen, werden diese kombinierten Daten zum Zeitpunkt der Abfrage an den externen Anbieter (Microsoft Azure) übertragen. Bei lokalen Modellen verbleiben die Daten innerhalb des KIT.

Drei-Stufen-Modell zur Klassifizierung von Daten (Orientierungshilfe)
Bitte halten Sie sich strikt an die folgende Einordnung:

- **Stufe 1: Öffentliche Daten**
  - **Beschreibung:** Daten, die ohne Einschränkungen öffentlich zugänglich sind.
  - **Einordnung:** Für diese Art von Daten können meist alle verfügbaren Modelle (lokal und Cloud) genutzt werden.

- **Stufe 2: Vertrauliche, aber halb-öffentliche Daten**
  - **Beschreibung**: Daten, die nicht für die Allgemeinheit bestimmt sind, deren Offenlegung aber kein großes Risiko darstellt (z.B. KIT-Intranet-Inhalte).
  - **Einordnung**: Hierfür können je nach dem die OpenAI-Modelle in der Azure Cloud verwendet werden. Durch vertragliche Regelungen besteht hier häufig ein adäquates Schutzniveau. Diese sind unter Azure in Open WebUI eingeordnet und folgen dem Namensschema *azure.[...]*.

- **Stufe 3: Streng vertrauliche oder geheime Daten**
  - **Beschreibung**: Daten, deren Offenlegung einen erheblichen Schaden verursachen könnte (sensible Forschungsdaten, personenbezogene Daten mit hohem Schutzbedarf).
  - **Einordnung**: Für diese Daten dürfen ausschließlich die lokalen Modelle verwendet werden. Diese sind unter KIT in Open WebUI eingeordnet und folgen dem Namensschema *kit.[...]*.

Verantwortungsvoller Umgang mit Ressourcen:
Nutzen Sie zunächst kleinere, kostengünstigere Modelle (z.B. GPT-4.1-nano) und greifen Sie nur bei Bedarf auf die leistungsstärkeren und teureren Modelle zurück.

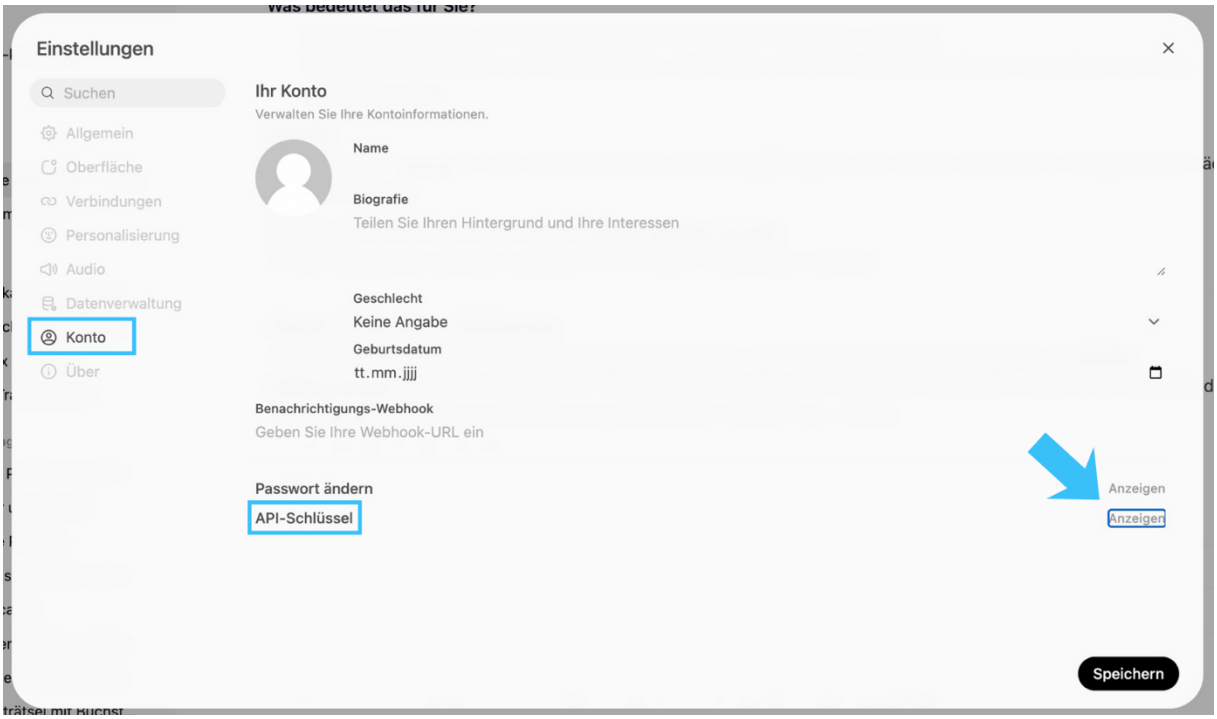
Anleitung zur API-Nutzung (Chat Completion)

Interaktive API-Dokumentation als Referenz
Die umfassendste und immer aktuelle Referenz für alle Endpunkte und Parameter finden Sie in der interaktiven API-Dokumentation (Swagger UI**)**. Dort können Sie Anfragen auch direkt testen.

- Erreichbar unter: https://ki-toolbox.scc.kit.edu/docs

Schritt 1: API-Schlüssel generieren

1. Melden Sie sich unter https://ki-toolbox.scc.kit.edu/ an.
2. Navigieren Sie über das Zahnrad-Symbol zu **Einstellungen > Konto**.
3. Klicken Sie auf "API-Schlüssel erstellen" und kopieren Sie den Schlüssel. Behandeln Sie diesen wie ein Passwort!



Schritt 2: Den OpenAI-kompatiblen API-Endpunkt verwenden
Der Basis-Endpunkt ist *`https://ki-toolbox.scc.kit.edu/api/v1`*. Der Endpunkt für Chat-Anfragen lautet: `https://ki-toolbox.scc.kit.edu/api/v1/chat/completions`

Schritt 3: Das passende Modell auswählen
Identifizieren Sie die ID des Modells (z. B. *`azure.gpt-4.1-mini`* oder *`gpt-oss:120b`*

Schritt 4: API-Anfrage stellen (Beispiele)

curl-Beispiel:

```
curl -X POST https://ki-toolbox.scc.kit.edu/api/v1/chat/completions \
     -H "Authorization: Bearer IHR_PERSÖNLICHER_API_SCHLÜSSEL" \
     -H "Content-Type: application/json" \
     -d '{
         "model": "azure.gpt-4.1-mini",
         "messages": [
             {
                 "role": "system",
                 "content": "Du bist ein hilfreicher Assistent am KIT."
             },
             {
                 "role": "user",
                 "content": "Erkläre das Prinzip der Rayleigh-Streuung in
drei Sätzen."
             }
         ]
     }'
```

Python-Beispiel mit der openai-Bibliothek:
(Voraussetzung: `pip install openai`)

```
from openai import OpenAI

# API-Schlüssel und Endpunkt konfigurieren
api_key = "IHR_PERSÖNLICHER_API_SCHLÜSSEL"
base_url = "https://ki-toolbox.scc.kit.edu/api/v1"

# Client initialisieren
client = OpenAI(api_key=api_key, base_url=base_url)

try:
    # Chat-Anfrage senden
    chat_completion = client.chat.completions.create(
        model="azure.gpt-4.1-mini",  # Wählen Sie hier das passende Modell
        messages=[
            {"role": "system", "content": "Du bist ein hilfreicher
Assistent am KIT."},
            {"role": "user", "content": "Erkläre das Prinzip der Rayleigh-
Streuung in drei Sätzen."}
        ]
    )

    # Antwort ausgeben
    print(chat_completion.choices[0].message.content)
```

```
except Exception as e:
    print(f"Ein Fehler ist aufgetreten: {e}")
```

Anleitung zur API-Nutzung (Chat Completion)

Mit Retrieval-Augmented Generation (RAG) können Sie dem Modell eigene Dokumente als Wissensgrundlage für seine Antworten zur Verfügung stellen.

Upload über die Web-Oberfläche (für den einfachen Einstieg)
Für eine Basisnutzung und den Upload einzelner Dokumente können Sie die Web-Oberfläche verwenden.

- **Wo?** Den Upload-Bereich finden Sie unter Werkzeuge > Wissen.
- **Was?** Diese Methode eignet sich gut für typische Dateiformate wie PDF, TXT etc.

Upload via API und Best Practices (für fortgeschrittene Nutzung)
Für die Automatisierung oder den Upload größerer Dokumentenmengen ist die Nutzung der API komfortabler. Die genauen Endpunkte hierfür finden Sie in der interaktiven API-Dokumentation

- **API-Endpunkte:** Die spezifischen Endpunkte zur Verwaltung von Dateien und Wissensbasen finden Sie in der interaktiven API-Dokumentation (https://ki-toolbox.scc.kit.edu/docs).
- **Empfehlung zur Datenaufbereitung:** Um die besten Ergebnisse mit RAG zu erzielen, empfiehlt es sich, Ihre Quelldokumente in ein sauberes, strukturiertes Markdown-Format aufzubereiten, bevor Sie diese hochladen. Dadurch kann das Modell die Inhalte zuverlässiger interpretieren. Werkzeuge wie Docling können bei dieser Konvertierung hilfreich sein.

Infos & Kontakt

Lizenzhinweis

Diese Anleitung des Zentrums für Mediales Lernen (ZML) am Karlsruher Instituts für Technologie (KIT) ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

Impressum

Herausgeber: Karlsruher Institut für Technologie (KIT) Kaiserstraße 12 76131 Karlsruhe

Kontakt: InformatiKOM Adenauer Ring 12 76131 Karlsruhe Deutschland Tel.: +49 721 608-48200 E-Mail: info@zml.kit.edu