2024
edition

TUTORIAL

DOKU-CHAT.DE

CHATGPT & DOKUMENTE

rechtssicher lokal nutzen

https://doku-chat.de

Anleitung: Wie kann ich ChatGPT mit Dokumenten nutzen?

Table of Contents

Einleitung

Die fortschrittlichen Sprachmodelle wie ChatGPT haben zweifellos beeindruckende Fähigkeiten, wenn es darum geht, menschenähnlichen Text zu generieren und auf Fragen zu antworten. Allerdings ist es wichtig zu verstehen, dass diese Modelle nur über Allgemeinwissen verfügen und keinen Zugang zu kontextspezifischen Informationen haben. Zudem sind sie oft auf einem älteren Wissensstand, was bedeutet, dass sie möglicherweise nicht über die aktuellsten Informationen verfügen.

Glücklicherweise gibt es eine Lösung für dieses Problem: Sie können Ihre eigenen Dokumente hochladen und ChatGPT dadurch mit eigenem Wissen erweitern. Indem Sie spezifische Informationen und Fachkenntnisse in Form von Dokumenten bereitstellen, können Sie sicherstellen, dass ChatGPT Zugriff auf aktuelle und kontextrelevante Informationen hat.

Durch das Hochladen von eigenen Dokumenten können Sie nicht nur die Fähigkeiten von ChatGPT erweitern, sondern auch maßgeschneiderte Antworten und Analysen erhalten, die auf Ihrem spezifischen Fachgebiet oder Ihren individuellen Bedürfnissen basieren. Das Hochladen von Dokumenten ist ein effektiver Weg, um ChatGPT an Ihre persönlichen Anforderungen anzupassen und ein maßgeschneidertes und präzises Texterlebnis zu gewährleisten.

Indem Sie ChatGPT mit eigenem Wissen erweitern, können Sie sicherstellen, dass Sie auf dem neuesten Stand der Dinge bleiben und aktuelle Informationen und Erkenntnisse nutzen. Es bietet Ihnen die Möglichkeit, das Potenzial von ChatGPT voll auszuschöpfen und es zu einem noch wertvolleren Werkzeug in Ihrem Arbeits- oder Forschungsumfeld zu machen.

Die Möglichkeit, eigene Dokumente hochzuladen und ChatGPT so mit individuellem Wissen auszustatten, ist ein bedeutender Schritt, um die Grenzen dieser Sprachmodelle zu überwinden. Es ermöglicht Ihnen, die Leistungsfähigkeit von ChatGPT zu maximieren und es zu einem effektiven Assistenten zu machen, der auf Ihre spezifischen Anforderungen zugeschnitten ist. Nutzen Sie diese Möglichkeit, um das volle Potenzial von ChatGPT auszuschöpfen und von einem personalisierten und erweiterten Texterlebnis zu profitieren.

Aber was ist mit Datensicherheit und DSGVO?

Aber was ist mit Datensicherheit? Diese Frage ist von großer Bedeutung, insbesondere wenn es um das Hochladen von eigenen Dokumenten und Informationen in ChatGPT geht. Es ist wichtig zu beachten, dass Unternehmen wie OpenAI und andere ähnliche Anbieter jeden Input, der ihren Modellen zugeführt wird, als Trainingsmaterial verwenden. Obwohl dies zur Verbesserung der KI-Fähigkeiten beitragen kann, besteht das Risiko, dass Firmengeheimnisse und vertrauliche Informationen offengelegt werden.

Tipp:

Um die Nutzung Ihrer Daten durch die amerikanischen AI Anbieter wie OpenAI zu vermeiden, sollten Sie ein „selbstgehostetes“ AI Modell betreiben.

Dazu benötigen Sie einen Server mit einer Nvidia GPU, und einem entsprechenden „torch“ setup. Für DIY empfiehlt sich z.B. das Open Source Projekt https://ollama.com.

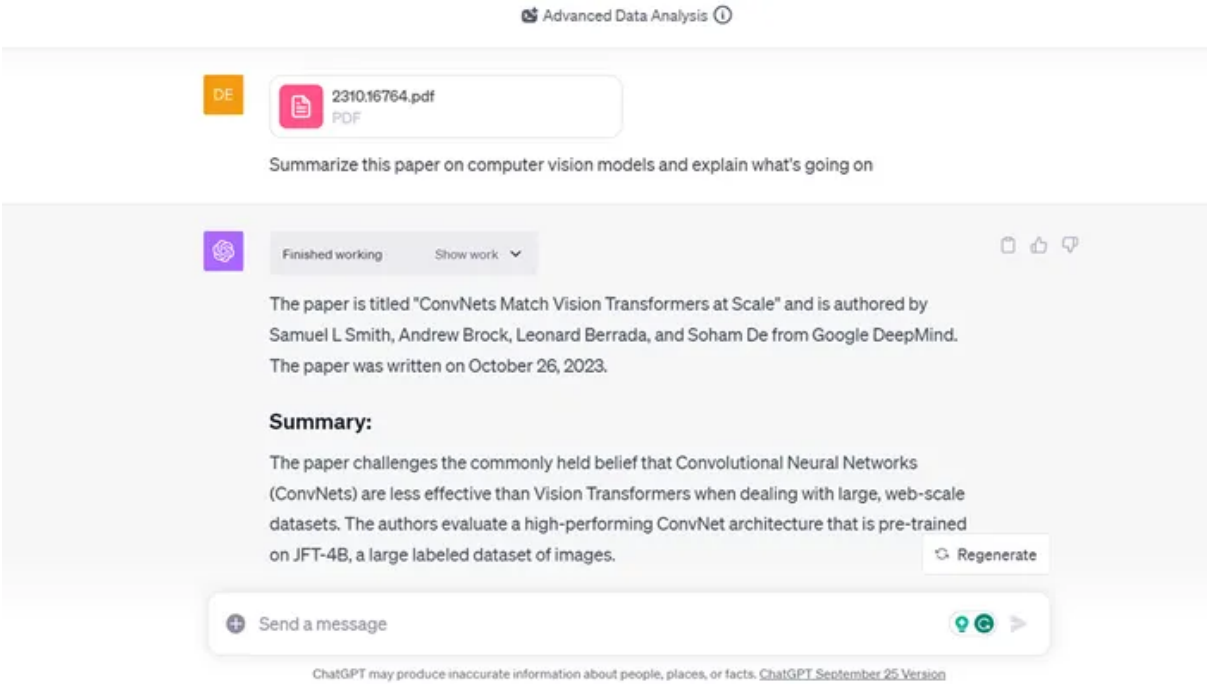
Alternativ ist ein deutscher KI Provider nötig, wie z.B. https://doku-chat.de oder https://datafortress.cloud. Doku-Chat.de bietet außerdem die Unterteilung nach Projekten, nützliche Sharing/Team Features und viel mehr!

Einige Wissensarbeiter geben an, dass die Verwendung von ChatGPT sie um ein Vielfaches produktiver macht. Dennoch blockieren Unternehmen wie JP Morgan und Verizon ChatGPT aufgrund der Risiken für vertrauliche Daten. Unsere Analyse zeigt, dass 4,7% der Mitarbeiter vertrauliche Daten in ChatGPT eingefügt haben.

Angesichts dieser Bedenken ist es entscheidend, die Datensicherheit und den Schutz vertraulicher Informationen zu gewährleisten. Bevor Sie Dokumente oder sensible Daten in ChatGPT hochladen, sollten Sie sicherstellen, dass Sie die Richtlinien und Sicherheitsmaßnahmen des Anbieters verstehen. Es ist ratsam, nur nicht vertrauliche Informationen oder verschlüsselte Daten in ChatGPT einzufügen, um das Risiko von Datenlecks zu minimieren.

Die Nutzung von ChatGPT bietet zweifellos viele Vorteile in Bezug auf Produktivität und Textgenerierung. Dennoch ist es wichtig, die damit verbundenen Risiken zu berücksichtigen und angemessene Sicherheitsvorkehrungen zu treffen, um vertrauliche Informationen zu schützen. Durch die verantwortungsvolle Nutzung von ChatGPT und die Einhaltung bewährter Sicherheitspraktiken können Benutzer das Potenzial dieses Tools nutzen, ohne Kompromisse bei der Datensicherheit einzugehen.

Tutorial: ChatGPT und Dokumente



In diesem Tutorial werden wir Ihnen zeigen, wie Sie ChatGPT nutzen können, um Dokumente hochzuladen, zu analysieren und mit ihnen zu interagieren. Folgen Sie einfach den Schritten unten, um loszulegen:

Schritt 1: Abonnieren Sie ChatGPT Plus oder nutzen Sie ChatGPT for Enterprise

Um die Funktion zur Verarbeitung von Dokumenten zu nutzen, benötigen Sie entweder ein Abonnement für ChatGPT Plus oder Zugang zu ChatGPT for Enterprise. ChatGPT Plus kostet $20 pro Monat und bietet Ihnen erweiterte Funktionen wie die Verarbeitung von Dokumenten.

Schritt 2: Wählen Sie das gewünschte Dokument aus

Sobald Sie über ein ChatGPT Plus-Abonnement oder ChatGPT for Enterprise verfügen, können Sie ein Dokument auswählen, das Sie hochladen möchten. Stellen Sie sicher, dass das Dokument im unterstützten Dateiformat vorliegt, wie z.B. PDF oder andere gängige Formate.

Schritt 3: Laden Sie das Dokument hoch

Klicken Sie auf die entsprechende Option in der ChatGPT-Benutzeroberfläche, um das ausgewählte Dokument hochzuladen. Warten Sie, bis das Dokument erfolgreich hochgeladen wurde.

Schritt 4: Interagieren Sie mit ChatGPT

Nachdem das Dokument hochgeladen wurde, können Sie direkt mit ChatGPT interagieren. Stellen Sie Fragen zum Inhalt des Dokuments, bitten Sie ChatGPT um eine Zusammenfassung oder fordern Sie Analysen oder Trends basierend auf den Daten im Dokument an.

Schritt 5: Nutzen Sie erweiterte Funktionen

ChatGPT bietet Ihnen auch erweiterte Funktionen wie die Möglichkeit, Python-Code zu schreiben und auszuführen, Datei-Uploads zu verarbeiten und bei der Datenanalyse zu helfen. Überprüfen Sie die Einstellungen, um sicherzustellen, dass Sie die erweiterten Funktionen aktiviert haben.

Schritt 6: Achten Sie auf Datensicherheit

Es ist wichtig, die Datensicherheit im Auge zu behalten, wenn Sie Dokumente in ChatGPT hochladen. Beachten Sie, dass OpenAI und ähnliche Anbieter jeden Input zu ihren Modellen als Trainingsmaterial verwenden können. Laden Sie daher nur nicht vertrauliche Informationen oder verschlüsselte Daten hoch, um das Risiko von Datenlecks zu minimieren.

Mit diesen Schritten können Sie ChatGPT effektiv nutzen, um Dokumente hochzuladen und mit ihnen zu interagieren. Das Hochladen von eigenen Dokumenten ermöglicht es Ihnen, ChatGPT mit individuellem Wissen zu erweitern und maßgeschneiderte Antworten und Analysen zu erhalten. Denken Sie jedoch immer daran, die Datensicherheit zu berücksichtigen und angemessene Vorsichtsmaßnahmen zu treffen.

Tipp:

Um die Nutzung Ihrer Daten durch die amerikanischen AI Anbieter wie OpenAI zu vermeiden, sollten Sie ein „selbstgehostetes“ AI Modell betreiben.

Dazu benötigen Sie einen Server mit einer Nvidia GPU, und einem entsprechenden „torch“ setup. Für DIY empfiehlt sich z.B. das Open Source Projekt https://ollama.com.

Alternativ ist ein deutscher KI Provider nötig, wie z.B. https://doku-chat.de oder https://datafortress.cloud. Doku-Chat.de bietet außerdem die Unterteilung nach Projekten, nützliche Sharing/Team Features und viel mehr!

- Consulting/Beratung: https://datafortress.cloud
- KI Anbieter: https://doku-chat.de/de/
- info@datafortress.cloud
- 01601136770

Selbstgehostete KI-Modelle

In der Welt der künstlichen Intelligenz (KI) haben sich große Sprachmodelle wie GPT (Generative Pre-trained Transformer) als leistungsstarke Werkzeuge für eine Vielzahl von Anwendungen etabliert. Unternehmen sollten jedoch vorsichtig sein, sich zu stark auf diese Modelle zu verlassen. Während sie Genauigkeit und Skalierbarkeit bieten, haben sie auch ihre Grenzen. Sie können aufgrund von Einschränkungen bei den Trainingsdaten voreingenommene oder falsche Antworten

liefern und sind nicht in der Lage, den menschlichen Kontext und Feinheiten vollständig zu verstehen.

In diesem Blog-Beitrag werden wir uns mit den Vorteilen von selbst gehosteten KI-Modellen auseinandersetzen. Selbst gehostete Modelle bieten Unternehmen eine größere Kontrolle über ihre KI-Infrastruktur, was zu einer besseren Leistung, Anpassungsfähigkeit und Datenschutz führen kann. Indem sie das Modell auf ihren eigenen Servern hosten, können Unternehmen sicherstellen, dass ihre proprietären Daten unter ihrer Kontrolle bleiben und Bedenken hinsichtlich Datensicherheit und Datenschutz adressiert werden.

Allerdings birgt die Selbst-Hosting auch Herausforderungen. Es erfordert erhebliche Investitionen in Hardware und Fachkenntnisse, um die Infrastruktur zu verwalten und zu warten. Darüber hinaus liegt die Verantwortung für die Aktualisierung des Modells, um neue Informationen zu berücksichtigen, und für den Schutz vor Sicherheitslücken vollständig beim Unternehmen. Angesichts des raschen Fortschritts in der KI und der Cybersicherheit kann dies eine anspruchsvolle Aufgabe sein.

Insgesamt repräsentieren große Sprachmodelle einen bedeutenden Fortschritt in der KI und bieten Unternehmen das Potenzial für Innovation und Effizienzsteigerung. Doch ihre Grenzen und die Komplexität von Feinabstimmung und Selbst-Hosting erfordern einen vorsichtigen Ansatz. Unternehmen müssen die Vorteile dieser leistungsstarken Werkzeuge gegen die Risiken einer übermäßigen Abhängigkeit, Ungenauigkeiten und ethischen Bedenken abwägen. Indem sie dies tun, können sie das Potenzial von selbst gehosteten KI-Modellen nutzen und gleichzeitig die Risiken im Zusammenhang mit diesen wegweisenden Technologien mindern.

Warum nicht einfach ein Cloud-Modell verwenden?

Es gibt mehrere Gründe, warum Unternehmen und Privatpersonen sich für das Selbsthosting von KI-Modellen entscheiden. Einer der Hauptgründe ist die Datensicherheit. Bei der Verwendung von Modellen von Anbietern wie OpenAI besteht die Möglichkeit, dass die eingegebenen Daten als Trainingsmaterial verwendet werden. Durch das Selbsthosting können Unternehmen sicherstellen, dass ihre sensiblen Daten unter ihrer Kontrolle bleiben und nicht für Trainingszwecke verwendet werden.

Tipp: Um die Nutzung Ihrer Daten durch die amerikanischen AI Anbieter wie OpenAI zu vermeiden, sollten Sie ein „selbstgehostetes“ AI Modell betreiben. Dazu benötigen Sie einen Server mit einer Nvidia GPU, und einem entsprechenden „torch“ setup. Für DIY empfiehlt sich z.B. das Open Source Projekt https://ollama.com . Alternativ ist ein deutscher KI Provider nötig, wie z.B. https://doku-chat.de oder https://datafortress.cloud . Doku-Chat.de bietet außerdem die Unterteilung nach Projekten, nützliche Sharing/Team Features und viel mehr!

Ein weiterer wichtiger Aspekt ist die Kostenersparnis. Selbst gehostete Modelle sind in der Regel kostenfrei oder erfordern nur geringe Lizenzgebühren im Vergleich zu den hohen Kosten, die mit der Nutzung von Modellen von Drittanbietern verbunden sein können. Dies macht selbst gehostete Modelle zu einer wirtschaftlich attraktiven Option für Unternehmen jeder Größe.

Der Schutz der Privatsphäre ist ein weiterer entscheidender Faktor. Durch das Selbsthosting behalten Unternehmen die Kontrolle über ihre Daten und können sicherstellen, dass sie nicht an Dritte weitergegeben werden. Dies ist insbesondere in Branchen mit strengen Datenschutzbestimmungen von großer Bedeutung.

Zusammenfassend bietet das Selbsthosting von KI-Modellen Vorteile wie verbesserte Datensicherheit, Kostenersparnis und Wahrung der Privatsphäre. Unternehmen können die volle Kontrolle über ihre KI-Infrastruktur behalten und gleichzeitig von den Vorteilen dieser leistungsstarken Technologie profitieren.

Was benötige ich um ein selbst gehostetes KI-Modell zu betreiben?

Um ein selbst gehostetes KI-Modell zu betreiben, benötigen Sie einige spezifische Ressourcen. Ein wichtiger Bestandteil ist ein Nvidia-GPU-Server, da GPUs für das Training und die Ausführung von KI-Modellen eine hohe Rechenleistung bieten. Es ist jedoch wichtig zu beachten, dass Nvidia-GPU-Server ziemlich teuer sein können und eine beträchtliche Investition erfordern können.

Glücklicherweise gibt es Open-Source-Projekte wie Ollama, die das Hosting von KI-Modellen erleichtern. Ollama ist ein Projekt auf GitHub (https://github.com/ollama/ollama) , das eine einfache und benutzerfreundliche Möglichkeit bietet, Modelle selbst zu hosten. Es bietet Tools und Ressourcen, um den Prozess des Selbsthostings zu vereinfachen und zu automatisieren.

Mit Ollama und einem geeigneten Nvidia-GPU-Server können Sie Ihr eigenes KI-Modell betreiben und von den Vorteilen des Selbsthostings profitieren, ohne die volle finanzielle Belastung eines dedizierten Servers tragen zu müssen.

Es ist wichtig zu beachten, dass das Betreiben eines selbst gehosteten KI-Modells weiterhin technisches Fachwissen erfordert. Es kann hilfreich sein, über Kenntnisse in den Bereichen Maschinelles Lernen und Serververwaltung zu verfügen oder Experten hinzuzuziehen, um Ihnen bei der Einrichtung und Wartung zu helfen.

Mit den richtigen Ressourcen und Tools wie Ollama können Sie Ihr eigenes KI-Modell betreiben und die Vorteile des Selbsthostings nutzen, um Kontrolle, Datensicherheit und Kostenersparnis zu gewährleisten.

Und was für einen Server brauche ich für die Modelle?

Zuerst sollte man abwegen, was man von dem Modell erwartet. “Coding Assistenten” á la Github Copilot brauchen weniger ressourcen und kommen mit einer GPU mit ~8GB VRAM aus. Für größere Modelle welche sich ähnlich ChatGPT verhalten braucht man schon mindestens 16 GB VRAM. Erwartet man eine Performance von GPT-4 muss man schon mehr Geschütze auffahren, und sollte sich an einem GPU Cluster mit mindestens 64 GB VRAM orientieren.

Eine kleine Übersicht der Modelle, und wie viel Speicherplatz sie benötigen findet man hier:

Model	Benötigte VRAM
Llama 2	8 GB
Mistral	8 GB
Dolphin Phi	8 GB
Phi-2	8 GB
Neural Chat	8 GB
Starling	8 GB
Code Llama	8 GB
Llama 2 Uncensored	8 GB
Llama 2 13B	16 GB
Llama 2 70B	64 GB
Orca Mini	8 GB
Vicuna	8 GB
LLaVA	8 GB
Gemma	8 GB
Gemma	8 GB

Und was für eine GPU ist das konkret?

Leider ist NVIDIA derzeit im Monopol was AI GPUs angeht. Man kann auch mit technischen "Hacks" eine AMD GPU verwenden, jedoch ist die Performance und Kompatibilität nicht mit allen Projekten gegeben, bzw. muss man die benötigte Arbeitszeit vs. die Kosten für eine NVIDIA GPU abwägen.

Die Server GPU Linie von NVIDIA kommt mit einem stolzen Preis daher, weshalb es durchaus Sinn machen kann die Gaming-GPU's von NVIDIA zu verwenden.

Da es nicht nur auf die VRAM, sondern auch die Tensor-Cores ("AI Cores") ankommt, macht es durchaus Sinn die 4000er Reihe von NVIDIA zu verwenden anstatt der 3000er Modelle.

NVIDIA GPU Name	VRAM	Geeignet für …	Preis
RTX 4070	12 GB	Coding Completion, sehr simple Tasks (fasse Zusammen, korrigiere, sortiere, …)	~648€
RTX 4080	16 GB	Schreibe einen Text über, generiere Code, Leistung ~60% von gpt-3.5	~1.100€
RTX 4090	24 GB	Schreibe einen Text über, generiere Code, Leistung ~80% von gpt-3.5	~1.900€

Das klingt zu stressig? Oder wäre besser als Hosted API? Nutzen Sie doku-chat.de oder lassen sie sich in einem kostenlosen Erstgespräch von uns beraten .

Die Vor- und Nachteile des selbst gehosteten KI-Modellen

Das selbst gehostete KI-Modell bietet eine Reihe von Vor- und Nachteilen im Vergleich zu anderen Hosting-Optionen wie Cloud-basierten Plattformen. Einer der Vorteile des selbst gehosteten Modells ist die Skalierbarkeit. Unternehmen haben die volle Kontrolle über ihre Infrastruktur und können das Modell an ihre spezifischen Anforderungen und Ressourcen anpassen. Dies ermöglicht eine bessere Skalierbarkeit und Leistungsoptimierung.

Ein weiterer Vorteil ist die Flexibilität. Durch das selbst gehostete Modell haben Unternehmen die Freiheit, eigene Anpassungen vorzunehmen und das Modell nach Belieben anzupassen. Dies ermöglicht eine größere Anpassungsfähigkeit an spezifische Anwendungsfälle und Geschäftsanforderungen.

Jedoch gibt es auch einige Nachteile beim selbst gehosteten Modell. Ein wesentlicher Aspekt ist der Wartungsaufwand. Unternehmen sind dafür verantwortlich, das Modell auf dem neuesten Stand zu halten, Patches und Updates einzuspielen und die Sicherheit zu gewährleisten. Dies erfordert technisches Fachwissen und Ressourcen für die Serververwaltung.

Ein weiterer Faktor ist die Kostenfrage. Obwohl das selbst gehostete Modell langfristig kosteneffizient sein kann, erfordert es zunächst eine erhebliche Investition in Hardware und Expertise. Dies kann für kleinere Unternehmen oder Start-ups eine finanzielle Herausforderung darstellen.

Es ist wichtig, die Vor- und Nachteile gründlich abzuwägen und die individuellen Anforderungen des Unternehmens zu berücksichtigen. Ein selbst gehostetes Modell bietet mehr Kontrolle und Anpassungsfähigkeit, erfordert aber auch mehr Verantwortung und Ressourcen. Cloud-basierte Plattformen können hingegen eine schnellere Bereitstellung und geringeren Wartungsaufwand bieten, jedoch mit weniger Kontrolle und Anpassungsmöglichkeiten. Es ist ratsam, die spezifischen Bedürfnisse und Möglichkeiten des Unternehmens zu analysieren, um die beste Hosting-Option zu wählen.

Das klingt zu stressig? Oder wäre besser als Hosted API? Nutzen Sie doku-chat.de oder lassen sie sich in einem kostenlosen Erstgespräch von uns beraten .

Tutorial: Selbstgehostete KI mit Datensicherheit mit Ollama und Chatbox

Das Problem: 75% der Firmen verbieten die Nutzung von ChatGPT

Trotz der anfänglichen Begeisterung für Generative KI-Tools wie ChatGPT, ziehen Unternehmen aufgrund wachsender Datenschutz- und Cybersicherheitsbedenken in Erwägung, deren Verwendung zu beschränken. Die Sorge besteht vor allem darin, dass diese KI-Tools Nutzerdaten speichern und aus ihnen lernen, was potenziell zu unbeabsichtigten Datenlecks führen könnte. Obwohl OpenAI, der Entwickler von ChatGPT, eine Opt-out-Option für das Training mit Nutzerdaten bietet, bleibt die Frage, wie die Daten innerhalb des Systems gehandhabt werden, unklar. Zudem fehlen klare gesetzliche Regelungen zur Verantwortung bei durch KI verursachten Datenverletzungen. Unternehmen sind daher zunehmend vorsichtig und warten ab, wie sich die Technologie und ihre Regulierung weiterentwickeln.

Microsoft hat seine Mitarbeiter davor gewarnt, sensible Daten mit ChatGPT, dem von OpenAI entwickelten Chatbot, zu teilen Quelle . Die Sorge ist, dass vertrauliche Informationen versehentlich mit dem Chatbot geteilt werden könnten, der diese Informationen dann möglicherweise mit anderen Nutzern teilen könnte. Microsofts vorsichtiger Ansatz ist bemerkenswert, wenn man bedenkt, dass das Unternehmen eine Partnerschaft mit OpenAI eingegangen ist und in diese investiert hat. Auch Amazon hat eine ähnliche Warnung an seine Mitarbeiter herausgegeben. Die Verantwortung für den Schutz vertraulicher Daten ist derzeit unklar, und es besteht ein Bedarf an klareren Richtlinien und Vorschriften für diese Situationen. Die Nutzungsbedingungen von OpenAI erlauben es dem Unternehmen, alle von den Nutzern und ChatGPT erzeugten Eingaben und Ausgaben zu verwenden, wobei personenbezogene Daten entfernt werden sollen. Es bestehen jedoch nach wie vor Bedenken hinsichtlich der Möglichkeit, dass private Unternehmensdaten durch geschickt gestaltete Eingabeaufforderungen extrahiert werden können.

Die Lösung: Selbstgehostete KI-Modelle

Eine Lösung für die Bedenken hinsichtlich der Datensicherheit und des Datenschutzes bei der Verwendung von KI-Modellen besteht darin, diese Modelle selbst zu hosten. Dies ermöglicht es Unternehmen, die Kontrolle über ihre Daten zu behalten und sicherzustellen, dass sie nicht in die falschen Hände geraten. Selbstgehostete KI-Modelle bieten eine sichere und vertrauenswürdige Möglichkeit, KI-Technologie zu nutzen, ohne sich um Datenschutz- und Cybersicherheitsbedenken sorgen zu müssen.

Wir werden dazu *Ollama* und *Chatbox* verwenden. Ollama ist ein optimiertes Tool zur lokalen Ausführung von Open-Source-Large Language Models (LLMs) wie Mistral und Llama 2. Chatbox ist eine Applikation, mit welcher die API Calls zu verschiedenen Modellen visualisiert werden können.
Das klingt zu stressig? Oder wäre besser als Hosted API? Nutzen Sie doku-chat.de oder lassen sie sich in einem kostenlosen Erstgespräch von uns beraten .

Schritt 1: Ollama Download und Installation

Sie benötigen einen Server oder Rechner, der eine GPU besitzt. Leider muss dies eine NVidia GPU mit mindestens 8GB VRAM sein, oder ein Macbook mit der M-Chip Reihe. Die genauen Anforderungen finden Sie im Post: Was sind die Vorteile von selbst gehosteten KI Modellen?
TLDR version:

1. https://ollama.com/download
2. ollama run llama2
3. ollama ist unter localhost:11434 erreichbar

Was ist Ollama?

Ollama ist ein optimiertes Tool zur lokalen Ausführung von Open-Source-Large Language Models (LLMs) wie Mistral und Llama 2. Ollama bündelt Modellgewichte, Konfigurationen und Datensätze zu einer einheitlichen Einheit, die von einem Modelfile verwaltet wird.

Das klingt zu stressig? Oder wäre besser als Hosted API? Nutzen Sie doku-chat.de oder lassen sie sich in einem kostenlosen Erstgespräch von uns beraten .
Ollama unterstützt verschiedene LLMs, darunter LLaMA-2, uncensored LLaMA, CodeLLaMA, Falcon, Mistral, Vicuna-Modell, WizardCoder und Wizard uncensored.

Ollama unterstützt eine Vielzahl von Modellen, darunter Llama 2, Code Llama und andere. Es bündelt Modellgewichte, Konfigurationen und Daten in einer einzigen Einheit, die von einem Modelfile definiert wird.

Die fünf beliebtesten Modelle auf Ollama sind:

- llama2: Das beliebteste Modell für den allgemeinen Gebrauch.
- mistral: Das 7B-Modell von Mistral AI, aktualisiert auf Version 0.2.
- codellama: Ein großes Sprachmodell, das Texteingaben verwendet, um Code zu generieren und zu diskutieren.
- dolphin-mixtral: Ein uncensoredes, feinabgestimmtes Modell auf Basis des Mixtral MoE, das sich besonders für Codieraufgaben eignet.
- mistral-openorca: Mistral 7b, feinabgestimmt mit dem OpenOrca-Datensatz.

Ollama unterstützt außerdem die Erstellung und Verwendung von benutzerdefinierten Modellen. Sie können ein Modell mithilfe eines Modelfiles erstellen, in dem Sie die Modelldatei übergeben,

verschiedene Schichten erstellen, die Gewichte schreiben und schließlich eine Erfolgsmeldung erhalten.

Einige der anderen Modelle, die auf Ollama verfügbar sind, umfassen:

- Llama2: Meta's grundlegendes "Open Source"-Modell
- Mistral/Mixtral: Ein 7 Milliarden Parameter-Modell, das auf dem Mistral 7B-Modell mit dem OpenOrca-Datensatz feinabgestimmt ist.
- Llava: Ein multimodales Modell namens LLaVA (Large Language and Vision Assistant), das visuelle Eingaben interpretieren kann.
- CodeLlama: Ein Modell, das sowohl auf Code als auch auf natürliche Sprache in Englisch trainiert ist.
- DeepSeek Coder: Von Grund auf auf 87% Code und 13% natürlicher Sprache in Englisch trainiert.
- Meditron: Ein Open-Source-Medizinmodell, das aus Llama 2 ins medizinische Umfeld adaptiert wurde.
-

Installation und Einrichtung von Ollama

- Laden Sie Ollama von der offiziellen Website herunter.
- Nach dem Download ist der Installationsprozess einfach und ähnlich wie bei anderen Softwareinstallationen. Für macOS- und Linux-Benutzer kann Ollama mit einem Befehl installiert werden: curl https://ollama.ai/install.sh | sh.
- Sobald Ollama installiert ist, erstellt es eine API, über die das Modell bereitgestellt wird, sodass Benutzer direkt von ihrem lokalen Rechner aus mit dem Modell interagieren können.
- Ollama ist mit macOS und Linux kompatibel, die Unterstützung für Windows wird in Kürze verfügbar sein. Es kann leicht installiert und verwendet werden, um verschiedene Open-Source-Modelle lokal auszuführen. Das gewünschte Modell können Sie aus der Ollama-Bibliothek auswählen.
-

Ausführen von Modellen mit Ollama

Das Ausführen von Modellen mit Ollama ist ein einfacher Prozess. Benutzer können Modelle herunterladen und mit dem Befehl "run" in der Terminalanwendung ausführen. Wenn das Modell nicht installiert ist, lädt Ollama es automatisch herunter. Zum Beispiel, um das CodeLlama-Modell auszuführen, verwenden Sie den Befehl "ollama run codellama".

Die Modelle werden in der ~/.ollama/models-Verzeichnisstruktur auf Ihrem lokalen Rechner gespeichert. Wenn Sie ein Modell mit dem Befehl "ollama pull" herunterladen, wird es im Verzeichnis ~/.ollama/models/manifests/registry.ollama.ai/library//latest gespeichert.

Ollama unterstützt auch die Verwendung des Modells über eine REST-API für Echtzeitinteraktionen.

Zusätzliche Features

Eine der einzigartigen Funktionen von Ollama ist die Unterstützung des Imports von GGUF- und GGML-Dateiformaten im Modelfile. Dies bedeutet, dass Sie ein Modell erstellen, darauf aufbauen, es iterativ verbessern und es hochladen können, um es mit anderen zu teilen, wenn Sie bereit sind.

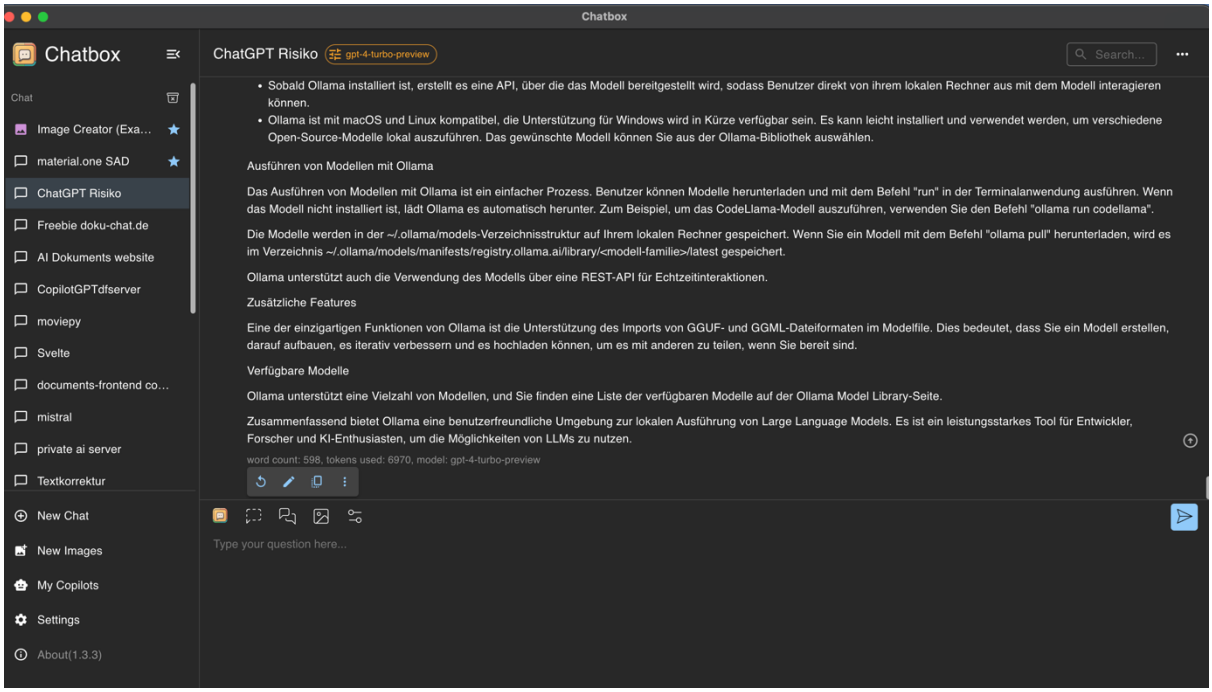
Verfügbare Modelle

Ollama unterstützt eine Vielzahl von Modellen, und Sie finden eine Liste der verfügbaren Modelle auf der Ollama Model Library-Seite.

Zusammenfassend bietet Ollama eine benutzerfreundliche Umgebung zur lokalen Ausführung von Large Language Models. Es ist ein leistungsstarkes Tool für Entwickler, Forscher und KI-Enthusiasten, um die Möglichkeiten von LLMs zu nutzen.

Schritt 2: Ansteuern und Nutzen von Ollama mit der Chatbox GUI

Chatbox ist eine Applikation, mit welcher die API Calls zu verschiedenen Modellen visualisiert werden können.

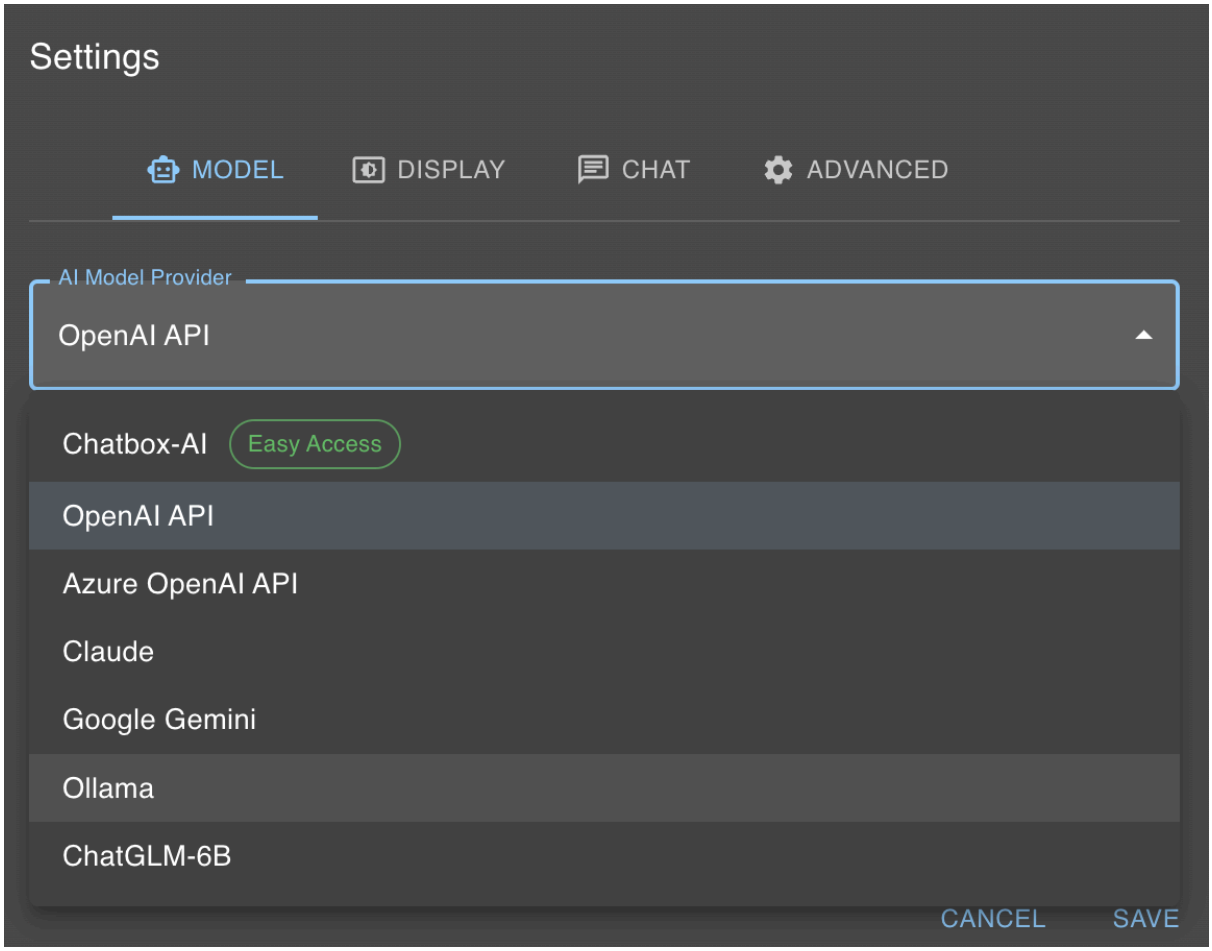


Chatbox Installation

Besuchen Sie https://chatboxai.app und installieren Sie die App auf Ihrem Rechner.

Chatbox und Ollama verbinden

Wie im vorherigen Abschnitt erwähnt, ist unser lokales KI Modell nun unter localhost:11434 erreichbar. Chatbox bietet glücklicherweise eine einfache Möglichkeit, um mit Ollama zu interagieren. Dazu klickt man auf “Settings”, und wählt im “AI Model Provider” Ollama aus.



Danach muss man nur aus dem Modelldialog das Modell auswählen, welches man mit dem olama run Befehl ausgewählt hat, also in unserem Fall llama2.

Das klingt zu stressig? Oder wäre besser als Hosted API? Nutzen Sie doku-chat.de oder lassen sie sich in einem kostenlosen Erstgespräch von uns beraten .

Tipp:

Um die Nutzung Ihrer Daten durch die amerikanischen AI Anbieter wie OpenAI zu vermeiden, sollten Sie ein „selbstgehostetes“ AI Modell betreiben.

Dazu benötigen Sie einen Server mit einer Nvidia GPU, und einem entsprechenden „torch“ setup. Für DIY empfiehlt sich z.B. das Open Source Projekt https://ollama.com.

Alternativ ist ein deutscher KI Provider nötig, wie z.B. https://doku-chat.de oder https://datafortress.cloud. Doku-Chat.de bietet außerdem die Unterteilung nach Projekten, nützliche Sharing/Team Features und viel mehr!

Sie haben noch offene Fragen zum Thema AI, Daten & co? Wir helfen Ihnen gerne!

- Consulting/Beratung: https://datafortress.cloud
- KI Anbieter: https://doku-chat.de/de/
- info@datafortress.cloud
- 01601136770