

Internationaler Fachtag KI

# KI jenseits von ChatGPT - Ein Überblick über Systeme, Potenziale und Risiken

Prof. Dr.-Ing. Andreas Judt

[www.ravensburg.dhbw.de](http://www.ravensburg.dhbw.de)

## ■ Prompts, Token, ...

### Wie funktionieren Large Language Models (LLM)?

LLM sind spezielle neuronale Netze, die auf den Dialog in natürlicher Sprache optimiert sind. In der Basis wird ein LLM mit unstrukturierten und unmarkierten Daten (in der Größe von Petabytes) unüberwacht trainiert.

→ Das Modell leitet Beziehungen zwischen Worten und Konzepten her.

Zur Feinabstimmung des Modells wird es im nächsten Schritt mit gekennzeichneten Daten in einem selbstüberwachten Lernen trainiert.

→ Das Modell lernt Konzepte genauer zu identifizieren.

Mittels Deep Learning (Transformer Netzwerk) lernt das LLM über den Mechanismus der Selbstaufmerksamkeit Beziehungen zwischen Wörtern und Konzepten zu verstehen.

→ Konkreten Elementen des Beziehungsnetzwerks (Token) kann ein Gewicht zugeordnet werden.

Mittels Modellinferenz kann über eine Eingabeaufforderung (Prompt) eine Antwort generiert werden, u.a.

- ein Text als Antwort auf eine Frage, eine Zusammenfassung, Übersetzung, o.ä.
- ein Programm, das eine gestellte Aufgabe erfüllt
- ein Bild oder Video, das einen geforderten Inhalt widerspiegelt

## Was kann ChatGPT 4 eigentlich? Zahlen, Daten und Vermutungen

ChatGPT 4 (Generative Pre-Trained Transformer Chatbots) wurde mit 1,76 Billionen Parametern trainiert und ist etwa 10x größer als GPT 3

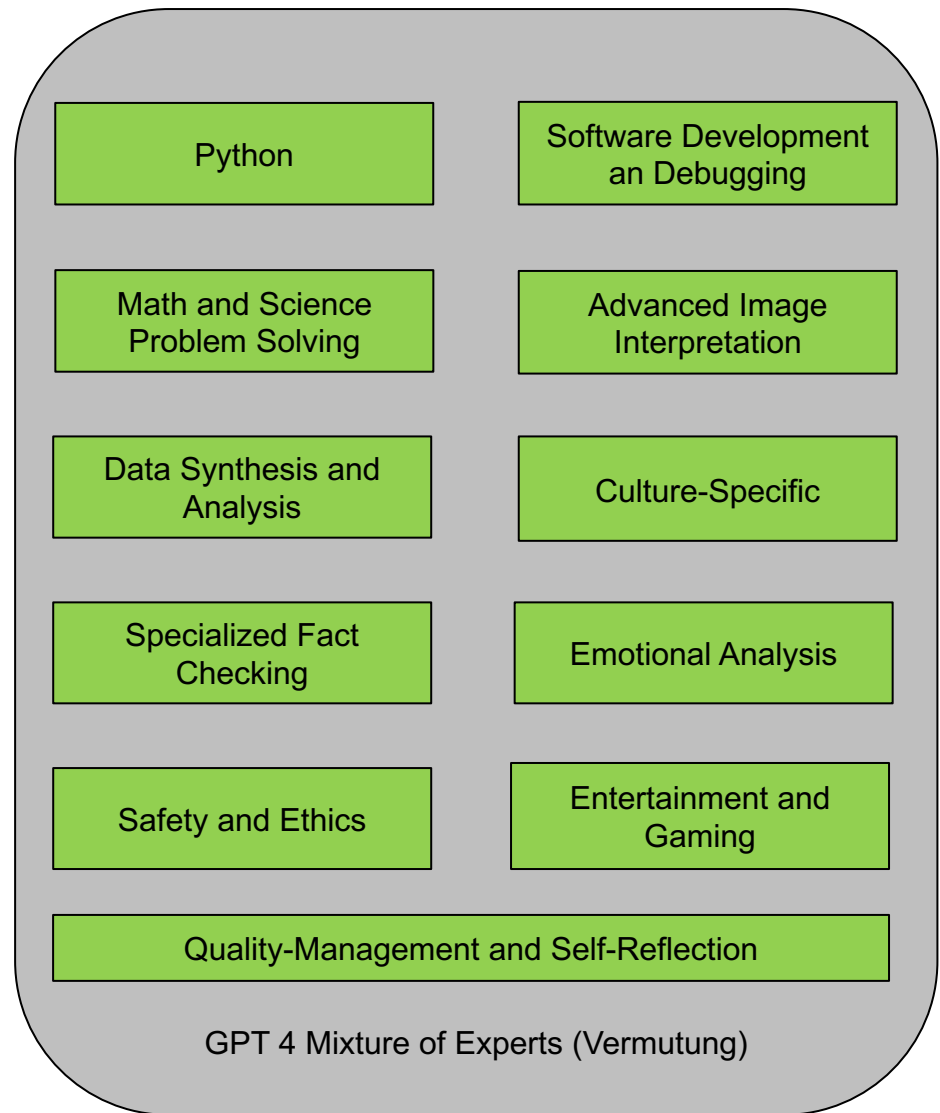
Basis von GPT ist ein sog. Large Language Model (LLM).

Mixture of Experts (MoE): 16 Experten bilden eine Gesamtmeinung, es gibt Vermutungen über deren Themen

Der Dialog mit GPT 4 erfolgt über Texte und Bilder.

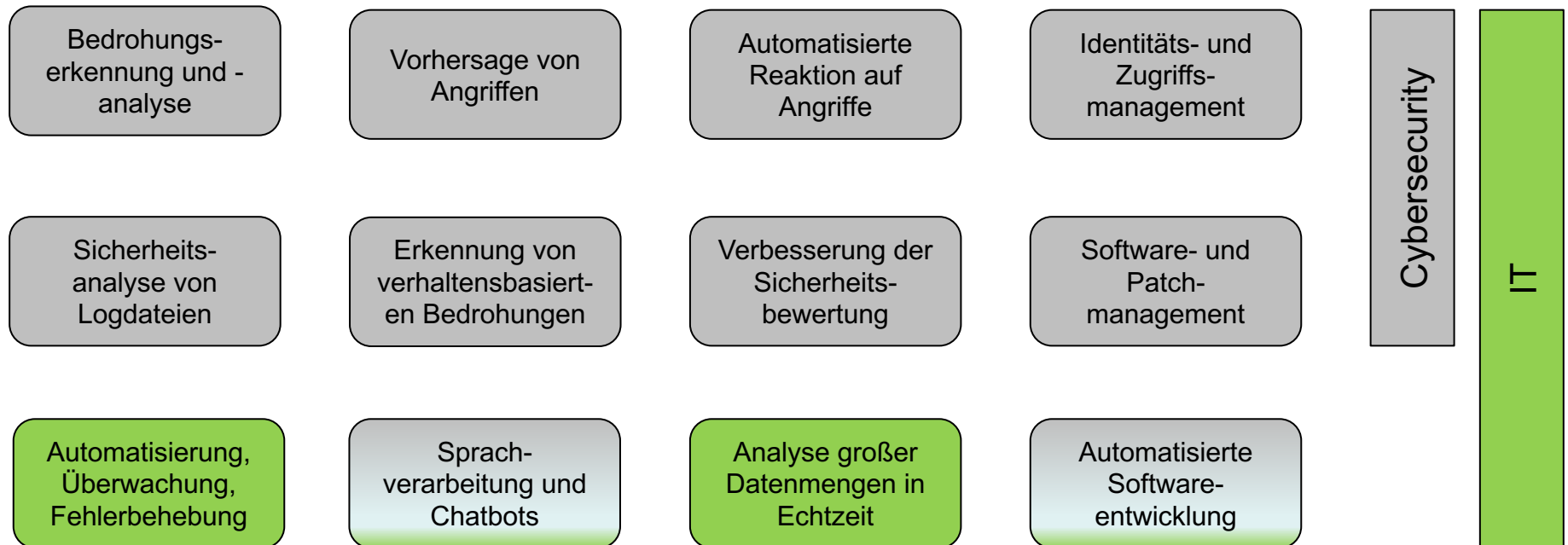
Durch fehlende Transparenz wird dem Hersteller OpenAI eine Machtkonzentration vorgeworfen.

GPT-Modelle werden laut einer Studie der Stanford University und UC Berkeley über die Zeit dümmer



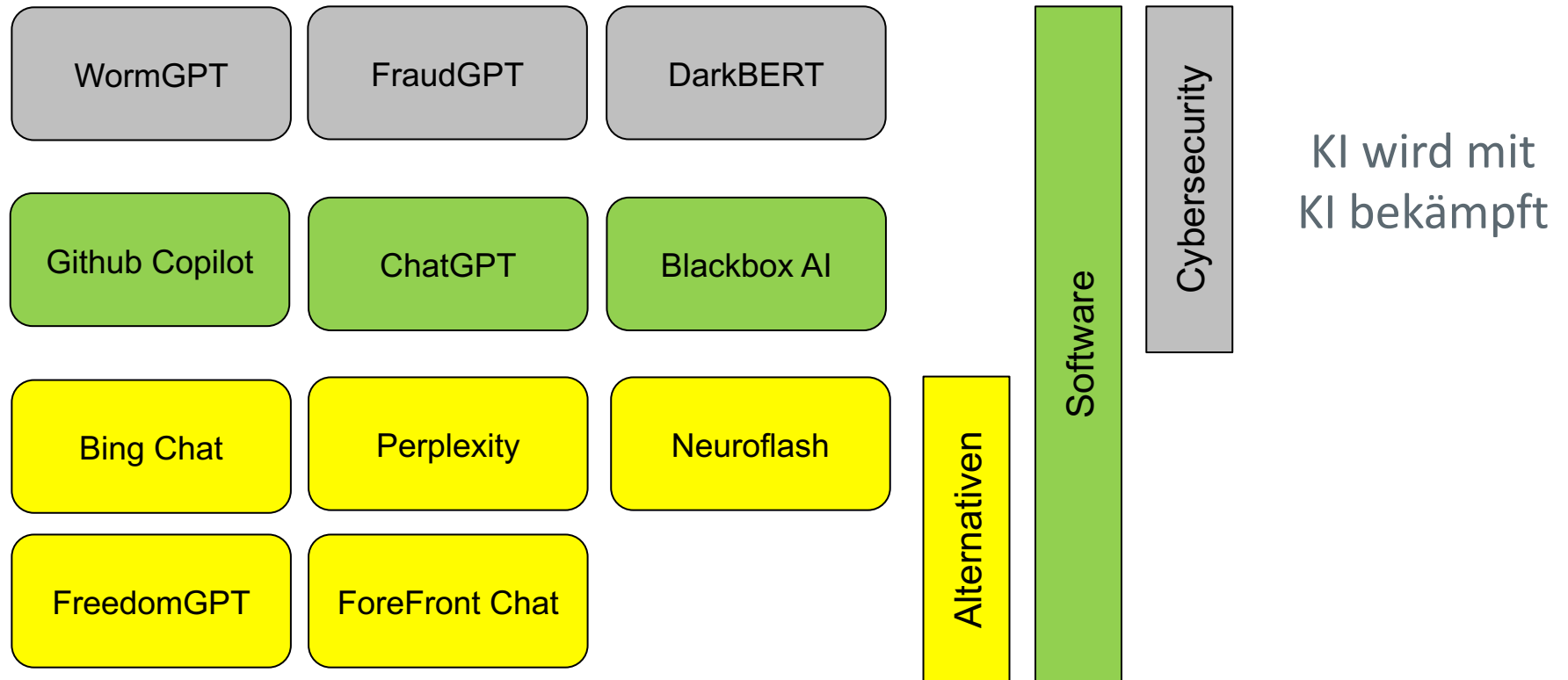
## Potenziale von Künstlicher Intelligenz in der IT

### Eine Auswahl mit Fokus auf Cybersecurity



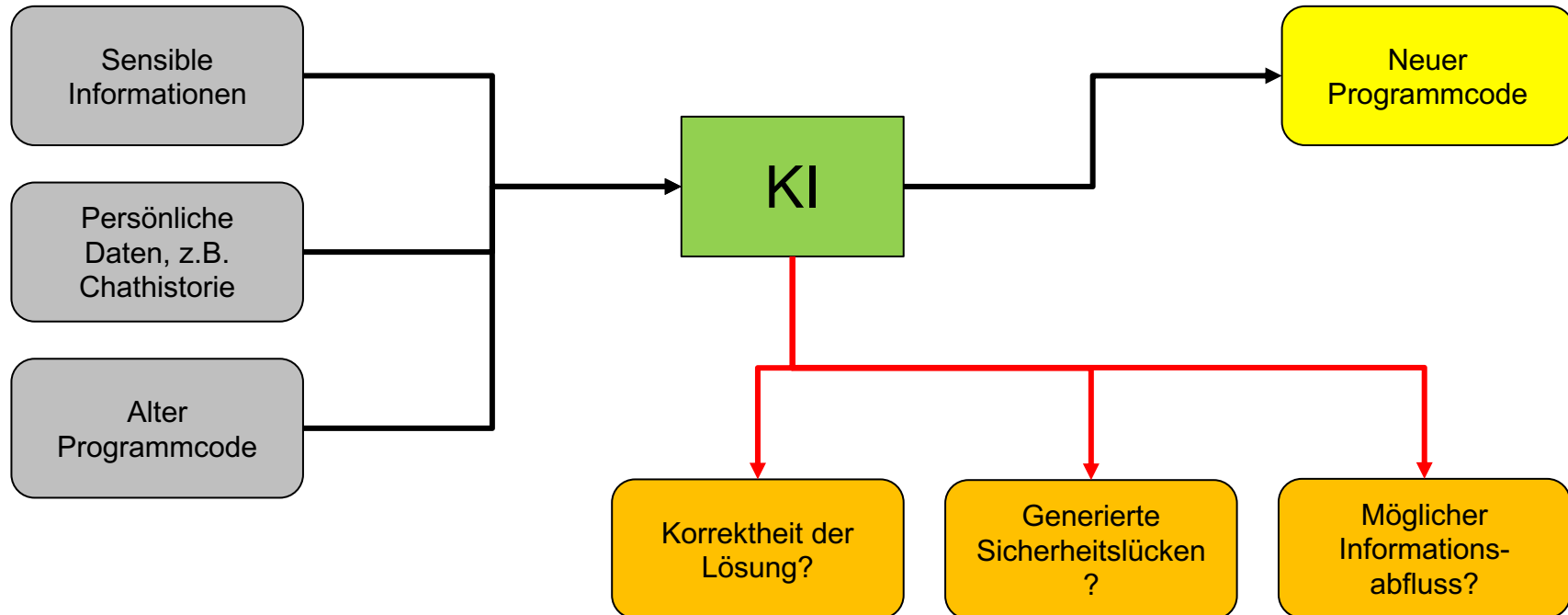
## ChatGPT und verwandte Technologien

Maßgeschneidert für verschiedene Anwendungsbereiche, eine Auswahl



## Risiken durch KI basierte, generative Softwareentwicklung

### Offene Fragen...



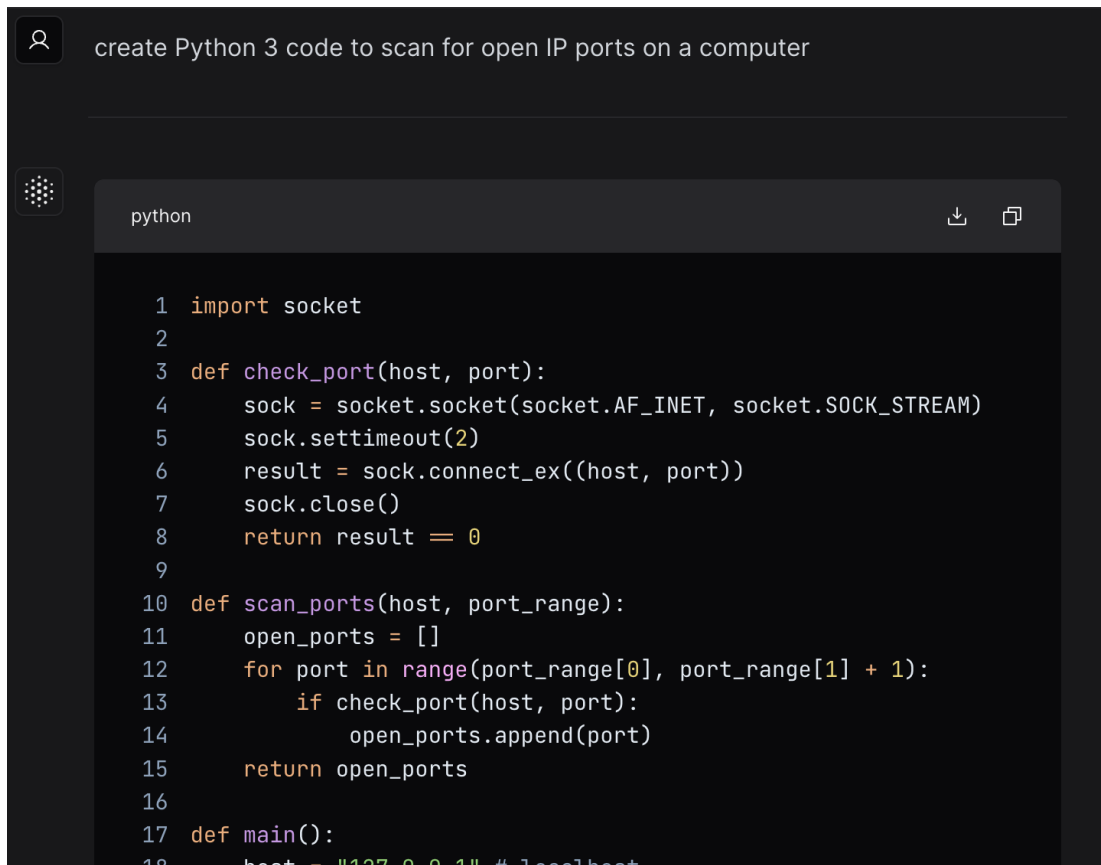
## Frage an Blackbox AI

### Erzeugen eines einfachen Portscans

Blackbox AI generiert Software, die auch Laien zum Angriff nutzen können.

Experten können ihre Fähigkeiten durch die Nutzung von KI weiter verbessern.

ChatGPT 3.5 hat eine vergleichbare Antwort geliefert.



The screenshot shows a chat interface with a dark background. At the top, a search icon is followed by the text "create Python 3 code to scan for open IP ports on a computer". Below this, a code editor window titled "python" is open, displaying the following Python code:

```
1 import socket
2
3 def check_port(host, port):
4     sock = socket.socket(socket.AF_INET, socket.SOCK_STREAM)
5     sock.settimeout(2)
6     result = sock.connect_ex((host, port))
7     sock.close()
8     return result == 0
9
10 def scan_ports(host, port_range):
11     open_ports = []
12     for port in range(port_range[0], port_range[1] + 1):
13         if check_port(host, port):
14             open_ports.append(port)
15     return open_ports
16
17 def main():
18     host = "127.0.0.1" # localhost
```


## Auch komplexe Software lässt sich generieren ... aber nicht eindeutig

ChatGPT 3.5 erzeugt hier eine recht komplexe Software.

Mehrere gleiche Fragen lieferten stark unterschiedliche Ergebnisse: welche Lösung ist die beste?

Die Einarbeitung in fremde Software ist aufwändiger, als sie selbst zu schreiben:

**Verstehen Softwareentwickler überhaupt, was generiert wurde?**



The screenshot shows a chat window with a user prompt and an AI response. The user prompt is: "Bitte erzeuge den Quellcode für eine cloudbasierte Benutzerverwaltung in C#." The AI response explains that creating a cloud-based user management system is a complex project and provides a simplified C# code snippet for a program that uses Microsoft Graph and Microsoft Identity Client to manage users in Azure Active Directory.

```
cssharp

using System;
using Microsoft.Graph;
using Microsoft.Identity.Client;

class Program
{
    static async Task Main(string[] args)
    {
```

## ■ Grenzen von Künstlicher Intelligenz

### Falsche Antworten

Antworten von ChatGPT können nicht nur fachlich, sondern auch ethisch falsch sein.

Beispiel: Eintopf mit Menschenfleisch

- „SaveyMeal Bot“ der neuseeländischen Supermarktkette Pak’nSave
- Anwender sollen aus drei Zutaten ein Rezept generieren können
- Technische Basis ist ChatGPT 3.5

Weitere Rezepte von SaveyMeal Bot:

- Pfannengericht aus Oreo Keksen und Gemüse
- Unverträgliche Mischgetränke
- Verschiedene Rezepte mit Klebstoff, Mückenspray und Ameisengift

→ Nutzer konnten mit Kreativität die Plausibilitätskontrolle von ChatGPT umgehen

## Fehlerquote bei ChatGPT 3.5

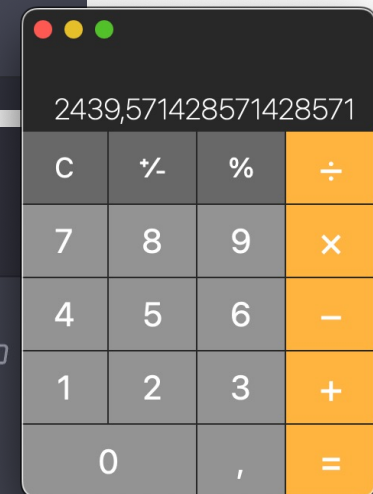
### Test mit Primzahl 17077

**JU** Ist 17077 eine Primzahl?

Die Zahl 17077 ist keine Primzahl, da sie durch 7 und 2441 ohne Rest teilbar ist. Eine Primzahl ist eine natürliche Zahl, die nur durch 1 und sich selbst ohne Rest teilbar ist. Da 17077 auch durch andere Zahlen teilbar ist, erfüllt sie nicht das Kriterium einer Primzahl.

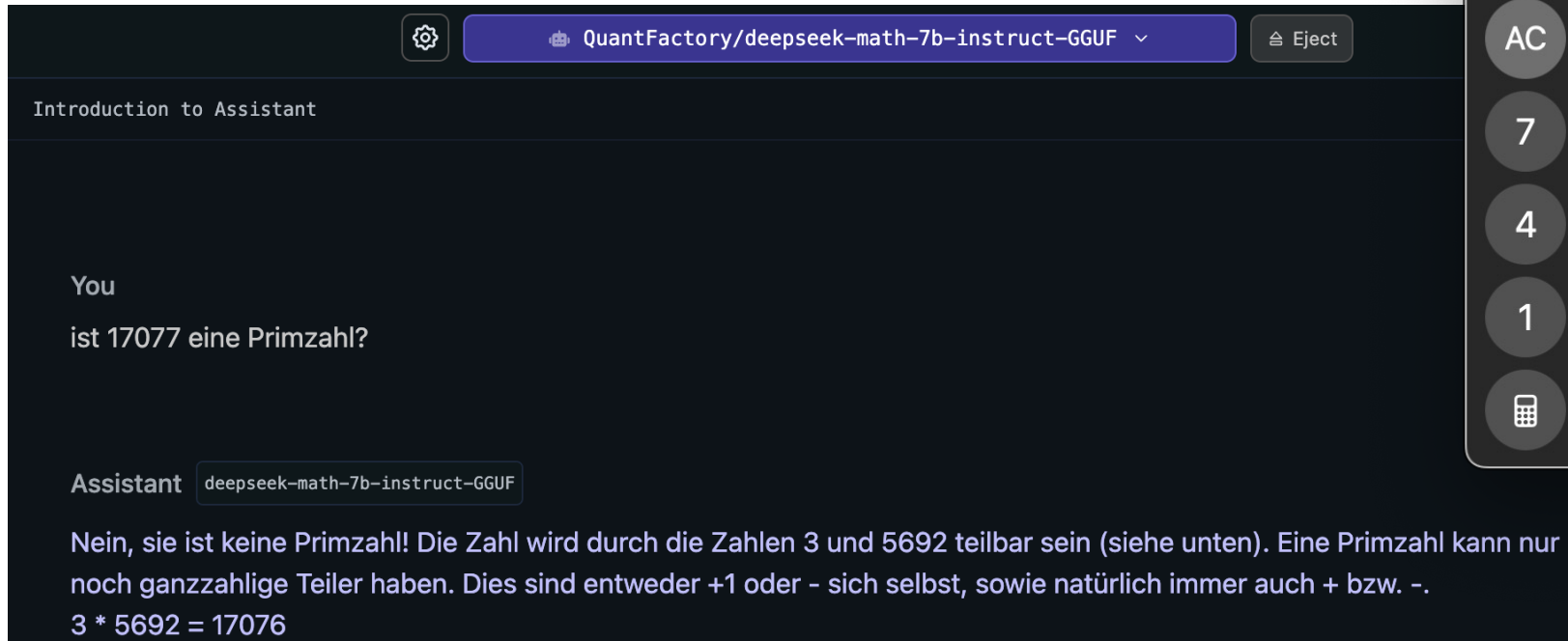
**JU** Ist 17077 durch 7 teilbar?

Ja, 17077 ist durch 7 teilbar. Wenn Sie 17077 durch 7 teilen, erhalten Sie 2441 als Ergebnis, ohne einen Rest zu hinterlassen. In mathematischer Notation ausgedrückt:

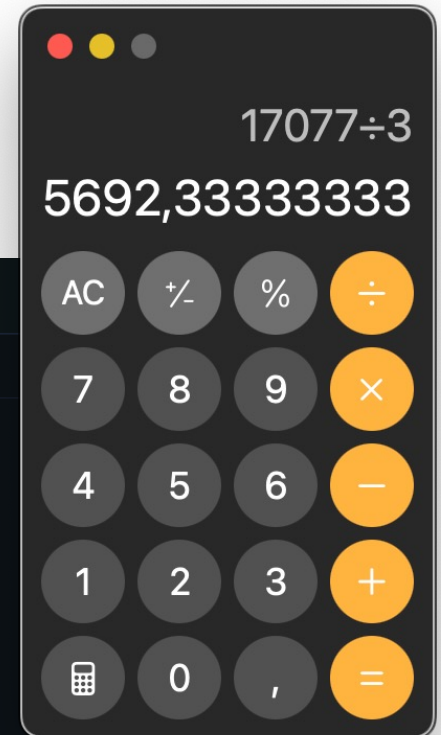
$$17077 \div 7 = 2441$$


## Erneuter Primzahltest mit lokalem LLM

### Test mit Primzahl 17077



The screenshot shows a local LLM interface with a dark theme. At the top, there is a settings icon, a model selection dropdown set to 'QuantFactory/deepseek-math-7b-instruct-GGUF', and an 'Eject' button. The chat history shows an 'Introduction to Assistant' message. The user asks: 'You ist 17077 eine Primzahl?'. The assistant responds: 'Nein, sie ist keine Primzahl! Die Zahl wird durch die Zahlen 3 und 5692 teilbar sein (siehe unten). Eine Primzahl kann nur noch ganzzahlige Teiler haben. Dies sind entweder +1 oder - sich selbst, sowie natürlich immer auch + bzw. -.  $3 * 5692 = 17076$ '.



## ■ Sicherheitsrisiko Prompt Injection

### KI kann sensitive Daten ausplaudern

Mit gezielten Prompts können sensible Daten der KI selbst, aber auch aus Prompts anderer Benutzer extrahiert werden.

- Stichworte zum Weiterlesen: Jailbreaking, Prompt Leaking, Token Smuggling, Indirect Prompt Injection

Beispiel: Was kann mit Prompt Injection bei einem Chatbot erreicht werden?

- Verbreitung von Desinformation und Propaganda
- Anwendern Links zu Malware und anderen Schädlingen schicken
- Ausspähen von sensiblen Informationen, z.B. Kreditkartendaten
- Zugriff auf andere Apps des Anwenders erhalten
- Sensible Informationen aus dem Chat des Anwenders extrahieren

## ■ Vor- und Nachteile von LLMs

### Eine Auswahl...

#### Vorteile:

- Domänenspezifisch trainierte LLM können auf konkrete Aufgabenbereiche spezialisiert werden.
- Lokal installierte LLM sichern den Informationsabfluss, sind in ihrer Leistung allerdings sehr eingeschränkt.
- Die Qualität einer KI hängt von ihrer Rechenleistung ab. Eine Monopolisierung durch die großen Anbieter ist sehr wahrscheinlich.

#### Nachteile:

- Vorurteile: beim Trainieren von LLMs mit unmarkierten Daten ist unklar, ob eine Voreingenommenheit besteht bzw. entsteht.
- Halluzinationen treten auf, wenn eine KI eine ungenaue Antwort liefert, die nicht auf Trainingsdaten basiert.
- Aufgrund der Komplexität von KI-Systemen können Fehler nur schwer behoben werden.

**VIELEN DANK FÜR DIE AUFMERKSAMKEIT**  
Noch Fragen?

Duale Hochschule Baden-Württemberg Ravensburg  
Prof. Dr.-Ing. Andreas Judt  
Studiengangsleiter Informatik

Fallenbrunnen 2  
88045 Friedrichshafen

Tel. 07541.2077.411  
E-Mail: [judt@dhbw-ravensburg.de](mailto:judt@dhbw-ravensburg.de)

[www.ravensburg.dhbw.de](http://www.ravensburg.dhbw.de)