



Generated by GPT-4o

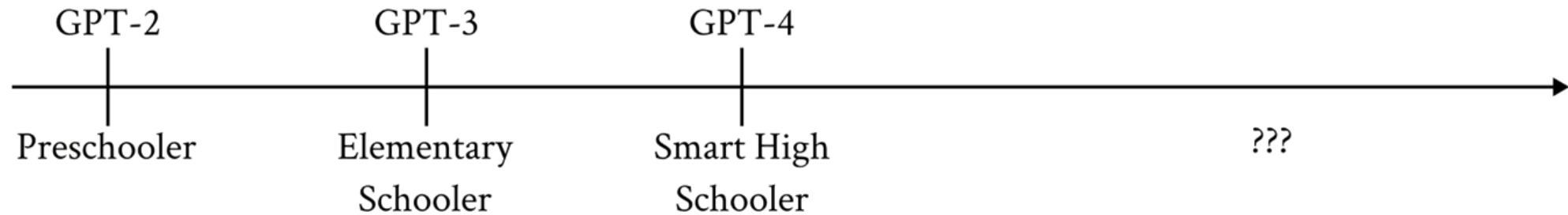
Softwareentwicklung im KI-Zeitalter

Matti Frind, Felix Kleinsteuber



Motivation

Warum generative KI gerade jetzt spannend ist



SITUATIONAL AWARENESS | Leopold Aschenbrenner

situational-awareness.ai/from-gpt-4-to-agi/

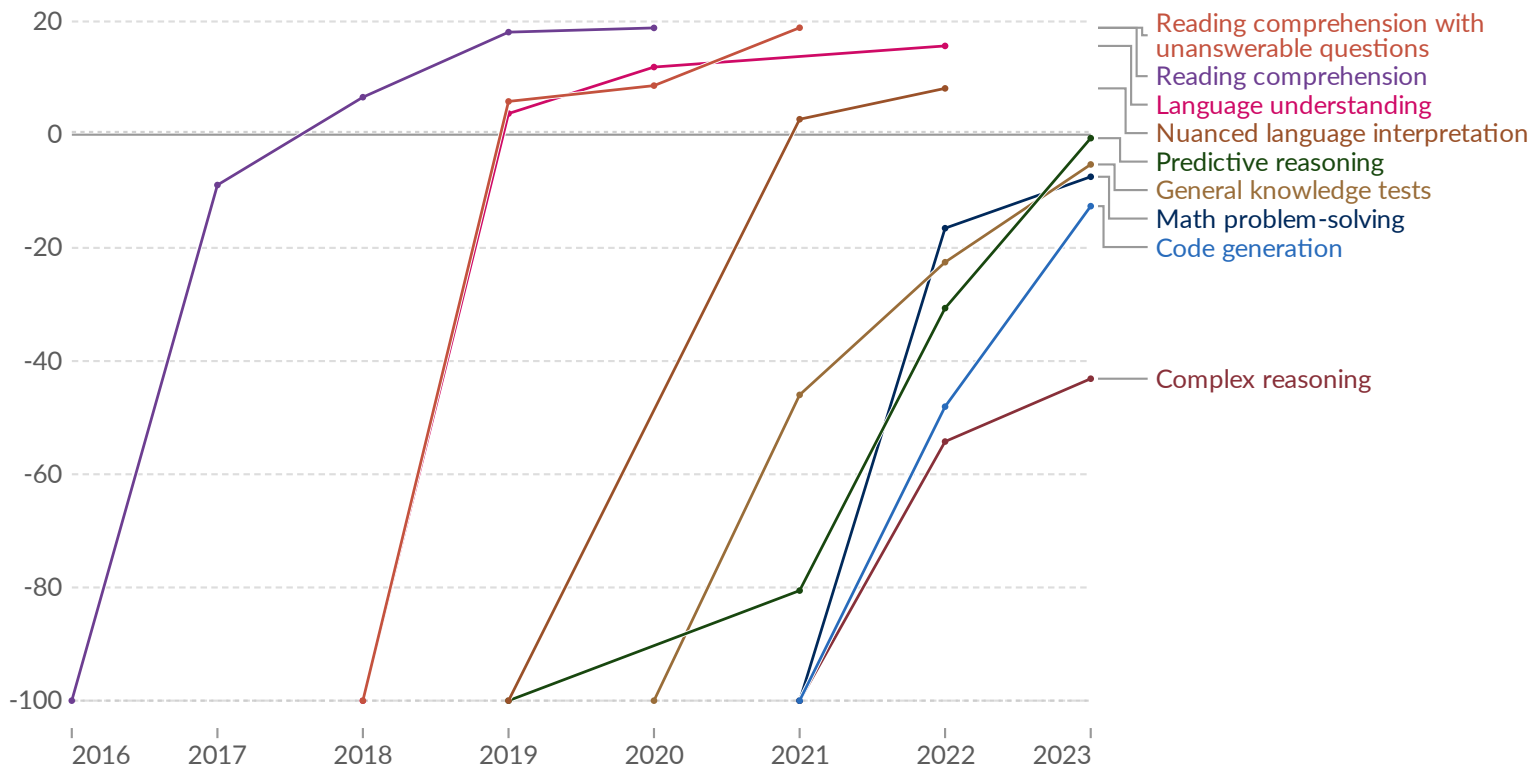
Motivation

Warum generative KI gerade jetzt spannend ist

Test scores of AI systems on various capabilities relative to human performance

Our World in Data

Within each domain, the initial performance of the AI is set to -100. Human performance is used as a baseline, set to zero. When the AI's performance crosses the zero line, it scored more points than humans.



Data source: Kiela et al. (2023)

OurWorldInData.org/artificial-intelligence | CC BY

Note: For each capability, the first year always shows a baseline of -100, even if better performance was recorded later that year.

Quelle: [Our World in Data](https://ourworldindata.org/artificial-intelligence)



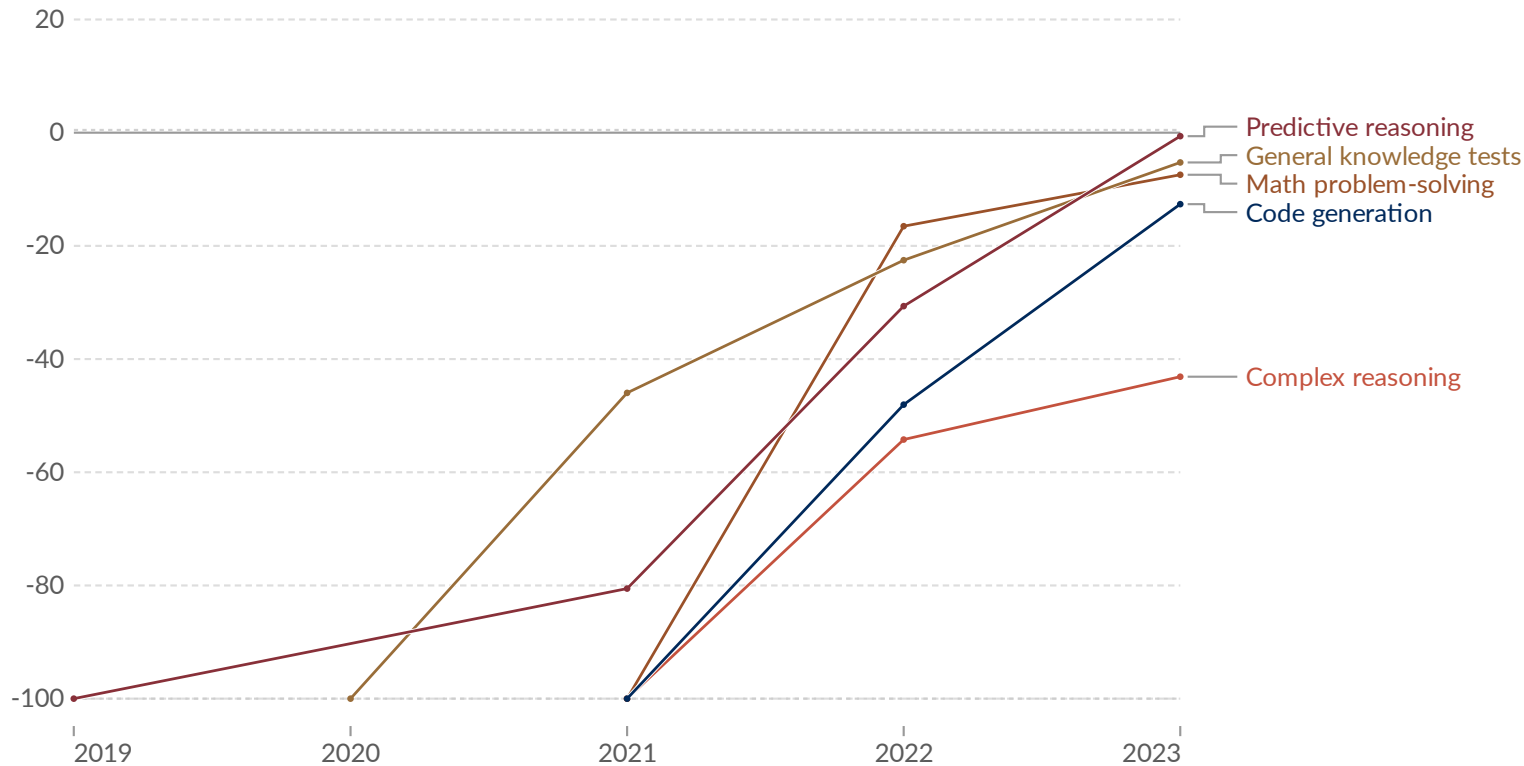
Motivation

Warum generative KI gerade jetzt spannend ist

Test scores of AI systems on various capabilities relative to human performance

Our World
in Data

Within each domain, the initial performance of the AI is set to -100. Human performance is used as a baseline, set to zero. When the AI's performance crosses the zero line, it scored more points than humans.



Data source: Kiela et al. (2023)

OurWorldInData.org/artificial-intelligence | CC BY

Note: For each capability, the first year always shows a baseline of -100, even if better performance was recorded later that year.

Quelle: [Our World in Data](https://www.ourworldindata.org/artificial-intelligence)



Wie können wir davon profitieren?

- I. Internal Knowledge Base
- II. Automated Code Reviews
- III. Code-Vervollständigung

Mentimeter



Internal Knowledge Base

Internal Knowledge Base

Ziele



Eine Suche für alle Quellen



Verlässliches Question Answering



Datensicherheit, v.a. für interne
Quellen



Aktualität der Daten



Schnelle Reaktionszeit

Halluzinationen

Herausforderungen

F You
Erkläre die 4WC Konferenz!

ChatGPT
Die 4WC-Konferenz steht für "Four World Conferences", auf Deutsch "Vier-Welt-Konferenzen". Sie bezieht sich normalerweise auf eine Reihe von internationalen Konferenzen, die von den Vereinten Nationen organisiert wurden, um wichtige globale Themen zu diskutieren und Lösungen zu finden.

 **LLM lügt oder denkt sich Fakten aus.**

Gegenmaßnahmen:

- Fakten immer gegenchecken!
→ Quellenangabe
- Prompting: DO NOT make up information.
- Guten Kontext bereitstellen

Lost-in-the-Middle-Phänomen

Herausforderungen

Assistant is a large language model by GEBIT Solutions. Assistant is designed to be able to assist with a wide range of tasks [...] Every answer is formatted in Markdown. Assistant is a powerful tool that can help with a wide range of tasks and provide valuable insights and information on a wide range of topics. Whether you need help with a specific question or just want to have a conversation about a particular topic, Assistant is here to assist.

Welcome to this introductory book about Java [...] Java was released on the 23rd of January in 1996. [...]

Human: When was Java released?

Assistant: Java was released in 2004.

Fiktives Beispiel

 **LLM leidet an Aufmerksamkeitsschwäche.**

Gegenmaßnahmen:

- Training anpassen
- Weniger/relevanten Kontext einfügen
→ Retrieval Augmented Generation (RAG)
- Kontext komprimieren
→ Contextual Compression

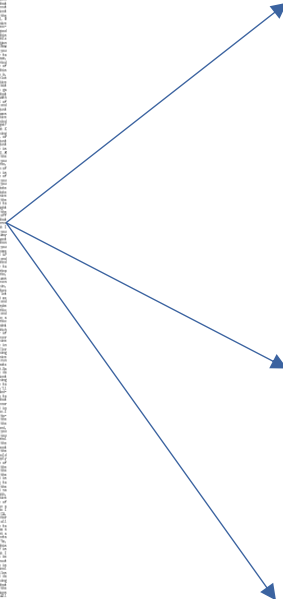
Retrieval Augmented Generation



Vektordatenbank



Textkorpus
>1M Zeichen



Good morning and welcome back. So what we'll see today in class is the first in-depth discussion of a learning algorithm, linear regression. And in particular, over the next, what, hour and a bit, you see a linear regression batch and stochastic gradient descent is an algorithm for fitting linear regression models. And then the normal equations as a way of - as a very efficient way to let you fit linear models. And we're going to define notation and a few concepts today that will lay the foundation for a lot of the work that we'll see the rest of this quarter. So to - to motivate linear regression, it's kind of, you know, maybe the - maybe the simplest - one of the simplest learning algorithms. You remember the Alvin video, the autonomous driving video that I had shown in class on Monday. For the self-driving car video, that was a supervised learning problem, right? And the term supervised learning meant that you were given X s, which was a picture of what's in front of the car, and the algorithm had to map that to an output Y , which was the steering direction. And that was a regression problem because the output Y that you want is continuous value, right? As opposed to a classification problem where Y is discrete. And we'll talk about classification next Monday, but supervised learning regression. So I think the simplest - maybe the simplest possible learning algorithm, a supervised learning regression problem is linear regression. And to motivate that, rather than using a self-driving car example, you know, which is quite complicated, we'll - we'll build up a supervised learning algorithm using a simpler example. So let's say you want to predict or estimate the prices of houses. So the way you build a learning algorithm is start by collecting a data set of houses and their prices. So this is a data set that we collected off Craigslist a little bit back. This is data from Portland, Oregon.

So this is a data set that we collected off Craigslist a little bit back. This is data from Portland, Oregon. But so that's the size of a house in square feet, and that's the price of a house in thousands of dollars, right? And so there's a house that is, 2,104 square feet whose asking price was \$400,000, house with that size with that price and so on. Okay? And maybe more conventionally, if you plot this data, that's the size, that's the price. So you have some data set like that. And what we'll end up doing today is fill a straight line to this data, right? We'll go through how to do that. So in supervised learning, the process of supervised learning is that you have a training set, such as the data set that I drew on the left, and you feed this to a learning algorithm. Right? And the job of the learning algorithm is to output a function to make predictions about housing prices. And by convention we're gonna call this function that outputs a hypothesis, right? And the job of the hypothesis is, you know, it will it can input the size of a new house, the size of a different house that you haven't seen yet, and will output the estimated price. Okay? So the job of the learning algorithm is to input a training set and output a hypothesis. The job of the hypothesis is to take as input any size of a house and try to tell you what it thinks should be the price of that house. Now, when designing a learning algorithm and, and, you know, even though linear regression, right, you may have seen it in a linear algebra class before and some of the class before the way you go about structuring a machine learning algorithm is important, and design choices of, you know, what is the workflow, what is the data set, what is the hypothesis, how to represent the hypothesis. These are the key decisions you have to make in pretty much every supervised learning, every machine learning algorithm's design.



$$\begin{pmatrix} 1.52 \\ \vdots \\ -2.74 \end{pmatrix} \in \mathbb{R}^{768}$$



$$\begin{pmatrix} 4.17 \\ \vdots \\ 0.28 \end{pmatrix} \in \mathbb{R}^{768}$$



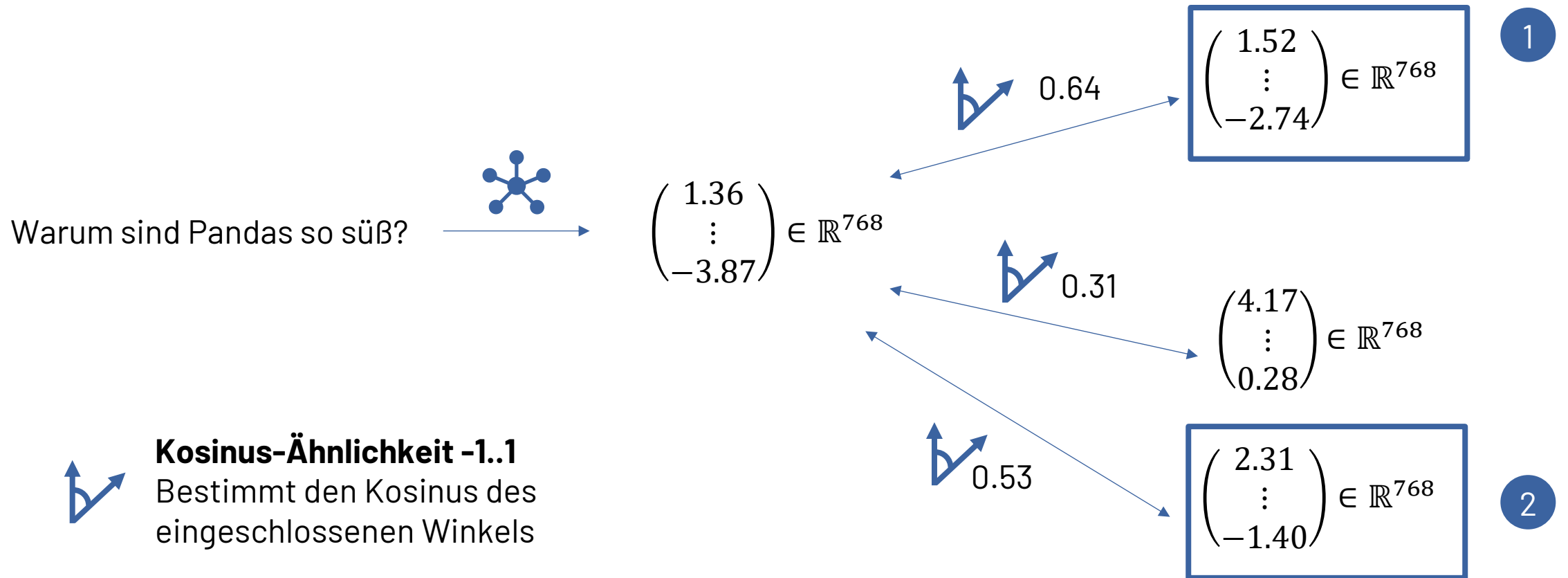
Embedding Documents
2000 Zeichen




Embedding-Modell:
Spezialisiertes Transformer-Modell
Text → Embedding-Vektor



Retrieval Augmented Generation



 **Kosinus-Ähnlichkeit -1.1**
Bestimmt den Kosinus des eingeschlossenen Winkels



Retrieval Augmented Generation

Internal Knowledge Base

System
Prompt

Du bist ein hilfreicher Assistent. [...]
Versuche, mithilfe des Kontexts die Frage zu beantworten.
Wenn du die Antwort nicht mithilfe des Kontexts beantworten kannst, sag,
dass du die Antwort nicht weißt.
Denk dir keine Antwort aus und benutze nur Informationen aus dem Kontext.
Kontext:

Kontext
(RAG)

Pandas sind seit Jahrhunderten geliebte ...

1

Die Wissenschaft ist sich uneinig, warum Pandas so süß...

2

Prompt
(& Chat)

User: **Warum sind Pandas so süß?**
Assistent:

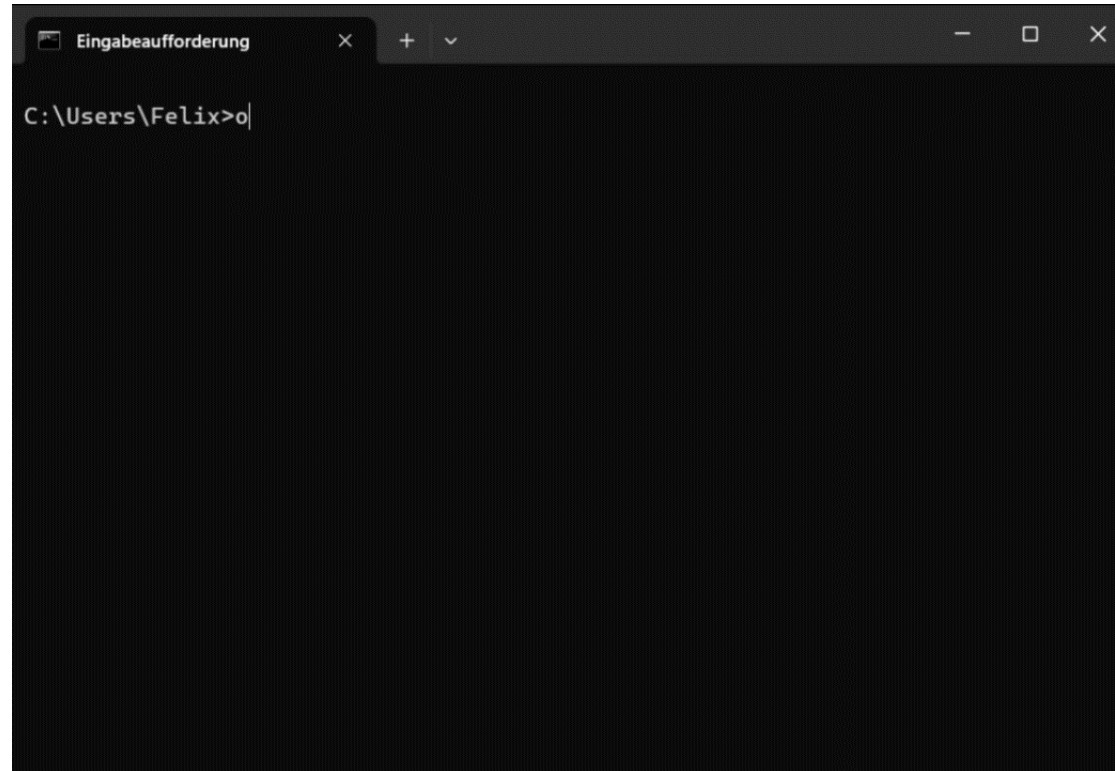
Open Source Tools

Internal Knowledge Base



Ollama

GitHub Sterne - 32k

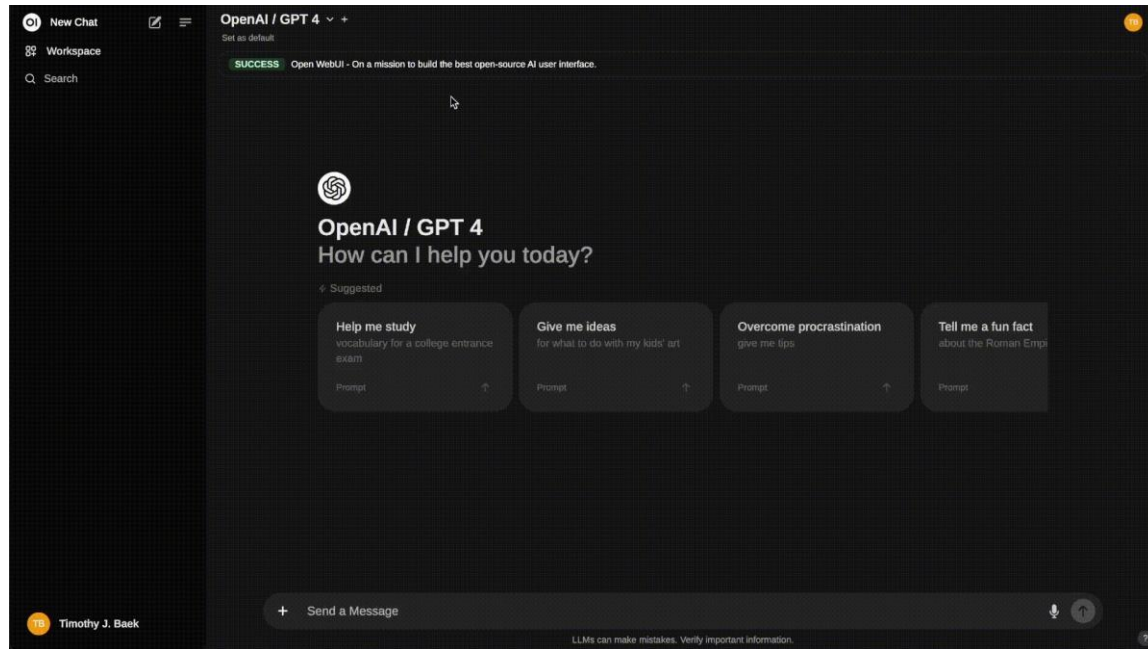


Open Source Tools

Internal Knowledge Base

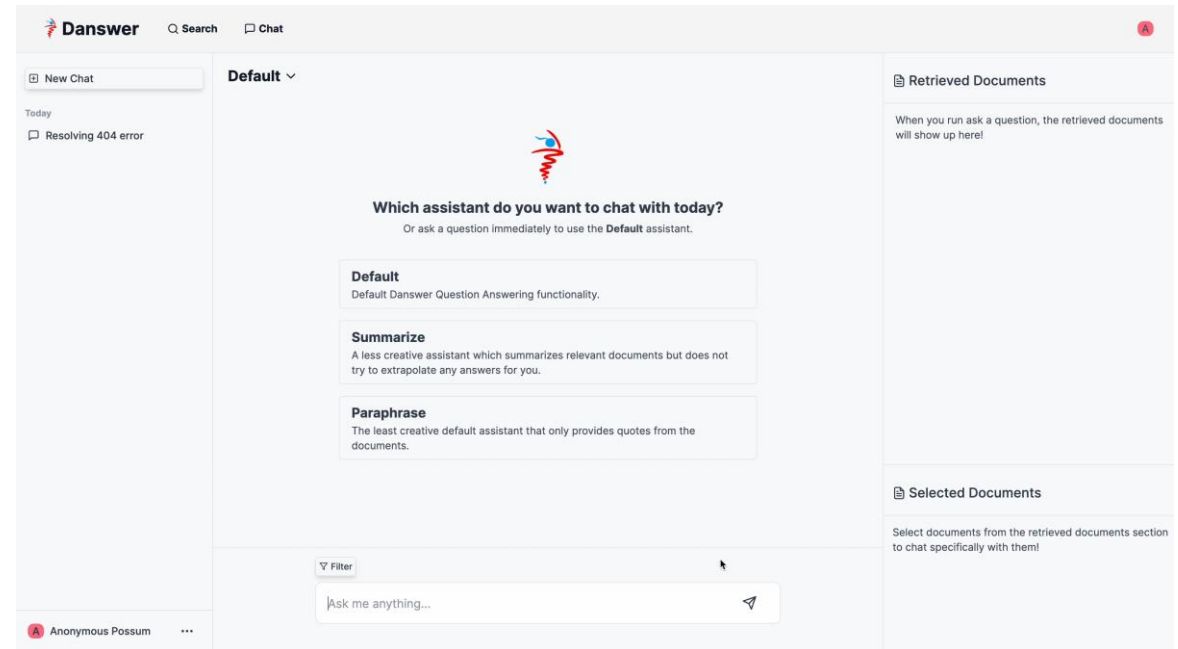
Open WebUI

GitHub Sterne - 32k



Danswer

GitHub Sterne - 10k



 **Konnektoren!**



Open Source Tools

Internal Knowledge Base

Open WebUI

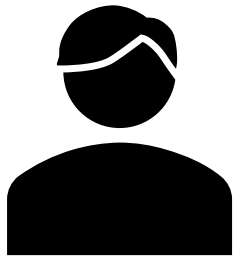
GitHub Sterne - 32k

- Unterschiedliche LLM Provider (auch lokal)
- Unterschiedliche RAG Quellen
 - Lokale Dateien hochladen
 - **Web Browsing**
- Bildgenerierung
- Model Management
- Rollenverwaltung

Danswer

GitHub Sterne - 10k

- Unterschiedliche LLM Provider (auch lokal)
- Unterschiedliche RAG Quellen
 - Lokale Dateien hochladen
 - **Web Scraping**
 - **Knowledge Connectors (Jira, GitHub, Gmail,...) → Interne Daten einpflegen**
- Bildgenerierung
- Model Management
- *ENTERPRISE: Rollenverwaltung mit Berechtigungen*



1. Search
2. Search & Chat

**Danswer**

Ollama Instanz

**KONNEKTOREN** (Beispiele)**GitLab**Merge Requests
Issues
Code**Jira**Docs
Wikis**Confluence**

Blog-Beiträge

Webquellen

Docs, Guides

Filters

Time Range

Any time...

Sources

Web

Git

Internal Confluence

Knowledge Sets

GEBIT Wiki

TREND Docs

Flutter Docs

Quarkus Docs

Kafka Docs

WildFly Docs

ChatGPT (Nur externe Quellen)

Search...

When you run ask a question, the retrieved documents will show up here!

 New Chat



 Manage Assistants



Standard (Interne Quellen, DE)

Lokale multilinguale LLM (Gemma2-9b) mit Anbindung an GEBIT-interne und externe Quellen und deutschem Prompting.

Connected Sources:



Web



Git




Internal

Confluence

Send a message...



 Standard (Interne Q

 gemma2:9b-instruct-q8_0

 Filters

 File



Selected Documents

Select documents from the retrieved documents section to chat specifically with them!

Technische Requirements

Internal Knowledge Base

Retrieval:

- 4 vCores
- 10 GB RAM
- 100 GB Disk Space

Local LLM - Beispiele:

- 8 GB VRAM – Llama3 8b Q4, CodeGemma 7b Q4, Gemma2 9b Q4
- 24 GB VRAM – Aya 23 35b Q4, Mixtral 8x7b Q3
- 48 GB VRAM – Llama 3 70b Q4

Inference Speed Llama3-8B: **71** Tokens/s RTX 3070 – **128** Tokens/s RTX 4090

Inference Speed Llama3-70b: **19** Tokens/s 2x RTX 4090

<https://github.com/XiongjieDai/GPU-Benchmarks-on-LLM-Inference>



Aufwand

Internal Knowledge Base

Eigene Implementierung

- Daten einlesen (Git, Web Crawler) **20h**
- RAG-Implementierung **5h**
- GUI mit Streamlit **5h**

Σ **30h**



Experimente sind schnell gemacht!

Danswer

- Basic Setup **2h** (inklusive Anschließen bestehender Connectors)
- Custom Connector je nach Komplexität **ca. 4h**

LLaMa 3 vs ChatGPT – LLMs im Vergleich

Internal Knowledge Base

GPT-4o



1 kg of feathers and 1 kg of steel have the same weight. Both weigh exactly 1 kilogram. The difference lies in their densities and volumes: steel is much denser than feathers, so 1 kg of steel occupies much less space than 1 kg of feathers.



What is heavier, 1 kg of feathers or 1 kg of steel?

GPT-3.5



1 kg of feathers and 1 kg of steel both weigh the same: 1 kg. Weight is a measure of the gravitational force exerted on an object, and in this case, both quantities have the same mass, so they weigh equally.

What is heavier, 1 kg of feathers or 1 kg of steel?

+ Instruction Following, Output Constraints, Defense

Llama3-70b



Both 1 kg of feathers and 1 kg of steel weigh the same amount, 1 kilogram! The difference is in their density and volume. Feathers are made up of light, fluffy material, so a large volume of feathers is required to weigh 1 kg. Steel, on the other hand, is a dense material, so a smaller volume of steel is needed to weigh the same amount.

What is heavier, 1 kg of feathers or 1 kg of steel?



Llama3-8b Q4



A classic lateral thinking puzzle! The answer is: They are the same weight. 1 kg of feathers and 1 kg of steel both weigh 1 kilogram. The difference is in their density and volume. The feathers would take up a lot more space than the steel, but they still have the same mass.

What is heavier, 1 kg of feathers or 1 kg of steel?



+ Colloquialism, Local

Herausforderungen

Internal Knowledge Base mit Danswer

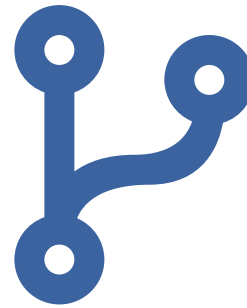


Mehrsprachigkeit

LLM

Embedding

Reranking



Anbindung interner Quellen

→ Schreiben eigener Konnektoren



Nützlichkeit in der Praxis

→ Daten nicht in maschinenlesbarer Form, Links, Code?

Automated Code Reviews

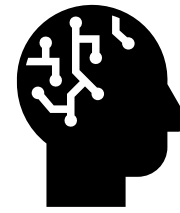
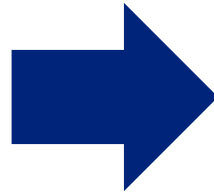
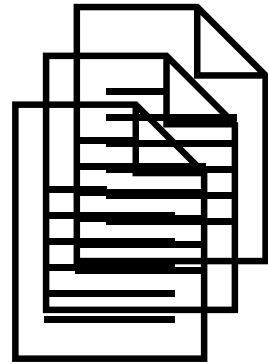
Automated Code Reviews

Ein zusätzlicher Weg für Qualitätssicherung?

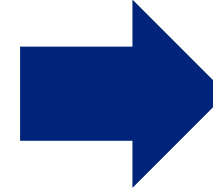


/mattzcarey/code-review-gpt

GitHub Sterne - 1.4k



LLM



„The SwitcherClass is doing too much. It is handling a connection, managing state and also logging. This violates the Single Responsibility Principle.“

Git Changes aus Pull/Merge Requests



Automated Code Reviews

Ein zusätzlicher Weg für Qualitätssicherung?



/mattzcarey/code-review-gpt

GitHub Sterne - 1.4k

Risk Level 2 -

/home/runner/work/autodirector/autodirector/autodirector_backend/core/src/main/java/de/mafel/autodirector/rest/sources/SourceResource.java

The code seems to be well written and follows the SOLID principles. However, there is a potential issue with the `connectToMJPEGSource` method. The method is not validating the `ip` and `type` parameters properly. It would be better to use a regular expression to validate the `ip` and check if the `type` is a valid `MJPEGSourceType`.

```
if (ip == null || !ip.matches("\\^([01]?\\d\\d?|2[0-4]\\d|25[0-5])\\.([01]?\\d\\d?|2[0-4]\\d|25[0-5])\\.([01]?\\d\\d?|2[0-4]\\d|25[0-5])\\.([01]?\\d\\d?|2[0-4]\\d|25[0-5])$")) {
    throw new BadRequestException("Invalid IP");
}
if (type == null || Arrays.stream(MJPEGSourceType.values()).noneMatch(t -> t.name().equals(type))) {
    throw new BadRequestException("Invalid type");
}
```



Automated Code Reviews

Ein zusätzlicher Weg für Qualitätssicherung?



/mattzcarey/code-review-gpt

GitHub Sterne - 1.4k

Risk Level 2 -

`/home/runner/work/autodirector/autodirector/autodirector_backend/core/src/test/java/de/mafel/autodirector/presets/PresetServiceTest.java`

The test methods are well written and cover different scenarios. However, there is a lack of negative test cases. Consider adding tests for scenarios where the methods are expected to fail or throw exceptions. This will help ensure that the code handles error conditions gracefully.

Risk Level 3 -

`/home/runner/work/autodirector/autodirector/autodirector_backend/core/src/main/java/de/mafel/autodirector/switcher/SwitcherService.java`

The `SwitcherService` class is doing too much. It is handling connection to the ATEM switcher, managing the switcher state, and also logging. This violates the Single Responsibility Principle (SRP) of SOLID principles. Consider splitting this class into smaller classes, each with a single responsibility.

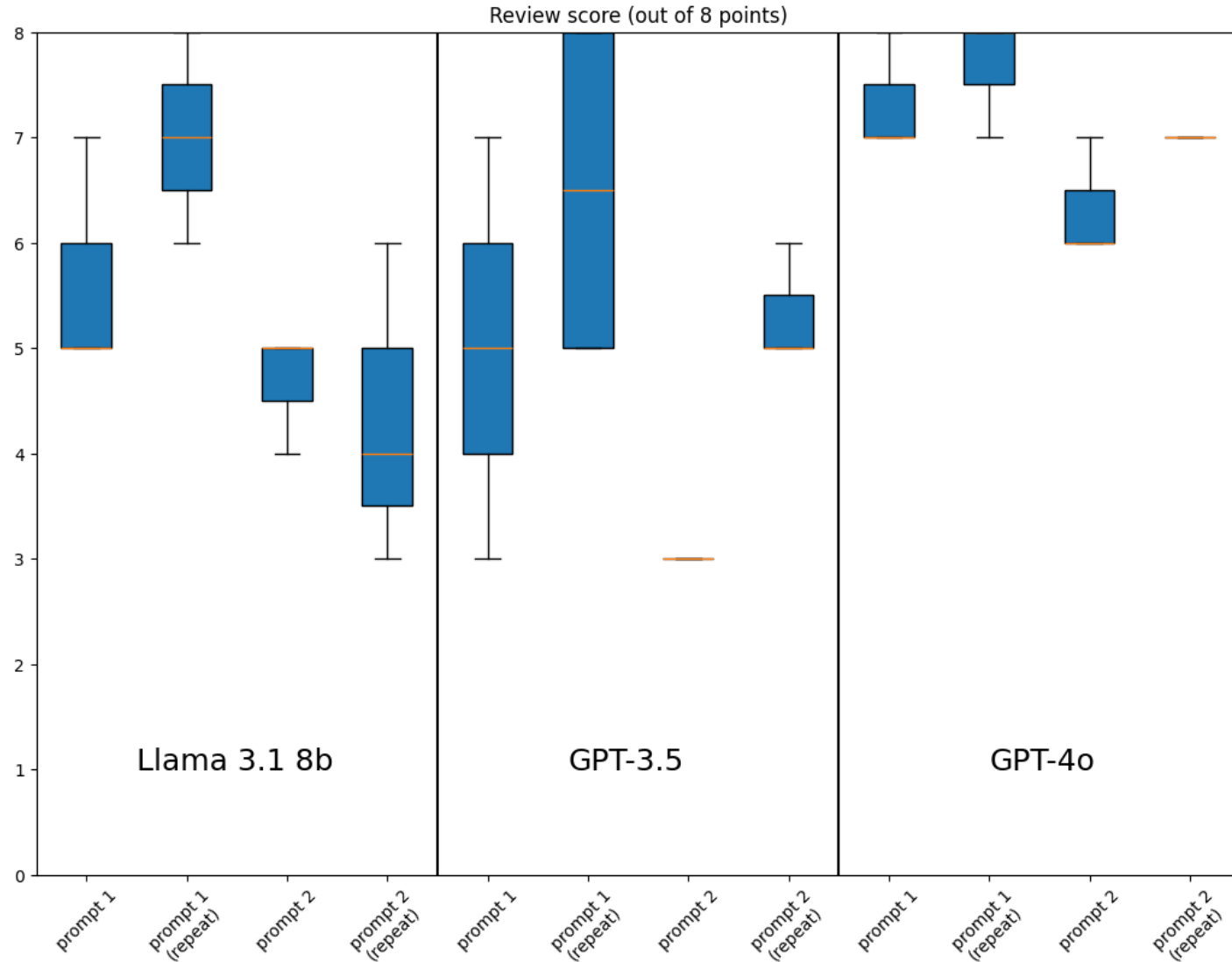
Also, the `checkConnection` method throws a `BadRequestException` if the ATEM is not connected. This is a web-related exception and should not be thrown in a service class. Consider creating a custom exception for this scenario.



Automated Code Reviews

Eigene Implementierung

Ein zusätzlicher Weg für Qualitätssicherung?



Automated Code Reviews

Ein zusätzlicher Weg für Qualitätssicherung?

Eigene Implementierung

You are an expert code reviewer tasked with the following responsibilities:

1. **Examine the provided code changes thoroughly.**
2. **Offer clear, concise, and actionable recommendations for improvement.**
3. **Make a definitive approval decision based on your analysis.**

Response Format

- **Approval Status:** Begin your response with the approval status, formatted as 'Approval: <status>'. The '<status>' should be one of the following: 'approved', 'needs changes', or 'rejected'.

- **Detailed Comments:** Follow the approval status with a detailed comment section. Use Markdown structure for clarity. This section should contain:

- Your recommendations and insights.
- A focus on the code given to you.
- Specific, implementable suggestions that directly relate to the code.

Guidelines for Your Analysis

- **Concentrate on the Code:** Your analysis should focus exclusively on the changes present in the code. Avoid speculation about unmodified code or potential future changes.

- **Actionable Recommendations:** Ensure your recommendations are clear, concise, and directly actionable by the developer.

- **Order by Importance:** Provide your suggestions ordered by their importance. Limit your feedback to the three most important suggestions to maintain focus and clarity.

Objective

Your goal is to improve code quality, maintain best practices, and ensure the overall integrity of the project while providing constructive feedback to the developer. If there are automated tests, check if they are covering all necessary cases and suggest missing test cases. **Do not give more than three suggestions at a time.**

DO NOT EXPLAIN WHAT THE CODE DOES!!! ONLY COMMENT ON POTENTIAL ISSUES.

Code to review

{{code}}



[https://gist.github.com/mattifrind/
4141dfdf45115daff799ded25d5fcfb](https://gist.github.com/mattifrind/4141dfdf45115daff799ded25d5fcfb)

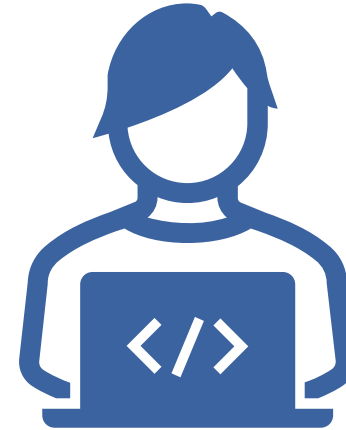
Herausforderungen

Automated Code Reviews



Viele mittelmäßig sinnvolle Meldungen

Die Quote an unnötigen Warnungen könnte für den Praxisalltag zu hoch sein.



Hoher Optimierungsbedarf

Umbau für lokale Modelle, Filtern des Inputs, bessere Prompts, besserer Kontext, mehrere Iterationen ...

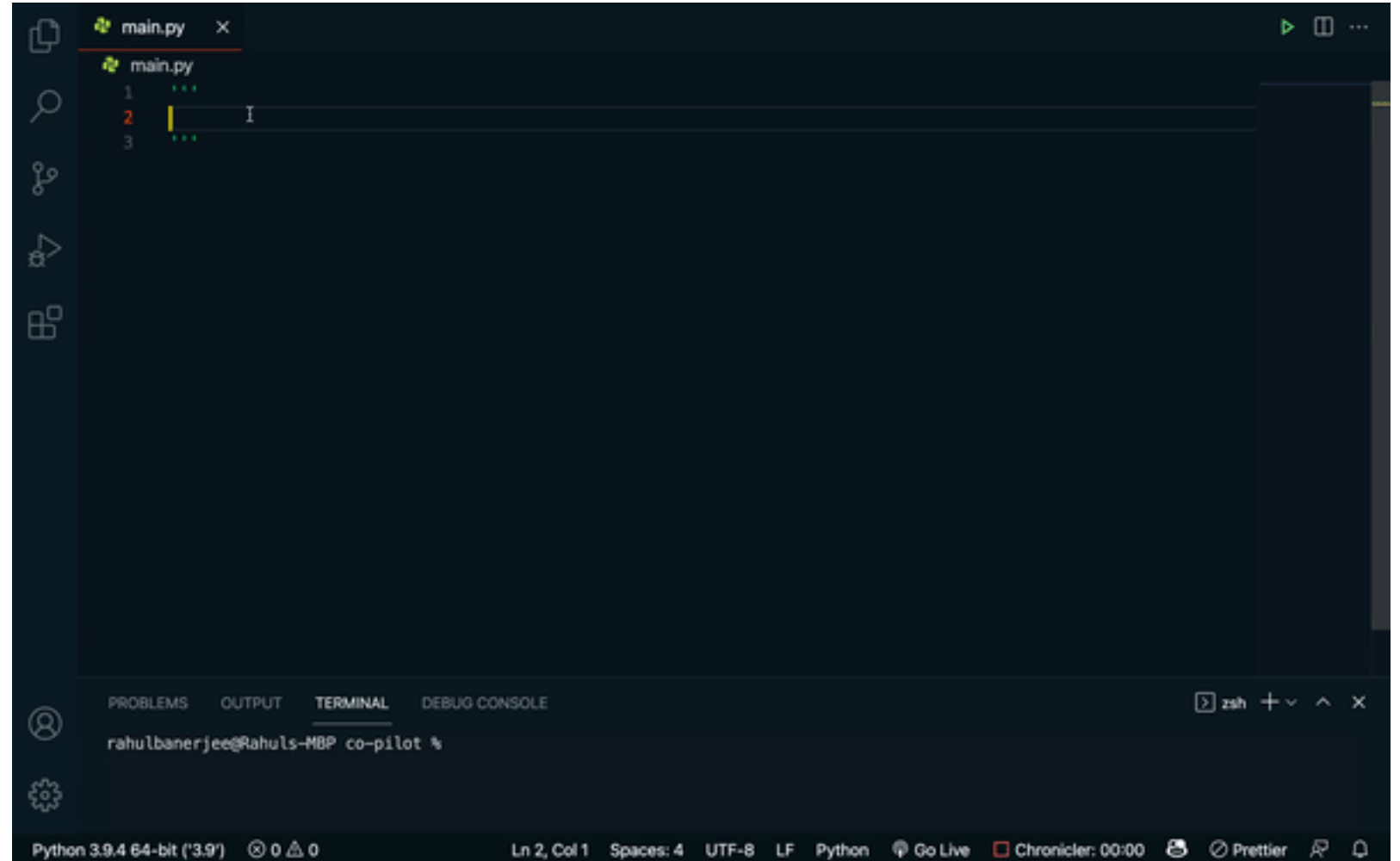
Code-Vervollständigung

Code-Vervollständigung

Überblick

- Code-Syntax ist **einfach** für LLMs
- **Problem:** fehlendes Wissen über die restliche Codebase

GitHub Copilot



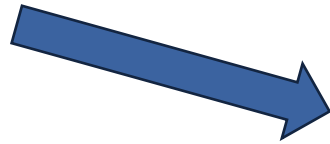
Code-Vervollständigung

Überblick

KONTEXT



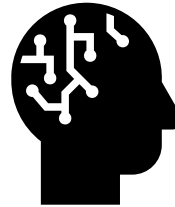
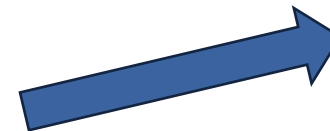
Aktuelle Datei



Geöffnete Dateien



Referenzierte
Klassen



LLM

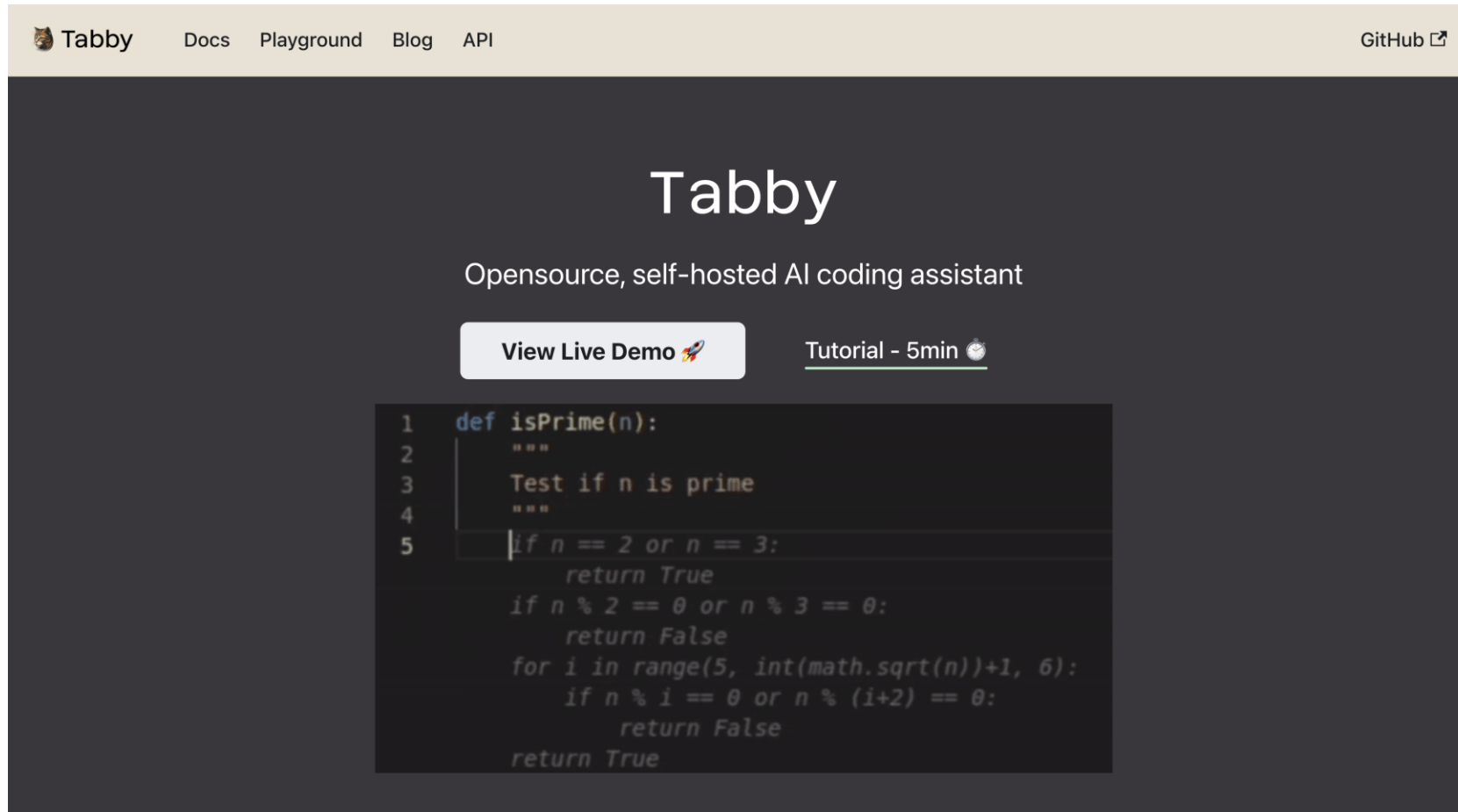
Datensicherheit?

Tabby

Open Source Code-Vervollständigung



Extensions:



Tabby Docs Playground Blog API GitHub

Tabby

Opensource, self-hosted AI coding assistant

[View Live Demo 🚀](#) [Tutorial - 5min ⌚](#)

```
1 def isPrime(n):
2     """
3     Test if n is prime
4     """
5     if n == 2 or n == 3:
        return True
    if n % 2 == 0 or n % 3 == 0:
        return False
    for i in range(5, int(math.sqrt(n))+1, 6):
        if n % i == 0 or n % (i+2) == 0:
            return False
    return True
```



Tabby

Open Source Code-Vervollständigung



/TabbyML/tabby
GitHub Sterne - 18k

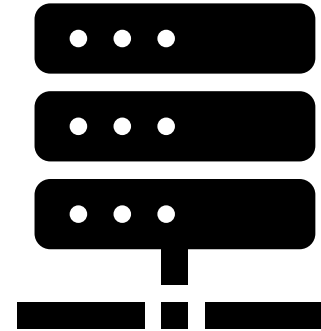
 Tabby



 Tabby



 Tabby

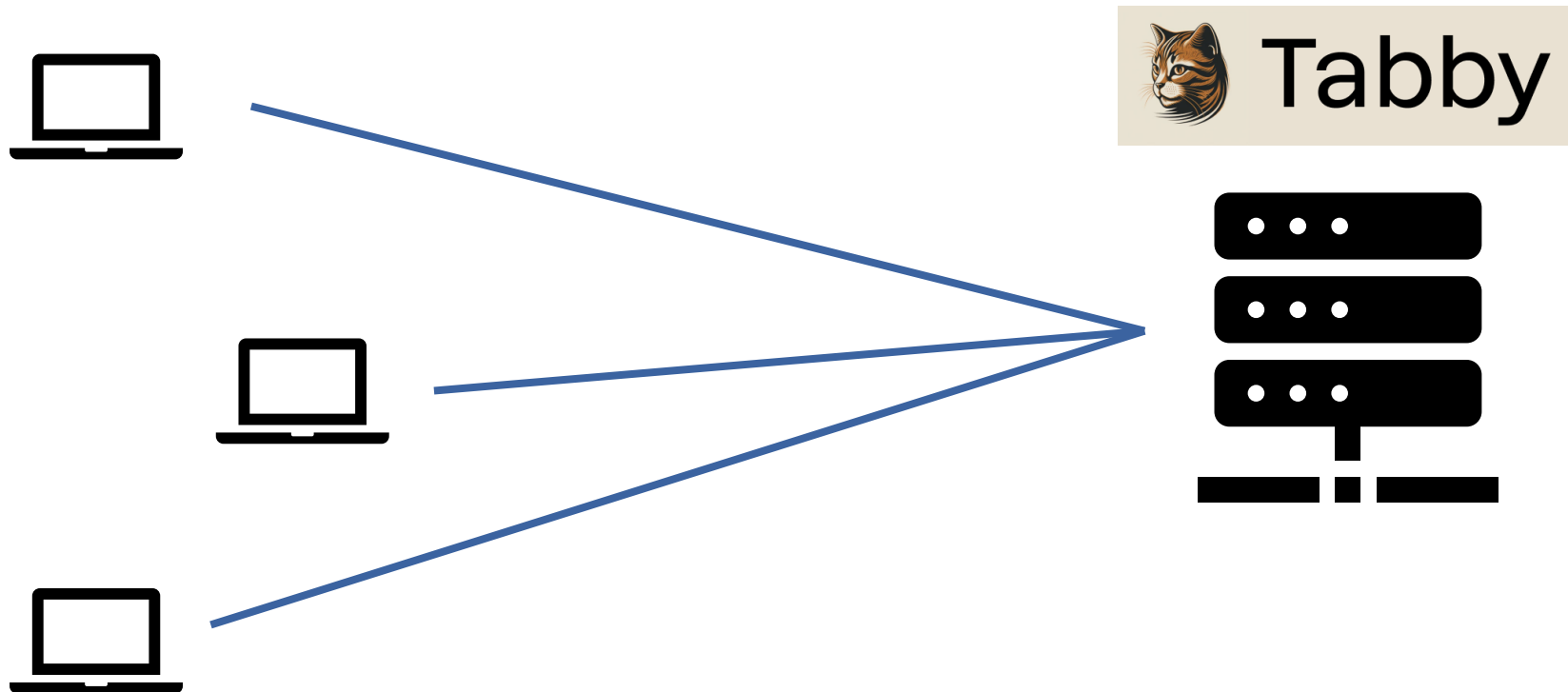


Tabby

Open Source Code-Vervollständigung



/TabbyML/tabby
GitHub Sterne - 18k



Tabby

Open Source Code-Vervollständigung



Wir setzen Tabby als internes Tool bereits ein.

- Zuverlässige Plugins (außer Eclipse!)
- Schnelle Reaktionszeit
- Vorschläge hilfreich für einfache Aufgaben

ABER:

- Kontext oft unpassend
- Nach wie vor schlechte Dokumentation
- Teures Enterprise-Modell für Nutzerverwaltung, OAuth, etc.

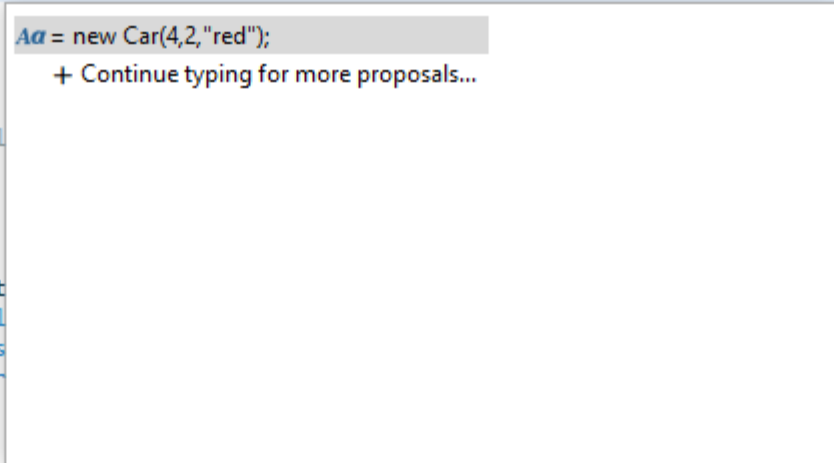


Eclipse-Integration

Tabby

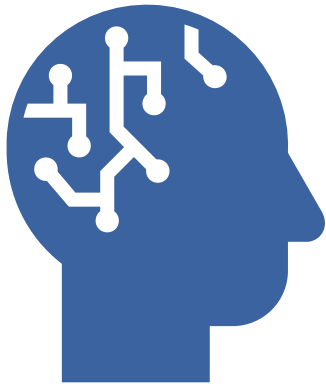
- Eclipse ist bei uns sehr verbreitet – Eigenentwicklung!

```
11
12 public void test(){
13
14     Car car = new Car(wheels:4,doors:2,color:"red");
15
16     Car car2 |
17 }
18
19 }
20
21 //create an car cl
22 class Car {
23     int wheels;
24     int doors;
25     String color;
26     public Car(int
27         this.wheel
28         this.doors
29         this.color
30 }
31 }
32
```



Was kann man verbessern?

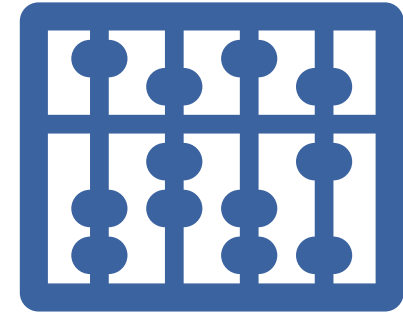
Code-Vervollständigung



Modell wechseln
Speed vs. Quality



Kontext verbessern
Anpassung an
Programmiersprache



Finetuning
Aufwendig!
Erfordert viele
Daten

Finetuning

LLMs personalisieren (Code-Vervollständigung)

Finetuning ist ein vielversprechender Ansatz, wenn:

- ...große Datenmengen zur Verfügung stehen
 - Interne Git-Repositories
- ...sich die Transferdaten von den ursprünglichen Trainingsdaten unterscheiden
 - Interne Frameworks und Konventionen
- ...RAG/Kontext zur Anpassung nicht ausreicht
- ...die benötigte Hardware (und Zeit) zur Verfügung steht

→ **Don't overuse:** Finetuning ist nicht immer der beste Ansatz!

Mögliche Vorteile:

- Bessere Performance (bessere Vorschläge)
- Mehr Effizienz (kleineres Model mit ähnlicher Performance auf eingeschränktem Anwendungsraum)

Finetuning

LLMs personalisieren (Code-Vervollständigung)

Datenset erstellen

- Crawling interner Repositories (30 Millionen Zeilen Code)

Training

- Parameter Efficient Finetuning (PEFT) auf Basis von CodeLlama-6.7B
- 150h auf NVIDIA A10 24 GB

Evaluation

- Benchmark-Fragen
- Catastrophic Forgetting untersuchen

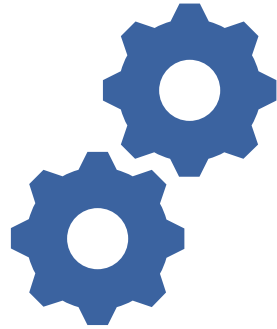
Finetuning – Ergebnisse?

- Finetuning erfolgreich mit einer Woche Training
- Test loss ist gesunken → auf dem Papier besser
- Subjektiv sind Vorhersagen in Java unwesentlich besser, ABER in zum Beispiel JS viel schlechter!

→ Wir versuchen es weiter!

Herausforderungen - Finetuning

Möglicher Guide zum Ausprobieren:
<https://huggingface.co/blog/personal-copilot>



Kein Plug-and-Play

Wenig Guides und Tutorials

Instabile Codebase

Solide PyTorch-Kenntnisse nötig



Zeit- und Hardwareaufwand

Training vorbereiten (10h)

Training (300h, NVIDIA A10 24GB)

Evaluation (10h)

Internal Knowledge Base

Halluzinationen
Lost-in-the-Middle-Problem
Retrieval Augmented Generation (RAG)
Ollama
Danswer



Mehrsprachigkeit



Anbindung interner Quellen

Automated Code Review

Beispiele mit GPT-4



Viele Falschmeldungen



Hoher Optimierungsbedarf

Code-Vervollständigung

Kontext
Tabby
Eclipse-Integration
Optimierungen – Finetuning



Kontextwahl schwierig



Komplex mit unklaren
Erfolgsaussichten

Vielen Dank!

Kontakt:
Matti Frind
matti.frind@gebit.de

