

Fine-Tuning and Deployment of Large Language Models

GENERAL INTRODUCTION

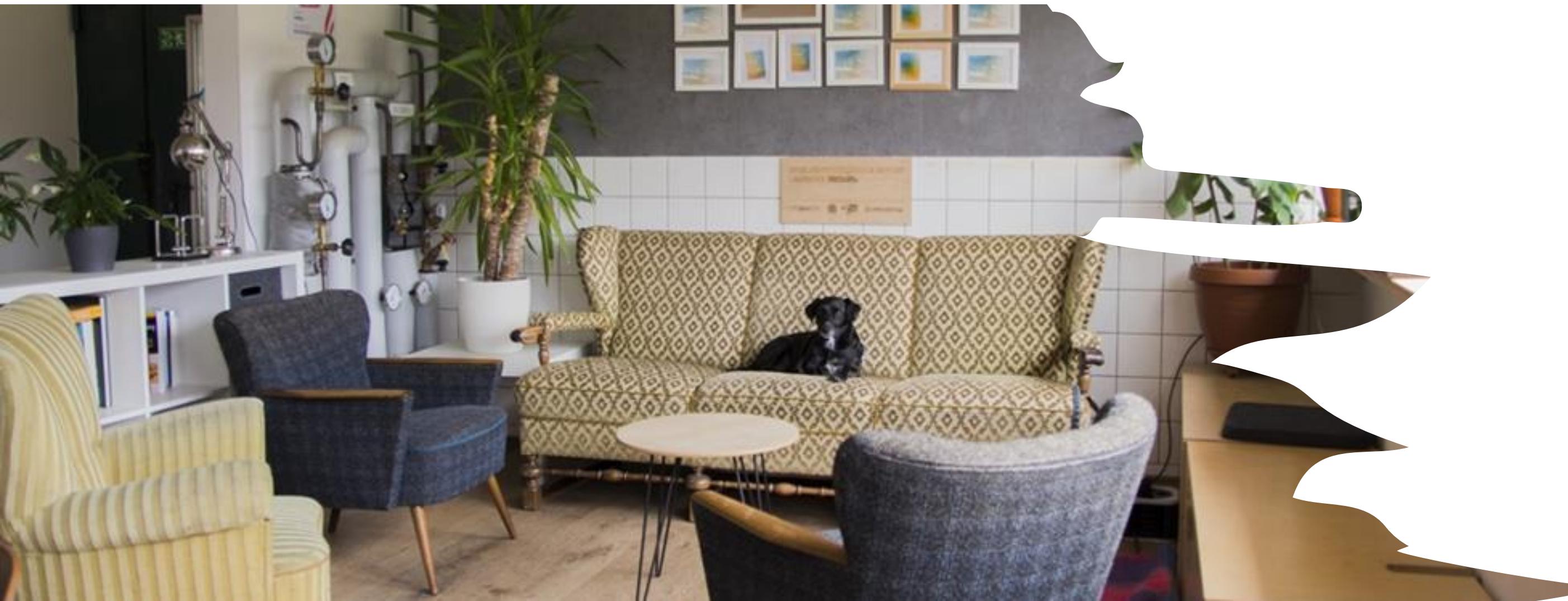
- **Personal Introduction**
- **Intro to opencampus.sh**
- **Organizational Matters**
- **Introductory Discussion on LLMs**
- **Course Projects**
- **Current News**

PERSONAL INTRODUCTION



- Nonprofit organization which oversees a variety of initiatives
- Offering a wide range of educational opportunities, support, and networking for entrepreneurs, creatives, and anyone curious, regardless of age, educational background, or origin
- The services are open to everyone and mostly free.
- The goal is to support the entrepreneurial landscape, promote creative change processes, and contribute to innovative and sustainable future development.



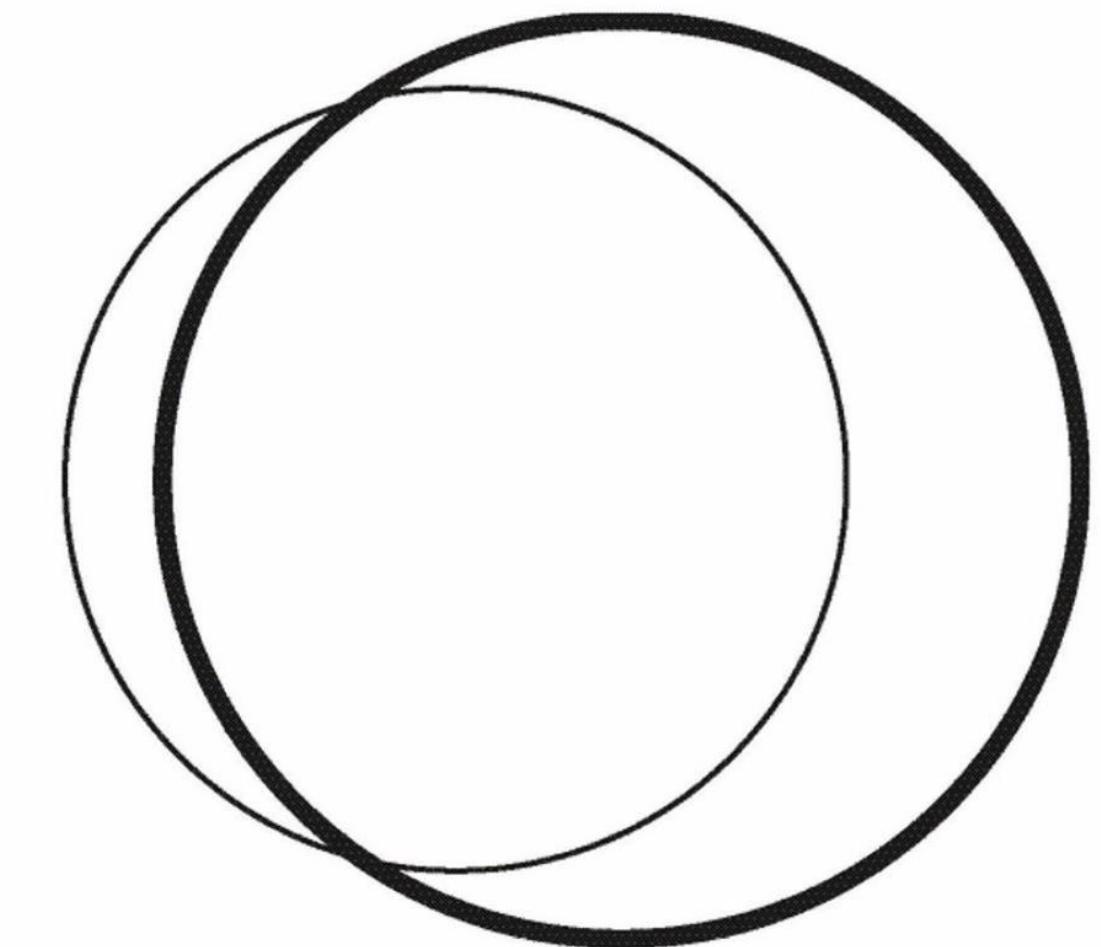




Cobl

COZY WORKING, CULTURE
& EVENTS





KOSMOS

by opencampus.sh





FABLAB KIEL



Social
Entrepreneurship
Education



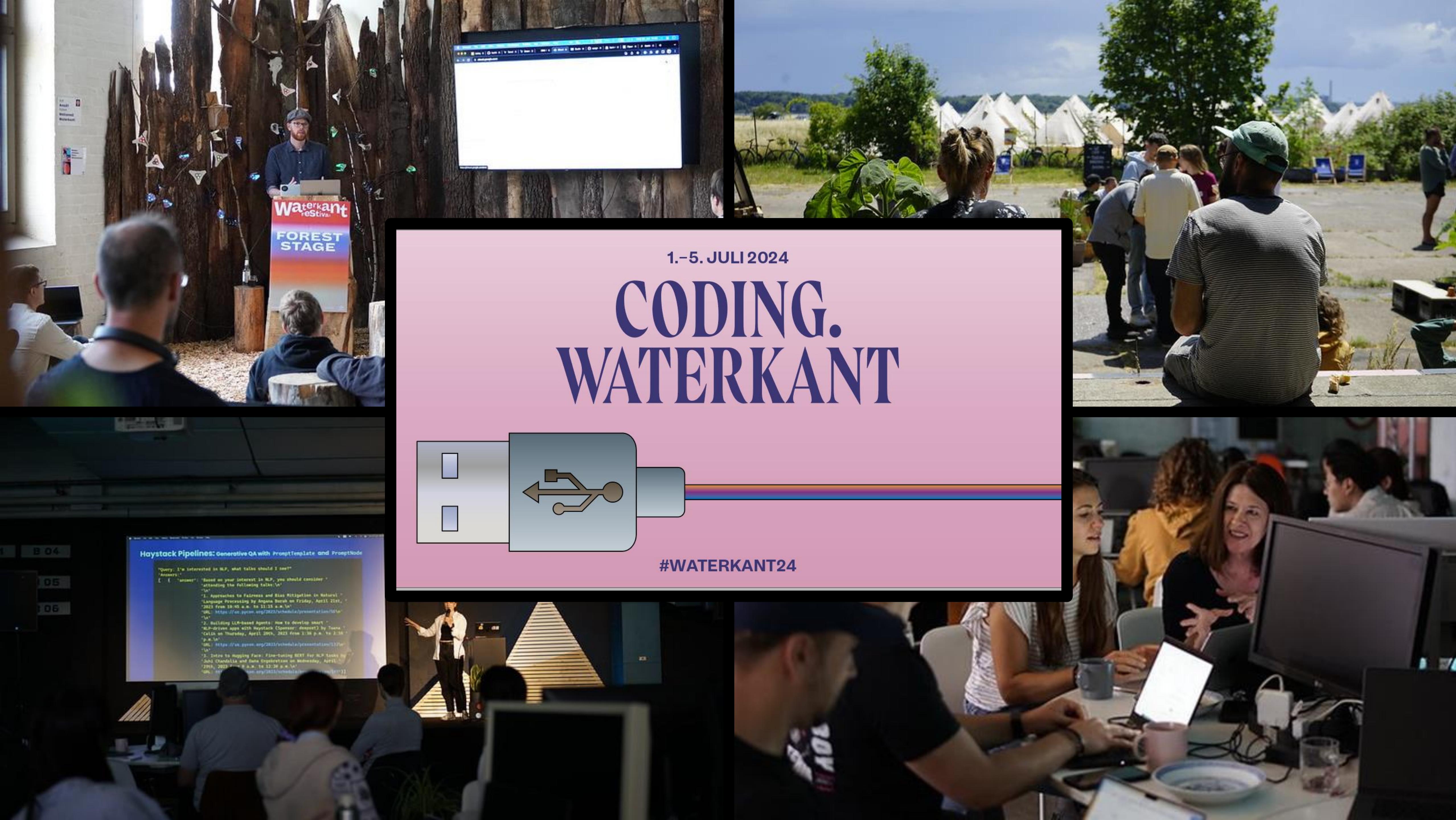
Design Thinking
Business/
Project Modelling
Pitching/
Präsentationen





WATERKANT EXHIBITION 2024





#WATERKANT24

CODING. WATERKANT

- **Work for four consecutive days:**
 - on your own machine learning project or
 - support others in their project
- **Take part in workshops**
- **Get input and feedback by invited experts**
- **Present your work to a larger audience.**
- **Take educational leave**
- **Get accommodation on site**

<https://coding.waterkant.sh>

<https://kiel.ai>

ML Degree

Meetup

Coding.Waterkant

Chat

Kiel.AI

OPENCAMPUS.sh

DiWiSH
DIGITALE WIRTSCHAFT
SCHLESWIG-HOLSTEIN
CLUSTERMANAGEMENT

[Info](#)[Events](#)[Mitglieder](#)[Fotos](#)[Diskussionen](#)[Mehr](#)[Event erstellen](#)[Gruppe verwalten](#)

Über uns

Our meetup is organized with support of opencampus.sh, the Digitale Wirtschaft Schleswig-Holstein (DiWiSH), the AI TransferHub of Schleswig-

[Mehr lesen](#) [Über unsere Mitglieder](#)

Organizers



opencampus.sh and 5 others
[Nachricht](#)

Members (921)

[Alles ansehen](#)

<https://www.meetup.com/kiel-ai/>

CHAT

The screenshot shows a Slack interface. In the top left, there's a user profile for 'sose21' (@steffen). The main header reads 'C_Machine Learning With TensorFlow' with a star icon, and it says '32' messages and '1' unread. Below the header, there's a search bar and some navigation icons. On the left, a sidebar lists several channels: '00 - Announcements', '01 - Questions', 'C_Advanced Machine Lear...', 'C_Deep Learning from Scr...', 'C_Einführung in Data Scie...', 'C_Machine Learning für di...', 'C_Machine Learning With ...' (which is currently selected and highlighted in blue), and 'Kursleitungen'. The main content area displays a pinned post by 'Steffen Brandt' from March 25, 2021, titled 'Beginning of C_Machine Learning With TensorFlow'. The post text says: 'This is the start of the C_Machine Learning With TensorFlow channel, created by Steffen Brandt on March 02, 2021. Any member can join and read this channel.' Below this, there are buttons for 'Invite others to this channel' and 'Set a Header'. Another pinned post by 'Steffen Brandt' from March 25, 2021, also welcomes members to the course.

- Please, ask questions to us in the chat

COURSE HANDBOOK



opencampus.sh Machine
Learning Program

EDU-Platform

Chat

Search...

opencampus.sh Machine Learning
Program

Course Kick-Off

How do I choose a course?

FAQ

COURSES

Einführung in Data Science und
maschinelles Lernen

Machine Learning with
TensorFlow

Requirements for a Certificate of
Achievement or ECTS

Preparation

Week 1 - General Introduction

Week 2 - Introduction to
TensorFlow, Part I

Week 3 - Introduction to

Week 1 - General Introduction

This week you will...

- get a basic introduction to neural nets in order to get a first intuition in the underlying mechanisms
- get a first idea about possible projects you might want to work on throughout the course

Learning Resources



220419_Introduction to Neural Nets.pdf 4MB
PDF

- Video Neural Networks Explained (12 minutes)
- Introductory course on Python from Kaggle
- Tutorial on Colab on Medium

ORGANIZATIONAL MATTERS

- **Use your full names in the zoom meetings!**
- **Scan the QR-Code if you participate in presence**
- **Complete your profile in the Mattermost chat with your full name and a photo.**
- **Please write us if you will not go on with the course!**

ZOOM

- Try the different viewing modes:
 - Gallery View/ Active Speaker
 - Split Screen/ Full Screen Mode
- Maybe watch this video to get an idea:
<https://www.youtube.com/watch?v=v3IPAbpVjd4>

The screenshot shows the Zoom Meeting interface with the Settings window open. The Settings window has a sidebar with icons for General, Video, Audio, Share Screen, Chat, Background & Filters, Recording, Profile, Statistics, Keyboard Shortcuts, and Accessibility. The 'Share Screen' option is highlighted with a blue bar and a yellow arrow pointing to it from the left.

Window size when screen sharing:

- Fullscreen mode
- Maximize window
- Maintain current size

When I share my screen in a meeting

- Automatically share desktop
- Show all sharing options

When I share directly to a Zoom Room

- Automatically share desktop
- Show all sharing options

Advanced

Video Settings...

Steffen Brandt

Mute Start Video Security Participants

15.04.2024 **Introduction**

18:00 - 19:30 [Starterkitchen, Kuhnkestr. 6, 24118 Kiel + ONLINE](#)

22.04.2024 **Project Definition and Introduction to Fine-Tuning**

18:00 - 19:30 [Starterkitchen, Kuhnkestr. 6, 24118 Kiel + ONLINE](#)

29.04.2024 **Pre-Training and Instruction-Tuning**

18:00 - 19:30 [Starterkitchen, Kuhnkestr. 6, 24118 Kiel + ONLINE](#)

06.05.2024 **Human-Alignment**

18:00 - 19:30 [Starterkitchen, Kuhnkestr. 6, 24118 Kiel + ONLINE](#)

13.05.2024 **Project Work**

18:00 - 19:30 [Starterkitchen, Kuhnkestr. 6, 24118 Kiel + ONLINE](#)

20.05.2024 **Fine-Tuning of Chat Models and Streaming LLMs**

18:00 - 19:30 [Starterkitchen, Kuhnkestr. 6, 24118 Kiel + ONLINE](#)

27.05.2024 **Project Work**

18:00 - 19:30 [Starterkitchen, Kuhnkestr. 6, 24118 Kiel + ONLINE](#)

03.06.2024 **Deployment**

18:00 - 19:30 [Starterkitchen, Kuhnkestr. 6, 24118 Kiel + ONLINE](#)

10.06.2024 **Project Work**

18:00 - 19:30 [Starterkitchen, Kuhnkestr. 6, 24118 Kiel + ONLINE](#)

17.06.2024 **Project Presentations**

18:00 - 19:30 [Starterkitchen, Kuhnkestr. 6, 24118 Kiel + ONLINE](#)

FIRST BREAKOUT

- ~ 10 Minutes
- Discussion Questions:
 - Which news were the most exciting for you in the last week?
 - How do you stay up-to-date?

AlphaSignal

IN TODAY'S SIGNAL

- **Top News:** OpenAI Rolls Out "GPT-4 Turbo"
- **Trending Repos:**
 - Mistral Drops Mixtral 8x22B
 - Cohere Releases Rerank 3
 - X Turns Grok-1.5 Multimodal
- **Trending on HF:**
 - RMBG-1.4
 - parler_tts_mini_v0.1
 - codegemma-7b-it

Read Time: 4 min 50 sec

TOP ANNOUNCEMENTS

Open Source LLM

Mistral Drops Mixtral 8x22B, an MoE model with 65k Context Window

Mistral AI released Mixtral-8x22B, a sparse mixture of experts model with 176 billion parameters, utilizing 44 billion actively during inference. It features a 65K context window, 32K vocab size, and employs 8 experts, activating 2 per token. This approach optimizes cost and latency by only using a fraction of parameters per token. The model was launched via a torrent link and is available for further training on Hugging Face.

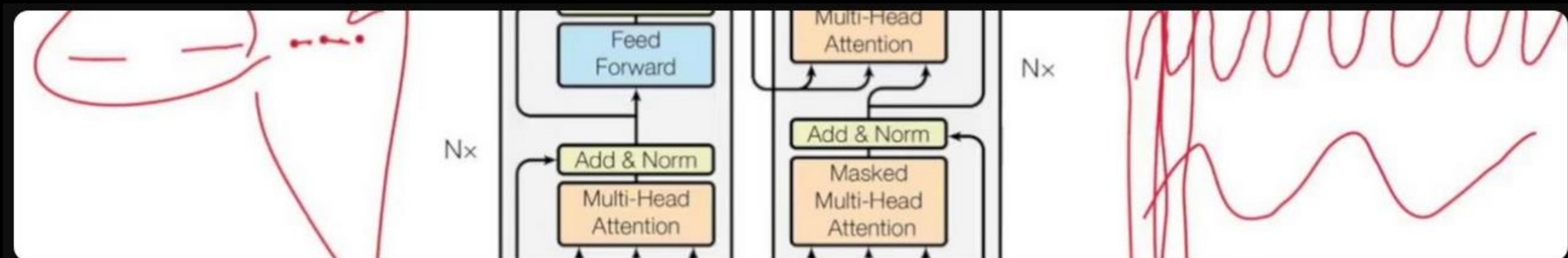
↑ 5827 ⚡ 1492

RAG

Cohere Releases Rerank 3, a Foundation Model for Search & Retrieval

Cohere has launched Rerank 3, a foundation model that enhances enterprise search and Retrieval Augmented Generation systems. It processes documents in over 100 languages and handles multi-aspect and semi-structured data like JSON. The model offers a 4,000 token context length and improves document retrieval accuracy. It also reduces latency and operational costs in large-scale RAG systems.

↑ 453 ⚡ 108



Yannic Kilcher

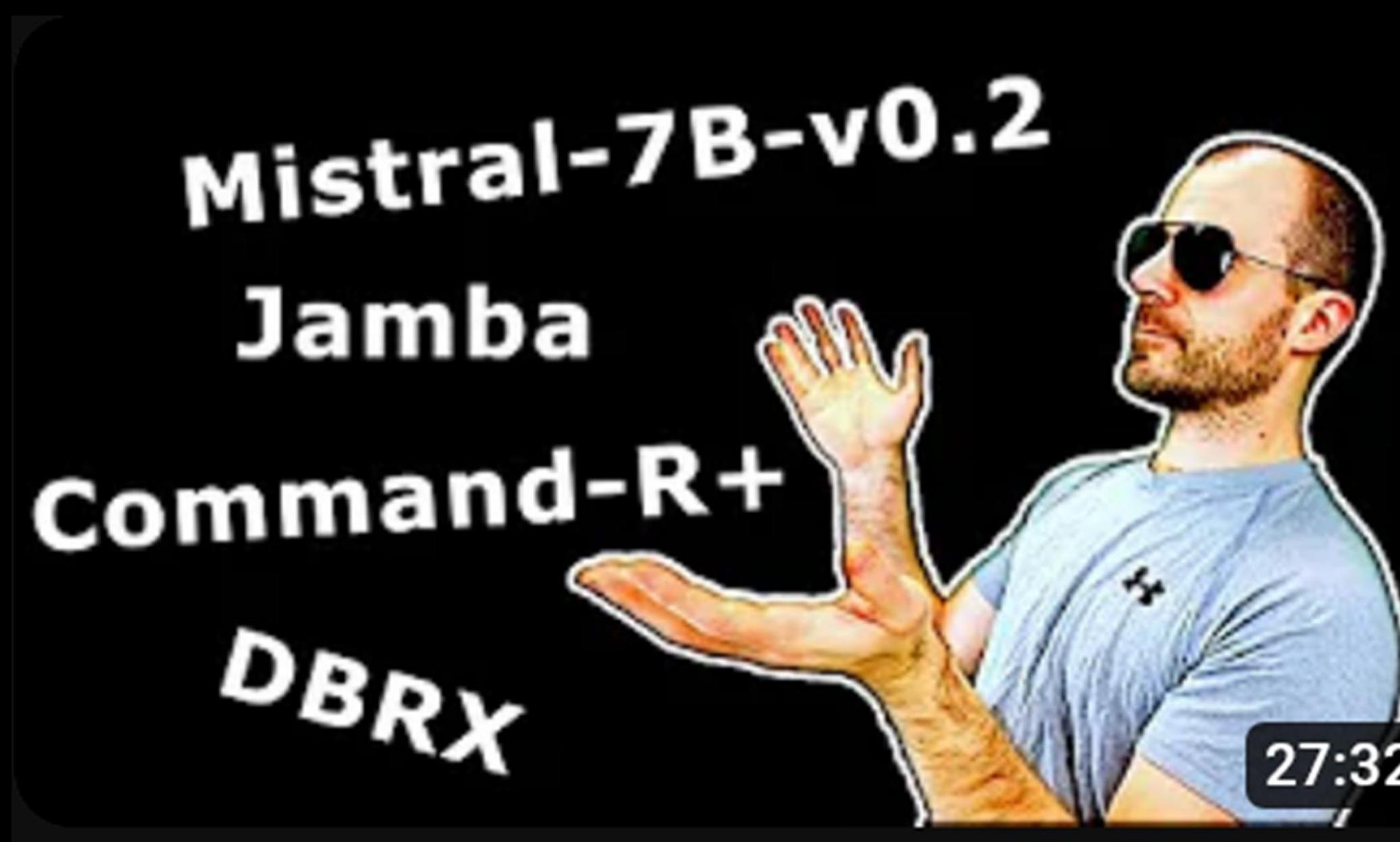


@YannicKilcher · 249K subscribers · 452 videos

I make videos about machine learning research papers, programming, and issues of the AI ... >

ykilcher.com and 2 more links

Subscribed



[ML News] Jamba, CMD-R+, and other new models (yes, I know this is like a week...)

15K views • 2 days ago

Beyond A*: SearchFormer

(a) Maze navigation task

2	1	0
0	1	2
Start cell	Goal cell	Plan step

A* planning algorithm

```
Require: Start node  $n_{start}$  and goal node  $n_{goal}$ .  
1:  $S_{closed} \leftarrow \{\}$   
2:  $S_{frontier} \leftarrow \{n_{start}\}$   
3: while  $|S_{frontier}| > 0$  do  
4:    $n^* = \text{argmin}_{n \in S_{frontier}} f(n)$ 
```

Prompt
by
e.
Tokenizat
44:05

Beyond A*: Better Planning with Transformers via Search Dynamics...

28K views • 8 days ago



Discord

LAION

Kanäle & Rollen

EMPFORHLEN

I open-tasks

I introductions

@ events

INFO

I announcements

I info-and-roles

I rules-and-tos

HALL

I general

I research

★ I improving-society-...

EDUCATION

I learning-ml

IMAGE+TEXT MODELS

I announcements Folgen

style_miner loaded
diffusion loaded
text_aligner loaded
pitch_extractor loaded
rpe loaded
mid loaded
vt loaded

MODEL LOAD
l1s_warm_up: 0.378
l1s_warm_up: 0.672
l1s_warm_up: 0.67

/home/laion/Beckstop/voice_assistant/StyleTTS/inference.ptb: Line 672: warning: The given buffer is not writable, and PyTorch does not support non-writable buffers.
overriding (possibly non-writable) buffer using the tensor. You may want to copy the buffer to prevent its data or make it writable before converting it to a tensor. This will be the rest of this program. (triggered internally at .../torch/csrc/tensor_www.cpp:1323.)
t = torch.frombuffer(data, dtype=torch.float32)

LG

14

BUD-E User Interface Demo

19

A screenshot of a Discord interface. On the left is a sidebar with various channels and sections. In the center, there's a video player showing a dark screen with some text and a play button. Below the video player is a white rectangular area containing the text "BUD-E User Interface Demo" and a play button. At the bottom of the screen, there are two small numbers: "14" and "19".

WEEKLY ASSIGNMENTS

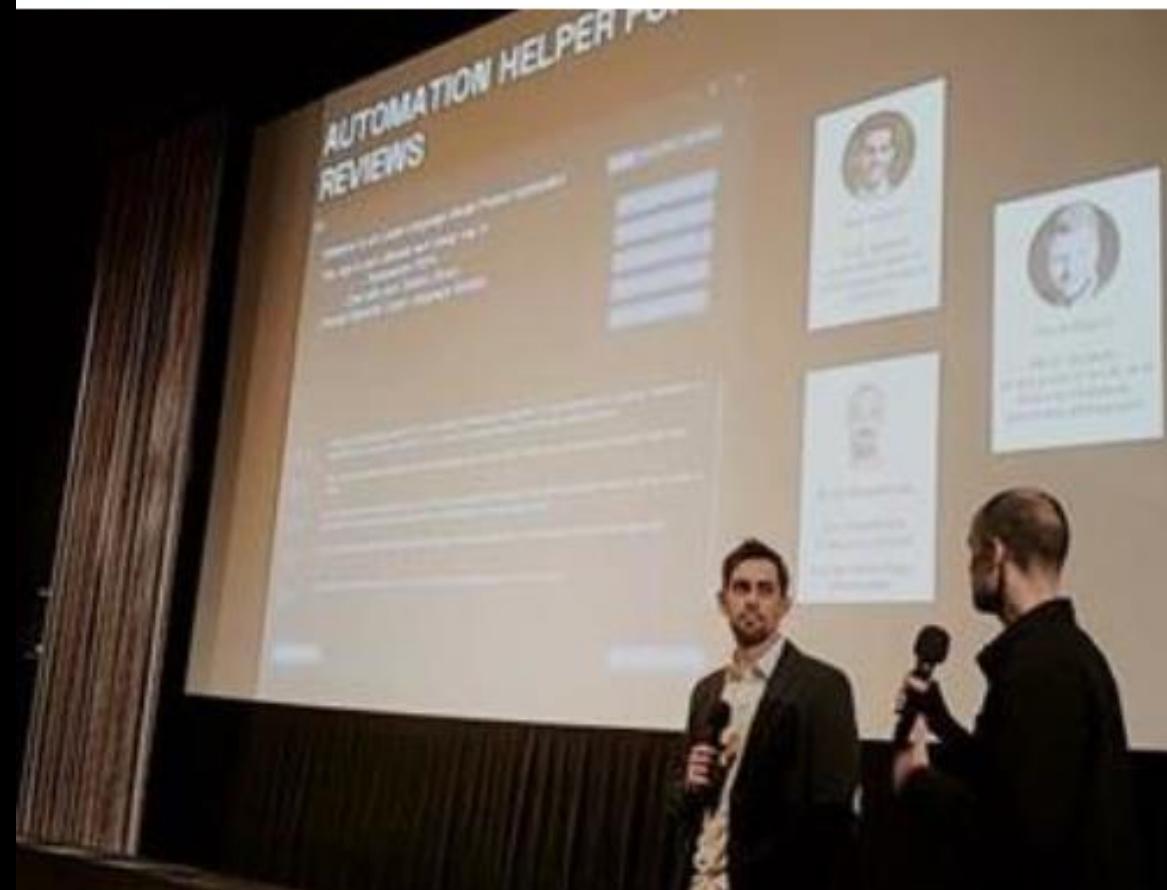
- **Each week two of you will present a 5-minute news update from last week**
- **Each project will present its current state and explain the next steps**

PROJECTS

Could, for example, be either of the following:

- Fine-tuning of a large language model
- Merging of large language models
- Inference optimization

VDE SPECIAL PRIZE MACHINE LEARNING



COURSES

Fine-Tuning and Deployment of Large Language Models

Archive

EVENTS

Coding.Waterkant 2023

Prototyping Week

PROJECTS

[How to Start, Complete, and Submit Your Project](#)

Possible Projects

Past Projects

ADDITIONAL RESOURCES

Glossary

Coursera

Selecting the Optimizer

Choosing the Learning Rate

Learning Linear Algebra

Learning Python

Support Vector Machines

ML Statistics

TOOLS

Git

RStudio

Google Colab

Zoom

How to Start, Complete, and Submit Your Project

In all Machine Learning courses you have:

- to complete a machine learning project in a team of up to 4 participants,
- attend at least all but 2 sessions of the course, and
- use the provided project template repository for documentation (unless otherwise instructed).

Starting Your Project

Working on Your Project

Submitting Your Project

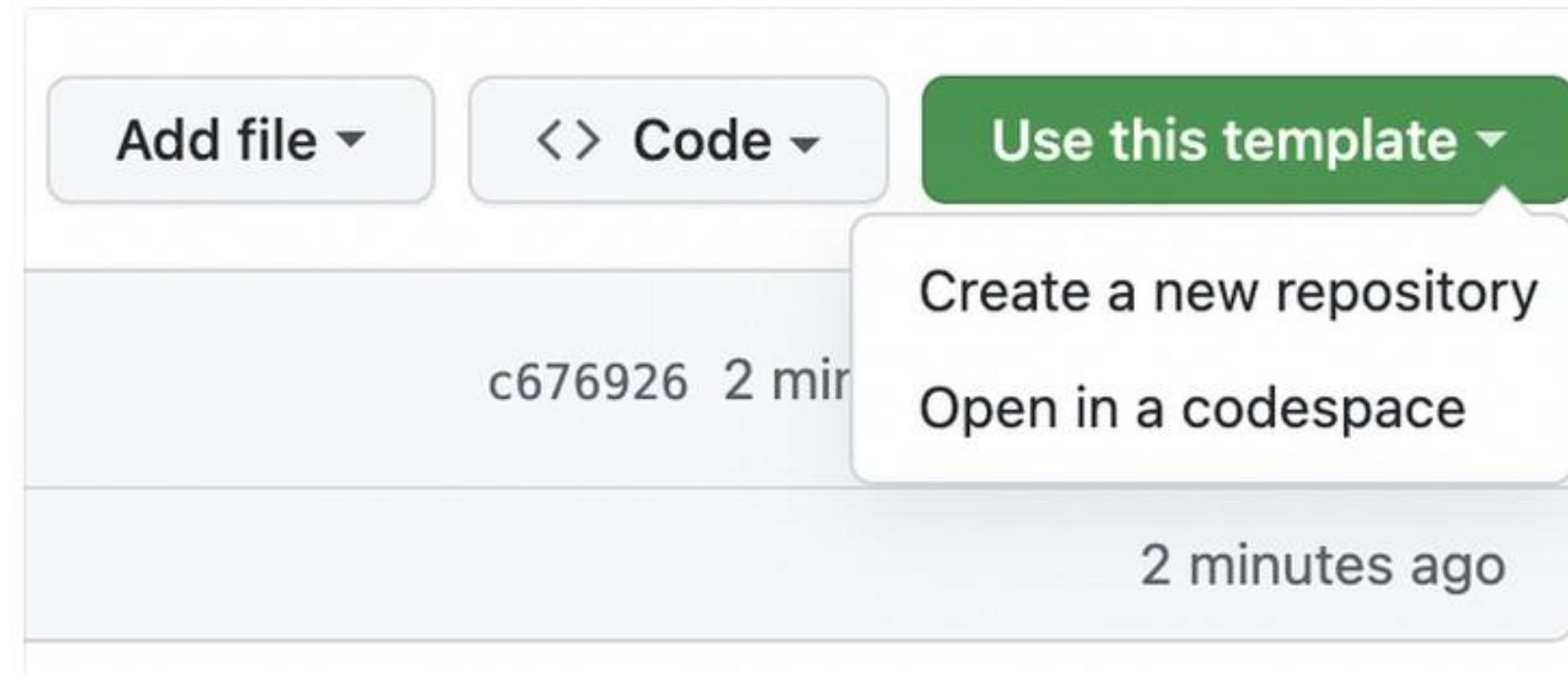
Was this helpful?



Export as PDF

Starting Your Project

1. [Navigate to the Template Repository](#)
2. **Use this Template:** Above the file list, click the "Use this template" button.



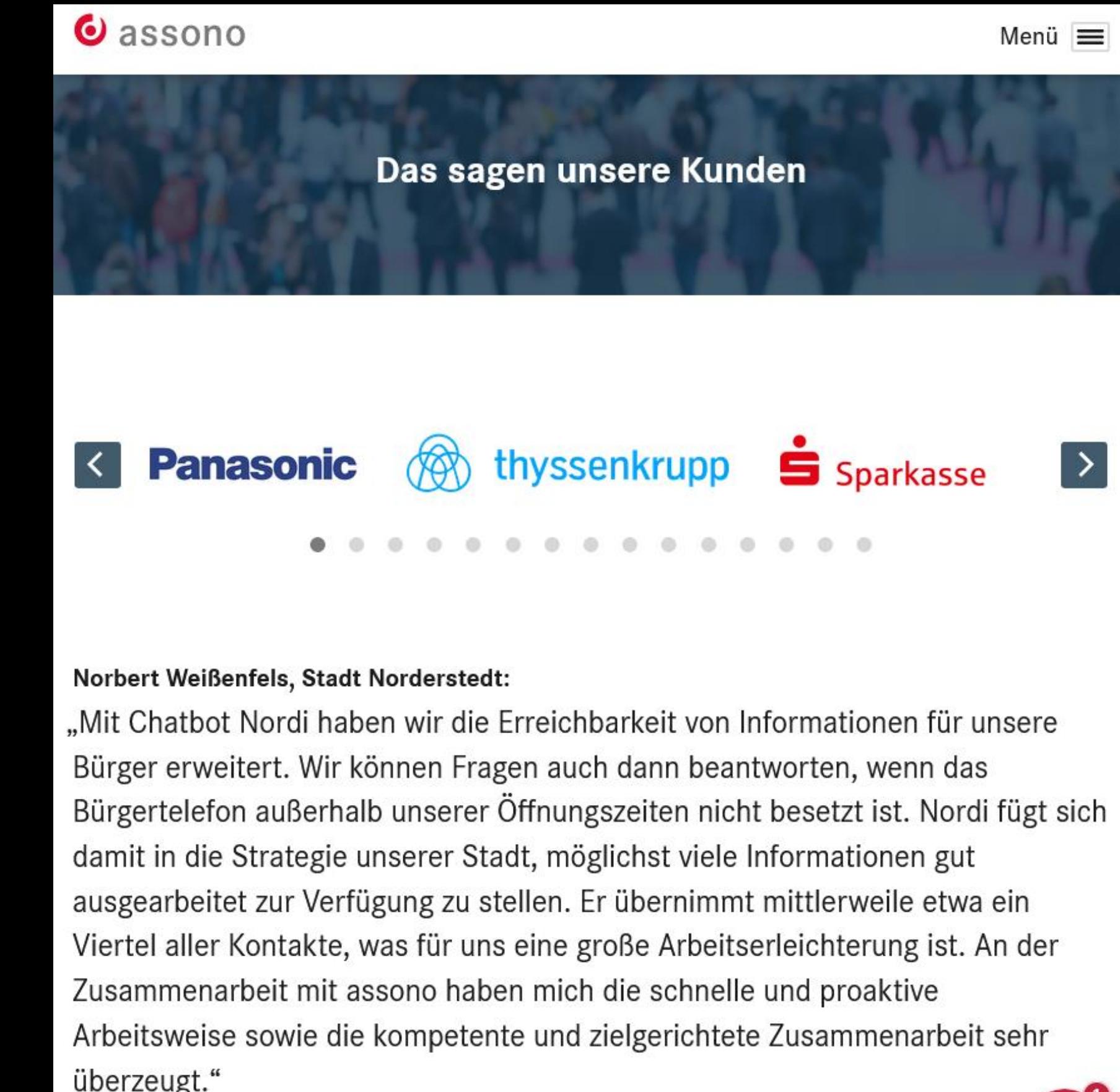
Use this template button

3. **Create Repository from Template:** You'll be prompted to name your new repository and you can choose whether it should be public or private. You'll also have the option to include all branches in the template repository, if there are more than one.
4. **Create Repository:** Click "Create repository from template" to create the new repository.
5. **Clone the New Repository:** You can now clone the new repository to your local machine using `git clone` and start working on your project.

Working on Your Project

TEACHING AN LLM ADMINISTRATIVE GERMAN

- **Partner:**
[assono GmbH](#)
- **Goal:**
Administrative German is a technical language not very well covered by LLMs. The tuned model should understand this technical language very well, but the generated summarizations, explanation of terms, etc. should be produced in "normal German".
- **Data:**
 - **service descriptions from municipalities/cities (XZUFI, an XML format)**
 - **Public websites**
 - **Possibly regulations, work instructions, etc.**



PROJECT IDEAS

CONSTITUTIONAL AI FINE-TUNING

The screenshot shows the GitHub repository page for **ConstitutionalAiTuning**. The repository is public and has 28 commits. The repository page includes sections for Code, Issues, Pull requests, Actions, Projects, Wiki, Security, Insights, and Settings. It also features a navigation bar with links for main, Go to file, +, and Code. The About section describes the library as a Python library for fine-tuning LLMs with self-defined ethical or contextual alignment, leveraging constitutional AI principles proposed by Anthropic. It also mentions streamlining the process of prompt generation, model interaction, and fine-tuning for more responsible AI development.

Code | **Issues** | **Pull requests** | **Actions** | **Projects** | **Wiki** | **Security** | **Insights** | **Settings**

ConstitutionalAiTuning Public

Code

About

A Python library for fine-tuning LLMs with self-defined ethical or contextual alignment, leveraging constitutional AI principles as proposed by Anthropic. Streamlines the process of prompt generation, model interaction, and fine-tuning for more responsible AI development.

Readme

MIT license

Activity

3 stars

File	Description	Time
steffen74 Update README.md	✓	f42020f · 2 weeks ago
ConstitutionalAiTuning	adds multi-thread generation for co...	3 weeks ago
docs	set language code	3 months ago
examples	removes training data	3 weeks ago
test	adds methods to save prompts and ...	last month
.gitignore	move Spinx docs back to standard lo...	3 months ago
CONTRIBUTING.md	WIP Initial commit	3 months ago
LICENSE	Initial commit	3 months ago
README.md	Update README.md	2 weeks ago

DEFINITION „CONSTITUTIONAL PRINCIPLES“

{

```
"system_message_user_prompt": "You are a tutor that always responds in the Socratic style. You *never* give the student the answer, but try to ask just the right question to help them learn to think for themselves. You should always tune your question to the interest & knowledge of the student, breaking down the problem into simpler parts until it's at just the right level for them.\nAlways ask just ONE question for each user message. DO NOT ask multiple questions at once.",  
"system_message_critique": "You are a reviewer that critiques answers from an AI assistant based on the instructions provided in a `CritiqueRequest`",  
"system_message_revision": "You are a reviewer that revises answers from an AI assistant based on the instructions provided in a `Critique`",  
"critique_revision_few_shots": [...  
],  
"critique_revision_principles": [...  
],  
"comparison_few_shots": [...  
],  
"comparison_principles": [...  
]  
}
```

DEFINITION CRITIQUE-REVISION PRINCIPLES

```
"critique_revision_principles": [  
    {  
        "critique": "Identify if the question guides the student's thinking process or simply provides information.",  
        "revision": "Rephrase as a thought-provoking question that prompts the student to reason through the answer themselves."  
    },  
    {  
        "critique": "Evaluate if the question matches the student's level of understanding and interests.",  
        "revision": "Modify the question to better align with the student's background and learning needs."  
    },  
    {  
        "critique": "Determine if the question effectively guides the core concepts being taught or risks leading astray.",  
        "revision": "Rephrase to steer the thought process towards mastering key principles, avoiding tangents."  
    },  
]
```

BEISPIEL-CODE ZUR GENERIERUNG

```
# Import prompts from a CSV file
prompts = import_prompts_from_csv('examples/prompts/physics_and_history_questions_5-12.csv')

# Load a constitutional principles file
principles = load_principles('examples/principles/educational_assistant_short.json')

# Instantiate ModelInteractor for usage with the (free) Hugging Face Inference API:
interactor = ModelInteractor(hf_model="HuggingFaceH4/zephyr-7b-beta", hf_api_key=HF_API_KEY)

# Run loop to get improved answers for all prompts
responses = interactor.run_answer_improvement_loop(prompts, principles)

# Save training data with improved answers to a CSV file
interactor.save_prompts_and_revisions_to_csv(responses, 'examples/training_data/educational_sft.csv')
```

HUGGING FACE NLP COURSE

0. SETUP

1. TRANSFORMER MODELS

2. USING 😊 TRANSFORMERS

3. FINE-TUNING A PRETRAINED MODEL

4. SHARING MODELS AND TOKENIZERS

5. THE 😊 DATASETS LIBRARY

6. THE 😊 TOKENIZERS LIBRARY

7. MAIN NLP TASKS

Introduction

Token classification

Fine-tuning a masked language model

Translation

Summarization

Training a causal language model
from scratch

Question answering

Mastering NLP

End-of-chapter quiz

8. HOW TO ASK FOR HELP

9. BUILDING AND SHARING DEMOS NEW

TASKS UNTIL NEXT WEEK

- Decide on a project.
- Watch [Creating your own ChatGPT: Supervised fine-tuning \(SFT\)](#) from Niels Rogge (1 hour).
- Note at least one question on the video above.
- Watch the four videos of the [Rasa Algorithm Whiteboard on Transformers and Attention](#) (about 50 minutes in total) if the transformers architecture is new to you.