15.04.24

OPENCAMPUS.sh

Fine-Tuning and Deployment of Large Language Models

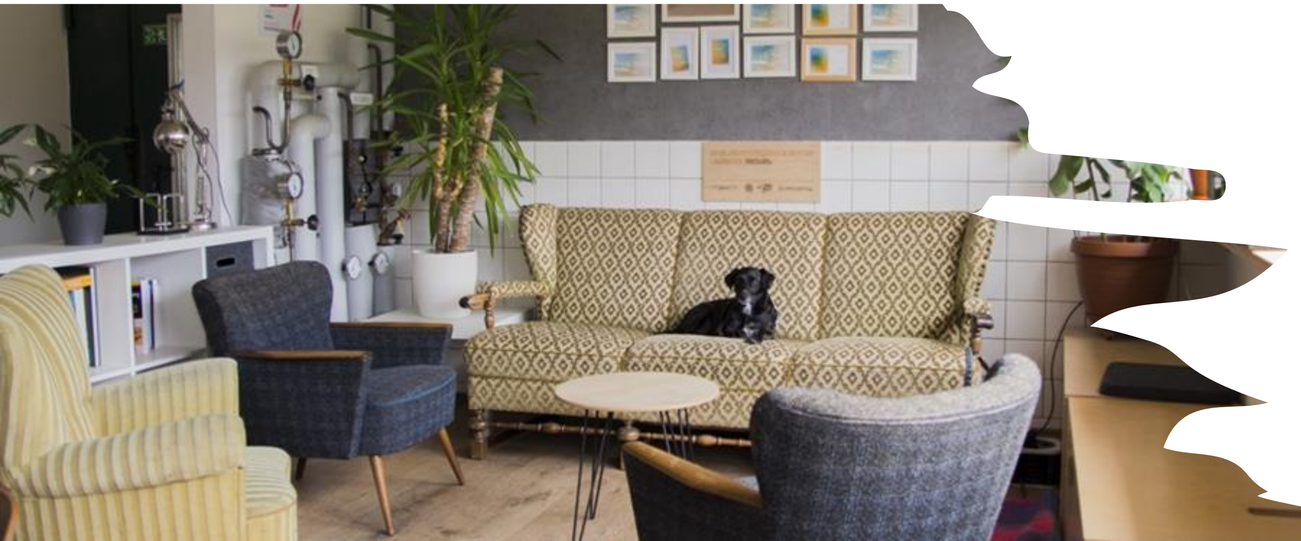
GENERAL INTRODUCTION

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

PERSONAL INTRODUCTION

OPENCAMPUS.sh

- 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.



starter kitchen.de

CoBL

COZY WORKING, CULTURE & EVENTS

KOSMOS
by opencampus.sh

FABLAB KIEL

SEEd | Social Entrepreneurship Education

#seed17

Design Thinking

Business/
Project Modelling

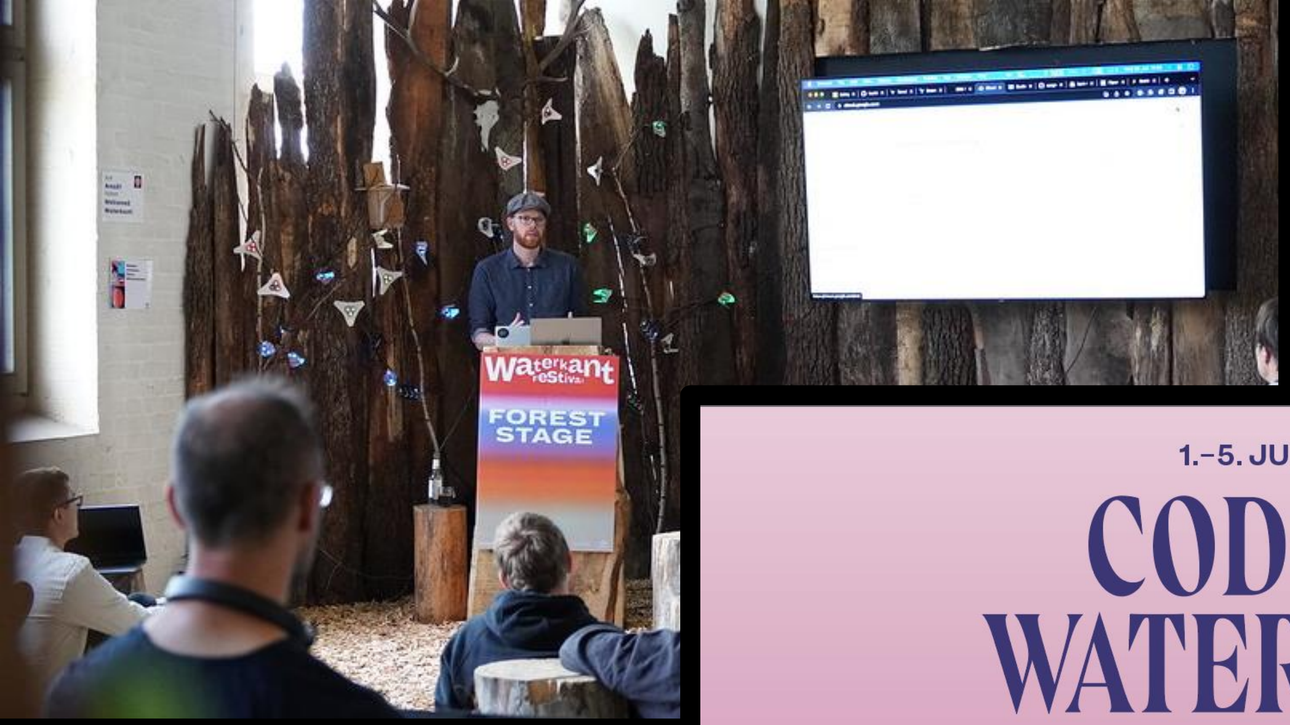
Pitching/
Präsentationen

MFG 5 KIEL WE CONNECT. 13.–14. JUNI 2024

WATERKANT FESTIVAL

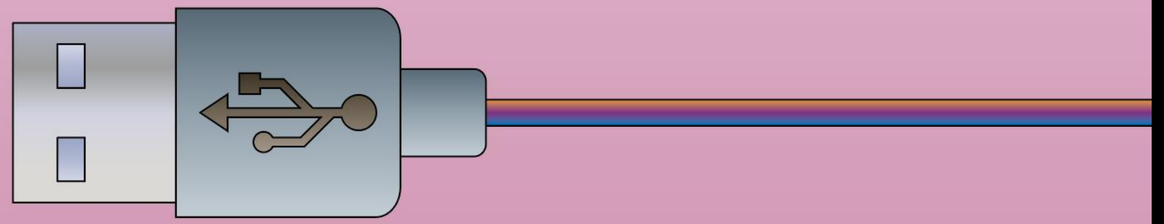
WATERKANT.SH

WATERKANT EXHIBITION 2024



1.–5. JULI 2024

CODING. WATERKANT



#WATERKANT24

1.–5. JULI 2024

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

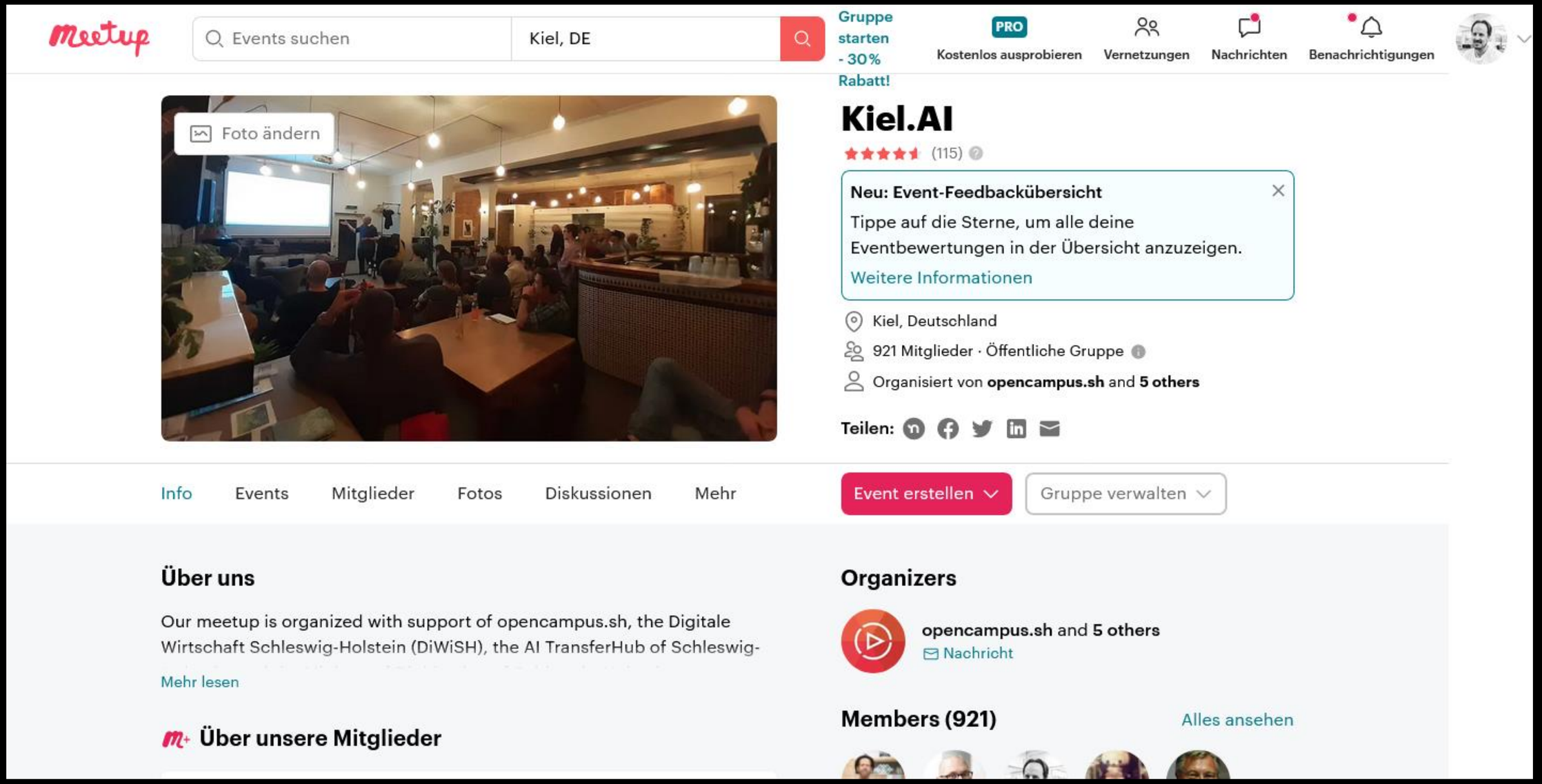
https://kiel.ai

ML Degree Meetup Coding.Waterkant Chat

Kiel.AI

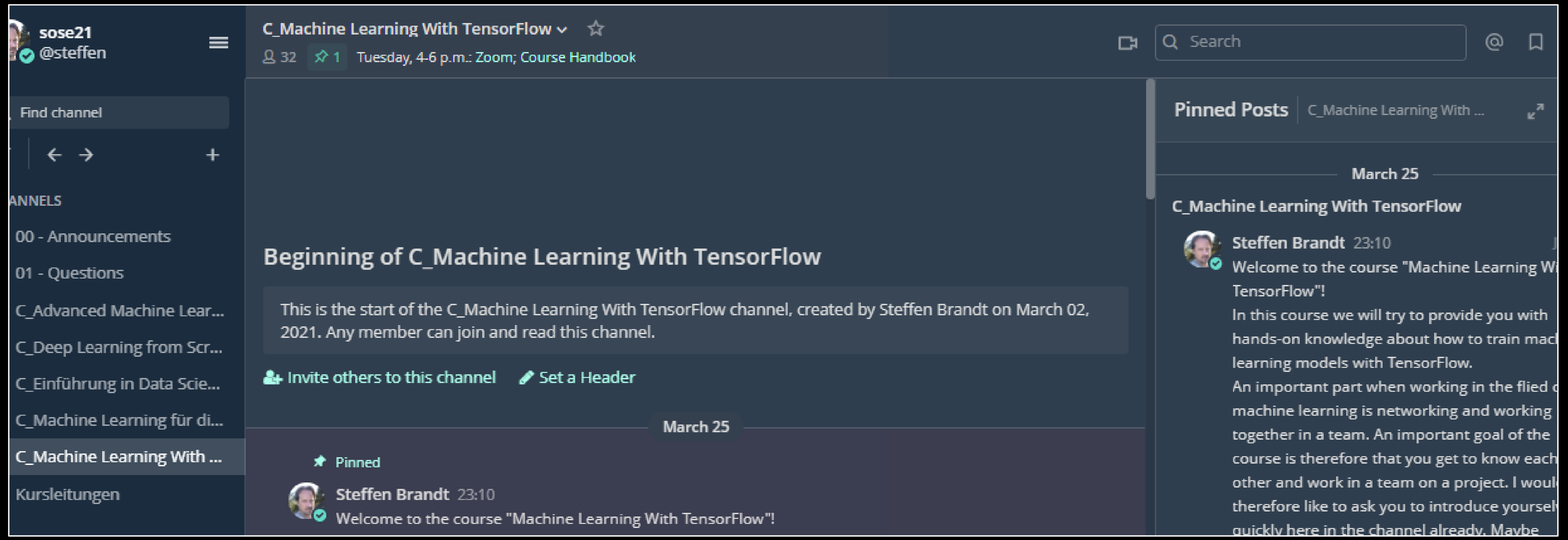
OPENCAMPUS.sh

DiWiSH
DIGITALE WIRTSCHAFT
SCHLESWIG-HOLSTEIN
CLUSTERMANAGEMENT



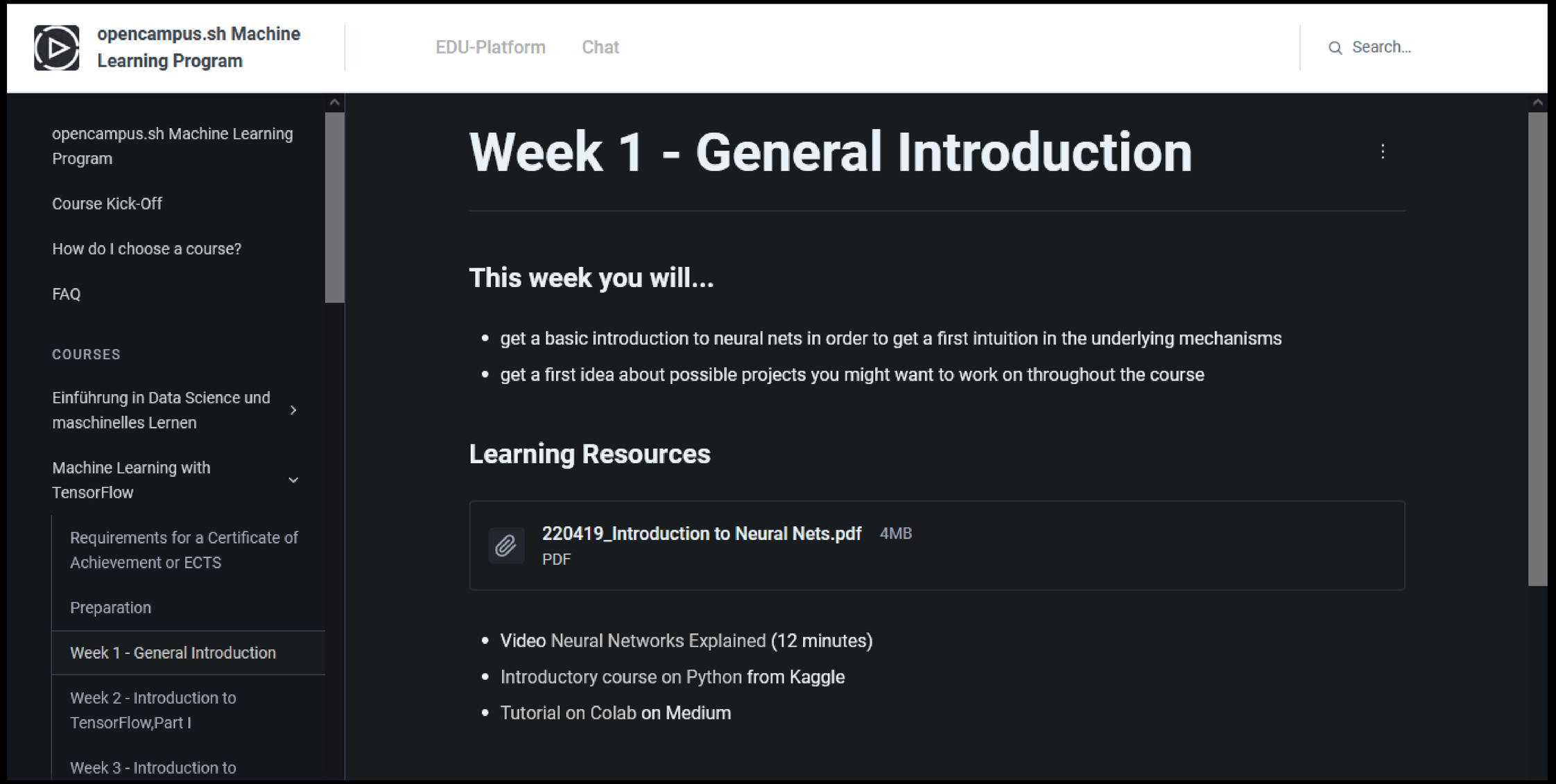
https://www.meetup.com/kiel-ai/

CHAT



- **Please, ask questions to us in the chat**

COURSE HANDBOOK



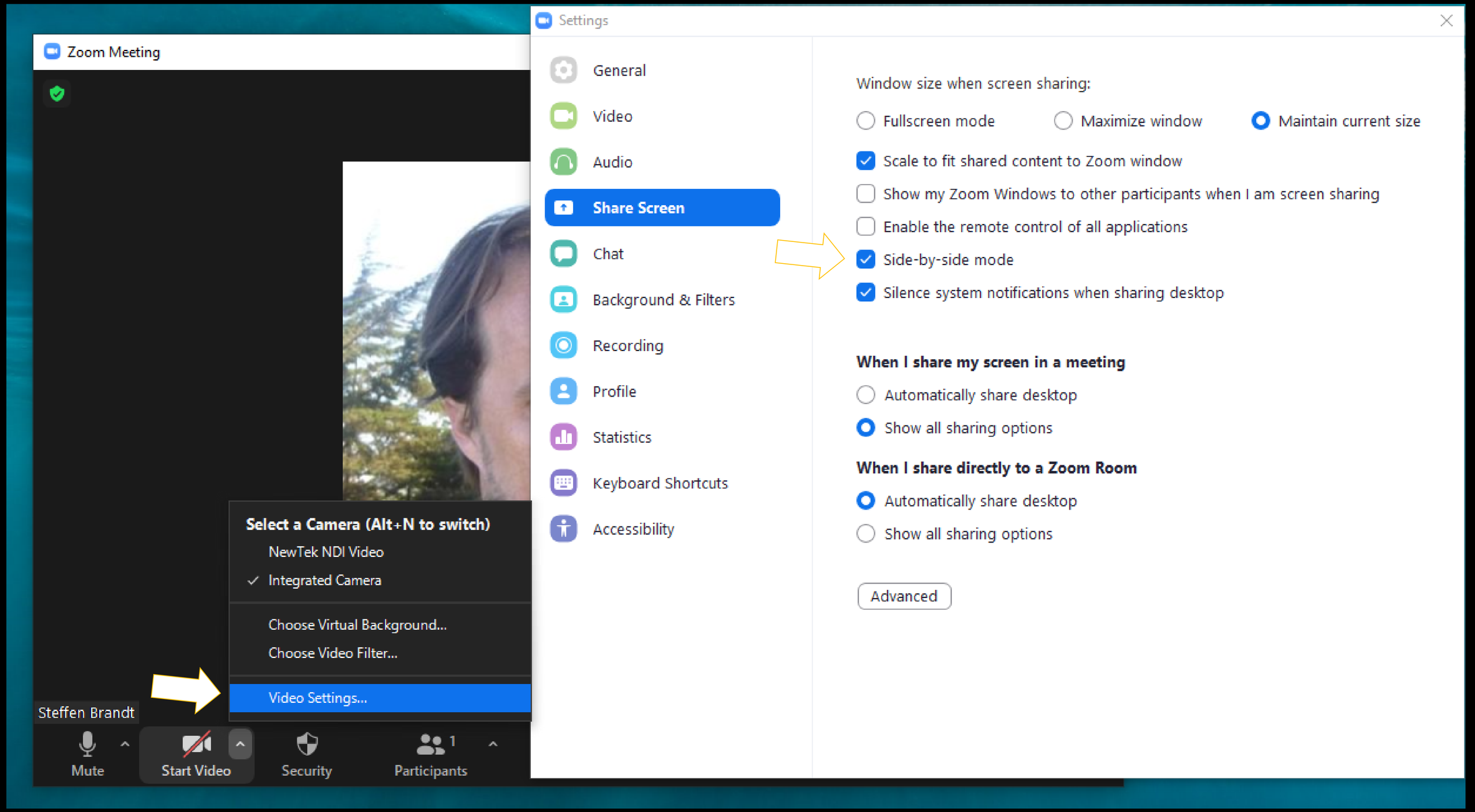
ORGANIZATIONAL MATTERS

- 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


Zoom Meeting
Steffen Brandt
Mute
Start Video
Security
Participants
1
Select a Camera (Alt+N to switch)
NewTek NDI Video
Integrated Camera
Choose Virtual Background...
Choose Video Filter...
Video Settings...
Settings
General
Video
Audio
Share Screen
Chat
Background & Filters
Recording
Profile
Statistics
Keyboard Shortcuts
Accessibility
Window size when screen sharing:
Fullscreen mode
Maximize window
Maintain current size
Scale to fit shared content to Zoom window
Show my Zoom Windows to other participants when I am screen sharing
Enable the remote control of all applications
Side-by-side mode
Silence system notifications when sharing desktop
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

15.04.2024 18:00 - 19:30	Introduction	Starterkitchen, Kuhnkestr. 6, 24118 Kiel + ONLINE
22.04.2024 18:00 - 19:30	Project Definition and Introduction to Fine-Tuning	Starterkitchen, Kuhnkestr. 6, 24118 Kiel + ONLINE
29.04.2024 18:00 - 19:30	Pre-Training and Instruction-Tuning	Starterkitchen, Kuhnkestr. 6, 24118 Kiel + ONLINE
06.05.2024 18:00 - 19:30	Human-Alignment	Starterkitchen, Kuhnkestr. 6, 24118 Kiel + ONLINE
13.05.2024 18:00 - 19:30	Project Work	Starterkitchen, Kuhnkestr. 6, 24118 Kiel + ONLINE
20.05.2024 18:00 - 19:30	Fine-Tuning of Chat Models and Streaming LLMs	Starterkitchen, Kuhnkestr. 6, 24118 Kiel + ONLINE
27.05.2024 18:00 - 19:30	Project Work	Starterkitchen, Kuhnkestr. 6, 24118 Kiel + ONLINE
03.06.2024 18:00 - 19:30	Deployment	Starterkitchen, Kuhnkestr. 6, 24118 Kiel + ONLINE
10.06.2024 18:00 - 19:30	Project Work	Starterkitchen, Kuhnkestr. 6, 24118 Kiel + ONLINE
17.06.2024 18:00 - 19:30	Project Presentations	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.

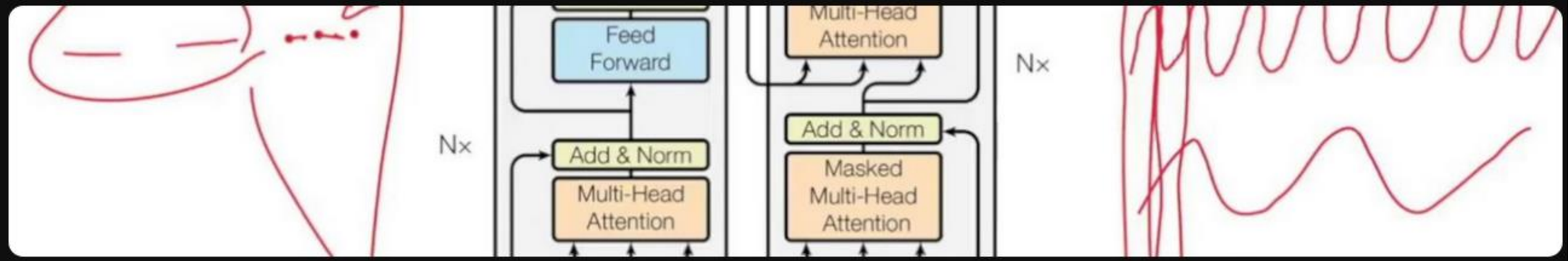
⇧ 5827 ⇆ 1492

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



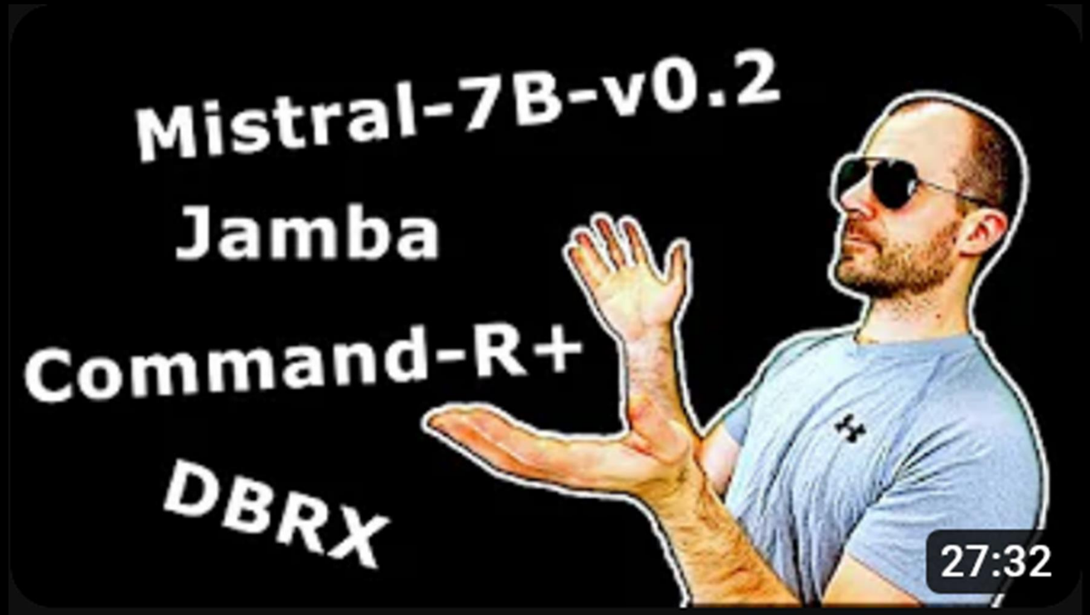
Yannic Kilcher

@YannicKilcher · 249K subscribers · 452 videos

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

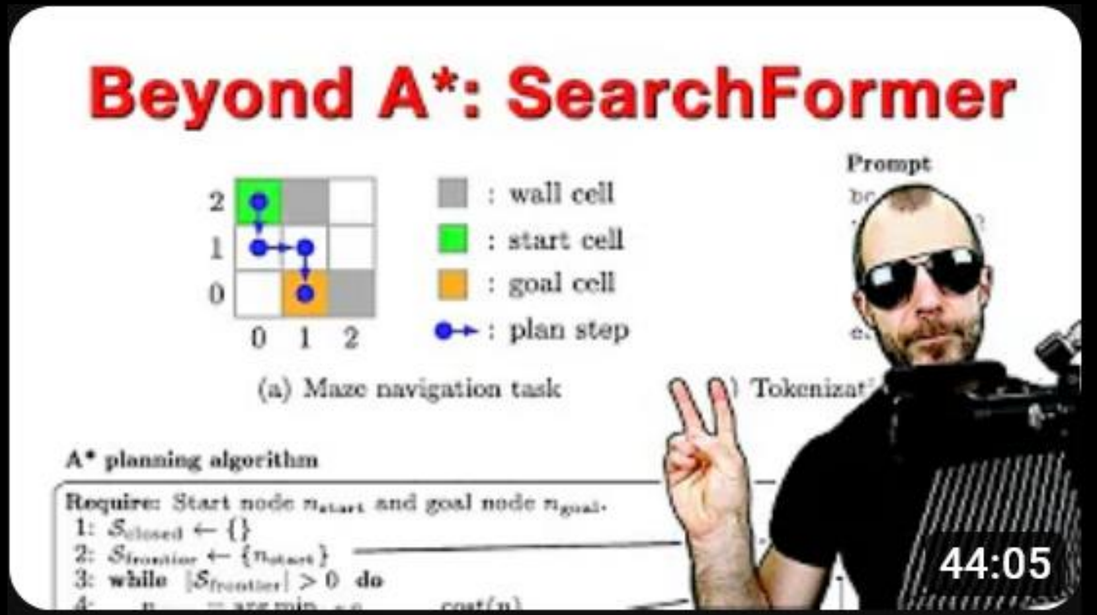
ykilcher.com **and 2 more links**

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

28K views • 8 days ago

Discord

LAION

Kanäle & Rollen

EMPFOHLEN

📝 | open-tasks

👋 | introductions

events

INFOS

| announcements

| info-and-roles

✍ | rules-and-tos

HALL

💬 | general

| research

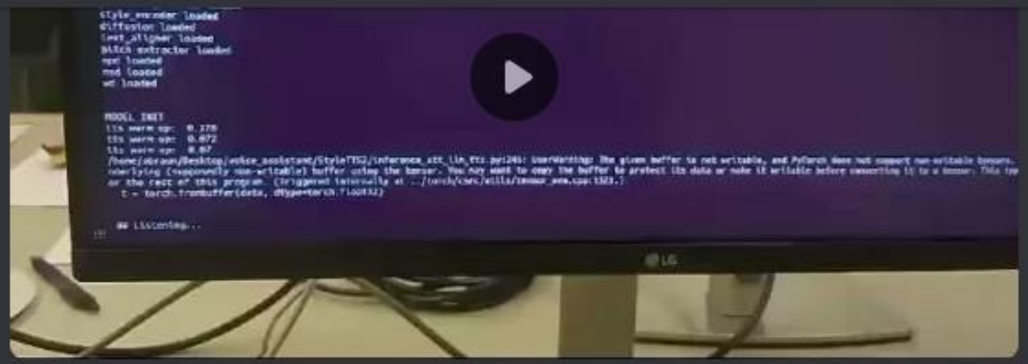
★ | improving-society-...

EDUCATION

| learning-ml

IMAGE+TEXT MODELS

| announcements Folgen



14

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

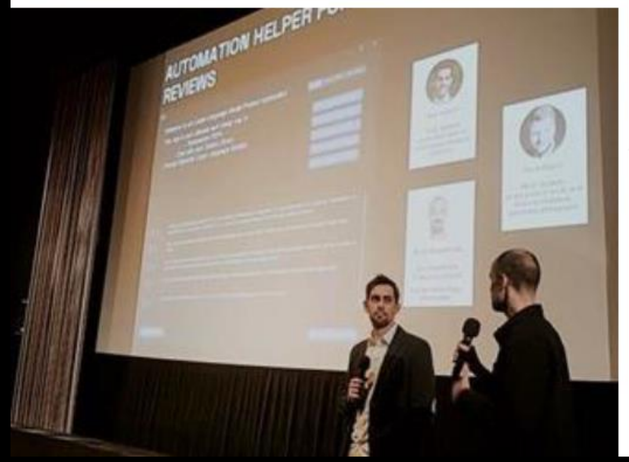
PROJECTS

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 RESOURSES

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

Starting Your Project

Working on Your Project

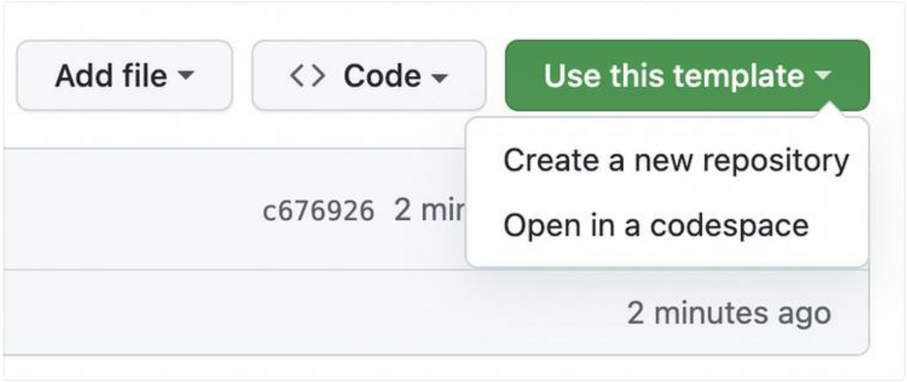
Submitting Your Project

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

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.

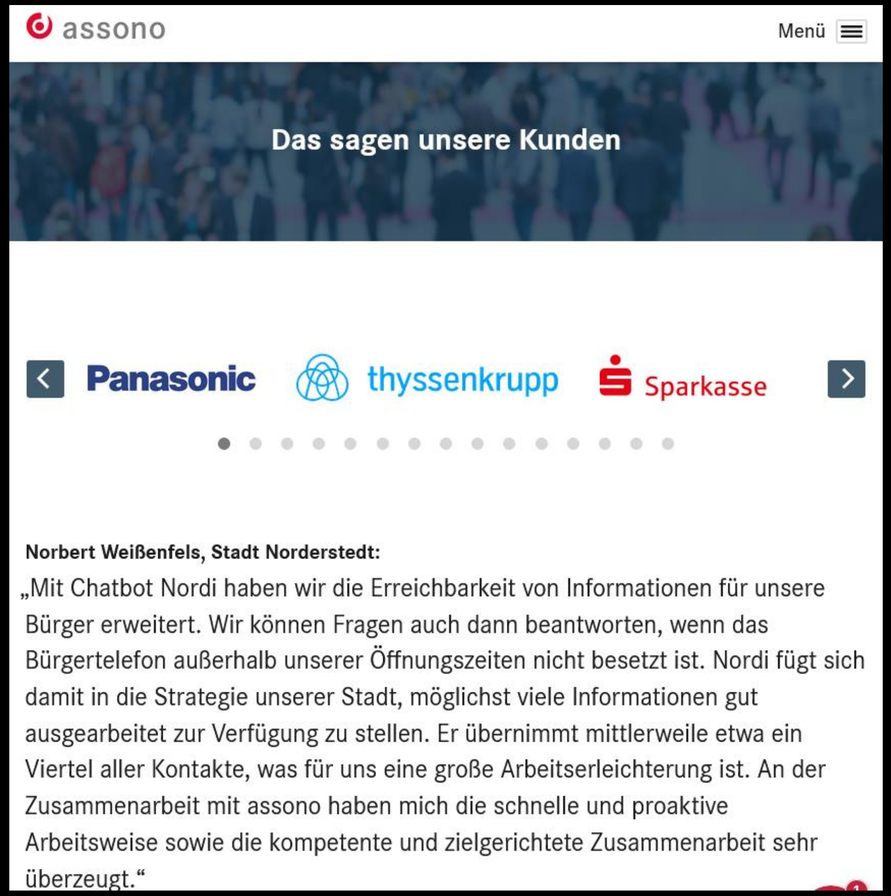
Working on Your Project

TEACHING AN LLM ADMINISTRATIVE GERMAN

- **Partner:**
  assono GmbH

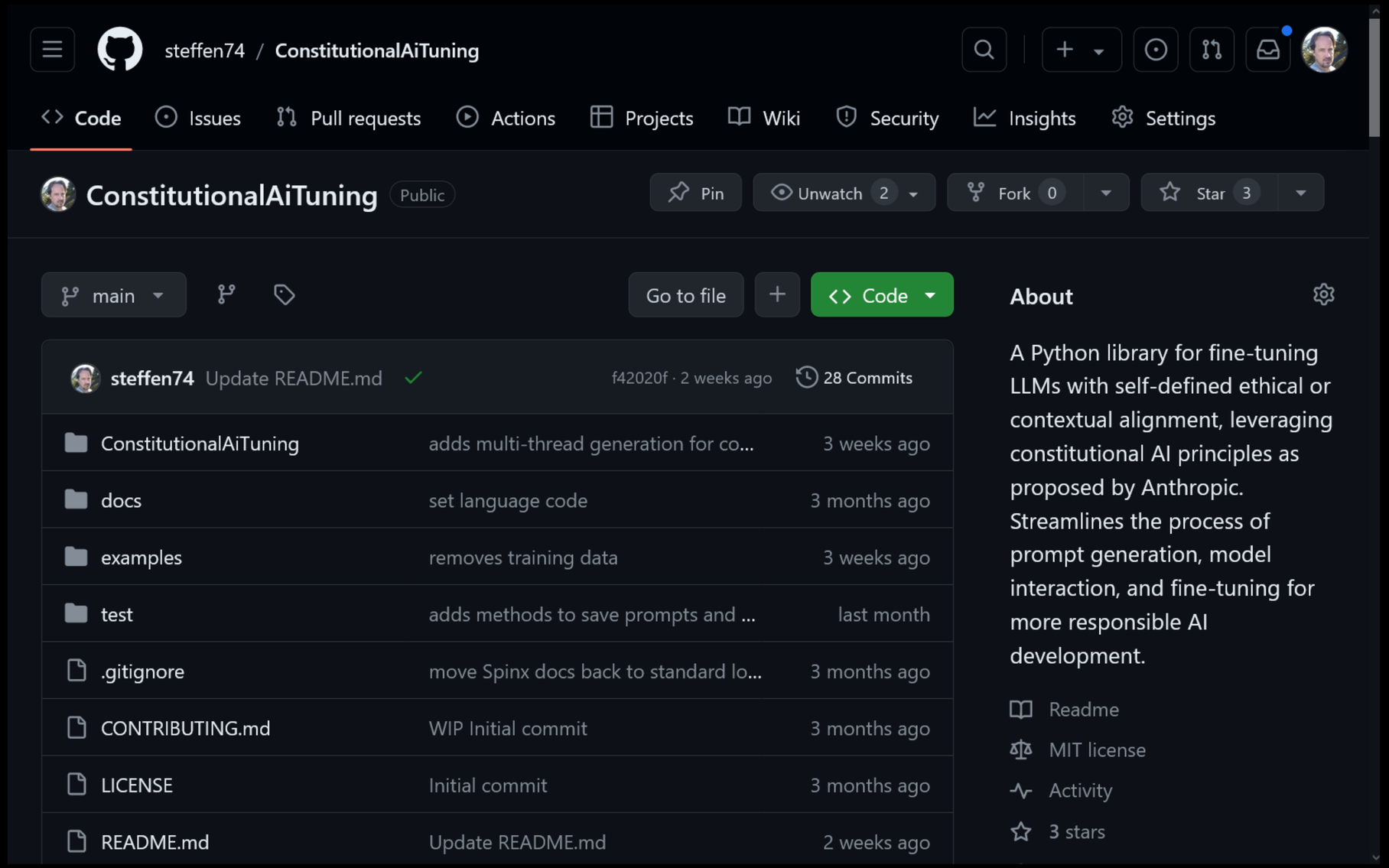
- **Goal:**
  Administrative German is a technical language not very well covered by LLMs. The tuned model should understand this technical language very well, but the generated summarizations, explanation of terms, etc. should be produced in "normal German".

- **Data:**
  - **service descriptions from municipalities/cities (XZUFI, an XML format)**
  - **Public websites**
  - **Possibly regulations, work instructions, etc.**



PROJECT IDEAS

CONSTITUTIONAL AI FINE-TUNING



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

```python
# 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')

# Instatiate 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_assistant_sft.csv')
```

Brandt, S. (2024). *Steffen74/ConstitutionalAiTuning* [Python]. Abgerufen von https://github.com/steffen74/ConstitutionalAiTuning

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 Transformes and Attention (about 50 minutes in total) if the transformers architecture is new to you.