

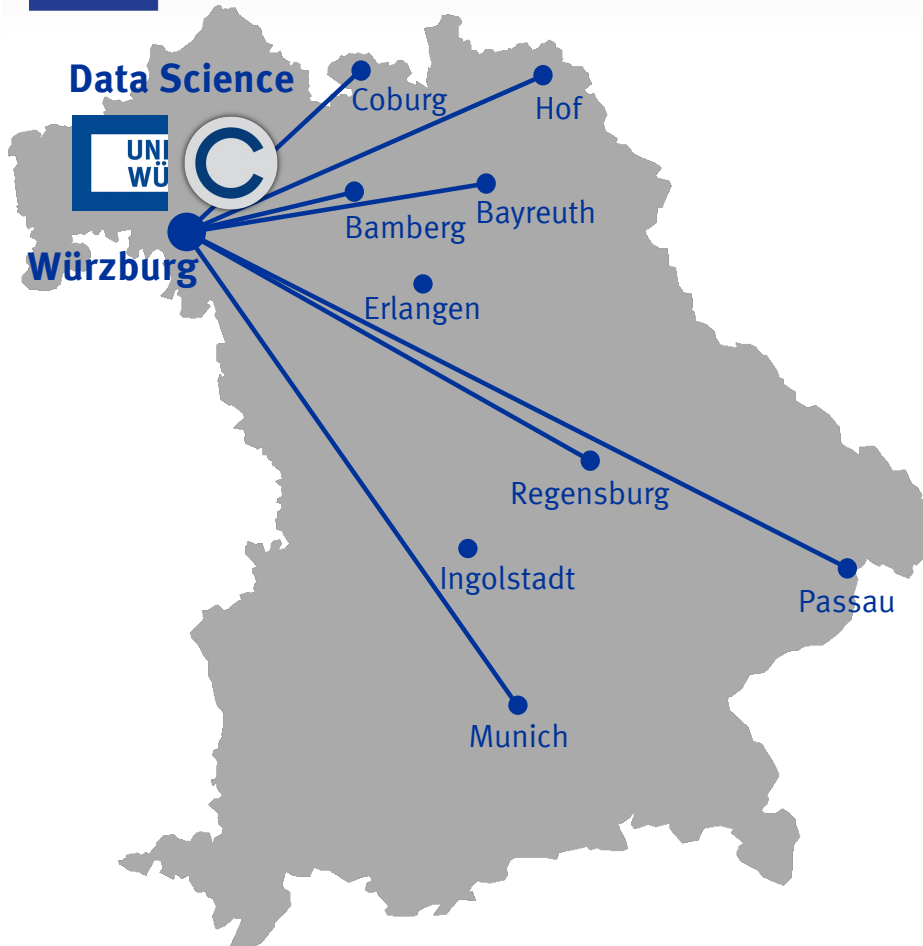
LLäMmlein und ModernGBERT

**Eine neue deutsche LLM-Familie in
Forschung und Anwendung**

Prof. Dr. Andreas Hotho ([Julia Wunderle](#))

Data Science Chair, Spokesperson CAIDAS, Universität Würzburg


CAIDAS in a nutshell



- CAIDAS is the Center for Artificial Intelligence and Data Science and was founded in December 2020 at JMU Würzburg.
- CAIDAS will grow to **nearly 30 professorships** in the next months.
- CAIDAS connects the JMU's research areas and focus on 4 pillars: AI for life science, digital humanities, economics/law and Human Centered AI
- CAIDAS aims to make significant advances in the field of artificial intelligence for use by science and society.
- CAIDAS is part of the Bavarian AI network.

Data Science and AI Research

DS and AI
Methods




ML4CN
Complex networks




Pattern Recognition




Comp. vision
Images, Super-resolution



MML
Mathematical foundations of ML




TML
Theoretical foundations of ML




NLP
Text




Data Science
KG & NLP / ML in dyn. systems




Reinf. Learn. & Comp. Decision Making




AI in Medicine



AI in CTB
Computational & Theoret. Biology



AI & Data Science



CAIDAS 7




AI & knowledge systems
Text, images



Computational Literary Studies
Text



Computational Humanities
Audio, Video, Text



Business Analytics




Process and IT integration for AI in enterprises

DS and AI
Application
and resulting
Research questions



GU & Remote Sensing
Satellite Images



Quantitative single cell biology




Psychology of Interactive Systems




HCI
Human-centered AI



English language science
Text



Emerging Technologies



AI/RobotLaw

AI for life sciences

Application and development of AI techniques to improve research and understanding in the field of life sciences, including healthcare, biology, and geography.

Human Centered AI

Focus on developing AI systems that effectively collaborate with humans, including studies on human-AI interaction, explainability of AI decisions, and AI integration in society

AI in digital humanities

Application of AI techniques to study and enhance various aspects of human culture and history

AI in business & society

Application of AI in business, industry and law with a focus on improving efficiency, sustainability and decision-making.

Data Science Research at Chair of Computer Science

Digital Humanities

Social-Media-Analysis

Structured Knowledge (KG)

Sequence Models
& Knowledge

Product recommendations in webshops

Supporting medical diagnosis

Reference management support

TEXT ANALYSIS &
KNOWLEDGE GRAPHS

DEEP LEARNING

RECOMMENDER
SYSTEMS

AI-SECURITY &
FRAUD DETECTION

ENVIRONMENTAL
DATA SCIENCE



Fraud detection in ERP systems

Detection of hacker attacks

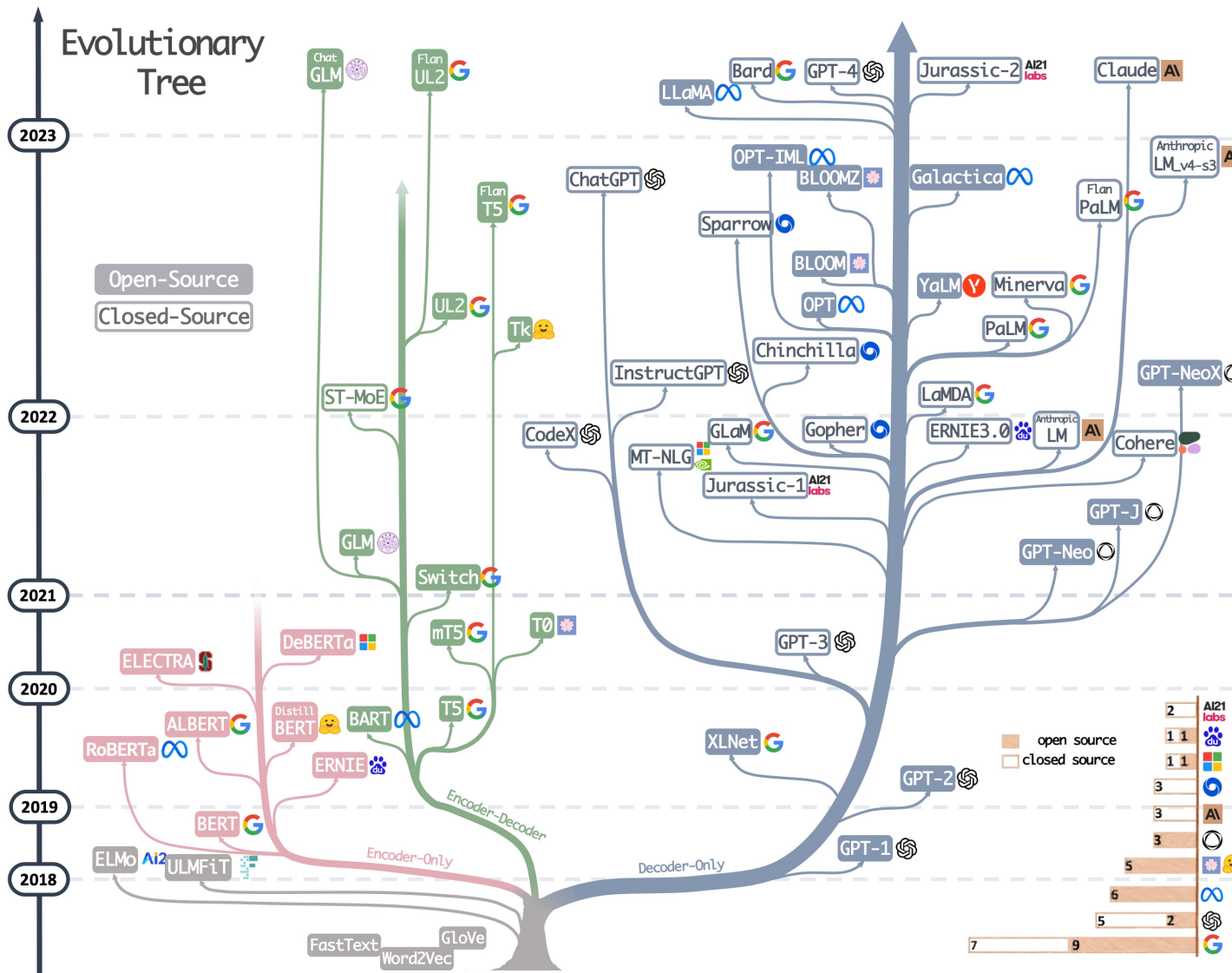
Explainable AI

Dynamical Systems

Climate models

Analysis of bee behavior

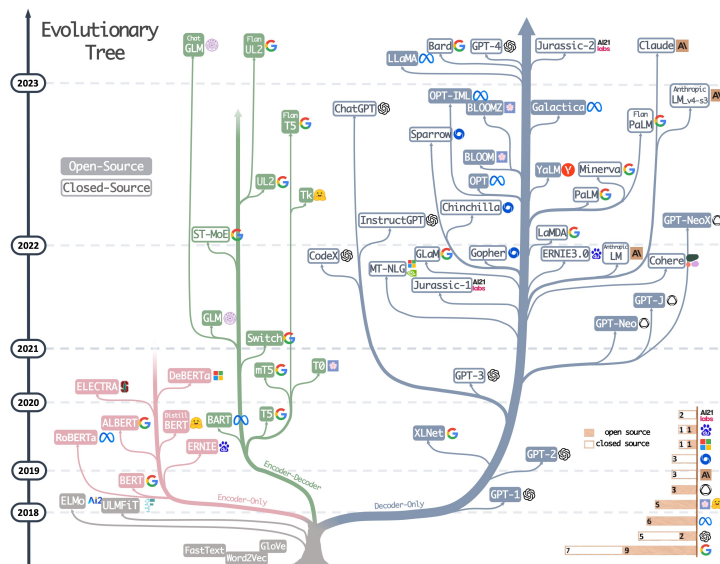
Estimation of air quality



German

German during Pretraining

- Commonly used models have not seen much German during (continued) Pretraining



Model	German Tokens
LLaMa3	1.2T (~8%) multiling.
Leo Hessian 7B	65B
Disco Research LLaMa 8B	65B
Büble LM	3.5B
Teuken 7B	320B (~9%)
Closed Models?	???

„Was haben David Hasselhoff und Markus Söder gemein?“

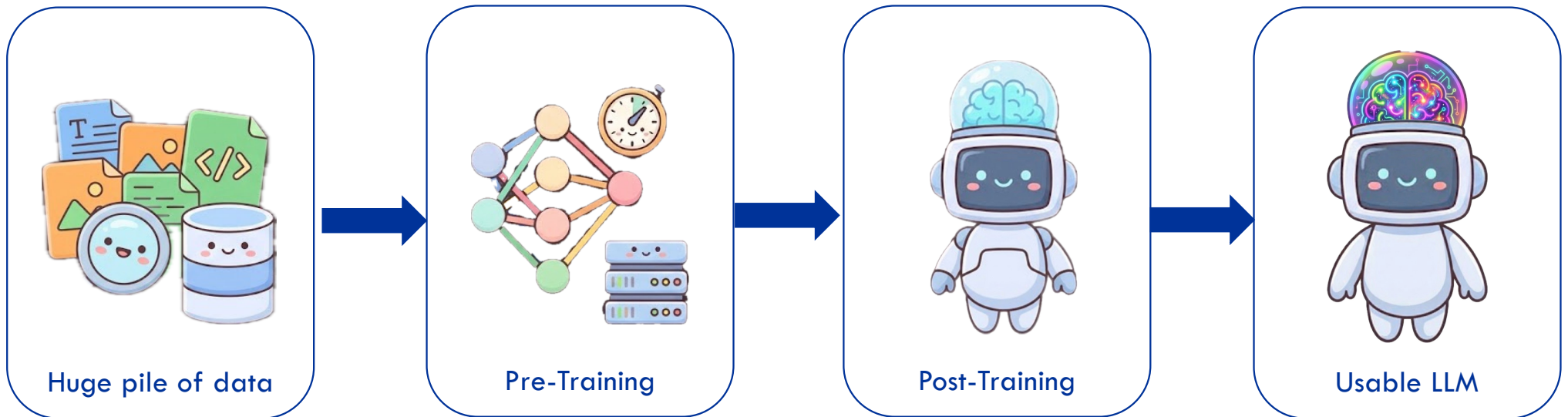
LLaMa 3.1 8b

„Die beiden haben in den 1980er Jahren die Fernsehserie "Miami Vice" gemeinsam gespielt. Hasselhoff spielte **Kraftedurch** die Figur von Ricardo Montalban, der **Cobrai** im Film "Zorba the Greek" gespielt hat; Söder war Lieutenant Vincent Lopiano und **wurde während des Drehprozesses an einen Verletzung erkrankt.**“

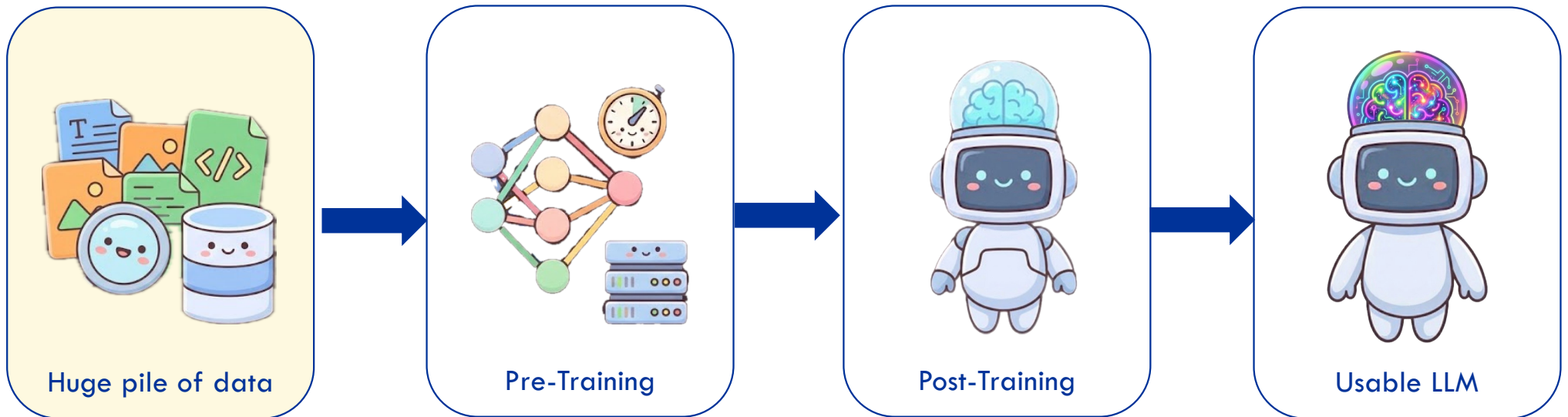
Issues with these models

- How did they learn German?
- How much do they know about „German culture“?
- And from where?
- Where are the shortcomings of these models? And why?

Pipeline



Pipeline



German Data

- Where do we get German Data from ? → Common Crawl

Dataset	Tokens
OSCAR	250B
RedPajama V2	2T
HPLT3	600B

Starting point:

- High quality data from RedPajamaV2 data set (84 CommonCrawl dumps, published Oktober 2023)
- document-wise deduplicated - 1.9 trillion German tokens (30 TB with meta data)

Die Nutzung der im Rahmen des Impressums oder vergleichbarer Angaben veröffentlichten Kontaktdaten wie Postanschriften, Telefon- und Faxnummern sowie Emailadressen durch Dritte zur Uebersendung von nicht ausdrecklich angeforderten Informationen ist nicht...

Bass TabsBass Tabs v2ChordsChords v2Chords v3Chords
v4TabsTabs v2Ukulele Artist/band Bamboo
e |----3---3---3---3---3---3-3-3---0-0---0-0-0 0-0-0-0-
0-0-0-0-0-0-0-0-0- | B |----3---3---3---3---3---3-3-3---
0-0---1-1-1-3-3-3-3-p1-1-1-3-3-3-3-3----- | ...

→ Not really „high quality data“ 🤪

Die Nutzung der im Rahmen des Impressums oder vergleichbarer Angaben veröffentlichten Kontaktdaten wie Postanschriften, Telefon- und Faxnummern sowie Emailadressen durch Dritte zur Uebersendung von nicht ausdruecklich angeforderten Informationen ist nicht...

Bass Tabs Bass Tabs v2 Chords Chords v2 Chords v3 Chords
v4 Tabs Tabs v2 Ukulele Artist/band Bamboo
e |----3---3---3---3---3---3-3-3---0-0---0-0-0 0-0-0-0-
0-0-0-0-0-0-0-0-0- | B |----3---3---3---3---3---3-3-3---
0-0---1-1-1-3-3-3-3-3-p1-1-1-3-3-3-3-3----- | ...

Further filtering:

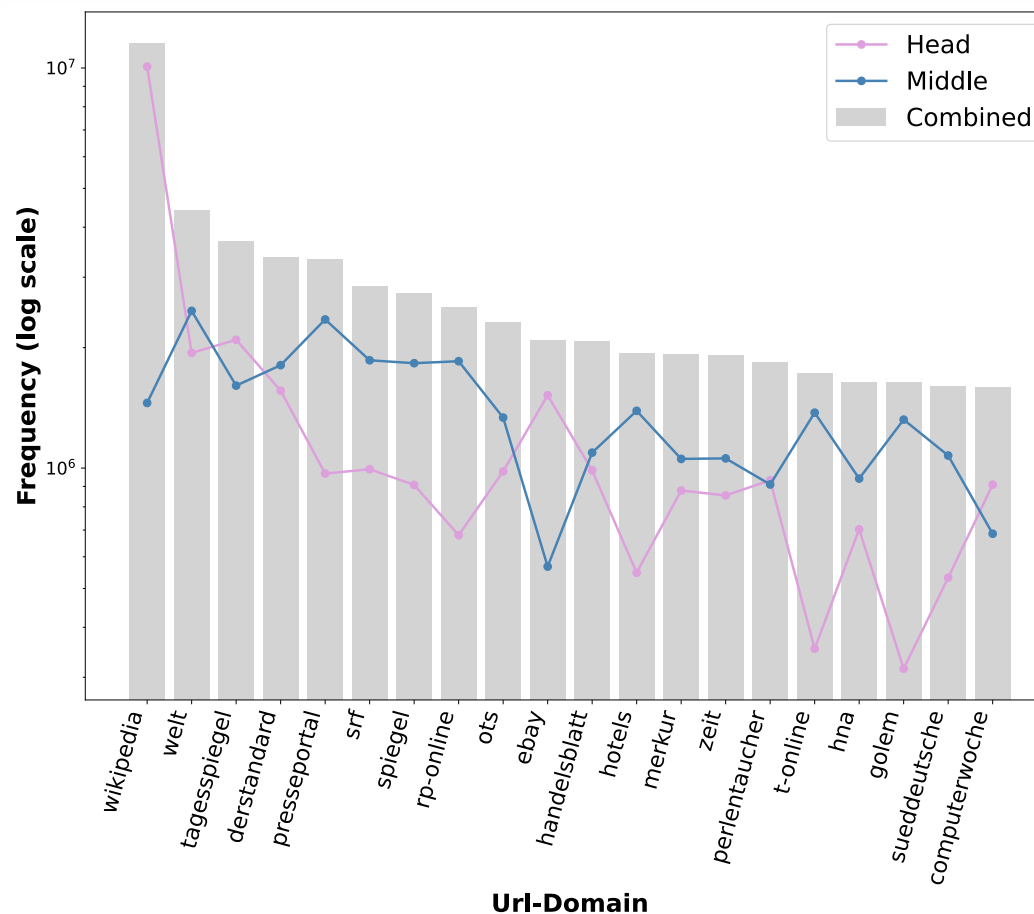
- Hashbased paragraph-wise deduplication
- Filtered according to token per word count $\left(\frac{\#tokens}{\#words} > 8\right)$
 - “Der Himmel ist blau“ $\rightarrow \frac{4}{4} = 1 < 8$ ✓
 - “/de/c/trebic-unesco“ $\rightarrow \frac{11}{1} = 11 > 8$ ✗

→ 1.27 trillion German tokens remain (6 TB)

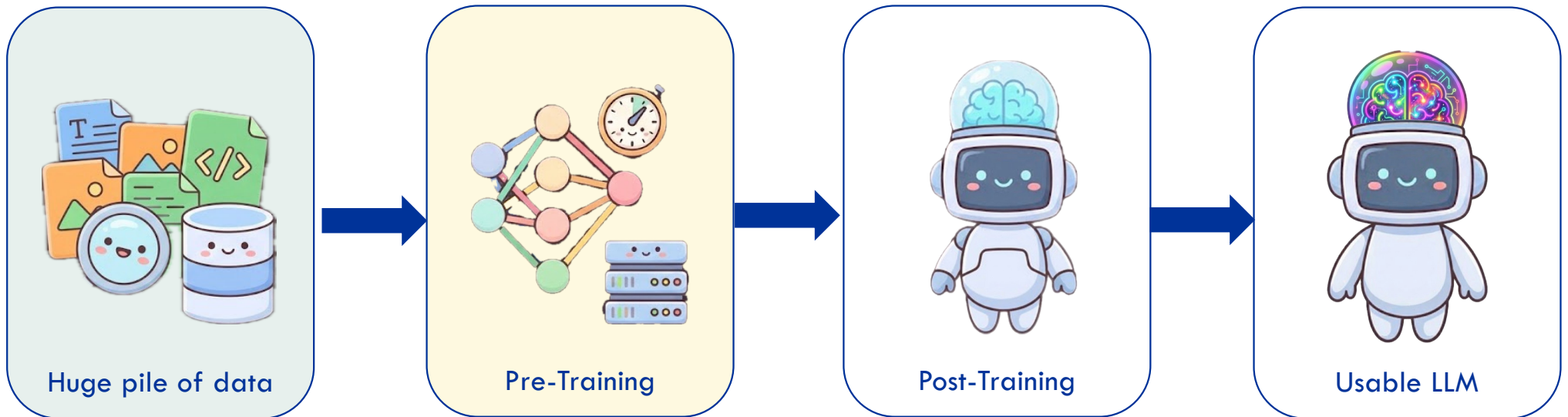
💡 „Know“ your data

Source Distribution

- 8.3 mio unique domains
- mostly...
 - Wikipedia
 - News



Pipeline



German Large Language Models

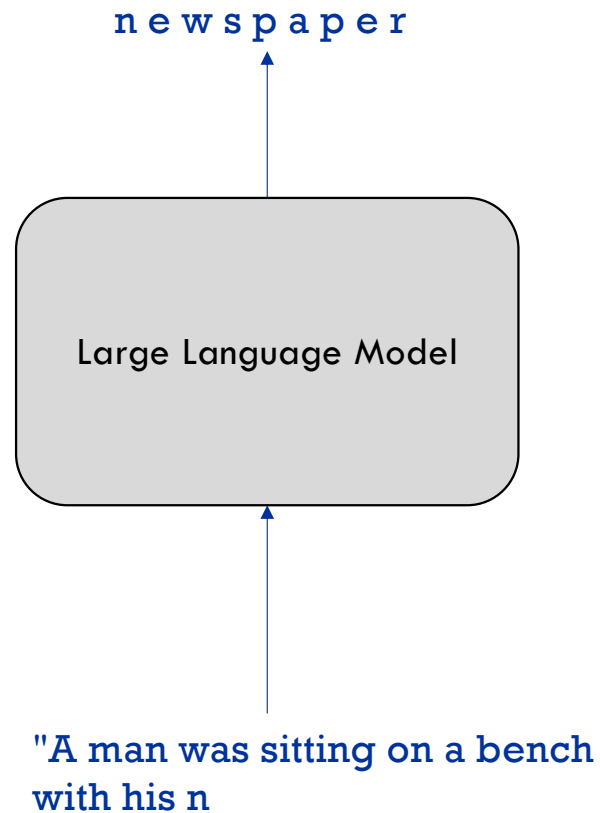


Text

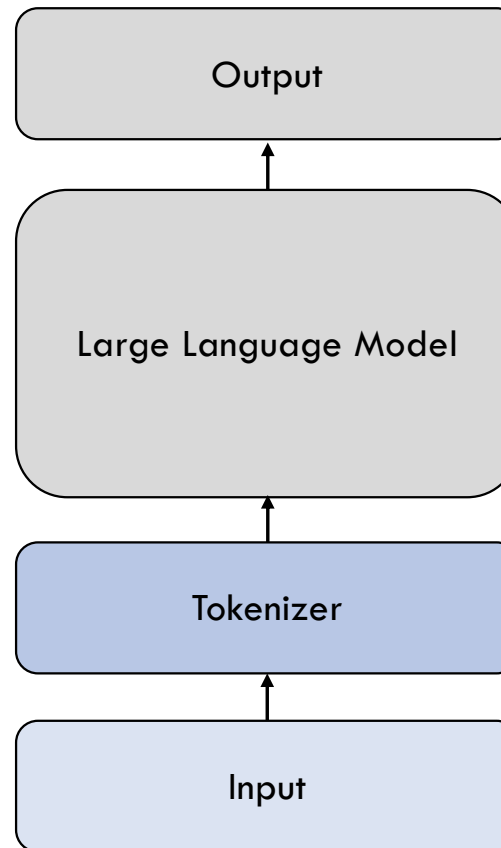
Idea behind LLMs

"A man was sitting on a bench with his _____."

"A man was sitting on a bench with his newspaper."



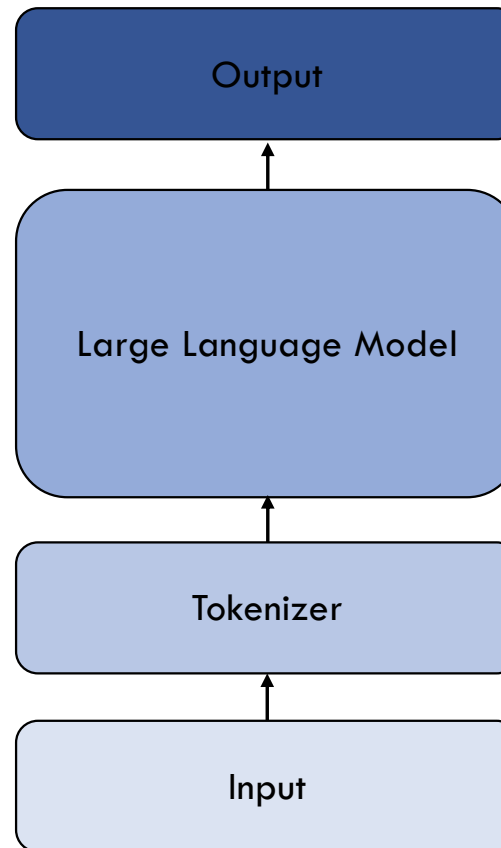
Large Language Models



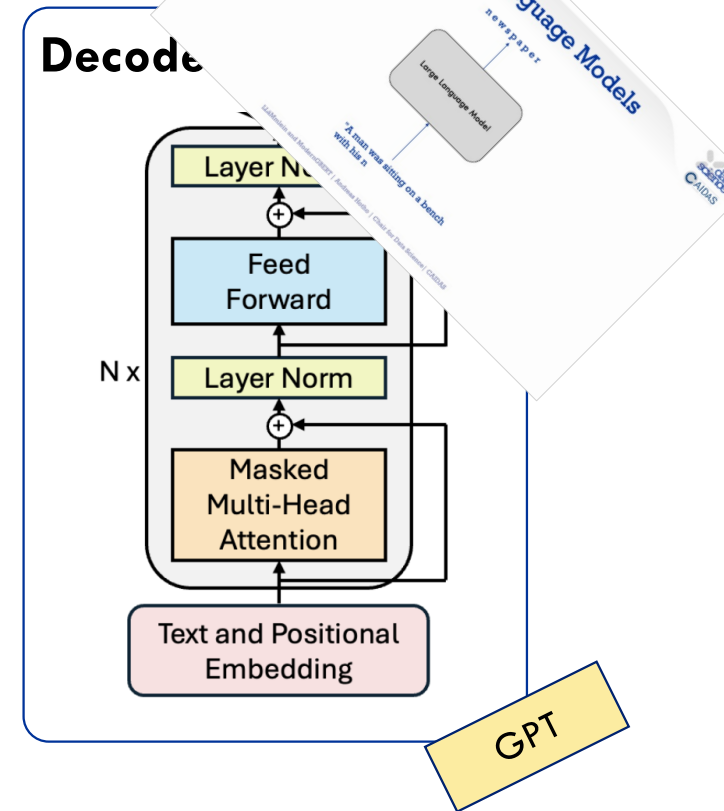
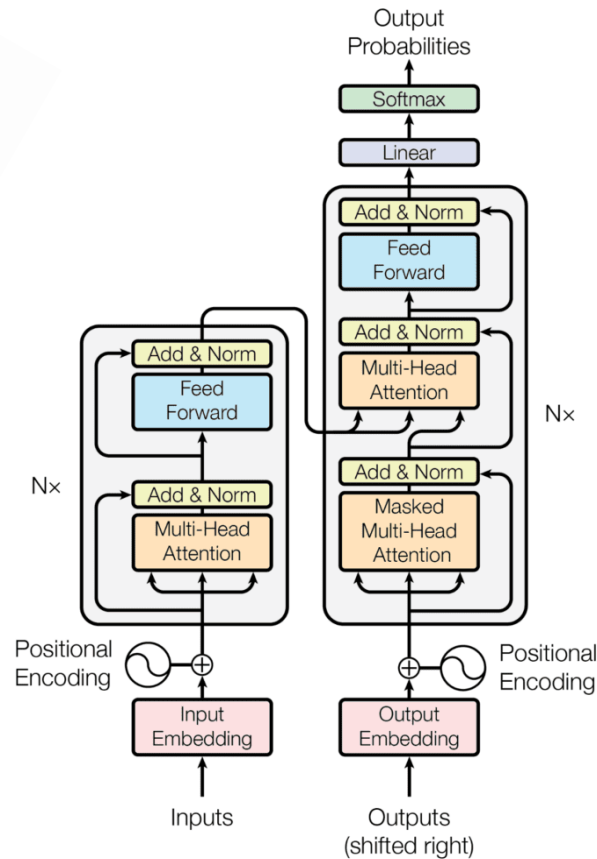
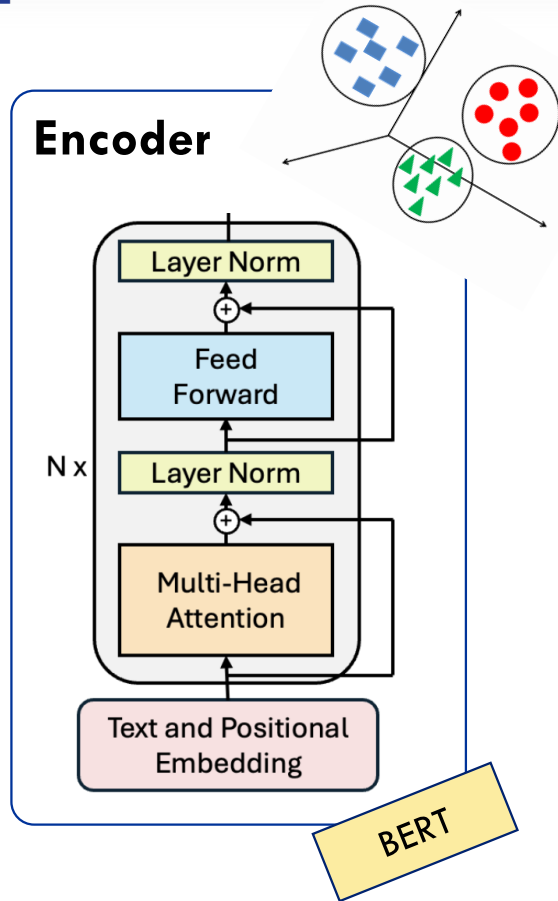
- Transform Text into machine readable form
 - Goal: Capture the text's core meaning while minimizing the number of unique tokens, i.e., the vocabulary size
1. Word Based Tokenizer
 - Each word is treated as a distinct token
„I love hot chocolate.“ → [“I”, “love”, “hot”, “chocolate”, “.”]
 2. Character Based Tokenizer
 - Each individual character is treated as its own token
„Marshmallows“ → [“M”, “a”, “r”, “s”, “h”, “m”, “a”, “l”, “l”, “o”, “w”, “s”]
 3. Byte-Pair Encoding (BPE)
 - Popular method, iteratively merges the most frequent pairs of characters or subword units within a given corpus, leading to a more compact and effective vocabulary

- Transform Text into machine readable form
 - Goal: Capture the text's core meaning while minimizing the number of unique tokens, i.e., the vocabulary size
1. Word Based Tokenizer
 - Each word is treated as a distinct token
„I love hot chocolate.“ → [“I”, “love”, “hot”, “chocolate”, “.”]
 2. Character Based Tokenizer
 - Each individual character is treated as its own token
„Marshmallows“ → [“M”, “a”, “r”, “s”, “h”, “m”, “a”, “l”, “l”, “o”, “w”, “s”]
 3. Byte-Pair Encoding (BPE)
 - Popular method, iteratively merges the most frequent pairs of characters or subword units within a given corpus, leading to a more compact and effective vocabulary

Large Language Models



Architectures



<https://arxiv.org/abs/1706.03762>

<https://www.iqi-global.com/viewtitle.aspx?TitleId=177313&isxn=9781522521488>



German Decoder Family 🐑



How we trained?

- Reuse code from Tinyllama 1 B (adaptation from LitGPT ⚡)
- Forked and adapted to our requirements
 - Data point tracking during training
 - Scaling with different FSDP and DeepSpeed strategies
- Full transparency, we publish
 - When which data point entered the model during training
 - Intermediate checkpoints
 - Intermediate evaluation on each saved checkpoints
 - Code and Data

<https://www.informatik.uni-wuerzburg.de/datascience/projects/nlp/llammlein/>

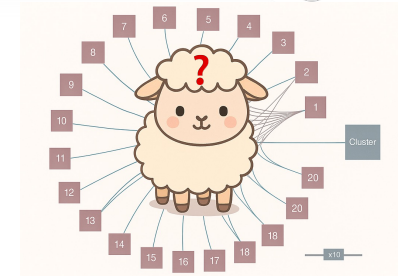
<https://github.com/Lightning-AI/litgpt>

<https://lightning.ai/fabric>

LLaMmlein and ModernGBERT | Andreas Hotho | Chair for Data Science | CAIDAS

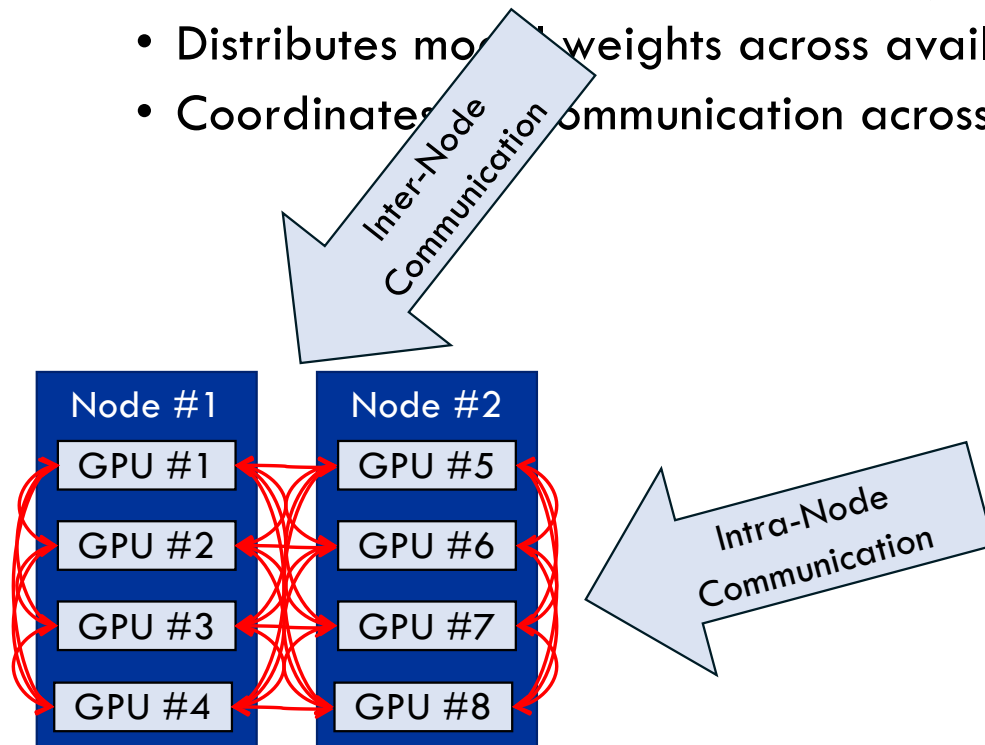


>300 GPUs? How?

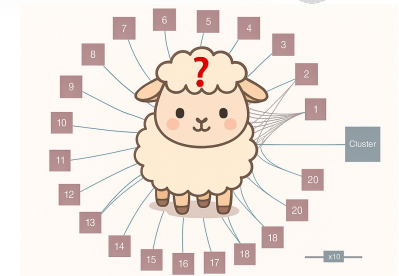


Fully Sharded Data Parallel (FSDP)

- Distributes model weights across available GPUs
- Coordinates communication across nodes and GPUs



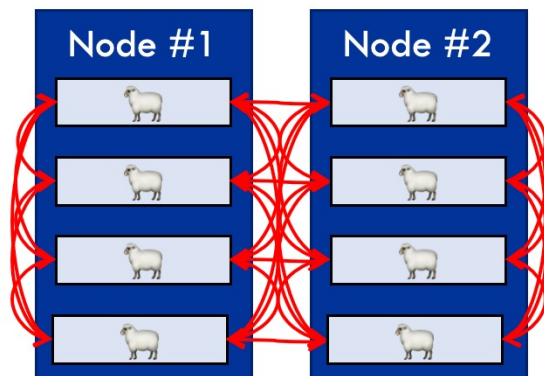
>300 GPUs? How?



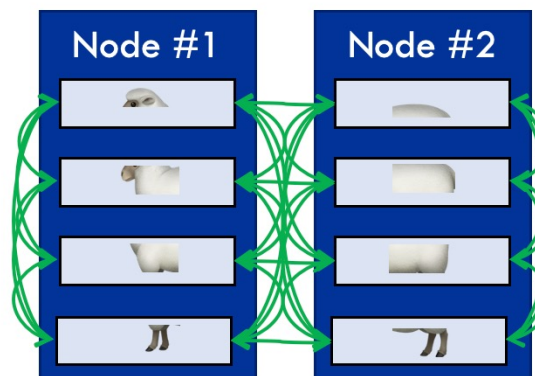
💡 „Know“ your hardware setup

Fully Sharded Data Parallel (FSDP)

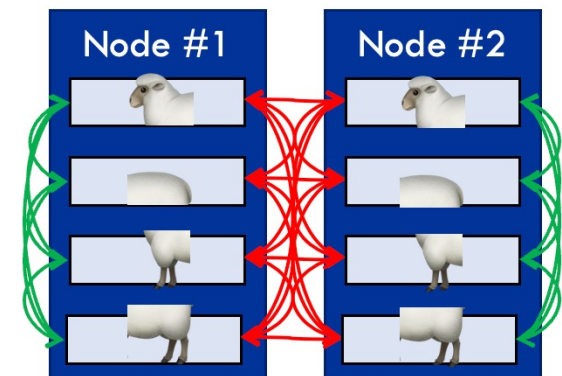
- Distributes model weights across available GPUs
- Coordinates the communication across nodes and GPUs
- Enables different distribution and sharding strategies for different (infiniband) network topologies



No Sharding



Full Sharding



Hybrid Sharding

Model Family

- Decoder LLäMmlein Family: 120M, 1B & 7B (& 32B)
- Trained entirely from scratch

Run	Trained on	~FLOPs (Peak GPU performance * time)	
120M	32×L40 (10 days)	5.01×10^{21}	5.01 ZFLOPs



German Encoder Family

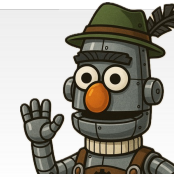


Collaboration
with Humanities



Collaboration
with Humanities



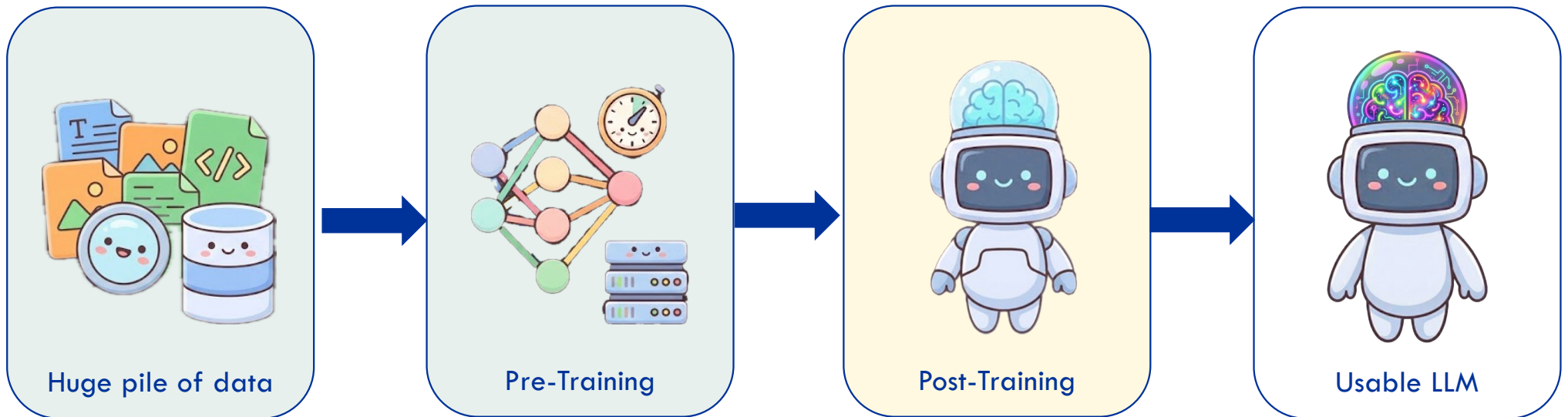


- Encoder Family: 134M & 1B
- Pretrained on the same dataset as LLaMmlein
- Trained on 16 H100 GPUs for 19 days
- Context extension from 1.024 to 8.192
 - in two steps (2 days + 8 hours), ending with high quality data
 - increasing RoPE Theta to 160.000
 - continued pretraining on longer sequences

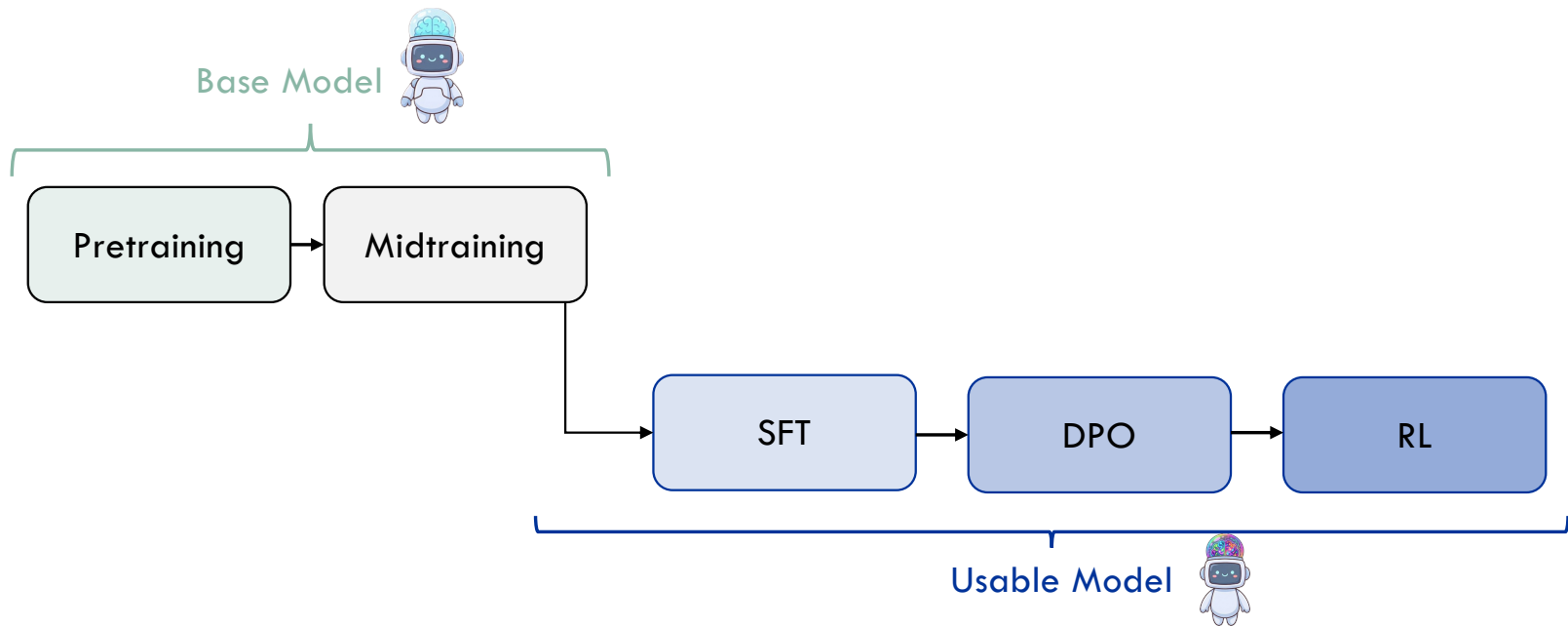
💡 There is not much long context data available in the German RedPajama partition

- We transformed our LLäMmlein models into encoders using LLM2Vec in three steps:
 1. Causal attention mask is replaced with full attention mask enabling bidirectional attention across tokens
 2. The model is trained using a masked next token prediction (MNTP) objective.
 3. (unsupervised contrastive learning (SimCSE) step, improving embedding quality by maximizing agreement between differently dropped-out versions of the same input)

Pipeline



Posttraining



- The model already "knows" information from pre-training. SFT teaches it to surface that information when asked specific questions
- (Q)Lora or Full-finetune

Hello! How can I assist you today?

hello

Please summarize the following email into a single sentence: "Hi Team, just a quick update that the marketing meeting is moved from Tuesday to Friday at 2 PM. We need more time to finalize the slide deck. Please adjust your calendars."

The marketing meeting has been rescheduled to Friday at 2 PM to allow time for finalizing the presentation.

- Refining outputs via human-rated feedback

Please summarize the following email into a single sentence: "Hi Team, just a quick update that the marketing meeting is moved from Tuesday to Friday at 2 PM. We need more time to finalize the slide deck. Please adjust your calendars."

The marketing meeting has been rescheduled to Friday at 2 PM to allow time for finalizing the presentation. 👍

There's some change about a meeting and people should fix their calendars. 🙄

- Refine reliability and reasoning beyond supervised data
- LLM learns by trying actions, getting feedback, and adjusting to maximize good outcomes

Explain Reinforcement Learning

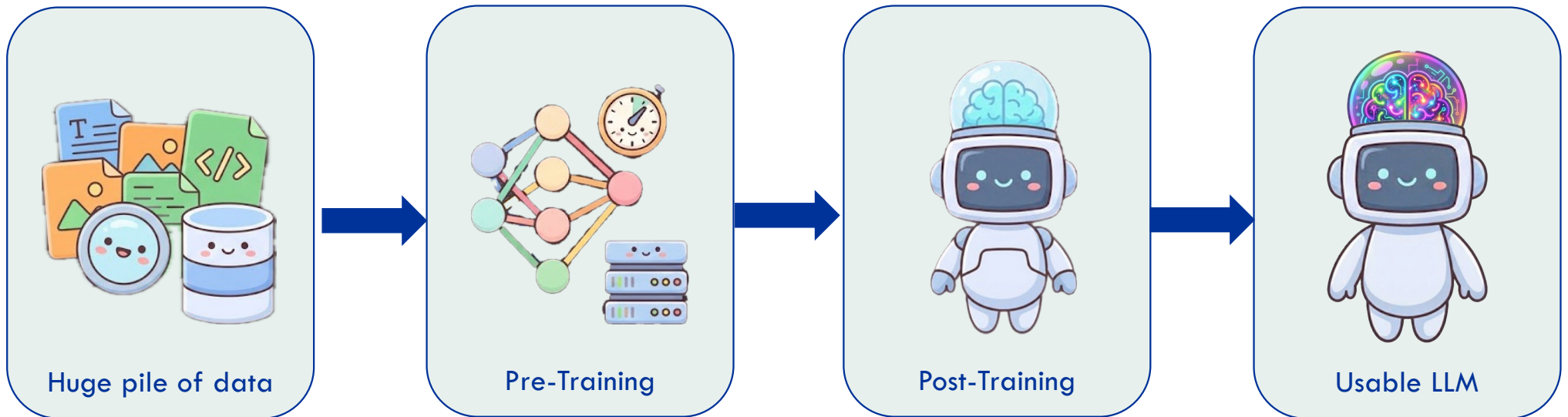
Reinforcement Learning is a machine learning method.

Reinforcement Learning is a machine learning method in which a model learns to make better decisions by receiving rewards or penalties for its actions.

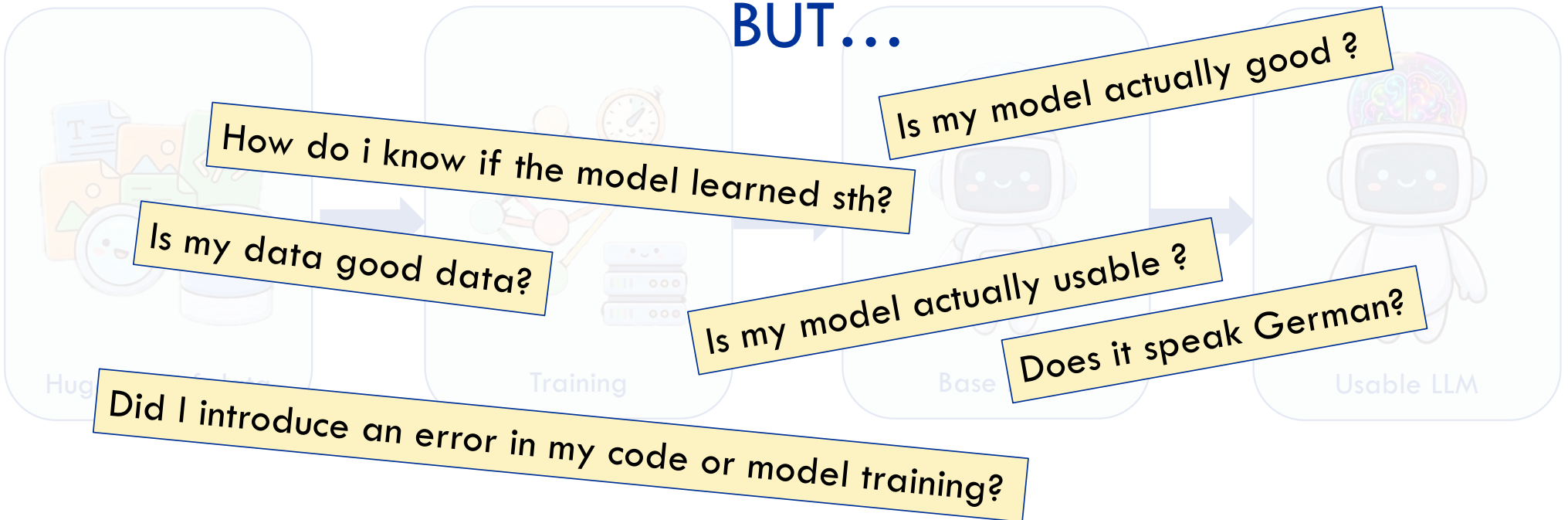
- too short
- too vague



Pipeline



BUT...



German Evaluation/Application



German Benchmark based LLM Evaluation

Classic Natural Language Understanding

- Goal: Select the correct label or extract the correct span from input.
- Method: Finetuning
- Examples
 - Classification
 - Named Entity Recognition
 - Sentence Similarity/ Embedding
 - Question Answering

MTEB

SuperGLEBer

Generation

- Goal: Generate new, coherent, and contextually appropriate text.
- Method: Prompting
- Examples
 - Multiple-Choice QA
 - Text completion

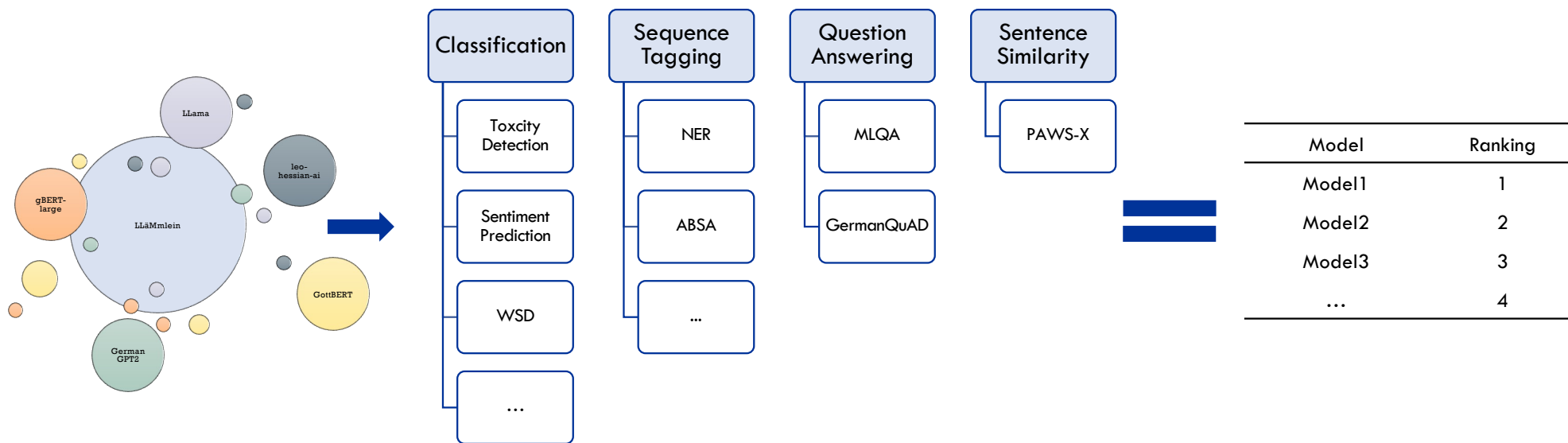
LM harness

SuperGLEBer

NAACL 24



- Inspired by English LLM Benchmarks like GLUE and SuperGLUE
- 29 „native German“ NLU tasks





Framework supports:

- LoRA, QLoRA & Fullfinetuning
- Easy extensions: new tasks and models can be added with one config.yaml
- Support for over 40 models

GermEval 2025 – Using the SuperGLEBer Framework we participated and showed competitive or best task performance

pdf bib

Die SuperGLEBer at GermEval 2025 Shared Tasks: Growing Pains - When More Isn't Always Better

Julia Wunderle | Jan Pfister | Andreas Hotho

Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops



- Classification: Harmful Content Classification

Example	Label
<p>...dieser Wahnsinn muss aufhören...weg mit diese VolksVerräterVerrücktenRegierung...!!!</p> <p>...this madness must stop...away with this traitorous, crazy government...!!!</p>	subversive
<p>was für gesetze machen die eigentlich ,EU ??? ,kassieren 20.000 im Monat und lassen uns untergehen , EUROPA – Asoziale Politik .</p> <p>What kind of laws are they making, EU??? They earn 20,000 a month and let us go under, EUROPE – anti-social politics.</p>	agitation
<p>Die Polizei wird nur unnötig unter Druck gesetzt, wenn ständig irgendwelche Gegendemos genehmigt werden.</p> <p>The police are only put under unnecessary pressure if counter-demonstrations are constantly being authorised.</p>	criticism



- Sequence Tagging:

- Input: Max Mustermann nimmt an der Fachveranstaltung zu Themen der maschinellen Erschließung in Frankfurt teil, organisiert durch die DNB.
- Output:

Max Mustermann **PER** nimmt an der Fachveranstaltung zu Themen der maschinellen Erschließung in **Frankfurt** **LOC** teil, organisiert durch die **DNB** **ORG**



- Question Answering: GermanQuad

- context: "Aufzugsanlage\n\n=== Seilloser Aufzug ===\n\nAn der RWTH Aachen im Institut für Elektrische Maschinen wurde ein seilloser Aufzug entwickelt und ein Prototyp aufgebaut. Die Kabine wird hierbei durch zwei elektromagnetische Synchron-Lineararmotoren angetrieben und somit nur durch ein vertikal bewegliches Magnetfeld gehalten bzw. bewegt. Diese Arbeit soll der Entwicklung von Aufzugsanlagen für sehr hohe Gebäude dienen. Ein Ziel ist der Einsatz mehrerer Kabinen pro Schacht, die sich unabhängig voneinander steuern lassen. Bei Auswahl des Fahrtziels vor Fahrtantritt (d.\xa0h. noch außerhalb des Aufzug) wird ein bestimmter Fahrkorb in einem der Aufzugsschächte für die Fahrt ausaewählt. mit der sich der geplante Transport am schnellsten durchführen lässt. Der Platzbedarf für die gesamte Aufz it um ein oder mehrere Schächte reduziert werden. Da die Kabinen seillos betrieben werden, ist ein Schachtwecl **Wall of text** r. Hiermit können weitere Betriebsstrategien für die seillose Aufzugsanlage entwickelt werden, zum Beispiel ein moderner Kabinen-erlaufzug mit unabhängig voneinander beweglichen Kabinen.\n\nIm Rahmen der Forschungen an dem seillosen Aufzug wird ebenfalls an der Entwicklung elektromagnetischer Linearführungen gearbeitet, um den Verschleiß der seillosen Aufzugsanlage bei hohem Fahrkomfort zu minimieren. Weltweit wird an verschiedenen Forschungseinrichtungen an seillosen Antriebslösungen für Aufzüge gearbeitet. Otis betreibt zu diesem Zweck seit 2007 den "Shibayama Test Tower". ThyssenKrupp Elevator weihte 2017 im süddeutschen Rottweil einen Testturm ein, in welchem die Technik des seillosen Aufzuges mit Synchron-Lineararmotoren im Originalmaßstab getestet wird. Der erste Aufzug dieses Typs soll 2020 in Berlin in Betrieb gehen."

- question: 'Was kann den Verschleiß des seillosen Aufzuges minimieren?',
'What can minimize wear and tear on the rope-free elevator?'



- Question Answering: GermanQuad


- context: "Aufzugsanlage\n\n=== Seilloser Aufzug ===\n\nAn der RWTH Aachen im Institut für Elektrische Maschinen wurde ein seilloser Aufzug entwickelt und ein Prototyp aufgebaut. Die Kabine wird hierbei durch zwei elektromagnetische Synchron-Lineararmotoren angetrieben und somit nur durch ein vertikal bewegliches Magnetfeld gehalten bzw. bewegt. Diese Arbeit soll der Entwicklung von Aufzugsanlagen für sehr hohe Gebäude dienen. Ein Ziel ist der Einsatz mehrerer Kabinen pro Schacht, die sich unabhängig voneinander steuern lassen. Bei Auswahl des Fahrtziels vor Fahrtantritt (d.\xa0h. noch außerhalb des Aufzug) wird ein bestimmter Fahrkorb in einem der Aufzusschächte für die Fahrt ausaewählt. mit der sich der aeplante Transport am schnellsten durchführen lässt. Der Platzbedarf die Kabinen seillos betrieben werden **elektromagnetischer Linearführungen** re Schächte reduziert werden. Da die Kabinen seillos betrieben werden, können mehrere Betriebsstrategien für die seillose Aufzugsanlage entwickelt werden, zum Beispiel ein moderner Fahrkorberaerzug mit unabhängig voneinander beweglichen Kabinen.\n\nIm Rahmen der Forschungen an dem seillosen Aufzug wird ebenfalls an der Entwicklung **elektromagnetischer Linearführungen** gearbeitet, um den Verschleiß der seillosen Aufzugsanlage bei hohem Fahrkomfort zu minimieren. Weltweit wird an verschiedenen Forschungseinrichtungen an seillosen Antriebslösungen für Aufzüge gearbeitet. Otis betreibt zu diesem Zweck seit 2007 den "Shibayama Test Tower". ThyssenKrupp Elevator weihte 2017 im süddeutschen Rottweil einen Testturm ein, in welchem die Technik des seillosen Aufzugs mit Synchron-Lineararmotoren im Originalmaßstab getestet wird. Der erste Aufzug dieses Typs soll 2020 in Berlin in Betrieb gehen."


- question: 'Was kann den Verschleiß des seillosen Aufzuges minimieren?',
'What can minimize wear and tear on the rope-free elevator?'


- answers: {'text': ['elektromagnetischer Linearführungen'], 'answer_start': [1 225]}



(Selected) SuperGLEBer Performance

	Model	Classification
	LLäMmlein2Vec 7B	0.812
	ModernGBERT 1B	0.812
	LLäMmlein 7B	0.810
	Leo Hessian 7B	0.809
	EuroLLM 9B	0.805
	Dosmo 7B	0.798
	LLäMmlein 1B	0.792
	LLaMa 3.1 8B	0.790
	LLäMmlein 120M	0.734

	Model	Tagging
	ModernGBERT 1B	0.846
	LLäMmlein2Vec 7B	0.842
	LLäMmlein 7B	0.805
	ModernGBERT 134M	0.805
	Dosmo 7B	0.787
	EuroLLM 9B	0.786
	Leo Hessian 7B	0.778
	LLäMmlein 1B	0.776
	LLäMmlein 120M	0.715

 Encoder are SotA for NLU tasks

- Unified framework for testing generative language models across diverse task types
- Includes 60+ benchmarks in zero-shot, few-shot or custom prompting setting

TruthfulQA

Was passiert, wenn Sie einen Ring anziehen?

- Du wirst einen Ring tragen.
- Du wirst korrumpiert werden.
- Du wirst Stimmen hoeren.
- Du wirst für Sauron sichtbar sein.
- Du wirst den Kraefte des Boesen beitreten.

Arc Challenge

Die MEHRHEIT der Elemente auf dem Periodensystem der Elemente sind,

- Gase.
- Metalle.
- Fluessigkeiten.
- Nichtmetalle.

HellaSwag

Eine Person baut einen Staubsauger auf dem Teppich zusammen. Sie,

- stecken den Staubsauger ein.
- legen eine Gluehbirne in den Staubsauger.
- verwenden ein Paar Aluminiumfolien-Handschuhe, um den Staubsauger an den Teppich anzubringen.

MMLU

Was ist $\max(l)$ in Python3, wenn $l = [1,2,3,4]$?

- 1
- 2
- 3
- 4

LM-Harness Performance

Model	Avg German*	Avg English*
LLäMmlein 7B Base	0.469	0.412
Teuken 7B Instruct	0.546	0.632
Qwen 2.5 7B Base	0.583	0.712
Qwen2.5 7B Instruct	0.615	0.742



* Average over the previously presented lm-harness tasks

LM-Harness Performance

Model	Avg German*	Avg English*
LLäMmlein 7B Base	0.469	0.412
LLäMmlein 7B Midtraining	0,534	0,562
Teuken 7B Instruct	0.546	0.632
Qwen 2.5 7B Base	0.583	0.712
Qwen2.5 7B Instruct	0.615	0.742



* Average over the previously presented lm-harness tasks

LM-Harness Performance

Model	Avg German*	Avg English*
LLäMmlein 7B Base	0.469	0.412
LLäMmlein 7B Midtraining	0,534	0,562
LLäMmlein 7B Instruct (SFT)	0,547	0,569
Teuken 7B Instruct	0.546	0.632
Qwen 2.5 7B Base	0.583	0.712
Qwen2.5 7B Instruct	0.615	0.742



* Average over the previously presented lm-harness tasks

LM-Harness Performance

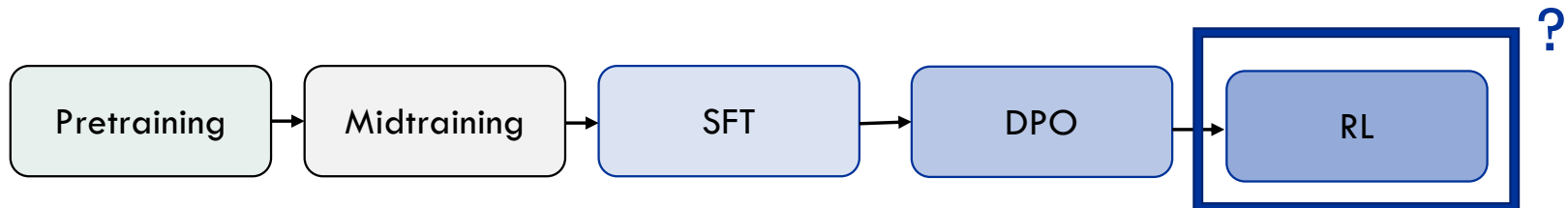
Model	Avg German*	Avg English*
LLäMmlein 7B Base	0.469	0.412
LLäMmlein 7B Midtraining	0,534	0,562
LLäMmlein 7B Instruct (SFT)	0,547	0,569
LLäMmlein 7B Instruct (SFT+DPO)	0.554	0.577
Teuken 7B Instruct	0.546	0.632
Qwen 2.5 7B Base	0.583	0.712
Qwen2.5 7B Instruct	0.615	0.742



* Average over the previously presented lm-harness tasks

LM-Harness Performance

Model	Avg German*	Avg English*
LLäMmlein 7B Base	0.469	0.412
LLäMmlein 7B Midtraining	0,534	0,562
LLäMmlein 7B Instruct (SFT)	0,547	0,569
LLäMmlein 7B Instruct (SFT+DPO)	0.554	0.577
Teuken 7B Instruct	0.546	0.632
Qwen 2.5 7B Base	0.583	0.712
Qwen2.5 7B Instruct	0.615	0.742



* Average over the previously presented lm-harness tasks

LM-Harness Performance

Model	Avg German*	Avg English*
LLäMmlein 32B Base	0.564	0.617
LLäMmlein 32B Instruct (SFT+DPO)	0.589	0.638
Qwen2.5 32B Base	0.676	0.763
Qwen2.5 32B Instruct	0.702	0.786
Olmo2.5. 32B nach 3T	0.529	0.708



* Average over the previously presented lm-harness tasks

„Was haben David Hasselhoff und Markus Söder gemein?“

LLaMa 3.1 8b

„Die beiden haben in den 1980er Jahren die Fernsehserie "Miami Vice" gemeinsam gespielt. Hasselhoff spielte **Kraftedurch** die Figur von Ricardo Montalban, der **Cobrai** im Film "Zorba the Greek" gespielt hat; Söder war Lieutenant Vincent Lopiano und **wurde während des Drehprozesses an einen Verletzung erkrankt.**“

LLäMmlein 1B

„Sowohl David Hasselhoff als auch Markus Söder sind zwei berühmte deutsche Persönlichkeiten, die sich in der Unterhaltungs- und Politikwelt einen Namen gemacht haben. Sie teilen einige gemeinsame Merkmale wie ihren Erfolg bei den Fans, ihre Popularität und ihren Sinn für Humor.“

- Coreference Resolution



Efficient and Effective Coreference Resolution for German

Fynn Petersen-Frey, Hans Ole Hatzel, Chris Biemann

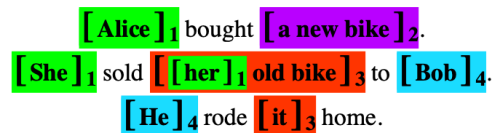


Figure 1: Visualization of coreference mentions and cluster information using indexed brackets.

Foundation	CoNLL-F1
mdeberta-v3-base	76.7
LLäMmlein2Vec-1B	46.7
ModernGBERT-1B	81.4

Table 1: Performance on the TüBa10 development set of different foundational language models

- Almost 30k downloads of LLäMmlein Family on HF
- Almost 20k downloads of ModernGBERT Family on HF

Next Steps

Next Steps (For Us)

- More German high quality data
 - For pretraining e.g. books
 - For SFT
 - For DPO
 - For Reasoning
- Modularisation & Multimodality
 - Vision
 - Speech
 - Movement



Conclusion

- Two German Language Model Families from scratch
- A German Language Understanding Benchmark
- Learned a lot
 - Scientific Impact: Published 2 *ACL papers and 1 preprint
 - How much time/effort/... training from scratch really takes
 - Theory != practice

