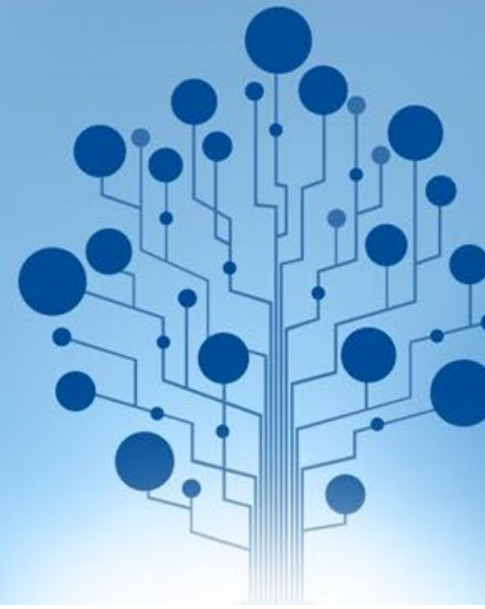




LLMs an der FernUni Hagen

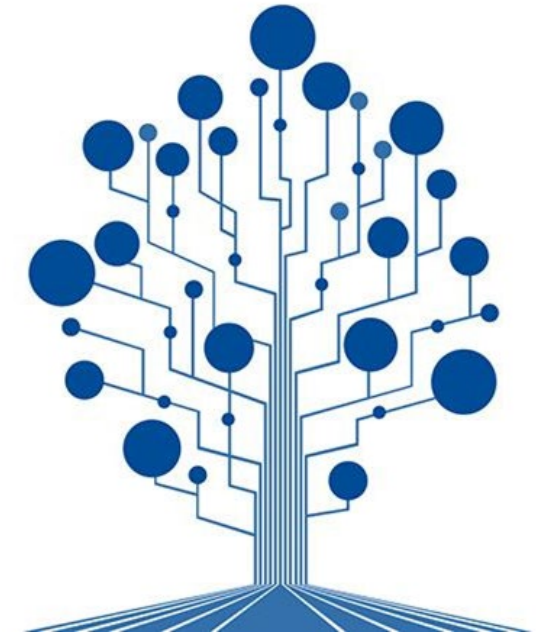
API-Lösung auf dem eigenen Server

M.Sc. Michael Hanses

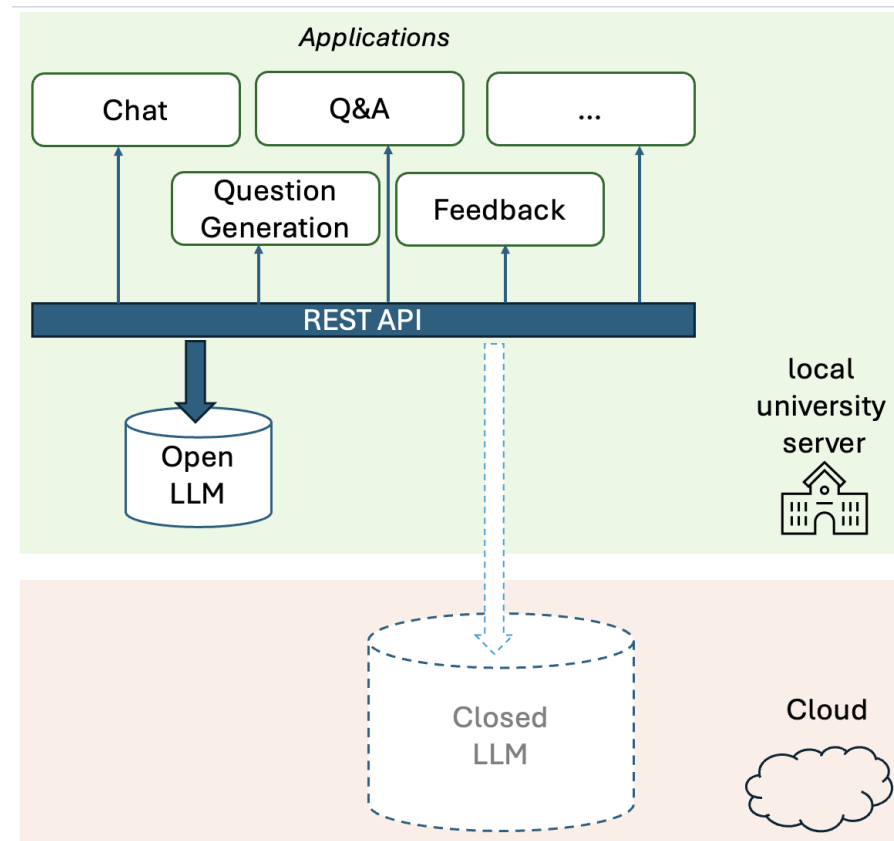
Prof. Dr.-Ing. Torsten Zesch

Motivation

- bisher kaum Evidenz zur Auswirkung von LLMs
- Experimente mit LLMs notwendig
 - insbesondere in der Lehre
- CATALPA Reallaboransatz
 - an der FernUni im echten Einsatz testen



FLEXI - Fernuni LLM Experimental Infrastructure



Was spricht für oder gegen den Betrieb von open-weight LLMs an einer Hochschule?

Pros

- Datenschutz
- Gezielte Modellauswahl
- Stabilität
- Kosten pro Anfrage

Cons

- Anschaffungskosten Server
- Installation & Wartung
- Know How

Zugangsmöglichkeiten

- API Access

```
> curl http://flexi.fernuni-hagen.de/api/tags
```

```
{
  "models": [
    {
      "name": "llama3:8b",
      "model": "llama3:8b",
      "modified_at": "2024-05-24T07:51:11.969810998Z",
      "size": 4661224676,
      "digest": "365c0bd3c000a25d28ddbf732fe1c6add414de7275464c4e4d1c3b5fcb5d8ad1",
      "details": {
        "parent_model": "",
        "format": "gguf",
        "family": "llama",
        "families": [
          "llama"
        ],
        "parameter_size": "8.0B",
        "quantization_level": "Q4_0"
      },
      "expires_at": "0001-01-01T00:00:00Z"
    },
  ],
}
```

Zugangsmöglichkeiten

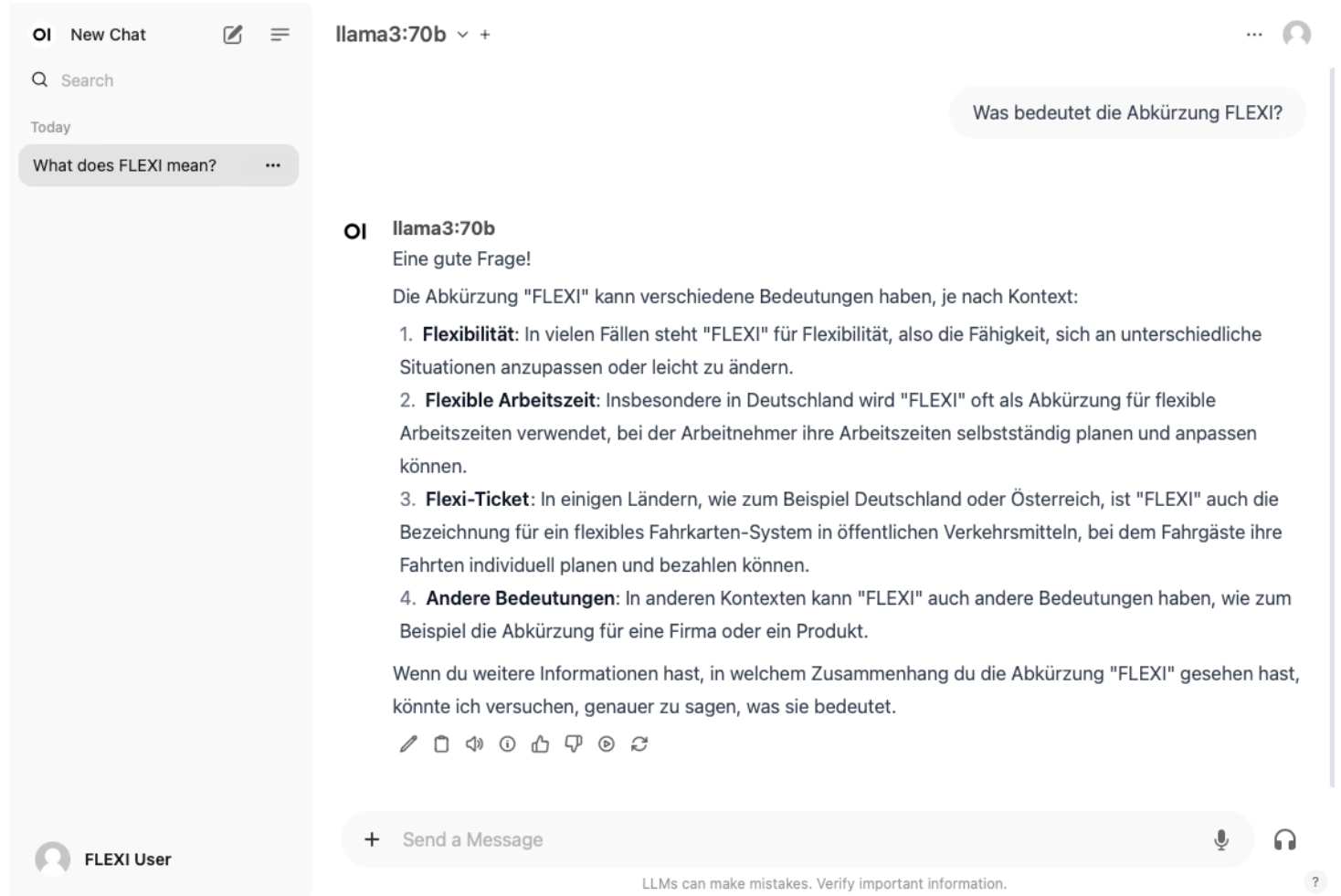
- API Access

```
> curl http://flexi.fernuni-hagen.de/api/generate -d '{  
  "model": "llama3",  
  "prompt": "Was bedeutet die Abkürzung FLEXI?"  
}'
```

```
{ "model": "llama3", "created_at": "", "response": "Die", "done": false }  
{ "model": "llama3", "created_at": "", "response": " Ab", "done": false }  
{ "model": "llama3", "created_at": "", "response": "k", "done": false }  
{ "model": "llama3", "created_at": "", "response": "ür", "done": false }  
{ "model": "llama3", "created_at": "", "response": "zung", "done": false }  
{ "model": "llama3", "created_at": "", "response": " F", "done": false }  
{ "model": "llama3", "created_at": "", "response": "LEX", "done": false }  
{ "model": "llama3", "created_at": "", "response": "I", "done": false }  
{ "model": "llama3", "created_at": "", "response": " steht", "done": false }  
{ "model": "llama3", "created_at": "", "response": " für", "done": false }  
{ "model": "llama3", "created_at": "", "response": " Flexible", "done": false }
```

Zugangsmöglichkeiten

- API Access
- Chatbot (generisch)



The screenshot displays a chat application interface. On the left, a sidebar shows a 'New Chat' button, a search bar, and a list of chat sessions under the heading 'Today'. One session titled 'What does FLEXI mean?' is selected. The main chat area shows a conversation with the model 'llama3:70b'. The user's message is 'Was bedeutet die Abkürzung FLEXI?'. The model's response explains that 'FLEXI' can have different meanings depending on the context, listing four examples: 1. **Flexibilität**: the ability to adapt to different situations. 2. **Flexible Arbeitszeit**: a work schedule where employees can plan and adjust their working hours. 3. **Flexi-Ticket**: a flexible travel ticket system for public transport. 4. **Andere Bedeutungen**: other meanings in different contexts, such as a company or product. The interface includes a 'Send a Message' button at the bottom and a footer note: 'LLMs can make mistakes. Verify important information.'

OI New Chat

Search

Today

What does FLEXI mean?

llama3:70b

Was bedeutet die Abkürzung FLEXI?

OI llama3:70b

Eine gute Frage!

Die Abkürzung "FLEXI" kann verschiedene Bedeutungen haben, je nach Kontext:

1. **Flexibilität**: In vielen Fällen steht "FLEXI" für Flexibilität, also die Fähigkeit, sich an unterschiedliche Situationen anzupassen oder leicht zu ändern.
2. **Flexible Arbeitszeit**: Insbesondere in Deutschland wird "FLEXI" oft als Abkürzung für flexible Arbeitszeiten verwendet, bei der Arbeitnehmer ihre Arbeitszeiten selbstständig planen und anpassen können.
3. **Flexi-Ticket**: In einigen Ländern, wie zum Beispiel Deutschland oder Österreich, ist "FLEXI" auch die Bezeichnung für ein flexibles Fahrkarten-System in öffentlichen Verkehrsmitteln, bei dem Fahrgäste ihre Fahrten individuell planen und bezahlen können.
4. **Andere Bedeutungen**: In anderen Kontexten kann "FLEXI" auch andere Bedeutungen haben, wie zum Beispiel die Abkürzung für eine Firma oder ein Produkt.

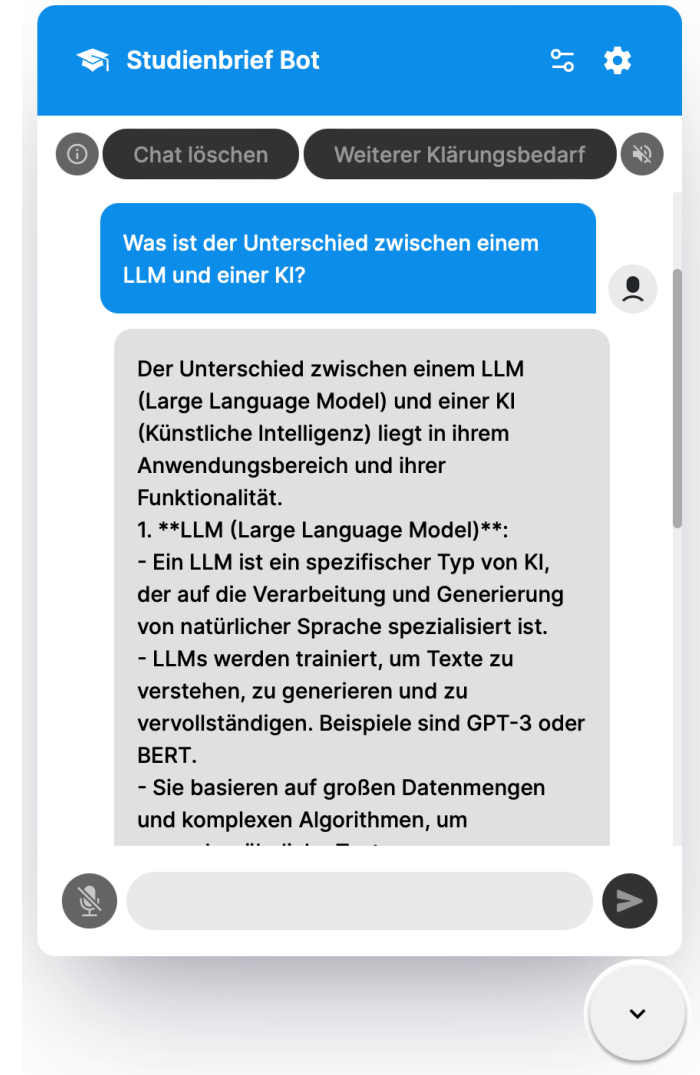
Wenn du weitere Informationen hast, in welchem Zusammenhang du die Abkürzung "FLEXI" gesehen hast, könnte ich versuchen, genauer zu sagen, was sie bedeutet.

+ Send a Message

LLMs can make mistakes. Verify important information.

Zugangsmöglichkeiten

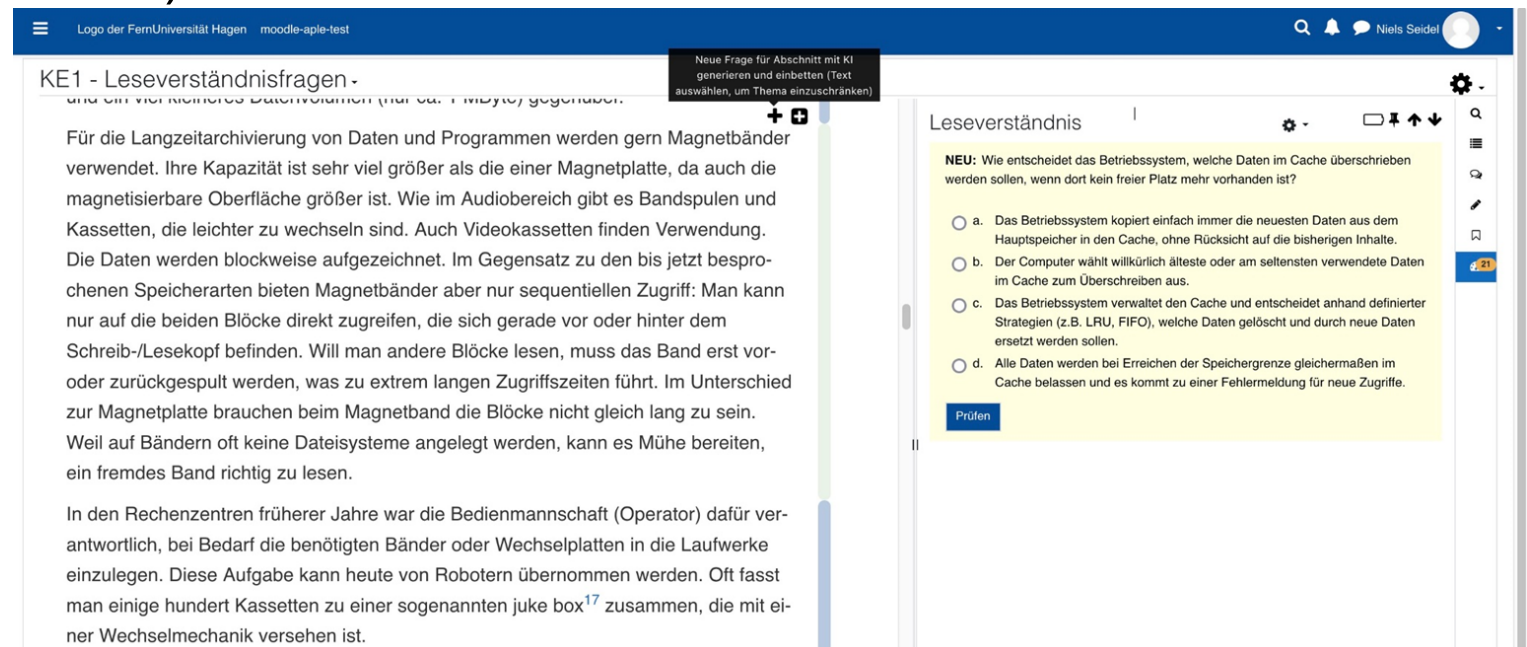
- API Access
- Chatbot (generisch)
- Chatbot (spezifisch für Kurse)



Zugangsmöglichkeiten

- API Access
- Chatbot (generisch)
- Chatbot (spezifisch für Kurse)
- Moodle Integration

https://github.com/CATALPAresearch/mod_longpage



The screenshot displays a Moodle course page titled "KE1 - Leseverständnisfragen". The main content area contains a long text question about magnetic tape storage. To the right, a quiz interface is visible with a question titled "LEU: Wie entscheidet das Betriebssystem, welche Daten im Cache überschrieben werden sollen, wenn dort kein freier Platz mehr vorhanden ist?". The quiz options are:

- ☐ a. Das Betriebssystem kopiert einfach immer die neuesten Daten aus dem Hauptspeicher in den Cache, ohne Rücksicht auf die bisherigen Inhalte.
- ☐ b. Der Computer wählt willkürlich älteste oder am seltensten verwendete Daten im Cache zum Überschreiben aus.
- ☐ c. Das Betriebssystem verwaltet den Cache und entscheidet anhand definierter Strategien (z.B. LRU, FIFO), welche Daten gelöscht und durch neue Daten ersetzt werden sollen.
- ☐ d. Alle Daten werden bei Erreichen der Speichergrenze gleichermaßen im Cache belassen und es kommt zu einer Fehlermeldung für neue Zugriffe.

A "Prüfen" (Check) button is located at the bottom of the quiz options. A small tooltip above the quiz area reads: "Neue Frage für Abschnitt mit KI generieren und einbetten (Text auswählen, um Thema einzuschränken)".

Zugangsmöglichkeiten

- API Access
- Chatbot (generisch)
- Chatbot (spezifisch für Kurse)
- Moodle Integration
- ...

FLEXI - Fernuni LLM Experimental Infrastructure

- Hardware & OS

Hardware	Server A	Server B
OS	Ubuntu Server 22.04 LTS	Ubuntu Server 22.04.4 LTS
Kernel	5.15	5.15
GPU	8x Nvidia RTX A5000 24GB	2x Nvidia A40 46GB
CPU	2x AMD EPYC 7402	2x Intel Xeon Gold 6442Y
RAM	256 GB DDR3 3200 MHZ	512 GB DDR5 4800 MT/s
Storage	8 TB SSD	6 TB SSD
Driver	Nvidia 535 Server	Nvidia 535 Server
Cuda	12.4	12.2

Table 2: OS and Hardware of FLEXI Servers

FLEXI - Fernuni LLM Experimental Infrastructure

- Hardware & OS
- LLM Software



Ollama

Get up and running with large language models.

Run [Llama 3](#), [Phi 3](#), [Mistral](#), [Gemma](#), and other models. Customize and create your own.

Download ↓

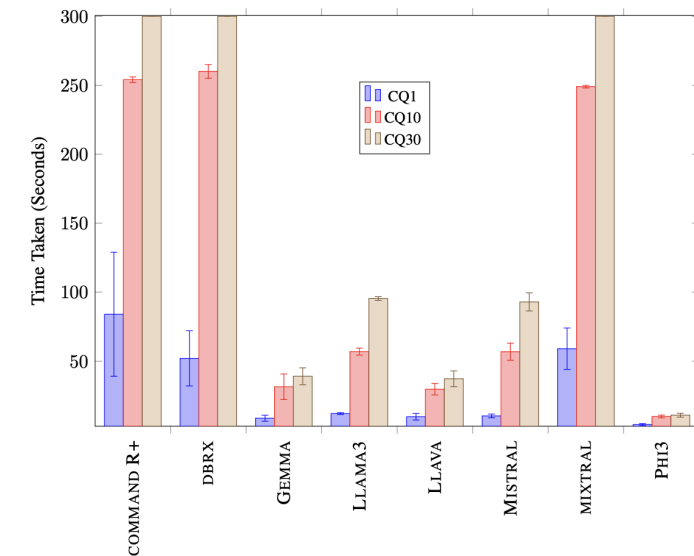
FLEXI - Fernuni LLM Experimental Infrastructure

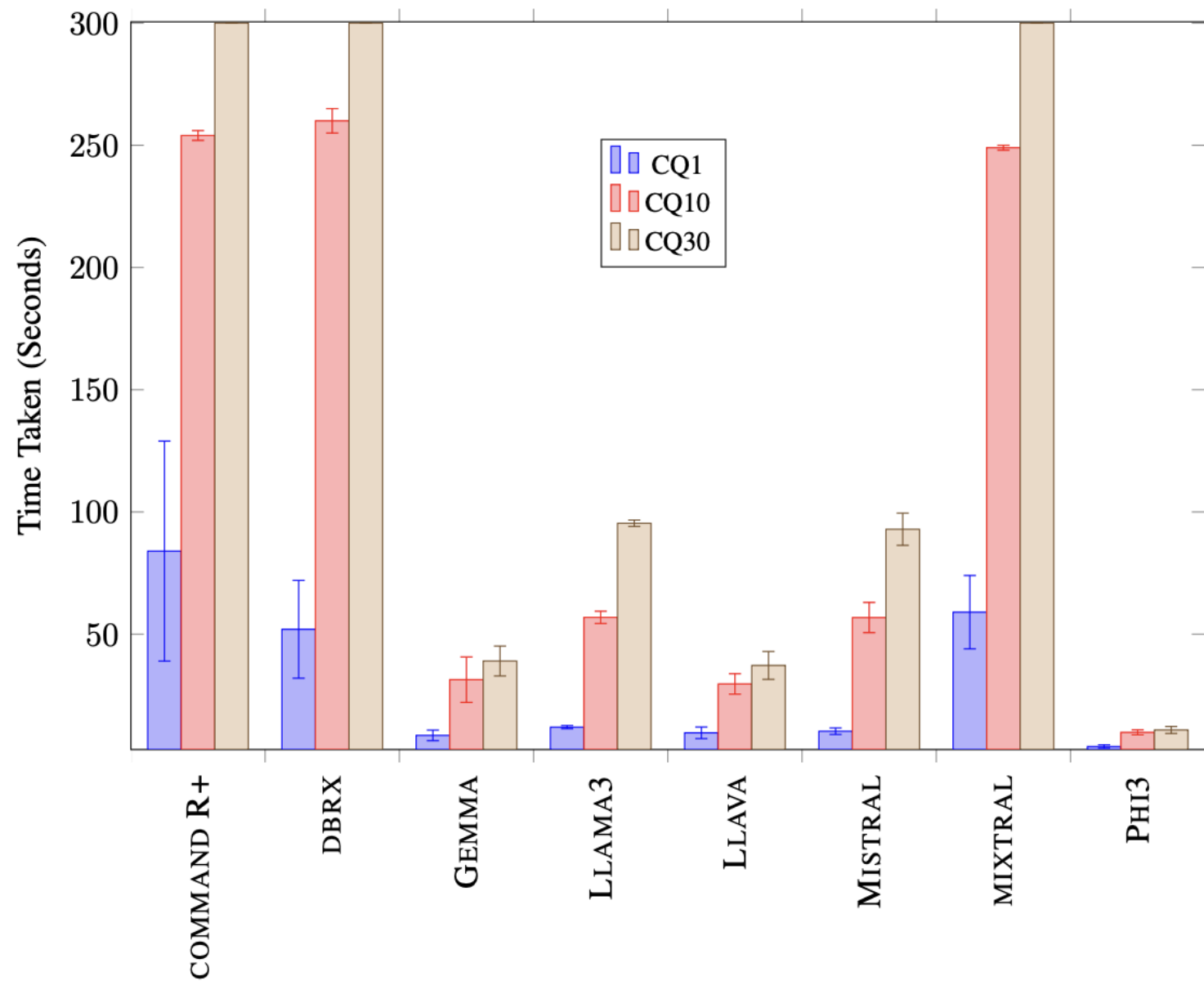
- Hardware & OS
- LLM Software
- LLM Modellauswahl
 - Qualität

Model Name	[token/s] ↑	ARC	HellaSwag	MMLU	TruthfulQA	WinoGrande	GSM8K	Ø
COMMAND R+	5	.71	.89	.76	.56	.85	.71	.75
DBRX	11	.68	.89	.74	.67	.82	.67	.75
GEMMA	77	.65	.81	.65	.55	.78	.73	.69
LLAMA3	82	.73	.86	.80	.64	.83	.88	.79
LLAVA	60	.53	.76	.52	.46	.72	.15	.52
MISTRAL	94	.73	.89	.64	.78	.85	.70	.77
MIXTRAL	11	.73	.89	.78	.68	.85	.82	.79
PHI3	127	.67	.86	.78	.58	.73	.80	.74

FLEXI - Fernuni LLM Experimental Infrastructure

- Hardware & OS
- LLM Software
- LLM Modellauswahl
 - Qualität vs. Geschwindigkeit





Einblicke in die Systeme



LIVE-DEMO

Erfahrungswerte

- Auch “kleine” Modelle nützlich (und sehr schnell)
- Zum Experimentieren benötigt es keine high-end-Server
 - aber Grafikprozessoren (GPU) zwingend erforderlich
- Technische Komplexität bei mehreren GPUs (Virtualisierung, Cuda, Driver, Linux)
- Viele Anwendungsfälle entstehen durch die Bereitstellung der API (LLM als “enabler”)

Vielen Dank für Ihre Aufmerksamkeit!

- FLEXI - Fernuni LLM Experimental Infrastructure
- datenschutzkonformer Zugang zu Open-weight Modellen
- Experimentierumgebung für LLM-basierte Lehr-Lernanwendungen

Torsten Zesch, Michael Hanes, Niels Seidel, Piush Aggarwal, Dirk Veiel, Claudia de Witt (2024)
FernUni LLM Experimental Infrastructure (FLEXI) - Enabling Experimentation and Innovation in
Higher Education Through Access to Open Large Language Models. Preprint
(<https://github.com/zesch/flexi>)

