FernUni LLM Experimental Infrastructure (FLEXI) – Enabling Experimentation and Innovation in Higher Education Through Access to Open Large Language Models

Torsten Zesch[1] and **Michael Hanses** [1, 2] and **Niels Seidel**[1]
Piush Aggarwal[1] and **Dirk Veiel**[1] and **Claudia de Witt**[1,2]
[1]CATALPA, FernUniversität in Hagen, Germany
[2] Institut für Bildungswissenschaft und Medienforschung

Abstract

Using the full potential of LLMs in higher education is hindered by challenges with access to LLMs. The two main access modes currently discussed are paying for a cloud-based LLM or providing a locally maintained open LLM. In this paper, we describe the current state of establishing an open LLM infrastructure at FernUniversität in Hagen under the project name FLEXI (FernUni LLM Experimental Infrastructure). FLEXI enables experimentation within teaching and research with the goal of generating strongly needed evidence in favor (or against) the use of locally maintained open LLMs in higher education. The paper will provide some practical guidance for everyone trying to decide whether to run their own LLM server.

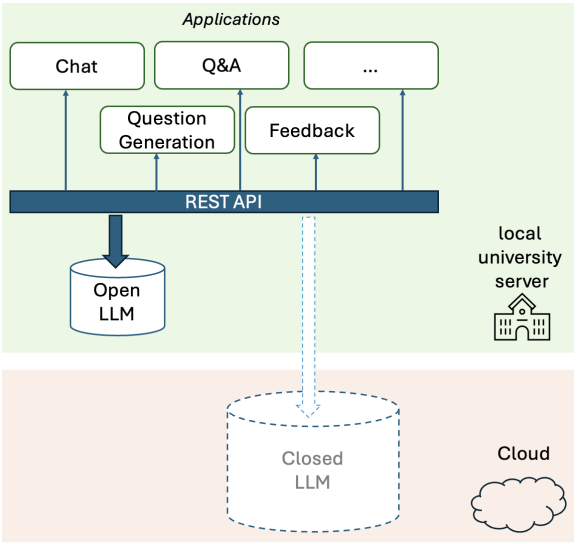

Figure 1: The FLEXI approach replacing a cloud-based LLM with a locally maintained open LLM

1 Motivation

While the potential of Large Language Models (LLMs) for higher education has been identified (Kasneci et al., 2023), as long as access is not provided by the university in some way, everybody is using the commercial service of their choice, leading to issues including potential data security problems, decreased educational equity, potentially high costs, etc.

Thus, there is an ongoing discussion about whether and how universities should provide LLM access (Salden et al., 2024). Two main modes are generally being discussed: paying for a cloud-based LLM or providing a locally maintained open LLM.[1] Figure 1 gives a high-level overview of how a closed, cloud-based LLM could be replaced with an open-source model. As it shows, replacing a closed LLM with an open LLM can be as easy as pointing the applications to a local REST endpoint once a local LLM is in place.

Both options, commercial and open, have pros and cons, as summarized in Table 1. Cloud-based LLMs are easy to set up and run the latest models, while open LLMs are currently more cost-effective overall, and data protection is much easier to achieve. Based on these considerations, we currently strongly favor open LLMs and decided to implement this setup at FernUniversität in Hagen.

In this paper, we describe the ongoing efforts at FernUniversität in Hagen to establish such an open LLM infrastructure to enable experimentation in teaching and research. This research is carried out by researchers from CATALPA (Center of Advanced Technology for Assisted Learning and Predictive Analytics) under the project name FLEXI (Fernuni LLM Experimental Infrastructure). The project is supposed to generate strongly needed evidence supporting (or challenging) the use of locally maintained open-source LLMs in higher education settings. The paper shall give some practical guid-

[1]Note that there is an ongoing discussion when a model could be called 'open' or 'open source' (Liesenfeld and Dingemanse, 2024). We decide to speak of 'open models' as soon as the weights are available so we can run them, but we acknowledge (and discuss later in the paper) that different levels of openness come with consequences for academic use.

		Closed LLM		FLEXI Open LLM
Setup costs	++	none	-	server storage; dedicated server
Operating costs	- -	pay per token	+	operating costs of server
Maintenance costs	++	included in the operating costs	- -	continued maintenance
Model quality	++	access to latest models	o	only open-weight models
Model stability	o	might change at any time; little control	++	under university control
Data protection	-	hard to ensure	++	everything stays within own infrastructure

Table 1: Pros and Cons of closed and open LLM provisioning

ance for everyone considering whether to run their own LLM server or not.

Note that the project's scope is currently limited to providing access to text-based, uni-modal LLMs but that other multi-modal services like image generation or speech recognition could be provided using a similar process.

2 FLEXI Concept & Realization

As we are aiming for an experimental proof-of-concept realization, we aim for a single server that is able to run most open-source models. However, our concept is based on a bare-metal Kubernetes[2] cluster, which could be extended by additional nodes and thus enable scalable and more robust operation. We currently ignore guaranteed uptime, redundancy, or other factors that would be central when moving from experimentation to central service delivery.

2.1 Hardware Setup

Our concept assumes that the university already operates a data center where the server can be housed. Consequently, the university can leverage existing processes for access control, hardware maintenance, network, security, or backup.

The bare-metal Kubernetes setup allows access to all server hardware settings without an additional (extra) virtualization layer. This is very helpful considering the configuration of GPU acceleration.

We first piloted the setup on one server (A) and then replicated it on another server (B). Hardware and software specifications of the servers can be found in Table 2. Both servers were purchased in 2023 for approximately 40,000 €.

[2] https://kubernetes.io

2.2 Software Setup

For serving the LLMs the open-source project Ollama is used.[3] To ensure optimal performance, GPU acceleration is necessary, but Ollama can also run the models without GPU support. NVIDIA and AMD graphic cards are supported. Our existing servers have NVIDIA GPU's build in. Thus, the combination of operating system, kernel, drivers, and software must match the CUDA version compatibility.[4] The operating system is Ubuntu 22.04 LTS on both of our servers with NVIDIA-535-Server and CUDA 12.4 on Server (A) and CUDA 12.2 on Server (B). On Server (A), both Ollama and Open WebUI[5] are deployed as Kubernetes pods. On server (B), the docker-compose service is used. For routing and load balancing, the open-source software traefik[6] is used as ingress-controller (Sharma et al., 2021). The usage of containerization allows us to quickly switch between Ollama versions and custom configurations, which is very helpful in this experimental setting.

2.3 Model Selection

The setup described so far allows us to install and serve any model publicly hosted in the Ollama library.[7] At the time of writing, there are over 90 models available, from which we must select a suitable subset. Additionally, the web interface enables experimentation with 16,848 models available in GGUF format[8] on Huggingface at the time of writing.

At the time of writing, we are testing the models

[3] https://ollama.com

[4] https://docs.nvidia.com/deploy/cuda-compatibility/

[5] A webui interface to interact with Ollama models available at https://github.com/open-webui/open-webui

[6] https://traefik.io/traefik/

[7] https://ollama.com/library

[8] https://github.com/ggerganov/ggml/blob/master/docs/gguf.md

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

Model Name	Size	Context Length	Licence
COMMAND R+	104 B	128 K	*model specific*
DBRX	132 B	32 K	*model specific*
GEMMA	7 B	8 K	*model specific*
LLAMA3	8 B	8 K	*model specific*
LLAVA	13 B	32 K	Apache 2.0
MISTRAL	7 B	32 K	Apache 2.0
MIXTRAL	22x8 B	8 K	Apache 2.0
PHI3	4 B	4 K	MIT

Table 3: Selected LLMs running on the FLEXI infrastructure (at the time of writing)

listed in Table 3.[9] We now discuss the dimensions that informed our selection of those models to experiment with.

2.3.1 Openness

Open-source large language models (LLMs) come in a variety of 'flavors' that significantly differ in how open they actually are (Liesenfeld and Dingemanse, 2024). The minimum requirement for our purposes is that the weights are available so we can run, modify, and improve models on our servers. Liesenfeld and Dingemanse (2024) lists several additional dimensions, including the availability of basic training data that is used for instruction tuning as well as open documentation and a permissive license.

2.3.2 Language

Various open models are available that can handle unilingual and multilingual queries. For instance, models like STABLELM2 are trained on multilingual data, including English, Spanish, German, Italian, French, Portuguese, and Dutch. The QWEN2 model supports 29 languages, including Chinese. Specific language models such as SAUERKRAUTLM-QWEN-32B are trained on German language datasets, making them ideal for general-purpose German queries. Additionally, the PHI3 MEDIUM model is designed for commercial and research use in English. However, German language support must be balanced with model quality.

2.3.3 Quality

A wide range of benchmarks are available on which LLM model quality can be evaluated. Benchmarks can have different specializations, e.g. focusing on language capabilities, world knowledge, common sense reasoning, or coding. We argue that some capabilities are more important in our educational settings than others. For example, medical knowledge might not be central at FernUniversität, while knowledge of German (Pfister and Hotho, 2024) or factual correctness seems more important. We discuss here our selection of benchmarks:

ARC (Clark et al., 2018) examine LLMs on 7,787 grade-school science questions. The test is challenging and demands extensive general knowledge and strong reasoning skills. It includes two sets: Easy and Challenge (with particularly difficult tasks).

GSM8K (Cobbe et al., 2021) is a set of 8,500 grade-school math problems, each requiring two to eight steps to solve using basic math operations. The questions are simple enough for a smart middle schooler to solve and are useful for testing LLMs' ability to handle multistep math problems.

HellaSwag (Zellers et al., 2019) This benchmark evaluates natural language inference (NLI) by

[9] While this list is certainly already outdated by the time you are reading this, we still think it serves to give an idea about the range of models being tested.

Model Name	[token/s] ↑	ARC	HellaSwag	MMLU	TruthfulQA	WinoGrande	GSM8K	∅
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

Table 4: LLM model quality based on established benchmarks covering different application areas. Normalized scores range from 0-1. Higher scores correspond to better models. For comparison, we also list the throughput of the models on our Server B in tokens per second.

prompting LLMs to complete a given passage. What adds to its difficulty is using adversarial filtering to create deceptive yet plausible incorrect answers for the tasks.

MMLU stands for Massive Multitask Language Understanding (Hendrycks et al., 2020) measures general knowledge across 57 different subject areas, spanning from STEM to social sciences. The difficulty levels range from elementary to advanced professional.

TruthfulQA (Lin et al., 2022) aims to determine if LLMs produce incorrect answers based on common misconceptions. The questions cover various categories, including health, law, fiction, and politics.

WinoGrande (Sakaguchi et al., 2021) is a massive set of 44,000 problems derived from the Winograd Schema Challenge (Levesque et al., 2012). These problems consist of nearly identical sentence pairs with two possible answers, where the correct answer depends on a trigger word. This tests the ability of LLMs to accurately understand context.

Table 4 illustrates the performance based on the Huggingface leaderboard.[10] The evaluation score has been normalized for each benchmark between 0 and 1, with higher scores indicating better performance. With the exception of LLAVA[11], most models perform well on average but have their specializations. For example, MISTRAL is especially good on TruthflQA but not on Math problems (GSM8K), whereas LLAMA3 is best.

We also show in the table throughput[12], as model quality has to be balanced with how fast the requests can be served. Combining these two metrics enables us to select the model best suited for a specific use case in FLEXI.

2.3.4 Safety

Applications of LLMs within higher education raise concerns about their security and potential vulnerabilities. Ensuring LLM security involves preventing misuse by malicious actors or avoiding unintentional errors, such as accidentally revealing email addresses. Unlike traditional cybersecurity, LLM security depends significantly on natural language processing (NLP) techniques because most attack strategies are language-based. Attacks can occur due to conflicts between application builders, end-users, and external tool outputs, especially when there is explicit knowledge about the builder's intentions or policies (Wei et al., 2024). Therefore, it is crucial for a secure model to undergo various vulnerability assessments before being deployed. LLM security evaluation frameworks, such as the 'Generative AI Red-teaming and Assessment Kit' (garak), facilitate this process (Derczynski et al., 2024). Through systematic probing, it helps users identify vulnerabilities in language models or dialog systems. Checks such as

[10] https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard

[11] LLAVA is focused on visual tasks, which our selection of benchmarks does not reflect.

[12] measured via https://github.com/aidatatools/ollama-benchmark

profanity, toxicity, encoding flaws, and jailbreaks are analyzed to evaluate model safety. The leaderboard scores are available[13], that can estimate these checks for models compatible with FLEXI.

2.3.5 Size

Finally, as FLEXI operates with the limited resources of a public university, model size is an issue, as bigger models might not run at all on our hardware, or throughput might be insufficient. As our main goal here is experimentation, we include a few models from the fringes of the size distribution but mainly focus on mid-distribution models. However, most currently available open models would run on our servers, but throughput and maximum concurrent queries might be insufficient (see sections 3 and 5).

2.4 Maintenance & Monitoring

We use Checkmk[14] to monitor the servers (A) and (B) and the resources they contain. The so-called Checkmk agent runs on our servers, which collects data from the local system via plugins and transmits it to the backend. The backend receives and manages the data and makes it available via dashboards. Our data center (Zentrum für Digitalisierung und IT, ZDI) operates the backend.

To monitor GPU utilization in particular, we use a special script as a local plugin. The script captures the GPU data (via nvidia_smi) and passes it to the backend so that the visualizations shown in Figure 2 can be viewed there. Using the dashboard, we can see the current utilization of the servers, especially the GPUs, and the trend over the past days and weeks. This allows us to identify both peak loads and average server utilization. The knowledge gained in this way is used to better determine the configuration of the servers for regular operation.

2.5 Data Protection

When using a cloud-based LLM, all requests are sent to the cloud provider, enabling user tracking and possibly exposing sensitive data. Projects like HAWKI[15] solve the tracking issue by bundling all requests from one university so that chats cannot be attributed to a specific person. Commercial providers like Microsoft offer services like 'Azure OpenAI' where the data is not sent to OpenAI (but to Microsoft), and the requests are guaranteed not to be used for model training. One key advantage of FLEXI is that requests never leave the premises of the university IT infrastructure (cf. Figure 1).

3 Experiences

In this section, we describe the most important experiences and takeaways from experimenting with FLEXI.

Load Test To analyze this, we attempted to test our server's load using a general-purpose laptop. We sent multiple REST API POST requests to FLEXI. To fully utilize the available GPU space, multiple models are initiated simultaneously. This approach will make efficient use of the GPUs and significantly reduce the server response time. In addition, multiple instances of smaller models can be created, allowing it to handle multiple requests simultaneously. Figure 3 illustrates the load in terms of time taken by different models on SERVER (B) while handling concurrent requests. On the one hand, models such as PHI3 can handle multiple requests with hardly an increase in latency time. On the other hand, models such as MISTRAL and LLAMA3 exhibit higher latency with increased concurrent queries. For synchronous tasks like ChatBots, the response times of larger models will probably be too high once this is scaled to many users. However, not all tasks require the largest models, especially since benchmark quality improvements are often marginal (cf. Figure 4).

Operating Costs Assuming that the data center itself already has fixed costs for the university, operating costs are dominated by energy demand[16]. In contrast to closed LLM servers, where very little information about energy usage is available, we can directly measure energy usage. For the 5-day period shown in Figure 2, approximately 26.7 kilowatt-hours (kWh) were consumed by 8 GPUs, i.e. about 5 kWh per day. The theoretical maximum, which we have not measured yet, would be around 44.16 kWh a day or 16 MWh a year. At 0.30 € per kWh, this translates to a maximal annual operating cost for the 8 GPUs of about 5,000 €.

This energy usage translates into 6 tons of emit-

[13] https://huggingface.co/spaces/AI-Secure/llm-trustworthy-leaderboard

[14] https://checkmk.com

[15] https://github.com/HAWK-Digital-Environments/HAWKI

[16] FernUniversität in Hagen is already using photovoltaic systems to meet some of the university's energy needs.

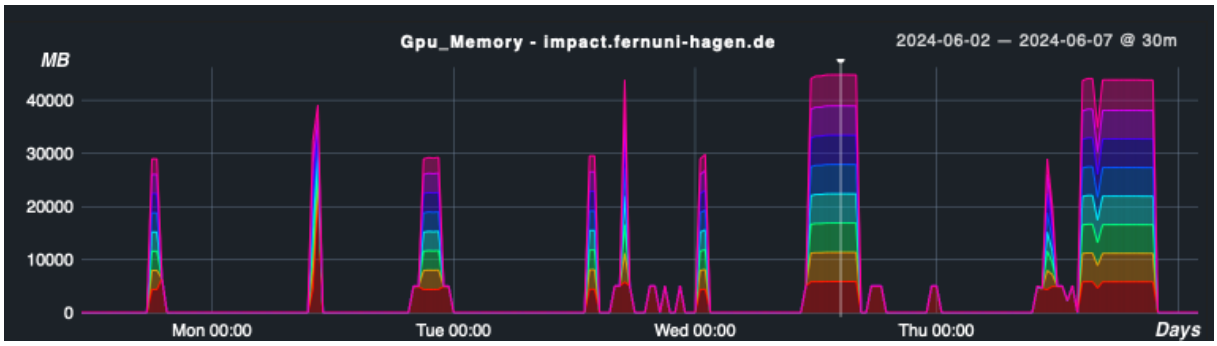

Figure 2: Usage spikes during a longer period of use of Server (A). The memory usage of all 8 GPUs is shown in color. You can see that usually, all 8 GPUs are used. However, as soon as it is sufficient to use only one GPU, e.g., for a smaller model (see smaller red spikes), the system implements this accordingly and in a resource-saving manner.

ted CO_2.[17] However, this number has to be put into perspective, as recent research (Tomlinson et al., 2024) indicates that LLM energy usage is dwarfed by the energy usage of the computers being used to access the LLMs (e.g. by students using Moodle to access their course contents).

4 Applications & Use Cases

While a setup as implemented through FLEXI can support a wide range of applications (Rashid et al., 2024), we list here some use cases that we believe to be of specific interest in higher education.

Chat Interface To assist the students and educators, as an entry point familiar to everyone who has used the web interface of, e.g., ChatGPT, we provide a chat interface based on OpenWebUI.[18] Figure 4 shows an example chat session.

RAG We are experimenting with retrieval-augmented generation (RAG) applications, where, e.g., lecture notes are indexed, and students are provided with access to a dedicated chatbot that can answer questions regarding the study material.

API Access Selected users who want to work with the API are granted direct API access and may implement their own applications. For example, the `https://what2study.de` project uses direct API access to experiment with their system.

LMS Integration We are developing middleware for LLM access in Moodle, a widely used open-source learning management system (LMS), under the name CAIPI (CATALPA AI Prompting Interface). At FernUniversität in Hagen, numerous students and lecturers use Moodle to access courses and interact with learning materials and each other. Integrating LLM access into Moodle is crucial for enhancing these interactions.

There already are Moodle plugins making use of LLMs directly, for use cases like question generation[19] or as chatbots.[20] They all implement LLM access directly. CAIPI creates an abstraction layer, establishing an authorized interface between user requests and the API acces provided by FLEXI. CAIPI structures requests, checks input parameters, regulates access based on user roles, and enables load balancing. Prompts, including parameters from the Moodle database, can be stored to be re-used.

5 Future work: University-wide scaling

FLEXI is an experiment aimed at learning more about the possible pitfalls of providing open LLM access. Access is thus currently limited to selected early adopters who know about models' possible shortcomings, who do not expect flawless operation, and who are giving us valuable feedback on how to improve the service.

Should we eventually want to drop the 'experimental' status and provide the same service on a university-wide level, we have some more chal-

[17] Germany 2023: 380 gCO2 per kWh, https://ember-climate.org/data/data-tools/data-explorer/

[18] https://github.com/open-webui/open-webui

[19] e.g. https://moodle.org/plugins/local_aiquestions, https://moodle.org/plugins/block_openai_questions.

[20] e.g. https://moodle.org/plugins/block_openai_chat, https://moodle.org/plugins/local_ulibot, https://moodle.org/plugins/block_ube_ta, https://moodle.org/plugins/block_openai_chat

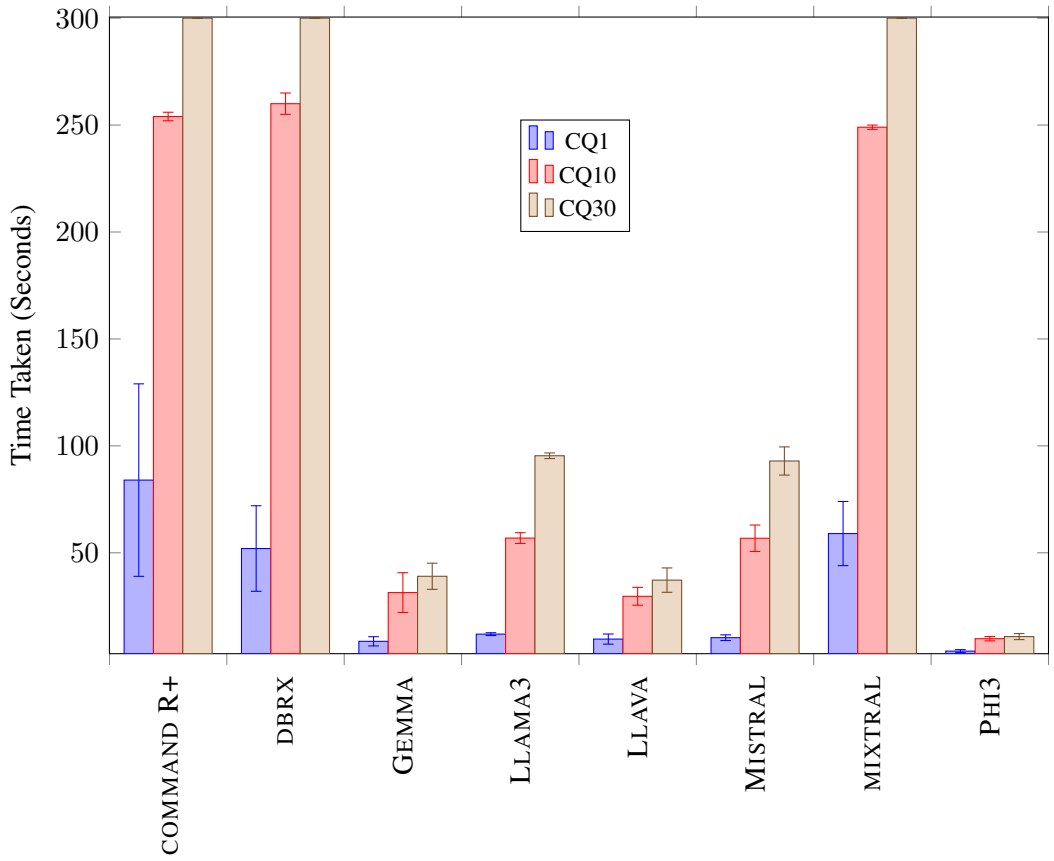

Figure 3: Load test results accessing Server B. We show results for a single query as well as 10 and 30 concurrent queries (CQ). There is a response timeout of 300s, so this is the maximum possible average.

lenges ahead of us. We would probably need a bigger **server** and an **operational concept** providing guaranteed uptime, redundancy, and load balancing. We would also have to look deeper into **legal issues** which we discuss here for the specific situation at our university in Germany:

First, the university's IT administration requires an operational concept for the regular operation of such a service. This includes the system architecture, security measures, maintenance routines, and responsibilities. Second, compliance requirements regarding the General Data Protection Regulation (GDPR) must be met. This includes a document titled "record of processing activities" detailing how personal data (e.g. server logs, access logs, and possibly user inputs) are processed. It specifically describes users' rights regarding information access, data retention periods, and deletion of personal data. Third, a user agreement is necessary, which users must accept upon their first access to an FLEXI-based application. This agreement includes information on the processing of personal data and disclaimers about the reliability of information generated by the LLMs.

Beyond legal issues, we are also facing **ethical questions**. Which level of freedom of speech should a university allow? The answer might differ depending on the use case, where research probably needs less censored models than teaching. Universities also may need to take a stance on which political orientation an LLM should express, as they have been found to vary quite a lot (Feng et al., 2023). However, recent research (Röttger et al., 2024) somewhat challenges this view, finding that "even small changes in situative context can substantially affect the values and opinions manifested in LLMs". Consequently, any decisions on which LLM is suitable in a higher education context should only be made if tested given salient use cases. It is the role of experiments like FLEXI to provide an environment where these kinds of tests can be conducted.

6 Summary

In this paper, we have argued that in an academic context, locally hosting open-source LLMs is cur-

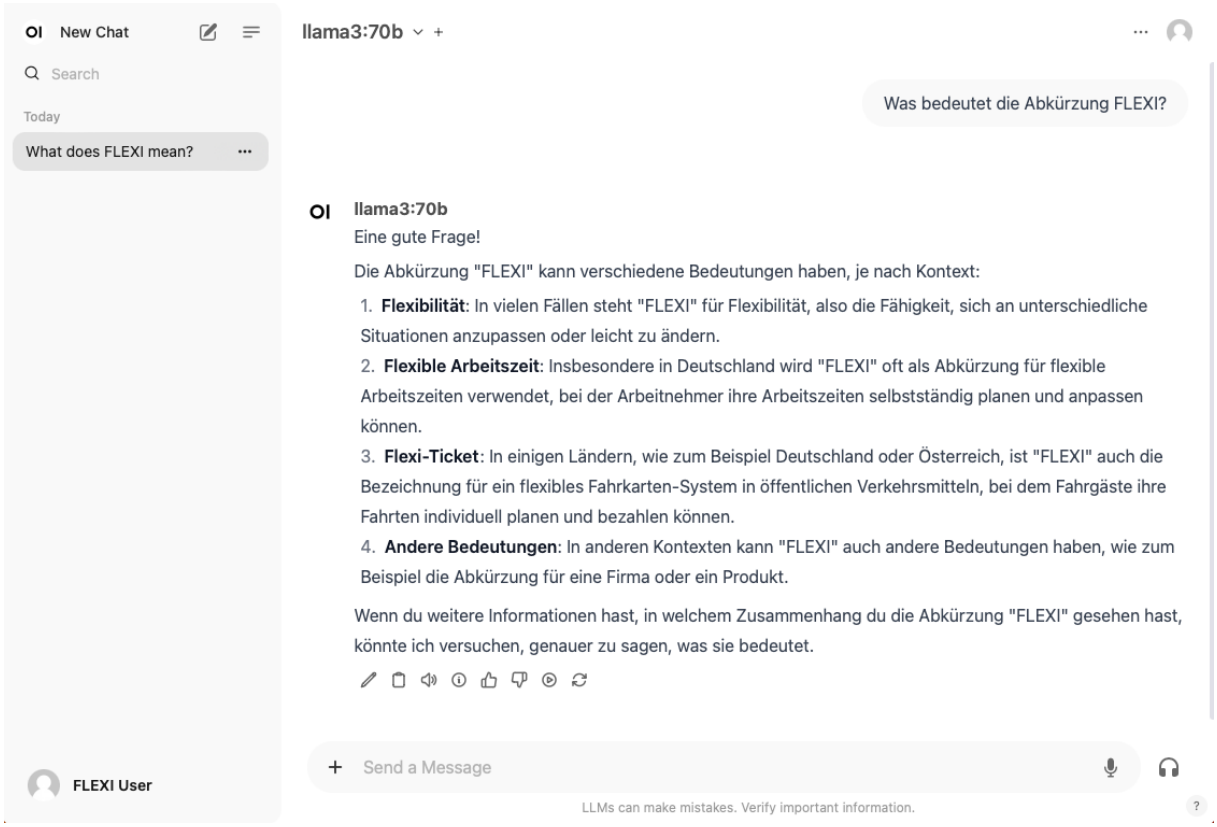

Figure 4: Screenshot of an example session with the OpenWebUI chat frontend using the Llamma3-70b model.

rently the best choice, balancing the pros (data protection and cost) with the cons (maintenance effort). To that end, we have set up FLEXI, a concrete implementation example that can serve as a reference point for others setting out on a similar endeavor. With moderate hardware, we have shown that it is possible to provide access to open LLM and thus support a wide range of educational applications. FLEXI provides maximal data protection, as no LLM request ever leaves the premises of our university. While our approach was designed and implemented for the higher education sector, it may be applied to other sectors or domains as well. A key open question remains model selection, as a wide variety is on offer and use cases might have different needs. A solution could be to run multiple models in parallel, as we are successfully doing right now.

Acknowledgments

This research was supported by the Center of Advanced Technology for Assisted Learning and Predictive Analytics (CATALPA) of FernUniversität in Hagen, Germany. Enabling access to LLMs at FernUniversität is a central activity organized by Zentrum für Lernen und Innovation (ZLI).

References

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Leon Derczynski, Erick Galinkin, Jeffrey Martin, Subho Majumdar, and Nanna Inie. 2024. garak: A Framework for Security Probing Large Language Models. https://garak.ai.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt.

2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Andreas Liesenfeld and Mark Dingemanse. 2024. Rethinking open source generative AI: open washing and the EU AI Act. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1774–1787, New York, NY, USA. Association for Computing Machinery.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods.

Jan Pfister and Andreas Hotho. 2024. SuperGLEBer: German language understanding evaluation benchmark. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7897–7916, Mexico City, Mexico. Association for Computational Linguistics.

Sheikh Faisal Rashid, Nghia Duong-Trung, and Niels Pinkwart. 2024. Generative ai in education: Technical foundations, applications, and challenges. In Dr. Seifedine Kadry, editor, *Artificial Intelligence for Quality Education*, chapter 2. IntechOpen, Rijeka.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schütze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Peter Salden, Malte Persike, and Jonas Leschke. 2024. Die Bereitstellung generativer KI in Hochschulen: Was ist möglich und was wünschenswert? Hochschulforum Digitalisierung Blog.

Rahul Sharma, Akshay Mathur, Rahul Sharma, and Akshay Mathur. 2021. Traefik as kubernetes ingress. *Traefik API Gateway for Microservices: With Java and Python Microservices Deployed in Kubernetes*, pages 191–245.

Bill Tomlinson, Rebecca W. Black, Donald J. Patterson, and Andrew W. Torrance. 2024. The carbon emissions of writing and illustrating are lower for AI than for humans. *Scientific Reports*, 14(1):3732.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2024. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence?

A Server Load Results

Model Name (Size)	Time [s] 1 CQ			10 CQ			30 CQ		
Prompt: Write a step-by-step guide on how to bake a chocolate cake from scratch.									
COMMAND R+	106	±	20	300	±	0	300	±	0
DBRX	61	±	7	300	±	0	300	±	0
GEMMA	10	±	3	35	±	4	53	±	42
LLAMA3	14	±	1	77	±	2	127	±	5
LLAVA	8	±	3	41	±	18	47	±	13
MISTRAL	11	±	1	66	±	11	104	±	7
MIXTRAL	76	±	12	300	±	0	300	±	0
PHI3	4	±	3	12	±	5	11	±	4
Prompt: Develop a python function that solves the following problem, sudoku game									
COMMAND R+	110	±	69	300	±	0	300	±	0
DBRX	66	±	12	300	±	0	300	±	0
GEMMA	9	±	8	23	±	6	28	±	12
LLAMA3	14	±	2	68	±	4	120	±	3
LLAVA	14	±	15	26	±	9	39	±	10
MISTRAL	12	±	1	86	±	15	144	±	12
MIXTRAL	83	±	48	300	±	0	300	±	0
PHI3	5	±	3	13	±	4	14	±	3
Prompt: Create a dialogue between two characters that discusses economic crisis									
COMMAND R+	101	±	58	300	±	0	300	±	0
DBRX	46	±	7	300	±	0	300	±	0
GEMMA	10	±	8	42	±	13	63	±	9
LLAMA3	13	±	1	64	±	4	103	±	6
LLAVA	6	±	1	28	±	11	30	±	10
MISTRAL	10	±	1	45	±	8	72	±	9
MIXTRAL	59	±	4	300	±	0	300	±	0
PHI3	3	±	1	7	±	1	8	±	1
Prompt: In a forest, there are brave lions living there. Please continue the story.									
COMMAND R+	80	±	71	300	±	0	300	±	0
DBRX	41	±	10	300	±	0	300	±	0
GEMMA	10	±	3	33	±	7	44	±	7
LLAMA3	13	±	9	52	±	7	92	±	6
LLAVA	13	±	9	41	±	15	56	±	17
MISTRAL	11	±	5	58	±	19	97	±	3
MIXTRAL	56	±	12	300	±	0	300	±	0
PHI3	4	±	1	10	±	3	10	±	3
Prompt: I'd like to book a flight for 4 to Seattle in U.S.									
COMMAND R+	21	±	8	69	±	8	300	±	0
DBRX	26	±	4	100	±	28	300	±	0
GEMMA	5	±	1	10	±	1	11	±	4
LLAMA3	8	±	1	23	±	3	35	±	3
LLAVA	7	±	5	8	±	0	12	±	1
MISTRAL	7	±	1	29	±	7	48	±	6
MIXTRAL	21	±	0	43	±	6	300	±	0
PHI3	4	±	2	9	±	1	11	±	1

Table 5: Concurrency test on Server B using multiple concurrent queries (CQ). There is a response timeout of 300s, so this is the maximum possible average.