LOCAL LLMS: SAFEGUARDING DATA PRIVACY IN THE AGE OF GENERATIVE AI. A CASE STUDY AT THE UNIVERSITY OF ANDORRA

A. Dorca Josa, M. Bleda-Bejar

Universitat d'Andorra (ANDORRA)

Abstract

The growing field of Generative Artificial Intelligence (GAI) presents an unprecedented opportunity for innovation across diverse sectors. The inherent nature of these models, trained on vast amounts of data often sourced from the public domain, raises critical concerns regarding data privacy and security. At the same time, the reliance on centralized servers hosted by large technology companies that have access and utilize these powerful Artificial Intelligence (AI) tools introduces a significant vulnerability. This paper argues about the importance of deploying local Large Language Models (LLMs) on on-premise servers to mitigate the risks associated with data leakage to GAI providers.

Also, it has to be taken into account that the very act of transmitting data to remote servers inherently introduces security vulnerabilities. Data breaches, cyberattacks, and unauthorized access can compromise the confidentiality of user and company information, potentially leading to identity theft, financial fraud, or other detrimental consequences.

Deploying local LLMs on on-premise servers presents a compelling solution to data security and accessibility challenges. The University of Andorra (UdA) has implemented a local LLM server using open technologies such as ollama, OpenWebUI, and Automatic1111. This infrastructure enables the hosting of most small to medium-sized open-source LLM models for the teaching and administrative staff community. By retaining both the AI model and the processed data within a controlled environment, organizations can establish and guarantee a strong data security framework.

At the same time, local LLMs empower organizations to exercise greater control over their data and AI capabilities. They can fine tune the model to their specific needs, ensuring alignment with their business objectives and ethical considerations.

Implementation of local LLMs, however, is not without its challenges. The initial investment in hardware infrastructure can be substantial, particularly for smaller organizations. Additionally, maintaining and updating local AI models requires technical expertise and ongoing resources.

This paper focuses on discussing the pros and cons of this kind of setup as well as the bare minimums needed to host an LLM on premises both to ensure the capabilities previously mentioned and, at the same time, ensure a successful user experience. Features like talking with documents or websites using Retrieval Augmented Generation (RAG), user data integrity and privacy, or image generation should be within the initial requirements.

Finally, a qualitative survey has been conducted among the teaching and administrative staff at the UdA to gather insights and use cases for improving the setup in the future. The results revealed that data privacy and open access are the most valued features, while the quality of responses compared to other private and closed online models is perceived as the least favorable aspect. Despite this, there is promising potential for local implementations of LLMs, with plans to extend the service to students in the near future to bridge any accessibility gap they may currently face.

Keywords: LLM, open, privacy, security, user experience.

1 INTRODUCTION

As GAI continues to evolve, transforming industries and changing how we interact with technology, it's crucial to ensure equitable access and ethical development practices when using LLMs. This paper explores using free and open-source software to deploy locally hosted LLMs. A qualitative survey of beta testers to understand their user experience has also been conducted. By taking advantage of open technologies like ollama, OpenWebUI, and Automatic1111, organizations can make sure that their data is safe within the organization limits, they can also customize AI models to meet specific needs, and build and promote a collaborative ecosystem. This, not only mitigates the risks associated with

centralized commercial AI platforms but also democratizes access to cutting-edge AI capabilities, promoting inclusivity and driving progress for all.

AI, and particularly LLMs, have been an outstanding growing topic in the last two years. The concept of AI has been a relevant subject for decades, tracing its roots back over 45 years when early discussions attempted to define its scope and implications. AI seeks to emulate human intelligence in machines, providing them with the ability to reason, learn, and adapt much like their human counterparts [1]. AI enables machines to tackle these tasks with unprecedented efficiency, revolutionizing industries and changing societal norms. Today, AI emerges into a multifaceted field spanning various disciplines such as Natural Language Processing (NLP), Machine Learning (ML), and speech recognition. The objective of NLP is to provide computers with the ability to comprehend, interpret, and generate human language in a nuanced and contextually relevant manner. Several key aspects are included: text processing, language understanding, and language generation. Ethical considerations have risen with concerns about bias in linguistic models. The need for fairness, transparency and accountability in AI systems highlights the importance of ethical discourse and responsible innovation in the field. Central to the effectiveness of these models is the underlying principle of probabilistic modelling, rooted in probability theory [2]. In 2017, the advent of the Transformer architecture marked a moment in the evolution of neural network design [3]. This framework has revolutionized sequence transduction tasks by taking into account attentional mechanisms and avoiding traditional recurrent or convolutional layers. This architecture has improved numerous tasks and fields, including automatic translation, delivering remarkable results across various standard datasets. At the time, experiments were conducted to dissect the intricacies of the components of the Transformer, providing invaluable insights into its inner workings. Notably, architectures like ChatGPT owe much of their efficacy to the foundational principles embodied by the Transformer.

AI applications in education have rapidly and continuously grown in the last two years. Entities like UNESCO recommend integrating AI, particularly LLMs, into educational practices [4]. However, ethical issues such as student plagiarism have arisen as well, highlighting the need for additional research to guarantee efficient use of tools like ChatGPT, Perplexity, or similar. Incorporating AI into education offers numerous advantages, including improved learning outcomes, increased efficiency and productivity [5]. It also brings greater accessibility to education, particularly for marginalized or underserved communities [6]. Nevertheless, there are potential issues, such as concerns regarding data privacy and security, the possibility of bias in AI algorithms, and the displacement of educators [7]. In this sense, one of the objectives of this paper is to provide alternatives to remotely hosted GAIs with a potential cost that not everyone may be able to assume. At the same time, it has to be taken into account that many organizations have started using LLMs to chat with private data, either from local databases or organization documents. Even if customer agreements with remote GAI providers may protect such data, the security and peace of mind that having the resources on-premise cannot be overstated [8].

In the end, it is imperative to ensure that the integration and deployment of AI adhere to principles of human rights and social justice [9]. Achieving this requires active engagement from institutions and authorities. The growing use of AI in personalized learning, analytics, and research assistance is perceived as advantageous for society.

The local LLM environment proposed in this paper has been provisioned at the University of Andorra (UdA, https://www.uda.ad). Initially, it has been offered only internally, in the scope of a beta program. The UdA is one of the smallest universities of the world, part of the Network of Universities of Small Countries and Territories (NUSCT, https://www.nusct.net). With approximately 1,500 students enrolled annually, the university has a core staff of around 50 individuals, including both teaching and administrative personnel. At the same time, while many external assistant teachers also take part in day-to-day teaching activities, these were not offered access to the platform to avoid overcrowding it at the beginning of the beta phase. This initiative is currently being evaluated as well in other Higher Education Institutions, like the project started at FernUniversität in Hagen, Germany [10].

The Principality of Andorra presents another interesting challenge. Its official language is Catalan, and at the UdA, it is the main form of communication, both spoken and written for administrative and day-to-day tasks. In this sense, choosing a suiting LLM that could write proficiently in this language proved to be one of the most difficult tasks. The model needed to excel in both conversational interactions and understanding/reasoning with documents written in Catalan. In this sense, numerous open-source models were evaluated before selecting the one that best accomplished the requirements, even if it did not always offer the most seamless user experience in terms of Catalan writing. The Barcelona Supercomputing Centre (BSC) has an evolving project called Projecte Aina (https://projecteaina.cat)

that trains specific models based on Catalan corpuses. Currently, their flagship model is called FLOR and shows promising results in terms of both generation and chat capabilities [11].

While conversational and reasoning abilities were considered among the most important features, image generation closely followed in popularity. Nowadays, commercial platforms offering LLMs services often include extra features like generating visuals and voice integration. This research also aimed to evaluate the use of image generation models that could seamlessly integrate into such environments while efficiently sharing computer resources with the conversational LLM.

2 METHODOLOGY

The methodology followed in this paper involved several distinct steps. First, both hardware and software options were explored to identify the optimal combination for both research and user needs. Second, a thorough analysis of user requirements was conducted to design an environment tailored to their specific needs. Finally, the evaluation of user experience and future improvements were assessed.

Since the beginning of the project, it was acknowledged that building a research environment dedicated to studying LLMs behavior differed significantly from building an environment designed to be used in everyday tasks. With this in mind, the methodology was focused on building an environment that could host both a conversational LLM and an image generation model, combined with a frontend web application that could simplify access to the models.

Different hardware options were analyzed. The simplest approach involved downloading open and freely available models and running them locally on readily available computers like PCs or Macs. Newer computers, particularly those equipped with consumer-grade GPUs (Graphics Processing Units), demonstrated good performance in terms of speed (tokens generated per second) and reliability (VRAM requirements), but were still far from production ready environments were multiple users had to access the system concurrently.

With an allocated budget for building a specific computing environment, specialized GPU cards from manufacturers like NVIDIA or AMD were considered. In the past, it has become increasingly common to utilize GPU gaming cards for computationally intensive tasks due to their powerful vector processing capabilities. A notable example is the PlayStation 3, which possessed a groundbreaking graphics power at its release and subsequently saw increased adoption in research labs for research applications [12]. The same can be said for AI and LLM. GPUs, or even Language Processing Units (LPU, see Groq: https://groq.com), power is necessary. Of course, having access to various different GPU cards to determine which is the best, in the small given lab at the UdA, was cost prohibitive and a compromise had to be taken.

The methodology also involved selecting both conversational and image-generating models. Today, numerous open models are available, originating from established AI companies or innovative startups. The selection criteria prioritized open access, manageable size to align with available computer resources, and speed at generating content.

Several software options have emerged to host web frontends for interacting with LLMs and image generation models. Some of these options have been evaluated based on various factors, including user management capabilities, the number of models they can interact with simultaneously, speed, ease of use, and research capabilities.

Having analyzed the different hardware and software options to make LLMs accessible to users, a user-friendly LLM environment was launched at the UdA in the form of a beta project. From there, the focus was set on two key tasks: first, writing a clear and concise use guide for the users, many of whom were working with LLMs for the very first time. Second, valuable feedback from users regarding the usability, reliability, speed, and potential applications of the LLM environment was gathered during the beginning of the 24-25 academic year.

2.1 Hardware

Implementing a local and private LLM solution presents many challenges, particularly regarding resource access and management. One key issue is VRAM usage, as both conversational and image generation models require substantial memory to operate effectively. The speed at which an LLM generates text is directly influenced, among other factors, to the amount of memory available and the possibility of loading the model completely in such VRAM. While it is technically possible to run an LLM using a standard CPU and RAM, this approach severely limits its real-world applications due to slow

processing speeds. A GPU, or an LPU, is essential for efficient LLM operation. Fortunately, the rest of the components do not need to be cutting-edge; even basic hardware can suffice when paired with a capable GPU. Open-source model providers are aware of hardware restrictions and offer different versions of their models designed to run on smaller GPUs. It is common to find open-source models with 6 to 12 billion parameters that can run efficiently on any GPU with 16 GB of VRAM.

The quality of a conversational model is directly influenced by several factors, including the number of parameters and the quantization techniques used to adapt it to smaller GPUs. Quantization involves reducing the precision of numerical representations within the model, ranging from full 32-bit or 16-bit precision down to Q2 with various options in between [13]. For example, an 8 billion parameter model like llama3.1 can be adjusted to fit into a 4GB VRAM GPU card using quantization. However, it is crucial to note that lower quantization levels directly impact the quality of the model's output. Finding the optimal balance between inference speed and generated quality often requires a compromise in the level of quantization used. In terms of reasoning capabilities and has been discussed that larger models, in terms of billions of parameters, perform better, but hosting these in low VRAM by heavily quantizing them may affect output quality [14].

NVIDIA ships different card models. Most of these are known to be game oriented, but drivers and interfaces have been developed to access the card capabilities from other environments (see CUDA: https://developer.nvidia.com/cuda-toolkit). NVIDIA GPUs are widely adopted for this purpose due to their high performance and parallel processing capabilities. The H100, NVIDIA's latest flagship GPU, excels in LLM inference with its Transformer Engine architecture specifically designed for accelerating transformer networks common in LLMs. It boasts significantly higher memory bandwidth and tensor cores compared to previous generations like the A100. While the A100 remains a powerful option, the H100 offers a notable performance leap for demanding LLM workloads. End-users seeking more accessible options, the RTX 3090 and 4090 consumer-grade GPUs can also be utilized for LLM inference. The RTX 4090 surpasses the 3090 in terms of CUDA cores and memory bandwidth, leading to faster inference times. However, they lack the specialized hardware optimizations found in the data center-focused H100 and A100, making them less efficient for large-scale LLM deployment.

Due to the prohibitive high cost of premium models, a consumer-grade RTX 4090 NVIDIA GPU with 24GB of VRAM was acquired and used in this research. This GPU card was chosen for its ability to accommodate both the conversational model and the image generation model while allowing access to different concurrent users without a huge impact on user experience.

2.2 Software

Interacting with an LLM requires two basic components: a conversational, multimodal or image generation model (albeit an open-source one) and an inference tool. The inference tool is responsible for loading the model, accepting user prompts, and returning the generated content – either text or images – back to the user. In recent years, as these models have become more accessible and manageable, a plethora of libraries, like llama.cpp (https://github.com/ggerganov/llama.cpp) or ollama (https://ollama.com), have emerged within the open-source community.

2.2.1 Choosing an open-source conversational model and an image generation model

Recently, numerous companies have released open-source versions of their flagship AI models. Prominent examples include Google's Gemma [15], Meta's Llama [16], Microsoft's Phi [17], or Mistral's Mixtral [18]. In the image generation field, the models released by Stability AI, i.e. SDXL, or Black Forest Labs' Flux.1 are among the most praised by the community [19]. These models, both conversational and image generation, are often available in various sizes, both in terms of the number of parameters and quantization, to accommodate diverse computational environments.

While the accessibility of these models simplifies evaluation and comparison, selecting a suitable model for specific tasks, particularly those involving languages like Catalan, remains challenging due to varying prerequisites. Models can be readily obtained from official company websites, platforms like HuggingFace, or public repositories such as the one ollama uses. Desktop applications, such as LM Studio, further streamline the process by providing search capabilities and tools for rapid inference.

Choosing a model involves considering factors beyond language proficiency, including size (measured in parameters), quantization techniques, and inferencing speed. For this particular project, several models were evaluated:

- Mistral: Mistral 7B, Nemo 12B and Mixtral 8x7B

- Google: Gemma2 9B and 27B
- Microsoft: Phi3 3B and 14B, Phi3.5 3B
- Meta: Llama3 8B and Llama 3.1 8B
- Projecte Aina: Aguila 7B and FLOR 7B

These models, with the exception of Aina's models, were available in various quantization flavors, with Q4_0 being the default and most common when using ollama. Other quantization variants could be easily accessed based on available VRAM and desired quality.

Extensive testing was conducted at the UdA to determine model compatibility with an RTX 4090 graphics card based on the number of parameters and quantization. Only Mistral's Mixtral 8x7B and Google's Gemma2 27B required GPU offloading, that is, that part of the model had to be run on the local CPU at high quantization values. The remaining models could be loaded even at fp16 quantization, which is arguably the best one. Notably, a fixed amount of VRAM had to be reserved for the image generation model.

To assess Catalan language proficiency, a self-evaluating test was developed involving the generation of brief (150-word) texts on diverse topics. The same models were also used to evaluate the correctness of the outputs on a scale of 1 to 10. Across all tests, Gemma2 consistently demonstrated superior performance, achieving high scores with both its 9B and 27B parameter configurations. Mistral's Mixtral 8x7B model closely followed Gemma2, which aligned with expectations given its size. However, resource constraints prevented deploying this larger model on the server.

Embedding models are also crucial for tasks like Retrieval Augmented Generation (RAG) on local documents or when accessing web pages. OpenAI's all-mini embedding model is widely used due to its low VRAM consumption and effectiveness. Alternative options include nomic-embed-text and mxbai-embed-text. While embedding models require significantly less VRAM than conversational models, their usage must also be accounted for.

At the time of writing, the image generation model SDXL 1.0 from Stability AI emerged as a cost-effective image generation model, consuming approximately 4GB of VRAM and producing high-quality results. Other evaluated models included Stability AI's SD3 or Black Forest Labs's Flux.1 [dev]. While Flux.1 [dev] delivered superior results compared to other local open-source models, its substantial VRAM requirements hindered the simultaneous loading of conversational models. Consequently, prioritizing a larger conversational model over image generation quality was deemed necessary.

2.2.2 LLM inference tools

Ollama was chosen as the LLM inference tool because it best met the needs of the proposed environment after evaluating different software solutions and their integration possibilities. Nowadays, this software is recognized as one of the most popular command-line tools for LLM inference. Its ease-of-use and simplicity are highly regarded features. The number of compatible models grows daily, with up-to-date open models running seamlessly. Ollama runs on GNU/Linux, macOS, and Windows, and it can be deployed either as a service or through command-line interaction.

Alternatives to ollama certainly exist for end users who wish to run models locally on their computers. Open-source solutions like LM Studio (https://lmstudio.ai), Jan (https://jan.ai), GPT4All (https://www.nomic.ai/gpt4all), or AnythingLLM (https://anythingllm.com) among other, are highly accessible due to their ease of installation and use. H2O.ai also offers an LLM solution called h2oGPT (https://github.com/h2oai/h2ogpt). These applications share several attractive features: chats remain private never leaving the device; models can run on standard consumer hardware without modifications; and they offer functionality for interacting with local files. However, these feature-rich applications are not easily adaptable for multiuser environments and, again, they lack image generation capabilities. A particular case of freely accessible open-source models on the web should also be mentioned. DuckDuckGo, commonly known for their focus on privacy and security, offers access to different models on their website, promising anonymity and never storing chat information (https://duckduckgo.com/chat).

API accessed solutions, while lacking the ease-of-use of desktop applications, provide centralized access that web applications can leverage. In this context, ollama falls into the category of server-based software. Like its desktop counterparts, ollama prioritizes privacy and local execution. However, it does not include advanced features such as document chat or handling; these have to be implemented using additional software.

Other tools comparable to ollama include llama.cpp, which offers minimal setup for LLM inference across devices. This C++ port of Llama2 supports GGUF format models, including multimodal ones like LLaVA. Its efficiency makes it suitable for consumer hardware and edge devices. Vllm is another viable option, serving as a high-throughput and memory-efficient engine for LLM inference. Finally, the HuggingFace and transformers library cannot be overlooked. This library empowers programmers and researchers to interact with LLMs using programming languages like C or Python, unlocking a wide range of model capabilities.

The chosen solution offered two key parameters that significantly impact user experience and response times: model loading capacity and response concurrency. Firstly, the maximum number of loaded models influences both speed and resource consumption. While having multiple models (each specialized for tasks like chatting or embedding) readily available improves response times, but it also increases VRAM usage. Secondly, adjusting the number of concurrent answers a model can provide also affects user experience. Allowing multiple users to receive responses simultaneously eliminates waiting queues, enhancing interaction fluidity. However, this benefit comes at the cost of increased VRAM consumption per concurrent user. In the proposed environment, due to hardware capabilities, both parameters had to be set to 1.

For the image generation Frontend/API software, two options emerged as popular choices at the time of writing: Automatic1111 and ComfyUI. Automatic1111 provides a straightforward gradio web interface capable of loading Stability AI's (i.e. SDXL or SD3) models. ComfyUI, on the other hand, offers a graphical workflow editor for fine-grained configuration. ComfyUI could, also, load the Flux.1 model. Both options were thoroughly assessed, with neither demonstrating a clear advantage. Ultimately, the software that best and easiest integrated with the chosen web application frontend was chosen - in this case, Automatic1111.

2.2.3 End user frontend web application

Configured with ollama as the LLM inference tool, Gemma2 27B as the conversational model, and Automatic1111 and SDXL as the image generation tools, the system required a suitable application for user interaction. This application, whether web-based or desktop, needed to fulfill several key requirements: bindings to both ollama and Automatic1111, user management options, and the capability to facilitate conversations not only with models but also with uploaded documents and web pages.

The best and only viable application that was found was OpenWebUI. This web application fits perfectly with all requirements. Namely, these include:

- User management and role-based access control
- Responsive design
- API Access to ollama and OpenAI, among other
- Integrated image generation supporting both Automatic1111 and ComfyUI
- Retrieval-Augmented Generation and web access directly from the chat
- Prompt templates support
- Concurrent multiple model evaluation
- Multilingual support, with Catalan included by default

2.2.4 Final implementation

To end this section, a summary of both hardware and software technologies that were used to build the final environment is presented below.

Hardware:

- Standard PC with an Intel i9 processor, 64 GB of RAM and 1TB of disk
- NVIDIA RTX 4090 with 24 GB of VRAM

Software:

- GNU/Linux Debian 12 *bookworm* with latest official NVIDIA drivers
- Current ollama (version 0.3.10 at the time of writing)
- Google's Gemma2 27B – Q4_0 LLM model (with a context size of 4096)

- Latest Automatic1111 version
- Stability AI - SDXL 1.0 image generation model
- OpenWebUI (version 0.3.21 at the time of writing)

2.3 Evaluating the user's experience

Having been built, installed, and thoroughly tested, the UdA local LLM environment was opened to beta testers. Of the 50 initial offered user slots, 33 were requested. After a three-month period, participants were surveyed with qualitative questions regarding their experiences. These included inquiries about how they were using the system (use cases), ease of use, response quality in both English and Catalan, web application features, and potential areas for improvement.

3 RESULTS

On the one hand, the evaluation of the server hosting ollama as the LLM inference tool revealed performance consistent with our projected benchmarks. The system demonstrates efficient and reliable handling of inference requests, delivering timely and accurate results for all user queries. It should be noted that concurrent user traffic was generally low. While the current setup performs admirably, we anticipate further performance gains with the integration of additional NVIDIA 4090 GPUs. This hardware upgrade would significantly enhance the server's processing power, allowing it to handle a larger volume of concurrent requests while maintaining rapid response times and, in general, improve user experience.

On the other hand, the qualitative analysis of the interviews revealed a range of perspectives on the use of the implemented local LLM. Participants included professional and administrative staff, teachers and researchers at the UdA. Out of a total of 33 potential user, 27 answered the questionnaire. The interviews focused on assessing users' familiarity with LLMs and particularly with the locally developed solution. Participants were asked about their usage experience, what they considered to be the key strengths of the implemented LLM, as well as any perceived weaknesses. Finally, they were encouraged to share suggestions to improve the solution.

The most frequently mentioned benefits in the interview responses regarding the use of the local LLM include its exceptional ability to translate texts accurately, making it a valuable tool for multilingual tasks. Additionally, it has been praised for its efficiency in generating text, which significantly reduces the time spent on tedious or mechanical tasks. Users also highlighted its usefulness in creating classroom activities, analyzing and summarizing shorter texts. Beyond these practical functions, many respondents noted that the local LLM serves as a creative catalyst, helping them generate new ideas and overcome creative blocks, particularly when starting a task or when they find themselves stuck. This combination of practical assistance and creative support has made the proposed environment a versatile tool in both academic and administrative contexts.

The interviews revealed several key themes regarding the use of this solution. In terms of frequency and use cases, a significant number of participants reported using the tool occasionally, often for professional tasks such as summarizing documents, generating content, or handling administrative duties. While some users were hesitant to adopt the tool fully, others noted its frequent use for specific work-related tasks, such as giving new ideas or text generation. Some respondents opted not to use it, preferring more powerful models like ChatGPT for tasks that do not involve sensitive data.

Many respondents felt that ChatGPT, or similar, outperformed the local LLM in terms of language accuracy, flexibility, and overall capabilities. The local LLM often required more specific prompts, leading to the perception that it was less powerful or refined. However, some users acknowledged that this LLM served well in tasks where security is a priority, such as text analysis, document drafting and evaluating some exams. Nevertheless, its image generation feature received some criticism for lacking sophistication compared to other LLMs.

Language performance emerged as a mixed point of feedback. While this solution performed well in Spanish, English and French, its support for Catalan was inconsistent. Some users expressed frustration with the model's tendency to switch languages unexpectedly, particularly when it shifted from Catalan to Spanish or English, which disrupted the experience for users working primarily in Catalan. Additionally, some users encountered issues with certain words in Catalan, but when using other languages, primarily Spanish or English, they did not experience any problems.

To improve the local LLM solution, users recommended to improve its accuracy and fluency in generating responses, aiming to match the performance of leading LLMs like ChatGPT. Investing in more advanced image generation algorithms would also make the tool more versatile, particularly for creative tasks. Addressing the issue of language switching would ensure more consistent language support. Additionally, expanding pre-built templates and use cases tailored to specific academic disciplines could encourage broader adoption among students and faculty.

Data security was one of the most praised features. Respondents appreciated that the tool processes data within the UdA infrastructure, offering a level of trust and security not found in any other LLMs. This local hosting of data was seen as a major advantage, and for some users, it justified their use even when its technical capabilities fell short of competitors like ChatGPT.

User experience and interface feedback were generally positive. Many respondents found the interface familiar and easy to navigate, especially for those with experience using similar tools. However, few users felt that the interface lacked certain intuitive features and polish, suggesting that further improvements could enhance user-friendliness, particularly for those less familiar with LLMs.

Some respondents proposed that granting students access to this local LLM could be highly advantageous, particularly in academic settings. They envisioned a scenario where the LLM could be integrated into the learning process by providing it access to all seminar materials. In this context, students could engage with the LLM to ask questions, clarify specific concepts, and explore topics in greater depth. By acting as a UdA assistant, the local LLM could enhance students' understanding of complex subjects, offering real-time feedback and personalized explanations. This would not only help students reinforce their knowledge but also encourage independent learning, critical thinking, and problem-solving. Moreover, such a tool could assist in study preparation and revision, making it easier for students to navigate difficult material and improving a more interactive and dynamic learning environment.

Finally, interviewees expressed a strong sense of pride in the fact that the UdA has its own local LLM. They see it as a significant achievement for the institution, reflecting both innovation and self-reliance in the field of AI. This homegrown technology not only reinforces the university’s commitment to data privacy and security but also highlights its ability to create advanced tools tailored specifically to the needs of the academic community. The existence of a locally developed LLM serves as a symbol of progress, enhancing the university's reputation and reinforcing a sense of ownership.

4 CONCLUSIONS

This paper discusses that open-source LLMs hosted locally and in an open manner is not only the best approach for educational settings, but also the one that encompasses best practices overall. It balances key benefits like data protection and cost savings against the challenge of maintenance effort. To illustrate this, an open LLM has been implemented at the UdA, a practical example that can serve as a guide for others wanting to implement similar systems. Using readily available hardware, it has been shown that providing access to open LLMs is achievable, enabling a wide range of applications in higher education settings. More importantly, it prioritizes data security because all LLM requests are processed entirely within the university network. This approach has the potential to be adapted for other sectors, even if it has been initially developed for educational institutions.

One ongoing challenge is determining the best LLM model for specific needs, as there are many options available. Different use cases may require different models. Examples of these may be fine-tuned LLMs for specific topics, multimodal LLMs to provide vision capabilities, or different image generation models depending on the needs of the user. A possible solution involves running multiple LLMs concurrently, an approach that is being currently evaluated. Recent versions of ollama and OpenWebUI already support concurrent models as well as the possibility of running the frontend in a high-availability scheme, lowering the resource necessities and being able to distribute the load accordingly.

The proposed environment was highly appreciated by the interviewed users for its strong data privacy and security, giving them confidence that their data remained within the university's secure network. Its ease of use was also widely praised, with users finding the interface intuitive and familiar, enabling quick adoption even for those experienced with other LLMs. However, one of its main weaknesses was its performance compared to other LLMs like ChatGPT, with users consistently noting it was less accurate and required more detailed prompts to generate acceptable results, often producing less polished outputs. The image generation feature was also considered underdeveloped, lacking the sophistication of other specialized models. Additionally, some users experienced occasional language errors, where

the LLM would unexpectedly switch to English despite being prompted in Catalan, impacting the overall user experience.

Finally, this research highlights a key observation: Catalan language models are still developing and have not quite reached the same level of sophistication as models for more common languages like English, Chinese, Spanish, or French.

ACKNOWLEDGEMENTS

We would like to express our sincere gratitude to the University of Andorra for allowing us to conduct the interviews and for providing us with the necessary resources. We also want to give a special thanks to the teaching and administrative staff members who participated in the interviews.

REFERENCES

[1] R. Kurzweil, R. Richter, R. Kurzweil, and M. L. Schneider, "The age of intelligent machines.," *MIT press Cambridge*, vol. 580, 1990.

[2] Y. Bengio, R. Ducharme, and P. Vincent, "A Neural Probabilistic Language Model," *Advances in neural information processing systems*, vol. 13, 2000. DOI: https://doi.org/10.1007/10985687_6 [Online].

[3] A. Vaswani, N. Shazeer, N. Parmar, *et al*., "Attention Is All You Need," *Advances in neural information processing systems*, vol. 30, 2017. DOI: https://doi.org/10.48550/arXiv.1706.03762. [Online].

[4] E. Sabzalieva and A. Valentini, "ChatGPT and Artificial Intelligence in higher education: Quick start guide," 2023.

[5] F. Miao, W. Holmes, R. Huang, H. Zhang *et al*., "AI and education: Guidance for policy-makers.," *UNESCO Publishing,* 2021.

[6] EU Parliament, "EU AI Act: first regulation on artificial intelligence," 2023.

[7] B. C. Stahl and D. Eke, "The ethics of ChatGPT – Exploring the ethical issues of an emerging technology," *International Journal of Information Management*, vol. 74, p. 102 700, 2024. DOI: https://doi.org/10.1016/j.ijinfomgt.2023.102700. [Online].

[8] S. Balloccu, P. Schmidtová, M. Lango, and O. Dusek, "Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs," *arXiv preprint arXiv:2402.03927*, 2024.

[9] UNESCO, "Beijing consensus on artificial intelligence and education," 2019.

[10] T. Zesch, M. Hanses, N. Seidel, P. Aggarwal, D. Veiel, and C. de Witt, "Fernuni LLM experimental infrastructure (FLEXI) – Enabling experimentation and innovation in higher education through access to open large language models," *arXiv preprint arXiv:2407.13013*, 2024. DOI: https://doi.org/10.48550/arXiv.2407.13013 [Online]. Available: https://arxiv.org/pdf/2407.13013

[11] S. Da Dalt, J. Llop, I. Baucells, M. Pàmies, Y. Xu, A. Gonzalez-Agirre, and M. Villegas, "Flor: On the effectiveness of language adaptation," in Proceedings of the *2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (LREC-COLING 2024), pp. 7377–7388, 2024.

[12] J. Kurzak, A. Buttari, P. Luszczek, and J. Dongarra, "The playstation 3 for high-performance scientific computing," *Computing in Science & Engineering*, vol. 10, no. 3, pp. 84–87, 2008.

[13] K. Marchisio, S. Dash, H. Chen, D. Aumiller, A. Üstün, S. Hooker, and S. Ruder, "How does quantization affect multilingual LLMs?," arXiv preprint arXiv:2407.03211, 2024.

[14] Z. Li, Y. Cao, X. Xu, J. Jiang, X. Liu, Y. S. Teo, S.-w. Lin, and Y. Liu, "LLMs for relational reasoning: How far are we?," arXiv preprint arXiv:2401.09042, 2024.

[15] G. Team, T. Mesnard, C. Hardin, *et al*., "Gemma: Open models based on gemini research and technology," *arXiv preprint arXiv:2403.08295*, 2024. DOI: https://doi.org/10.48550/arXiv.2403.08295. [Online]. Available: https://arxiv.org/abs/2403.08295

[16] H. Touvron, T. Lavril, G. Izacard, *et al*., “Llama: Open and efficient foundation language models,” *arXiv preprint arXiv:2302.13971*, 2023. DOI: https://doi.org/10.48550/ arXiv.2302.13971. [Online]. Available: https://arxiv.org/abs/2302.13971

[17] S. Gunasekar, Y. Zhang, J. Aneja, *et al*., “Textbooks are all you need,” *arXiv preprint arXiv:2306.11644*, 2023. DOI: https://doi.org/10.48550/arXiv.2306.11644. [Online]. Available: https://arxiv.org/abs/2306.11644

[18] A. Q. Jiang, A. Sablayrolles, A. Mensch, *et al*., “Mistral 7b,” *arXiv preprint arXiv:2310.06825*, 2023. DOI: https://doi.org/10.48550/arXiv.2310.06825. [Online]. Available: https://arxiv.org/abs/2310.06825

[19] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Múller, J. Penna, and R. Rombach, “Sdxl: Improving latent diffusion models for high-resolution image synthesis,” *arXiv preprint arXiv:2307.01952*, 2023.