Legal and Ethical Considerations that Hinder the Use of LLMs in a Finnish Institution of Higher Education

Mika Hämäläinen
Metropolia University of Applied Sciences
Helsinki, Finland
mika.hamalainen@metropolia.fi

Abstract

Large language models (LLMs) make it possible to solve many business problems easier than ever before. However, embracing LLMs in an organization may be slowed down due to ethical and legal considerations. In this paper, we will describe some of these issues we have faced at our university while developing university-level NLP tools to empower teaching and study planning. The identified issues touch upon topics such as GDPR, copyright, user account management and fear towards the new technology.

Keywords: GDPR, privacy, copyright, GenAI, policies, AI ethics, LLM

1. Introduction

In this paper, we will describe our practical experience in building university-level NLP solutions in Metropolia University of Applied Sciences in Finland. We are developing two tools, a Moodle plugin and a tool for planning curricula, both of which rely heavily on Vertex AI[1] (see Kharlashkin et al. 2024), which lets us use PaLM 2 (Anil et al., 2023), a Large Language Model (LLM) provided by Google.

While Vertex AI makes the development process easy, it is not free of legal and ethical hurdles. What makes maters more difficult are the rigid organizational practices related to data safety and user account control that do not scale to the requirements of modern AI.

In 2023, GenAI emerged seemingly out of nowhere (see García-Peñalvo and Vázquez-Ingelmo 2023), at least this was the case for people who were outside of the scientific discipline of NLP. The IT departments of many organizations have been caught off guard with great many staff members asking for ready-made tools like ChatGPT or Copilot (cf. Uren and Edwards 2023). Even these off-the-shelf tools cannot be taken into use in a big organization without planning, let alone custom-made NLP solutions the likes of which we are developing.

This paper presents our experience on the challenges we have faced while striving for a usable AI solution that can provide teachers and study planners alike with the added value of NLP based analysis and content creation.

2. Why can we not host an LLM locally?

The simplest solution to most of our legal issues would be hosting our own LLM instead of relying on Vertex AI. With the fast development of open LLMs (Groeneveld et al., 2024; Chen et al., 2024; Penedo et al., 2023) and their remarkable benchmark performance, together with their availability through tools like Hugging Face Transformers (Wolf et al., 2020), makes this question a valid one that requires serious consideration.

The issue we have with this is not only related to the computational requirements such models have, but it is more related to the fact that our tools must be able to understand and generate Finnish. Based on our trials, many of the open models either do not support Finnish at all or they struggle with the Finnish grammar. All models we have tried and that do support Finnish fail to produce grammatical Finnish. The commercial models PaLM 2 and ChatGPT are capable of producing mostly fluent Finnish due to their sheer size both in terms of training data and parameters. However, even these models make occasional mistakes.

There is one open LLM under development that is tailored for the Finnish language. This LLM is called Poro[2], but, at the time of writing, the model is not yet fully trained and it has not been fine-tuned to handle ChatGPT-like prompting. This means that, for the time being, our only viable solution is to use either PaLM 2 or GPT-4.

[1] https://cloud.google.com/vertex-ai?hl=en

[2] https://huggingface.co/LumiOpen/Poro-34B

LEGAL2024 @LREC-COLING-2024, pages 24–27
20 May, 2024.

3. GDPR and preventing personal data leak

Because we are bound to use an LLM provided by an external provider, the first question that we, as an organization, need to tackle is the GDPR law[3]. OpenAI has already faced legal troubles in one of the member states, namely Italy[4], due to their failure to meet the regulatory requirements on data protection as established by the GDPR law.

This unfortunate situation only leaves us with one alternative: Google Cloud. According to their terms and conditions, Google both promises that our data will not be use for training their LLM[5] and that the output of their model will not violate the copyrights of any author[6]. Of course, it is impossible for us to know the degree to which this is true, but they are the only option we have.

Our Moodle plugin is designed to analyze teachers' slides using an LLM and provide teachers with useful feedback on how to incorporate sustainable development goals into their teaching, what kind of assignments they might give and so on. As slides may very well contain personal information such as students' names or email addresses, we need to anonymize them locally before analyzing them using Vertex AI.

Our anonymization is as simple as running a Finnish named entity recognition model (Luoma et al., 2020) trained on the Finnish BERT model (Virtanen et al., 2019). We use the entity tag "PERSON" to remove all names from the slides. In addition, we use regular expressions to remove emails and phone numbers. We see that this is a necessary step in protecting the privacy of our students and staff regardless of what Google promises us in their terms of service. This approach, however, is not free of problems. Many slides will have citations to authors, which will get removed as well because they are recognized as names.

4. Copyright, work contract and teachers' rights

The first line of organizational resistance we faced was the question of copyrights. Can we analyze teachers' slides using Vertex AI without violating their copyrights? This might sound strange given that teachers are members of our staff. According to the Finnish copyright law[7], copyright is something a person will have when they produce something that is copyrightable. This means that an organization, such as a university, will not automatically get copyrights to the creative work of its staff.

Our organization has a statement of copyright transfer in new work contracts, but this statement was absent in older work contracts. This means that we have several lecturers and senior lecturers who have never given the university any rights for their copyright protected material, such as course slides.

Copyright only protects the form, not the idea behind any creative work. This means that we can, legally, use an LLM to analyze copyrighted material for as long as we don't reproduce that material to a significant degree. Because GenAI as such a new thing and the wording in the Finnish copyright law has never been meant to cover such a use-case, we have taken a more ethical approach and opted for protecting the copyrights more than what the law protects.

In practice, this means that we will not analyze teachers' materials automatically, but we instead let the teachers decide which slides they want to analyze and let them be in charge of deciding whether they want to even use our NLP tool or not.

5. User account management and access rights

Another issue that AI brings to the table is that of access rights. The question of what type of data can be passed to an existing solution like Copilot or ChatGPT is one part of the discussion. Another, larger part is the question whether the organization has the user rights management on such a nuanced level that AI tools can be given only the rights they need and nothing more.

Moodle makes it easy to create an "AI user" that can only access slides. However, our university also uses another learning environment named Peppi[8], which does not have any nuanced access right management. A user can either have access to everything in the system, including student's grades, essays and so on, or individual student or teacher access for a given course.

An individual access means that the NLP tool should be added manually to each course. Even this would not solve this issue because the AI would need to have a teacher access to be able to access slides before they are made available for the students. Teacher rights also entail access to students' grades. Even though we are developing the NLP

[3] https://www.consilium.europa.eu/en/policies/data-protection/data-protection-regulation/

[4] https://techcrunch.com/2024/01/29/chatgpt-italy-gdpr-notification/

[5] https://cloud.google.com/terms/data-processing-addendum/

[6] https://cloud.google.com/blog/products/ai-machine-learning/protecting-customers-with-generative-ai-indemnification

[7] https://www.finlex.fi/fi/laki/ajantasa/1961/19610404

[8] https://www.peppi-konsortio.fi

tool by ourselves, we do not want to give rights to grades to any additional systems because this will increase the attack surface (see Schuster and Holz 2013) and potential leak of data.

6. Fear or ethical thinking?

AI ethics is a serious consideration, and there is a growing body of work that highlights issues such as bias on gender (Shrestha and Das, 2022), ethnicity (Garrido-Muñoz et al., 2021) and sexual orientation (Felkner et al., 2022). Especially in the education sector, attempts to grade students or profile them fully automatically are not ethically sustainable. Our goal is not to diminish any of these real ethical considerations, but to highlight the difference between these AI expert driven ethical considerations and those of non-technical people.

We have faced that "dropping the ethics bomb" is a way for certain actors within the organization to hinder the organization from embracing AI. Many times people cannot even verbalize what the exact ethical problems are, but oftentimes it is fear that tools like ChatGPT might produce erroneus output. Or, in the case of image generation systems, such as Adobe Firefly, that using such machine generated images is somewhat morally wrong. Curiously, the teachers who are most strongly against GenAI, are also the ones that demand that all student work be passed through a GenAI detector. Using ChatGPT is seen as unethical, but failing students because of an AI detector's analysis is seen as ethical. Despite the fact that recognizing AI generated text is not an accurate practice (Chaka, 2023).

Much of the fear towards the new technology is thus masked as an ethical consideration. The way non-technical people approach AI ethics in our organization is strikingly different from the way NLP researchers approach the problem.

7. Conclusions

In this paper, I have presented some of the legal and ethical issues we have faced in our organization when embracing LLMs. Despite these problems, many members of the staff are eager about the possibilities our NLP tools bring to teaching and study planning. However, the road to production is long and bureaucratic.

8. Bibliographical References

Chaka Chaka. 2023. Detecting ai content in responses generated by chatgpt, youchat, and chatsonic: The case of five ai content detection tools. *Journal of Applied Learning and Teaching*, 6(2).

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv*.

Virginia K Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2022. Towards winoqueer: Developing a benchmark for anti-queer bias in large language models. *arXiv preprint arXiv:2206.11484*.

Francisco García-Peñalvo and Andrea Vázquez-Ingelmo. 2023. What do we mean by genai? a systematic mapping of the evolution, trends, and techniques involved in generative ai.

Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L Alfonso Ureña-López. 2021. A survey on bias in deep nlp. *Applied Sciences*, 11(7):3184.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. 2024. Olmo: Accelerating the science of language models. *Preprint*.

Lev Kharlashkin, Melany Macias, Leo Huovinen, and Mika Hämäläinen. 2024. Predicting sustainable development goals using course descriptions–from llms to conventional foundation models. *arXiv e-prints*, pages arXiv–2402.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Felix Schuster and Thorsten Holz. 2013. Towards reducing the attack surface of software backdoors. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 851–862.

Sunny Shrestha and Sanchari Das. 2022. Exploring gender biases in ml and ai academic research through systematic literature review. *Frontiers in artificial intelligence*, 5:976838.

Victoria Uren and John S Edwards. 2023. Technology readiness and the organizational journey towards ai adoption: An empirical study. *International Journal of Information Management*, 68:102588.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: Bert for finnish. *arXiv preprint arXiv:1912.07076*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

9. Language Resource References

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Jouni Luoma, Miika Oinonen, Maria Pyykönen, Veronika Laippala, and Sampo Pyysalo. 2020. A broad-coverage corpus for Finnish named entity recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4615–4624.