Teaching LLMs at Charles University: Assignments and Activities

Jindřich Helcl and **Zdeněk Kasner** and **Ondřej Dušek** and **Tomasz Limisiewicz**
and **Dominik Macháček** and **Tomáš Musil** and **Jindřich Libovický**
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics, Charles University
V Holešovičkách 2, 180 00, Prague, Czech Republic
`${surname}@ufal.mff.cuni.cz`

Abstract

This paper presents teaching materials, particularly assignments and ideas for classroom activities, from a new course on large language models (LLMs) taught at Charles University. The assignments include experiments with LLM inference for weather report generation and machine translation. The classroom activities include class quizzes, focused research on downstream tasks and datasets, and an interactive "best paper" session aimed at reading and comprehension of research papers.

1 Introduction

Reflecting contemporary trends in education is a challenging task. The teachers often need to decide which promising topics to cover in their course and which topics are better to leave for discussion in reading groups. The unstable nature of research progress also means that courses that are not updated regularly lose their relevance in time. However, when the trend becomes as prominent as large language models (LLMs) have become, the need for a systematic overview in the form of a specialized course gets increasingly urgent.

This paper presents one of these efforts – a new course taught at Charles University composed of a series of LLM-related lectures and interactive sessions. During its first year in 2024, 51 students enrolled in the optional course, mainly attending local BSc or MSc study programmes.

The course is composed of 13 sessions with various levels of interactivity, including lectures, directed discussions on the current topic, quizzes, as well as practical work with LLMs (Section 2) and other classroom activities (Section 3). After a broader discussion in the first session, the course focused on the following topics: The Transformer model (Vaswani et al., 2017), LLM training and inference, data collection and evaluation, LLM applications, efficiency, multilinguality, speech processing, translation, meaning/understanding, and ethics of LLM training and use.

All course materials, including slides, recordings, and assignments, are available on the course website.[1]

2 Assignments

We organized two assignment-based sessions focused on (1) generating weather reports using LLMs and (2) using LLMs for machine translation (MT). For each task, we ran instances of different models on our GPU cluster with an API provided by the `text-generation-webui`[2] package. The API allowed the students to access and configure the models without the need to access specialized hardware or rely on commercial platforms.

In both tasks, the students worked in small teams (up to 5 people), and were provided with a starter code[3] that would call the model API with a specified set of parameters. Besides choosing the appropriate prompts, the teams experimented with various decoding parameters, including the sampling temperature, the k and p parameters for top-k and top-p sampling, and the beam size.

Text Generation. In this task, the students were asked to generate weather reports in natural language using an LLM. The students were provided with a selection of JSON files retrieved from the openweathermap.org API for various cities. The assignment was divided into 4 subtasks: (1) generating a report about the current weather, (2) generating a 5-day forecast, (3) generating a report in a language other than English, and (4) changing the forecast style (e.g., for specific target groups).

The four models the students experimented with

---

[1] https://ufal.mff.cuni.cz/courses/npfl140
[2] https://github.com/oobabooga/text-generation-webui
[3] https://github.com/kasnerz/npfl140

were Mistral 7B,[4] Mistral 7B Instruct (Jiang et al., 2023),[5] Phi-2 (Javaheripi et al., 2023),[6] and Aya-101 (Üstün et al., 2024).[7] The students reported on the difficulties of generating factually accurate outputs from the models, confirming recent findings (Kasner and Dušek, 2024). They also proposed improved data preprocessing, prompt formatting, and decoding parameters.

Machine Translation. In the MT assignment, the teams were given paragraphs of text in 21 (unknown) languages and instructed to translate them using an LLM into English and then into a language of their choice. Again, the students experimented with prompt engineering and decoding parameters.

For the first part, we created a simple web app[8] for submitting the English translation, which computed the Character F-score (Popović, 2015) and showed a leaderboard of the 10 best-scoring teams per language during the session. After the assignment, the leaderboard can be configured to show the source language and the reference translations. The leaderboard then shows all submissions made to the app except those marked as debug submissions. In the second part of the assignment, the students were asked to experiment with translation into a language of their choice. They should submit a report, which is due a week after the hands-on session.

We used a slightly different set of models compared to the previous assignment. Mistral 7B Instruct and Aya-101 remained, and we added translation-specific models Tower Instruct (Alves et al., 2024) and ALMA-R (Xu et al., 2024).

When translating into English from medium-resourced languages, the students could generally match the quality of commercial MT systems. However, translating into their languages (e.g., Slovak, Ukrainian, Georgian, Serbian) appeared challenging.

3 Classroom Activities

Discussions. Discussion among the students was a recurring activity in many of the sessions. To encourage as many students to participate, we either let them discuss in small groups and then present their position, or we used interactive slides to collect and show their input in real-time,[9] which also encouraged less self-confident students to share their opinion.

Discussion is an effective method for teaching non-technical topics. In the final session on this course, we focused on two primary areas. The first area involves the question of whether LLMs can truly understand language. We recommend engaging students in discussions about various thought experiments (e.g. Searle, 1985; Bender and Koller, 2020) and exploring both sides of the debate: those who argue that it is impossible (e.g. Bender et al., 2021) and those who believe it is possible to some extent (e.g. Andreas, 2022; Søgaard, 2022). The second area covers ethical considerations. Here, students discussed environmental and labor issues related to training LLMs (e.g. Bender et al., 2021) and the broader challenges associated with the development and deployment of language technologies (e.g. Jørgensen and Søgaard, 2023).

Class Quizzes. Every session began with a short multiple-choice quiz based on the topics from the previous class. These quizzes were implemented using a simple web app[10] that shows a QR code to join the quiz, and after a certain amount of time, it shows the results. Each question can be answered multiple times until the correct choice is selected, providing immediate feedback to the students. When the time is up, the app shows the correct answers and the number of unsuccessful attempts for each incorrect choice.

At the final session, students complete a similar immediate-feedback test in the form of scratch cards (Epstein et al., 2001).

Downstream tasks and datasets. During the class on LLM fine-tuning, we asked the students to split into groups and assigned downstream tasks (summarization, code generation, hate speech detection, and machine translation). The students were supposed to find suitable datasets and evaluation metrics. The groups presented their findings to the class and then discussed the potential drawbacks of the benchmarks and evaluation metrics.

Reading research papers. One of the goals we set for the course was to teach the students to responsibly assess the quality and trustworthiness of recent research papers. We organized an activity

---

[4] https://hf.co/mistralai/Mistral-7B-v0.1
[5] https://hf.co/mistralai/Mistral-7B-Instruct-v0.1
[6] https://hf.co/microsoft/phi-2
[7] https://hf.co/CohereForAI/aya-101
[8] https://github.com/jlibovicky/llm-mt-assignment

[9] Slido: www.slido.com
[10] https://github.com/jlibovicky/class-quiz

where the students role-played a best-paper committee, partially inspired by the Role-Playing Paper-Reading Seminars (Jacobson and Raffel, 2021).

We selected five papers to encourage critical assessment of the values of model descriptions (Jiang et al., 2023; Schick et al., 2023) and analytical works (Basmova et al., 2023; Balloccu et al., 2024; Yoon et al., 2024). Each student was randomly assigned one of the papers to read thoroughly.

During the class, we first divided the students into groups containing at least one student per article, where the students explained the papers to each other. Then, the students re-grouped by their assigned article, where they discussed the paper again and nominated an advocate for and against it. Then, the advocates presented their final one-minute speeches. Finally, students secretly voted for the best paper using an online form.

4 Conclusion

We presented teaching materials and class activities for a new LLM course taught at Charles University. In the first year, 51 students enrolled course of which around 30 were actively participating. We expect a larger attendance in the following years after the course is upgraded from optional to elective.[11] All classroom activities can be applied in larger cohorts as well. Scaling up the LLM-based assignments for larger number of students might pose an issue for institutions with limited access to computing resources. However, we used four model setups (for each we needed one GPU) that were available for all the students, and so we used up only a relatively small portion of the resources available in our GPU cluster. We therefore expect no severe issues with scaling up to a few hundreds of active students. The availability of enough teaching assistants to ensure proper feedback to the students will potentially become a more significant issue.

Acknowledgments

---

[11]Enrolling in a subset of elective courses is mandatory in the study programme.

References

Duarte M. Alves, José Pombal, Nuno Miguel Guerreiro, Pedro Henrique Martins, João Alves, M. Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *CoRR*, abs/2402.17733.

Jacob Andreas. 2022. Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Victoria Basmova, Yoav Goldberg, and Reut Tsarfaty. 2023. Chatgpt and simple linguistic inferences: Blind spots and blinds. *CoRR*, abs/2305.14785.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Michael L Epstein, Beth B Epstein, and Gary M Brosvic. 2001. Immediate feedback during academic testing. *Psychological reports*, 88(3):889–894.

Alec Jacobson and Colin Raffel. 2021. Role-playing paper-reading seminars. *Colin Raffel's Blog*.

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Caio César, Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, Suriya Gunasekar, Piero Kauffmann, Yin Tat Lee, Yuanzhi Li, Anh Nguyen, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Michael Santacroce, Harkirat Singh Behl, Adam Taumann Kalai, Xin Wang, Rachel Ward, Philipp Witte, Cyril Zhang, and Yi Zhang. 2023. Phi-2: The surprising power of small language models. *Microsoft Research Blog*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud,

Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Anna Katrine Jørgensen and Anders Søgaard. 2023. Rawlsian ai fairness loopholes. *AI and Ethics*, 3(4):1185–1192.

Zdeněk Kasner and Ondřej Dušek. 2024. Beyond reference-based metrics: Analyzing behaviors of open llms on data-to-text generation. *arXiv preprint arXiv:2401.10186*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

John R. Searle. 1985. *Minds, brains, and programs*, page 282–307. MIT Press, Cambridge, MA, USA.

Anders Søgaard. 2022. Understanding models understanding language. *Synthese*, 200(6):443.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. *CoRR*, abs/2401.08417.

Dongkeun Yoon, Joel Jang, Sungdong Kim, Seungone Kim, Sheikh Shafayat, and Minjoon Seo. 2024. Langbridge: Multilingual reasoning without multilingual supervision. *CoRR*, abs/2401.10695.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.