Evaluating Poro-34B-Chat and Mistral-7B-Instruct-v0.1: LLM System Description for ELOQUENT at CLEF 2024

Vasumathi Neralla[1], Sander Bijl de Vroe[1]

[1]*Silo AI*

Abstract

In the following discussion, we demonstrate the implementation of the multilingual Large Language Model (LLM) Poro-34B-Chat and Mistral-7B-Instruct-v0.1 across different tasks under the ELOQUENT test suite. Poro-34B is currently the leading LLM for Finnish and has been further fine-tuned on a collection of English instruction datasets and automatically translated Finnish instructions. In this article, we delve into the processing of user and system prompts tailored to each task, focusing on topical proficiency, robustness, and Voight-Kampff tests. Furthermore, we provide an explanation of the inference engine responsible for generation, including all relevant sampling parameters to ensure reproducibility of our results.

Keywords

Large Language Models, Evaluation, Multilingual, Low-Resource

1. Introduction

LLMs have recently made rapid advances, achieving state-of-the-art results in virtually every NLP task as well as proving their usefulness in widespread practical applications. However, evaluation of LLMs still leaves much to be desired, given the challenging nature of evaluating free-form generated output, especially when compared to the traditional closed-form output requirements common to NLP classification tasks. In support of development of more strategically aimed generative evaluation efforts we submit Poro-34B-Chat to the ELOQUENT evaluation shared task.

At the time of publication Poro 34B was the state-of-the-art Finnish LLM, even at a small fraction of its training tokens, constituting clear evidence that multilingual LLM training can provide a substantial benefit over monolingual training for under-resourced languages.

As far as we are aware, this is the first evaluation of this model in a shared task, and the first evaluation of its interaction with other LLM outputs. We also evaluate the Mistral-7B-Instruct-v0.1 in the same shared task, providing comparison to another open-source model. In the following section we provide some background on both Poro 34B, Poro-34B-Chat and Mistral-7B-Instruct-v0.1, before sharing details of our model inference setup per task.

2. Background

2.1. Poro 34B

The base model Poro 34B uses a decoder-only architecture closely matching BLOOM [1] and FinGPT [2]. The model uses 56 attention heads, 54 layers and a hidden dimension of 7168. As the positional encoding method it uses AliBi [3], as well as additional layer normalization after the input embedding layer for more robust training.

Poro 34B is pre-trained with 1T tokens. The majority of its English tokens are derived from SlimPajama [4], while the Finnish tokens were collected from Finnish portions of ParseBank, Wikipedia and Reddit, and the Finnish media company YLE, among other sources. The StarCoder [5] dataset provides training tokens for code. A sequence length of 2048 as well as a batch size of 2048 are used. A decaying cosine learning rate scheduler was employed, with a 10B token linear warmup, a maximum of 1.5e-4 and a

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

vasumathi.neralla@silo.ai (V. Neralla); sander.bijldevroe@silo.ai (S. Bijl de Vroe)

minimum of 1e-5 for the last 10B training tokens. Training was performed on the LUMI supercomputer using the FinGPT [2] fork of Megatron-DeepSpeed, compatible with the cluster's AMD MI250X-GPUs. A 128 node configuration was chosen, matching a data parallel degree of 128. LUMI runs on fully renewable electricity, so that the total carbon emissions for GPU usage amounted to 0 tCO2eq.

At the time of publication Poro-34B was the strongest Finnish language model of its size, outperforming FinGPT on FIN-Bench [2] after just 10% of training tokens. It also shows strong performance on English in spite of its Finnish training, and due to its relatively large proportion of code training tokens, exhibits favorable performance on programming tasks.

2.2. Poro-34B-Chat

The base model alone does not consistently manage to follow the various task instructions without further training, so we evaluate the fine-tuned version of Poro 34B, Poro-34B-Chat [6, 7] (unpublished). Poro-34B-Chat was instruction fine-tuned [8] on both Finnish and English instruction-response pairs, using code built on the Alignment Handbook [9]. Full-parameter supervised fine-tuning (SFT) was employed.

Since Finnish is under-resourced when it comes to instruction data, a machine translation approach was used, with the Poro 34B base model itself translating instruction datasets from English to Finnish. The English instruction data was compiled from OASST2 [10], the Dolly dataset [11], and an Argilla SFT dataset [12].

The Finnish and cross-lingual data [13] were constructed using a variety of methods. The instruction data was obtained by automatically translating a curated set of the English instructions with Poro 34B (alongside a number of heuristics and filtering steps). Furthermore, Poro 34B was used to generate English-Finnish translation data and language identification data, and the collection was supplemented with Finnish paraphrase data, see [13] for details. Combined these sources result in a data mixture of 40% English, 40% Finnish, and a further 20% cross-lingual data.

2.3. Mistral-7B-Instruct-v0.1

We also assessed Mistral-7B-Instruct-v0.1 [14], a fine-tuned version of the Mistral-7B base model, on the same tasks. Mistral does not publicly share the specific training data composition for either model. However, they state that Mistral-7B-Instruct-v0.1 was fine-tuned without proprietary data, using only instruction datasets publicly available on HuggingFace.

3. Methods and Tasks

To produce model responses we use an internal model service based on OpenAI compatibility. The service utilizes vLLM (v0.2.4) as the inference engine.

A large number of the sampling parameters for vLLM can be kept constant between tasks:

```
SamplingParams(
    n=1, best_of=1, presence_penalty=0.0, frequency_penalty=0.0,
    repetition_penalty=1.0, temperature=0.7, top_p=1.0,
    top_k=-1, min_p=0.0, seed=None, use_beam_search=False,
    length_penalty=1.0, early_stopping=False, stop=[], stop_token_ids=[],
    include_stop_str_in_output=False, ignore_eos=False, max_tokens=1630,
    logprobs=None, prompt_logprobs=None, skip_special_tokens=True,
    spaces_between_special_tokens=True
    )
```

Then the main parameters of interest per task are the `user_prompt` and the the `system_prompt`. Below we describe handling of these input parameters per task. Note that we choose not to participate in the HalluciGen task.

3.1. Topical Competence

There were two steps in the Topical Competence task:

1. Generate questions on the specified topic
2. Generate answers for the questions generated in Step 1

When generating questions (Step 1) we found that the system benefited from using both the title and description fields. Our final user_prompt was defined as the following string:
`user_prompt` = "Title:" + `topic_title` + "\n" + "Description: " + `topic_description`

We include an example of the task below:
Step 1
`system_prompt`:
"Create a set of questions that can be used to assess if someone knows about the following topic:"
`user_prompt`:
"Title: risks and benefits of keeping a pet cat
Description: A set of questions to help decide if one should get a pet cat or not."
`model_output`:
"Sure, here are some questions that could be used to assess someone's knowledge about the risks and benefits of keeping a pet cat: `list_of_output_questions`

Step 2
`system_prompt`:
"Answer the following questions:"
`user_prompt`:
`model_output_from_step_1`

3.2. Robustness

For the robustness `user_prompt` we utilized the prompt provided with the entry. Since Poro-34B-chat was trained on English, Finnish, and code tokens, we evaluated the model only for the English variants of entries. We found that the model was able to provide sensible outputs with an empty system prompt.

3.3. Voight Kampff

The output of this task will serve as source material for assessing the capability to discern between text authored by humans and text generated mechanically. About 500 words are generated based on the given summary or topic. Again for Voight-Kampff we choose to use the provided prompt, and do not apply a system prompt.

4. Conclusion

We have presented here a straightforward application of Poro-34B-Chat and Mistral-7B-Instruct-v0.1 to the tasks in the ELOQUENT test suite. Poro 34B is the state-of-the-art multilingual LLM for Finnish, and its fine-tuned version Poro-34B-Chat was created using a mixture of English instruction datasets and automatically translated Finnish instructions. Mistral-7B-Instruct-v0.1 provides an open-source model for comparison. With a simple prompting approach Poro-34B-Chat is able to follow instructions on the chosen tasks.

Acknowledgments

We would like to thank Jonathan Burdge, Kai Hakala, Mark van Heeswijk, Andrey Ivannikov, Jussi Karlgren, Aku Rouhe, Mittul Singh and Elaine Zosa for their insightful comments.

References

[1] T. Le Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al., Bloom: A 176b-parameter open-access multilingual language model (2023).

[2] R. Luukkonen, V. Komulainen, J. Luoma, A. Eskelinen, J. Kanerva, H.-M. Kupari, F. Ginter, V. Laippala, N. Muennighoff, A. Piktus, T. Wang, N. Tazi, T. Scao, T. Wolf, O. Suominen, S. Sairanen, M. Merioksa, J. Heinonen, A. Vahtola, S. Antao, S. Pyysalo, FinGPT: Large generative models for a small language, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 2710–2726. URL: https://aclanthology.org/2023.emnlp-main.164. doi:10.18653/v1/2023.emnlp-main.164.

[3] O. Press, N. A. Smith, M. Lewis, Train short, test long: Attention with linear biases enables input length extrapolation, arXiv preprint arXiv:2108.12409 (2021).

[4] D. Soboleva, F. Al-Khateeb, R. Myers, J. R. Steeves, J. Hestness, N. Dey, Slimpajama: A 627b token cleaned and deduplicated version of redpajama, 2023.

[5] R. Li, L. B. Allal, Y. Zi, N. Muennighoff, D. Kocetkov, C. Mou, M. Marone, C. Akiki, J. Li, J. Chim, et al., Starcoder: may the source be with you!, arXiv preprint arXiv:2305.06161 (2023).

[6] Silogen, Poro 34b chat is here, 2024. URL: https://www.silo.ai/blog/poro-34b-chat-is-here, accessed on May 28th 2024.

[7] LumiOpen, Poro-34b-chat, 2024. URL: https://huggingface.co/LumiOpen/Poro-34B-chat, accessed on May 28th 2024.

[8] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, R. Lowe, Training language models to follow instructions with human feedback, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, volume 35, Curran Associates, Inc., 2022, pp. 27730–27744. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

[9] L. Tunstall, E. Beeching, N. Lambert, N. Rajani, S. Huang, K. Rasul, A. M. Rush, T. Wolf, The alignment handbook, https://github.com/huggingface/alignment-handbook, 2023.

[10] A. Köpf, Y. Kilcher, D. von Rütte, S. Anagnostidis, Z. R. Tam, K. Stevens, A. Barhoum, D. Nguyen, O. Stanley, R. Nagyfi, et al., Openassistant conversations-democratizing large language model alignment, Advances in Neural Information Processing Systems 36 (2024).

[11] M. Conover, M. Hayes, A. Mathur, J. Xie, J. Wan, S. Shah, A. Ghodsi, P. Wendell, M. Zaharia, R. Xin, Free dolly: Introducing the world's first truly open instruction-tuned llm, Company Blog of Databricks (2023).

[12] Argilla, instruction-collection-fin, 2024. URL: https://huggingface.co/datasets/argilla/10k_prompts_ranked_mistral_large_responses, accessed on May 28th 2024.

[13] LumiOpen, instruction-collection-fin, 2024. URL: https://huggingface.co/datasets/LumiOpen/instruction-collection-fin, accessed on May 28th 2024.

[14] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b, arXiv preprint arXiv:2310.06825 (2023).