Project Work II. - 2024/25/1

Large Language Models on Hungarian Tasks

Author: Sándor Zsombor
Supervisor: Lukács András

Introduction

Large language models are shaping our everyday lives. They have become part our routine: we use them as enhanced search engines, general knowledge bases, virtual assistants and more. It is increasingly important to make sure these models understand the small intricacies of the languages we use in our everyday lives, and to have insight of how different LLMs process text and "think".

Evaluating the linguistic capabilities of a model (or of a person for that matter) is no easy task. It is already difficult to define tasks which indicate competence in a language, and even more so when we have to take the possibility of automatic evaluation into account.

In our first project work last semester, we experimented with some fine tuning techniques, which provide a compromise between ease of use and resource intensity. In this project work, we aimed to measure similar experiments, except with a more complete set of tasks which better describe linguistic competence.

Problem setting

The goal of this project is to try to evaluate LLMs in Hungarian language understanding. For the evaluation of English language understanding, a tried and tested set of benchmarks is the General Language Understanding Evaluation dataset (GLUE for short) [1], and it's successor, SuperGLUE [2]. Both provide essential tasks which mimic parts of language understanding. With the rapid advancement of LLMs, most benchmarks in GLUE became "too easy". Nevertheless, SuperGLUE – the creation of which was motivated by this observation – collects tasks which are still challenging. As a Hungarian "translation" of SuperGLUE, we have HuLU [3], which contains Hungarian alternatives of selected tasks.

As for models, we decided we would continue to mainly experiment with the LLMs created by Meta AI. We ran measurements on the 3 latest smaller models: the lightweight Llama 3.2 1B and 3B [4], and the Llama 3.1 8B [5]. With these models, the problem is especially interesting, since they have no official support for the Hungarian language.

We also evaluated our methods on the smaller members of the Qwen 2.5 [6] family, on Llama 2 13B [7] (which provided the optimal balance of performance and size in our previous project work), and on the English-Hungarian bilingual Llama model created by SambaNova Systems [8].

HuLU - Hungarian Language Understanding Benchmark Kit

As mentioned previously, HuLU is a collection of several benchmarks found in SuperGLUE, translated to Hungarian. It was created by the Hungarian Research Centre for Linguistics (Nyelvtudományi Kutatóközpont in Hungarian, NYTK for short). Some examples are exact translations of their English counterparts, some are adapted into a Hungarian context, and there are a couple of brand new entries.

Let's break down the benchmark kit. The exact task definitions can be found on the website of HuLU and SuperGLUE. We will focus on the practical side of the problems.

HuCB - Commitment Bank

We are given a short paragraph which either has an entity speaking/writing, or contains a description of the opinions/acts of an entity. We receive a statement related to said said entity. Our job is to decide whether the statement is contradictory to the paragraph, aligns with it, or they're logically independent.

HuCOLA - Corpus of Linguistic Acceptability

We are given a sentence. Our task is to decide whether the sentence is grammatically acceptable or not, roughly meaning whether it "sounds natural".

HuCoPA - Choice of Plausible Alternatives

We are given a premise along with two alternative statements (both either causes or effects). Our task is to select the alternative that logically and causally relates to the situation described in the premise.

HuRTE - Recognizing Textual Entailment

We are given a paragraph, and a one-sentence hypothesis. Our task is to determine whether the paragraph logically entails the hypothesis or not (that is, they are independent or contradictory).

HuSST - Stanford Sentiment Treebank

We are given a sentence. Our task is to classify the sentiment expressed in the sentence. The categories are the following: positive, neutral and negative.

HuWNLI - Winograd Natural Language Inference

We are given a short paragraph and a sentence. The paragraph contains a pronoun, which is ambiguous from a grammatical point of view, but the entity it refers to can be inferred from the context. The sentence formulates a statement, which incorporates a piece of information, that is described with the pronoun in the paragraph. The statement uses an entity name instead of the pronoun. Our job is to decide whether the sentence is implied by the paragraph, or in other words, is the pronoun properly bound to the corresponding entity.

Model evaluation strategy

We evaluated every model on every task. Since we opted to use the foundation models instead of the ones pretrained for chat, we are unable to achieve the best score without training (prompt engineering is unreliable, when the model is not trained to reply, only to generate). Nevertheless, we measured the Llama 3 models' performance.

For a more insightful score, we trained every model for the given tasks. Using the experience from our first project work, we decided to utilize low rank adaptation (LoRA) [9]. We used a learning rate of $5 \cdot 10^{-4}$ with linear scheduling. The number of epochs is dependent on the task and the model size, since different benchmarks contain different magnitudes of training data, and larger models can overfit when trained on many epochs.

As opposed to last semester, when we trained with a rank of 8, these tasks proved to be more difficult than the semantic categorization of movie reviews. Therefore we (heuristically) increased the LoRA rank to 64 (and the α to 128).

For the 0.5B and 1.5B Qwen models we used the same number of epochs as for Llama 3.2 1B. Likewise, for the 3B Qwen model we used the same configuration as for Llama 3.2 3B. For the models with size above 6B parameters, we used the Llama 3.1 8B configurations.

Finally, we were curious whether we can achieve competence in all task simultaneously. For this, we unified the training sets of the different tasks (with multiplicity defined by the number of epochs), and train models on this unified dataset for one epoch.

	1B	3B	8B
HuCB	7	7	3
HuCOLA	5	5	3
HuCoPA	5	5	5
HuRTE	4	4	4
HuSST	2	2	2
HuWNLI	5	5	3

Table 1: Number of epochs for each model size and task, Llama 3 models

	Metric
HuCB	weighted F1
HuCOLA	Accuracy Matthews corr.
HuCoPA	Matthews corr.
HuRTE	Matthews corr.
HuSST	Accuracy
HuWNLI	Accuracy

Table 2: Metrics used for each task

The benchmark datasets are split into train, validation and test sets, however for most of them the test set contains no labels (although it is possible to evaluate a result on the benchmark kit's website). Therefore we used the train set for training, and the validation set for evaluation.

We used a custom prompt for every task, and assign the appropriate labels to the generated outputs. Each model performed inference on the corresponding tasks 5 times. The results shown are the averages of the 5 attempts.

Results

Table 3 contains a few selected results from the official website of the benchmark kit, for comparison. PULI Llumix 32K Instuct – based on Llama 2 7B – is the currently most performant Hungarian model of NYTK. PULI BERT-large [10] and ParancsPULI [11] were the results of earlier research. The latter is based on GPT-NeoX, a 7B parameter GPT variant. ChatGPT[1] and text-davinci-001 are closed source models by OpenAI. Note that our results were measured on the validation set, while NYTK measured on the test set.

Table 4 collects the measurements from the Llama 3 base models, while Table 5 shows the metrics form the task specific LoRA trains, we denoted these models with the "LoRA" postfix. Lastly, the evaluations of the models trained on the unified datasets are presented in Table 6, here we used the "Ens" postfix (for "ensemble").

	HuCB (F1)	HuCOLA (ACC)	HuCOLA (MCC)	HuCoPA (MCC)	HuRTE (MCC)	HuSST (ACC)	HuWNLI (ACC)
PULI Llumix 32K Instruct	0.661	0.911	0.703	0.736	0.670	0.801	0.731
PULI BERT-large	-	-	0.711	0.414	0.517	0.799	0.657
ParancsPULI	0.582	0.891	0.631	0.445	0.592	0.791	0.649
ChatGPT	-	0.818	0.277	-	-	0.718	-
text-davinci-001	-	0.805	0.316	-	0.798	0.528	-

Table 3: Official measurements by NYTK for selected models [12]

Llama 3.2 1B	0.193	0.700	-0.033	-0.117	-0.241	0.182	0.417
Llama 3.2 3B	0.242	0.693	0.013	0.040	-0.012	0.030	0.380
Llama 3.1 8B	0.273	0.685	0.019	0.142	0.073	0.266	0.453

Table 4: Our measurements on the base Llama 3 models

Llama 3.2 1B LoRA	0.384	0.664	0.213	0.015	0.066	0.670	0.473
Llama 3.2 3B LoRA	0.644	0.783	0.377	-0.054	0.712	0.751	0.490
Llama 3.1 8B LoRA	0.669	0.835	0.501	0.783	0.748	0.755	0.477
Qwen 2.5 0.5B LoRA	0.435	0.684	0.207	-0.086	0.236	0.464	0.417
Qwen 2.5 1.5B LoRA	0.385	0.670	0.124	0.095	0.395	0.573	0.450
Qwen 2.5 3B LoRA	0.616	0.658	0.172	-0.060	0.598	0.664	0.467
Qwen 2.5 7B LoRA	0.707	0.755	0.321	0.620	0.735	0.700	0.550
SambaLingo Hun LoRA	0.459	0.863	0.576	0.000	0.751	0.770	0.400
Llama 2 13B LoRA	0.688	0.810	0.398	0.183	0.740	0.754	0.463

Table 5: Our measurements on the task-specific LoRA trains

Llama 3.2 1B Ens	0.396	0.674	0.214	-0.017	0.148	0.673	0.407
Llama 3.2 3B Ens	0.665	0.818	0.438	0.359	0.608	0.732	0.393
Llama 3.1 8B Ens	0.609	0.825	0.491	0.589	0.704	0.753	0.463
Qwen 2.5 0.5B Ens	0.539	0.636	0.109	-0.098	0.219	0.448	0.367
Qwen 2.5 1.5B Ens	0.439	0.712	0.226	0.047	0.379	0.595	0.400
Qwen 2.5 3B Ens	0.538	0.762	0.280	0.179	0.571	0.634	0.383
Qwen 2.5 7B Ens	0.618	0.767	0.356	0.543	0.675	0.707	0.367
SambaLingo Hun Ens	0.591	0.865	0.599	0.469	0.751	0.767	0.417
Llama 2 13B Ens	0.620	0.769	0.334	0.272	0.611	0.713	0.447

Table 6: Our measurements on the LoRA trains with the tasks' unified training data

[1] We do not know which version of ChatGPT was evaluated. According to the date of the measurement, it is either 3.5 or 4.

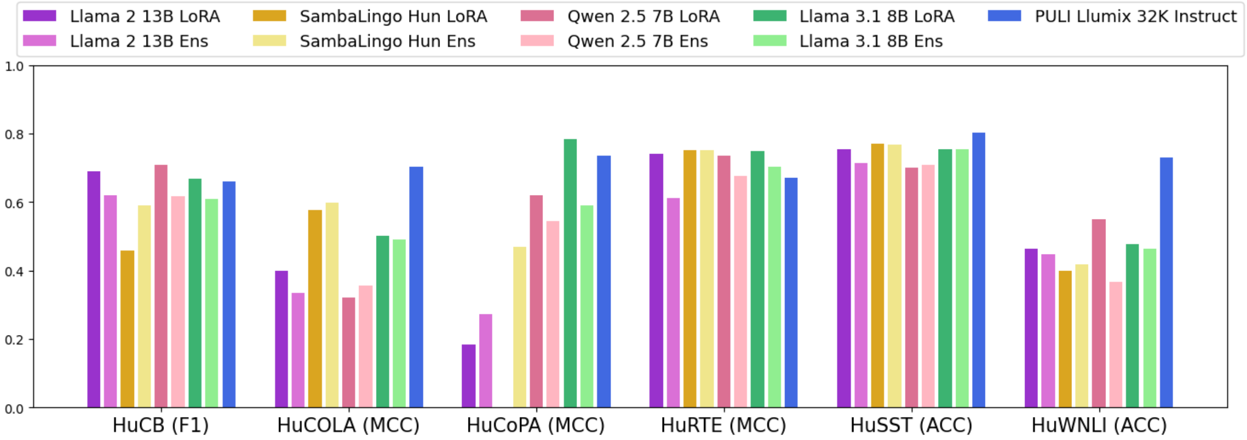

Figure 1: Performance of our larger models compared to PULI Llumix

Base models

As expected, the untrained models underperform on most tasks. On every benchmark, the models' performance is close to random guessing: the Matthews correlations are close to 0, while the accuracies approximately the ratio of one class.

Task specific LoRA models

With the trained models, we were able to closely match the PULI Llumix 32K Instruct model on almost all tasks (even beating it in some cases). For example, on the HuCoPA and HuRTE tasks our larger models perform exceptionally well in comparison. Llama 3.1 8B is the model that seems to match the competence of the PULI Llumix model the closest.

The two outliers are the linguistic acceptability and Winograd schema benchmarks. Observing the problem descriptions however, we may shed some light on this result: HuWNLI and HuCOLA are the 2 tasks which require the deepest level of understanding of the inner workings of a language, with the first one requiring "native-level intuition", and the second capable of even stumping people.

With this in mind, we remark, that while WNLI is undoubtedly a hard challenge, our trained models all perform close to random guessing. This suggest a fault in our training methods. We could probably attain better results with alternative formulations of the prompt used for training, or with focusing more on hyperparameter optimization specific to each model with the task.

LoRA models on unified training data

Training with the unified datasets, we were able to achieve similar metrics. However these models also suffer from the same problems: the performance on the HuCOLA and HuWNLI tasks is subpar. Furthermore, the HuCoPA scores also seem to drop a little, a result which may be explored in our future work.

Nevertheless, the experiment suggests, that if we train a model on enough tasks in a given language, we can possibly improve it's language understanding capabilities as a whole.

Remarks about the smaller models

The smaller models' performance does not show any general trend across all benchmarks. There are tasks, where the model competence is proportional to the parameter size (loosely speaking), like in the case of the Qwen family and the HuSST problem, or the Llama 3 family and the HuCOLA dataset.

On the flip side, in the case of Qwen and HuCOLA, the smallest model seems to outperform it's mid-size counterparts (though this could probably be resolved with a careful hyperparameter optimization).

However, we are able make an interesting observation looking at the Llama 3 and Qwen families with the HuCoPA benchmark. Evaluating the task specific LoRA models, we see a score close to 0 in case of models up to 3B parameters. When looking at the larger models on the other hand, we notice the score now rivals the values achieved by PULI. This may indicate, that there is an emergent behaviour in case of the Hungarian language in this range of model complexity[2].

Conclusion and future work

In this semester's project work, we explored the Hungarian linguistic capabilities of several large language models. We found, that while the LLMs we tested do not officially support Hungarian (with the exception of the SambaLingo model), they are able to preform quite well on numerous benchmarks designed for language understanding. However there are aspects, where they still fall behind when compared to models trained specifically for the language.

This work is possibly the precursor of a larger project, where we aim to bridge this gap using large Hungarian corpora, utilizing PEFT methods, and continuing the models' pretraining focusing on the language.

Furthermore, there are many interesting questions regarding the model structure to improve linguistic understanding. These include dimensional extensions, deconstruction of models, mimicking layer behaviours, and exploration of the latent states to try to find the model's "language independent thoughts" (if such exists).

References

[1] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding, 2019.

[2] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems, 2020.

[3] Noémi Ligeti-Nagy, Gergő Ferenczi, Enikő Héja, László János Laki, Noémi Vadász, Zijian Győző Yang, and Tamás Váradi. HuLU: Hungarian language understanding benchmark kit. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8360–8371, Torino, Italia, May 2024. ELRA and ICCL.

[4] Llama 3.2: Revolutionizing edge AI and vision with open, customizable models — ai.meta.com. `https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/`, 2024. [Accessed 15-12-2024].

[5] Llama 3 Team. The Llama 3 herd of models, 2024.

[6] Qwen Team. Qwen2.5: A party of foundation models, September 2024.

[7] Llama 2 Team. Llama 2: Open foundation and fine-tuned chat models, 2023.

[8] Zoltan Csaki, Bo Li, Jonathan Li, Qiantong Xu, Pian Pawakapan, Leon Zhang, Yun Du, Hengyu Zhao, Changran Hu, and Urmish Thakker. SambaLingo: Teaching large language models new languages, 2024.

[9] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models, 2021.

[10] Zijian Győző Yang, Réka Dodé, Gergő Ferenczi, Enikő Héja, Kinga Jelencsik-Mátyus, Ádám Kőrös, László János Laki, Noémi Ligeti-Nagy, Noémi Vadász, and Tamás Váradi. Jönnek a nagyok! BERT-Large, GPT-2 és GPT-3 nyelvmodellek magyar nyelvre. In *XIX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2023)*, pages 247–262, Szeged, Hungary, 2023. Szegedi Tudományegyetem, Informatikai Intézet.

[11] Zijian Győző Yang, Réka Dodé, Laki, Enikő Héja, László János, Noémi Ligeti-Nagy, Gábor Madarász, and Tamás Váradi. ParancsPULI: Az utasításkövető PULI-modell. In *XX. Magyar Számítógépes Nyelvészeti Konferencia*, pages 61–72, Szeged, Magyarország, 2024. Szegedi Tudományegyetem.

[12] HuLU leaderboard – hulu.nytud.hu. `https://hulu.nytud.hu/leaderboard`, 2024. [Accessed 15-12-2024].

[2] Note that the Llama and Qwen architectures are quite similar, therefore this might be specific to this archetype of models.