Project Work II. - 2024/25/1

Large Language Models on Hungarian Tasks

Author: Sándor Zsombor
Supervisor: Lukács András

Introduction

- We interact with LLMs every day
- How well do they actually understand us?
- Important: models can decipher linguistic intricacies
- Problem: how well does a human know a language?

How much ’language’ do Large Language Models model?

Evaluation of language understanding

- Designing tests is hard
- They need to be
  - linguistically relevant: propose a problem which requires deep conceptual and contextual understanding of given sentences
  - hard: results need large variance to differentiate between various models (and humans)
  - automated: for mass evaluation and exact reproducibility
- if not designed properly, problems can either be too easy for LLMs or too hard for humans

The English gold standard – GLUE

General Language Understanding Evaluation (GLUE) benchmarks

- Contains various essential tasks which mimic parts of English language understanding
- While the tasks are well thought-out, for modern LLMs the benchmark is obsolete (since BERT, most of them are „too easy”)

SuperGLUE

- more challenging tasks than its predecessors’
- carried over the harder benchmarks from GLUE (e.g. WNLI)

Note: for state-of-the-art LLMs, even SuperGLUE is simple; instead: MMLU, HellaSwag, etc...

Hungarian Language Understanding Dataset

Introducing: HuLU

Our aim: evaluate how well open-source LLMs „know Hungarian”
Used benchmark kit: HuLU

- created by the Hungarian Research Centre for Linguistics (NYTK)
- serves as a „translation” of SuperGLUE
- contains 6 task (all found in SuperGLUE)

Tasks in HuLU

- Recognizing entailment, causality: HuCB, HuRTE, HuCoPA
- Linguistic acceptability: HuCOLA
- Sentiment analysis: HuSST
- Winograd scheme: HuWNLI

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

Experimentation strategy

Models

We evaluated the following models:

- Llama 3.1 8B, Llama 3.2 1B and 3B
- Llama 2 13B
- Qwen 2.5 0.5B, 1.5B, 3B, 7B
- SambaLingo-Hungarian-Chat (Llama 2 7B tuned to Hungarian)

Training and inference

During last semester's project work, we experimented with PEFT techniques $\longrightarrow$ we aimed to utilize the experience $\Longrightarrow$ LoRA

Training parameters:

- $5 \cdot 10^{-4}$, linear scheduling
- LoRA: $r = 64$, $\alpha = 128$
- Number of epochs: task and model size dependent

Each task inference was run 5 times $\longrightarrow$ average as result

Note: last semester we used $r = 8$, that proved to be insufficient now

Is a general improvement possible?

Question: is it possible to finetune the model in such a way that it improves on all tasks simultaneously?

Experiment:

- try to train on all tasks at the same time
- training dataset $=$ merge all train datasets with multiplicity

$\Longrightarrow$ Combined models

Results

Task specific models

	HuCB (F1)	HuCOLA (MCC)	HuCoPA (MCC)	HuRTE (MCC)	HuSST (ACC)	HuWNLI (ACC)
PULI Llumix 32K Instr	**0.661**	**0.703**	**0.736**	**0.670**	**0.801**	**0.731**
Llama 3.2 1B	0.384	0.213	0.015	0.066	0.670	0.473
Llama 3.2 3B	0.644	0.377	-0.054	0.712	**0.751**	0.490
Llama 3.1 8B	0.669	0.501	**0.783**	**0.748**	**0.755**	0.477
Qwen 2.5 0.5B	0.435	0.207	-0.086	0.236	0.464	0.417
Qwen 2.5 1.5B	0.385	0.124	0.095	0.395	0.573	0.450
Qwen 2.5 3B	0.616	0.172	-0.060	0.598	0.664	0.467
Qwen 2.5 7B	**0.707**	0.321	0.620	**0.735**	0.700	**0.550**
SambaLingo Hun	0.459	**0.576**	0.000	**0.751**	**0.770**	0.400
Llama 2 13B	**0.688**	0.398	0.183	**0.740**	**0.754**	0.463

Combined models

	HuCB (F1)	HuCOLA (MCC)	HuCoPA (MCC)	HuRTE (MCC)	HuSST (ACC)	HuWNLI (ACC)
PULI Llumix 32K Instr	**0.661**	**0.703**	**0.736**	**0.670**	**0.801**	**0.731**
Llama 3.2 1B	0.396	0.214	-0.017	0.148	0.673	0.407
Llama 3.2 3B	**0.665**	0.438	0.359	0.608	0.732	0.393
Llama 3.1 8B	0.609	0.491	**0.589**	0.704	**0.753**	**0.463**
Qwen 2.5 0.5B	0.539	0.109	-0.098	0.219	0.448	0.367
Qwen 2.5 1.5B	0.439	0.226	0.047	0.379	0.595	0.400
Qwen 2.5 3B	0.538	0.280	0.179	0.571	0.634	0.383
Qwen 2.5 7B	0.618	0.356	0.543	0.675	0.707	0.367
SambaLingo Hun	0.591	**0.599**	0.469	**0.751**	**0.767**	0.417
Llama 2 13B	0.620	0.334	0.272	0.611	0.713	**0.447**

The positive

PULI Llumix 32K Instruct is the state-of-the-art model by NYTK

Strengths:

- Larger models are on par with PULI Llumix on most tasks
- Llama 3 8B seems to match the closest
- Exceptional performance in HuCoPA and HuRTE tasks
- Combined models perform similarly, only a slight drop despite „knowing all tasks”

The latter suggests: training on a well defined set of tasks instead of an enormous corpora $\longrightarrow$ possible to achieve similar linguistic competence

The negative

Weak spots: 2 outlier tasks:

- HuCOLA: significant difference in MCC scores
- HuWNLI: no models seem to be better than random guessing
- Combined models suffer the same fate

Possible explanation:

- These are the most linguistically challenging tasks out of the 6
- HuCOLA: requires high level understanding of intricacies
- HuWNLI: needs a „deep contextual intuition”

Note: the HuWNLI scores are close or even worse than a coin-flip
$\Longrightarrow$ Our methodology needs to be revised, it may be at fault

Remarks about smaller models

- Smaller models' performance doesn't show any general trend
- There are tasks where they underperform, on some they're on par, and in a few cases they outperform their larger counterparts
- Interesting observation with HuCoPA:
  - Up to 3B parameters: near 0 score
  - Llama 8B and Qwen 7B: PULI Llumix like score

$\Longrightarrow$ Possible emergent behaviour

Model	HuCoPA (MCC)
Llama 3.2 1B	0.015
Llama 3.2 3B	-0.054
Llama 3.1 8B	0.783
Qwen 2.5 0.5B	-0.086
Qwen 2.5 1.5B	0.095
Qwen 2.5 3B	-0.060
Qwen 2.5 7B	0.620