klmbr

Ivan Charapanau 0009-0000-5894-0087
Independent Researcher
Warsaw
Email: av@av.codes

***Abstract*—Tokenization is an essential part of LLM training and inference. It has implicit impact on the model's performance and generalization capabilities, as the model learns from the tokenized data. At the same time, many reinforcement learning techniques are often including randomization of the inputs to improve the model's robustness. We show that there is evidence of this approach being a common practice in many LLM training routines, we speculate that this also includes the closed-source models.**

We introduce a library that can be used to induce different tokenization schemes to the input data for arbitrary LLMs. We demonstrate that it can improve the model's performance and generalization capabilities.

I. Introduction

- Tokenization is an essential part of LLM training and inference.
- Overview of tokenization techniques.
- Tokenization has implicit impact on the model's performance and generalization capabilities, as the model learns from the tokenized data.
- Experiment: model outputs for the same input, but with different tokenization schemes.
- Question: How much does the tokenization scheme affect the model's performance and generalization capabilities?

II. Retokenization

- Retokenization is a simple pre-processing technique to work around the embedded biases in the input data.
- Inducing Retokenization for arbitrary LLMs.
- Overview of the technique.

III. Conclusions

- We introduce a library that can be used to induce different tokenization schemes to the input data for arbitrary LLMs.
- We demonstrate that it has an impact on the model's performance and generalization capabilities.