

M1 DELIVERABLES

This webapp gives users ability to search for GitHub repos in an intuitive and friendly matter. The results can then be exported as csv files which can then be used as training data to fine-tune code LLMs. The code is also available from page and our python package have more flexibility in specifying the query parameters or uploading the extracted code to HuggingFace datasets. The Python package also gives filtering option after the extraction based on factors like code licensing , code quality , relevance and date of issue. The full package will be available to download as pip package by milestone 4 , but the code is openly available on the project repositories now.

A web based demo for our tool for licensed code extraction from GitHub

<https://customizable-code-assistant.streamlit.app/>

M2 DELIVERABLES

In this milestone we opted to deliver our report about best practices in finetuning code LLMs in more interactive and applied way instead of a purely academic report. The application is divided into two parts both based on academic papers from top AI labs (Deep mind and OpenAI). The first is about what we refer to as "Compute Optimal LLM Training". This section allows user to interact with the factors of the scaling law in a highly customizable manner controlling the hardware setup , training time, Model parameters and dataset tokens. Then based on these factors we approximate if the compute provided is optimal for performance given the selected model and dataset. In the second we investigate the optimal parameters for finetuning code LLMs for optimal perplexity loss in terms of finetuning parameters and baseline LLMs choices. We provide case specific graphs and variables for the user to adjust and interact with. This is a more specialized case as Fine-tuning is highly task and data specific, so to give meaningful estimations while keeping the application user friendly and accessible we focused on our specific use case.

Similar to milestone 1, the python package gives more flexibility , is available on GitHub, and will be officially released in Milestone 4

A web based Demo of our Fine-tuning math calculator:

<https://customizable-code-assistant.streamlit.app/>