



## More Efficient Transformers with high-level multi-token prediction

LLMs such as ChatGPT, Gemini or Llama have demonstrated remarkable capabilities in recent years. While post-training steps such as RLHF have contributed significantly to their performance, pre-training has stayed mostly the same. Models are simply trained using a next-token prediction loss. Previous work (e.g. <https://arxiv.org/pdf/2404.19737>) has found that training a model to predict multiple token improves performance and generation speed. In this project, we plan to create a novel multi-stage stage LLM. The first stage should predict a sentence-level token, providing guidance on what the next tokens should be. The second stage should then generate the actual text, given a small context window and the sentence-level token. We will build the model in a way to leverage pre-trained LLMs. Additionally, we will execute many ablations to get an in-depth understanding of the factors needed to make this work.

We will have weekly meetings to address questions, discuss progress and think about future ideas.

### Requirements

Strong programming skills (Python, etc.) and a excellent knowledge of machine learning. Previous experience with PyTorch and other common deep-learning libraries is a plus.

### Contact

Interested? Please reach out with a brief description of your motivation in the project, along with any relevant courses or prior projects (personal or academic) that demonstrate your background in the area.

- Frédéric Berdoz: [fberdozn@ethz.ch](mailto:fberdozn@ethz.ch), ETZ G60.1
- Benjamin Estermann: [besterma@ethz.ch](mailto:besterma@ethz.ch), ETZ G60.1

