Swiss AI Initiative - Collecting Large Amounts of High-quality Data for Training Large Language Models

Jan Deriu and **Maud Ehrmann** and **Emanuela Boros**
Maximilian Böther and **Christiane Sibille** and **Ihor Protsenko**
Marta Brucka and **Imanol Schlag** and **Elliott Ash**
deri@zhaw.ch

Abstract

The Swiss AI Initiative, a consortium led by ETH Zurich and EPFL, consists of over 70 professors in Switzerland. Its goal is to develop and research large-scale large language models (LLM) for the Swiss population, leveraging the CSCS's Alps supercomputer. Central to this initiative is the commitment to curating high-quality datasets reflective of Swiss cultural and linguistic diversity. High-quality data is crucial for the effective pre-training of LLMs. Current open-source LLMs mostly utilize extensive datasets compiled from web sources, often subjected to minimal quality control. This approach results in datasets containing low-quality texts from social media and other unreliable platforms, embedding significant biases within the LLMs that necessitate alignment. Empirical evidence suggests that integrating high-quality texts into the training regimen enhances LLM performance across various parameter scales [1]. A further complication arises from the reliance on copyrighted content, including newspapers and books, which often embroils open-source initiatives in legal complexities, hindering the release of models under permissive licenses. This environment also contributes to publishers' reluctance to grant research access to their content. This talk will present the overall vision of the Swiss AI Initiative with a particular emphasis on the challenges in data acquisition, storage, and curation.