After the release of ChatGPT in 2022, the number of papers published every day about Large Language Models (LLMs) increased more than 10-fold. The number of parameters in these LLMs jumped from 340 million in implementations such as BERT to billions of parameters in models like GPT or LLaMA. A large part of these parameters come from the word-embedding layers which are used to represent finite vocabularies of characters sets or characters or words. Apart from increasing model complexity, a fixed vocabulary also responsible for simple models, which cannot deal with out-of-vocabulary inputs and cannot generalize to new languages. A