

After the release of ChatGPT in 2022, the number of papers published every day about Large Language Models (LLMs) has increased more than 20-fold. The number of parameters in these LLMs jumped from 340 millions in implementations such as BERT to billions of parameters in models like GPT-3 or LLaMA. A large part of these parameters come from the word-embedding layers which are used to represent a finite vocabulary of characters, sets of characters or words. Apart from increasing model complexity, a fixed vocabulary is also responsible for brittle models, which cannot deal with out-of-vocabulary inputs and cannot generalize to new languages. As a