

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 millions in implementation such as BERT to billions of parameters in models like GPT-3 and LLaMA. A large part of these parameters come from the word-embedding layer which are used to represent finite vocabulary of characters, bits of characters or words. Apart from increasing model complexity, a fixed vocabulary also responds for brittle models which cannot deal with out-of-vocabulary inputs and cannot generate new languages.