

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 of GPT to billions of parameters in models like LLaMA. A large part of the parameters are in the word-embedding layer and the hidden layers which are used to represent the vocabulary of characters, sets of characters or words. Apart from increasing model complexity, a fixed vocabulary also results in brittleness which cannot deal with out-of-vocabulary inputs and cannot generate new languages.