Mask Rate = 80%

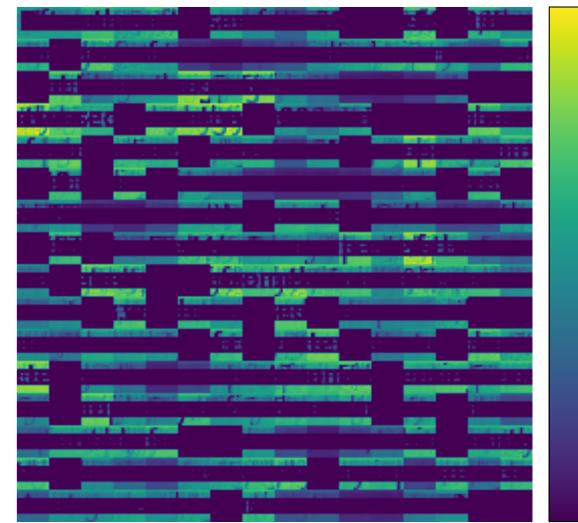
Original

After the release of ChatGPT in 2022, the number of papers published every day al pout Large Language Models (LLMs) has in creased more than 20-fold. The number of of parameters in these LLMs jumped from 340 millions in implementations such as B ERT to billions of parameters in models like e GPT-3 or LLaMA. A large part of these pa arameters come from the word-embeddin ig layers which are used to represent a fin ite vocabulary of characters, sets of chara cters or words. Apart from increasing mod del complexity, a fixed vocabulary is also re esponsible for brittle models, which cannot t deal with out-of-vocabulary inputs and ca nnot generalize to new languages. As a

Original + SD

Large Language Mode

Mean Prediction + SD



T 1.0

- 0.8

- 0.6

- 0.4
