

Heaps' law is an empirical law relating the number of unique words to the number of total words in corpora of textual documents. While this law has been validated in diverse human-authored text corpora, its applicability to large language model generated text remains unexplored. To address this gap, we investigate the validity of Heaps' law in GPT-Neo large language model emulated documents. In particular, we emulated corpora of PubMed abstracts using three distinct parameter sizes of the GPT-Neo model (125 million, 1.3 billion, and 2.7 billion parameters). Using simple linear regression, we validated Heaps' law on the GPT-Neo emulated corpora. Notably, as GPT-Neo model complexity increased, emulated vocabulary growth rate increasingly approached that of the human-authored PubMed abstracts. However, training GPT-Neo models with ever-increasing parameters did not yield linear improvements in adherence to Heaps' law as observed in the human-authored PubMed abstracts. This finding has practical implications for reducing the energy cost of large language model queries. For instance, we are presently investigating how to leverage Heaps' law to minimize large language model energy consumption while still accurately mimicking human-authored text. This underscores the imperative for future advancements in large language models to prioritize energy-efficient strategies, possibly through architectural enhancements or optimization, over mere parameter inflation.