https://osf.io/preprints/osf/xjmtq

The basic concept of "attention heads" in Transformers matches with partial features of "islets of inter-LINKed spine heads", a model of nervous system functions.

- *Kunjumon Vadakkan*

Summary

The semblance hypothesis has put forward a model of operations of the nervous system. Here, spatial relations between specific signals arriving from input neurons are registered in “islets of inter-LINKed spine heads”, whose spines belong to different output neurons. In response to a prompt (cue stimulus), operation of these inter-LINKed spine heads generates both first-person inner sensations and motor outputs such as speech and behavior. The field of artificial intelligence uses artificial neural networks where neurons are connected with each other in different configurations. The configuration in large language models (LLMs) has succeeded in showing features of good generalization - the ability of a trained system (nervous or artificial) to perform well on new input data unseen during training. This prompted us to examine the basic concept of operations in LLMs. This shows that "attention heads" within the hidden layer located in between input and output neuronal layers of Transformer in LLMs is equivalent to a linear algebraic treatment of one segment of operation of "islets of inter-LINKed spines" that generates motor outputs.