IT2026 Supplementary Note 08

November 20, 2026

1 Different Computation Models in Different LLMs

Recall that a so-called LLM is in essence a number-in-number-out computation model in combination with a word embedding module adding in front of and a word embedding module adding after the computation model. Once a prompt has been entered, the word embedding module in front converts each word or punctuation in the prompt to a vector. The vectors are then fed in the computation model to generate a sequence of vectors. Each vector generated by the computation model is finally converted to its corresponding word. The sequence of words is then output in response to the prompt.

Roughly speaking, the main differences among the LLMs are their computation models. Figure 1 shows the early LLM design in which a single computation model is developed for and applied in generating text for any prompt

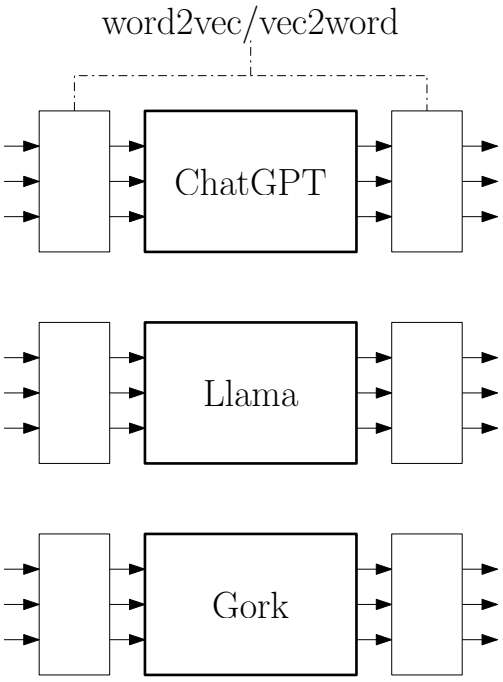

Figure 1: An LLM generally consists of three important modules: two embedding modules and a computation model. The main differences among the LLMs are their computation models.

2 Mixture of Experts (MoE) Architecture

The early LLM architecture could be treated as a single expert design. Each LLM acts like a language expert who is good in text generation. So, those early LLMs are *general* text generation

computation models. In response to some area specific prompts, those LLMs are clearly inadequate to generate precise text contents to the prompts. In this regard, an intuitive approach is to develop multiple area specific LLMs to handle multiple area specific prompts. This is the idea behind the recent approach in the development of an LLM – mixture of experts (MoE), see Figure 2.

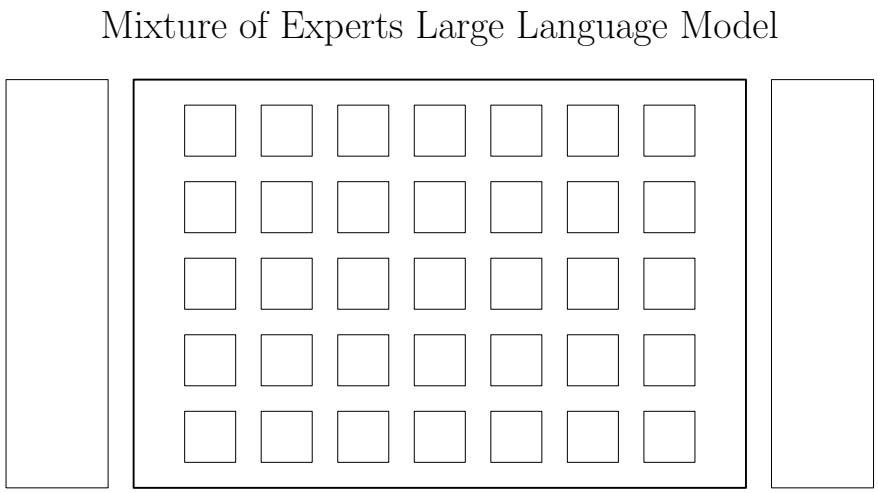

Figure 2: An mixture of experts (MoE) architecture for a large language model. In it, multiple small language models each specialized in an expertise area are developed. In response to a prompt, only a small number of small language models are selected to generate the text output. Therefore, the number of parameters being accessed in run-time is comparatively much smaller than the total number of parameters in the LLM.