* 關於**llama3 8B**的decode output dimension

**logits**、**attentions**的部分:

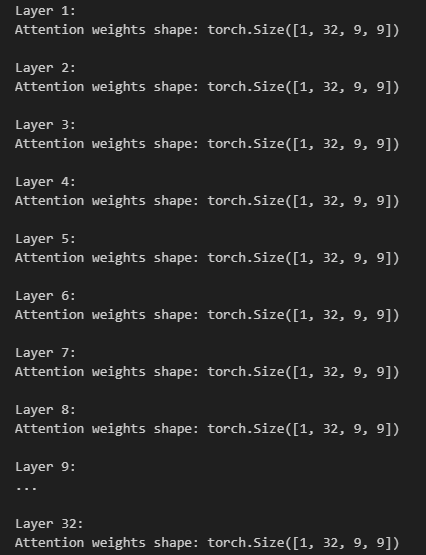
* + outputs.logits.shape的輸出

**(batch\_size, sequence\_length, vocab\_size)**



* attentions的輸出(llama3 layer=32)

**(batch\_size, num\_head, sequence\_length, sequence\_length)**



其中一層的一個attention head