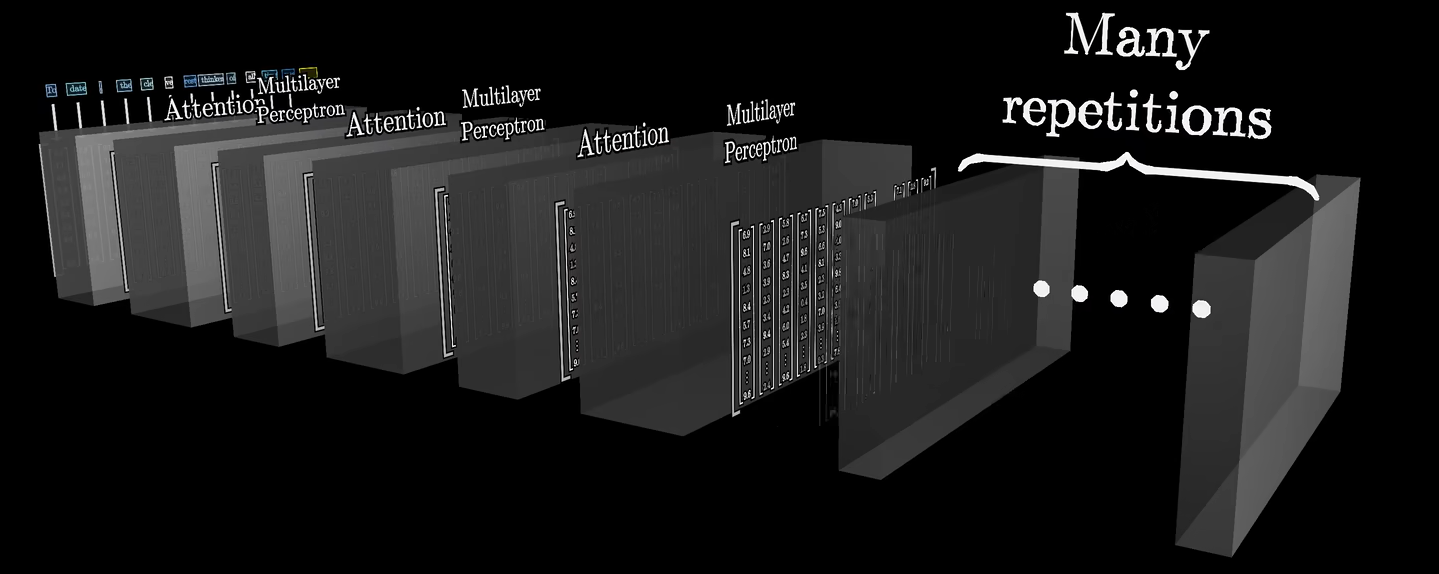
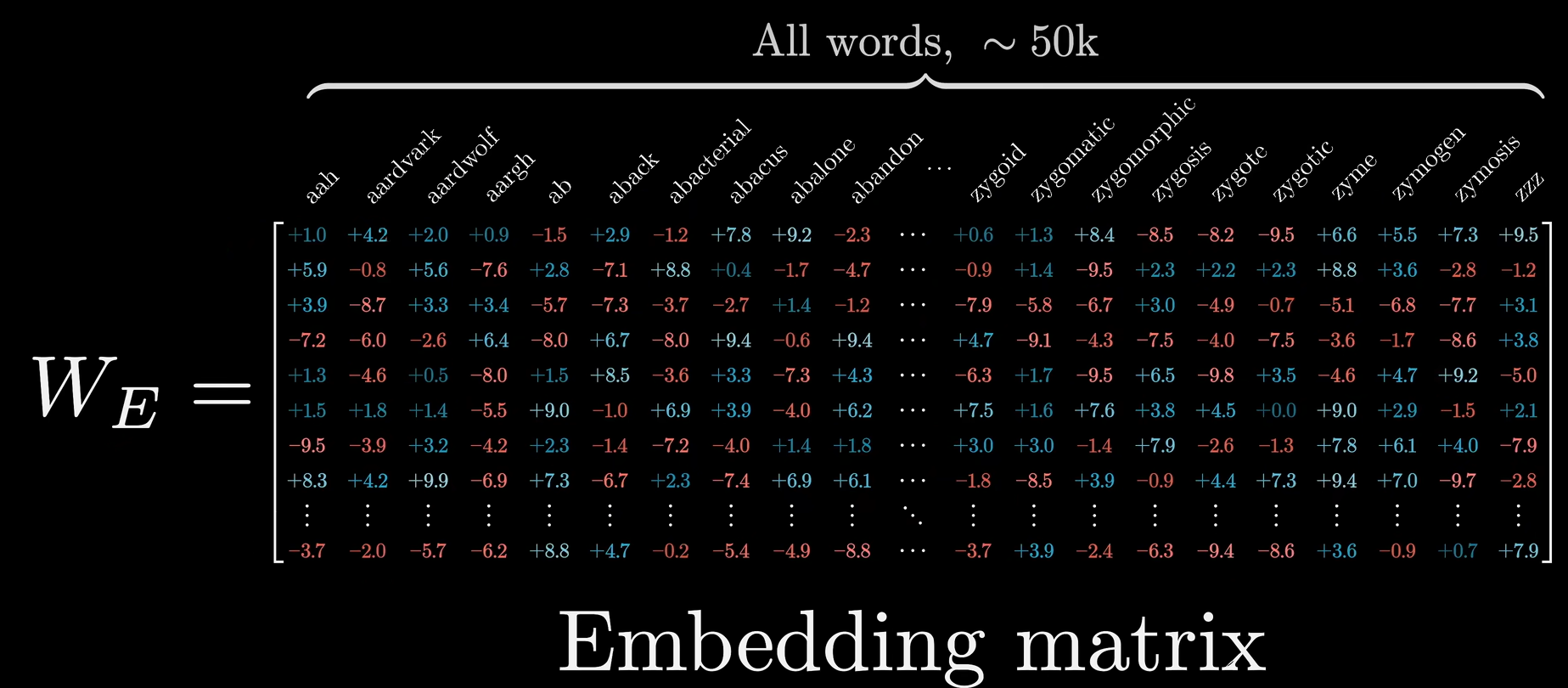
01



重复 attention + MLP 的过程。

02

1. GPT-3有1750亿个权重。
2. Embedding Matrix



GPT-3 的词汇库里有 50257 个词（token），每个词对应 12,288 维度



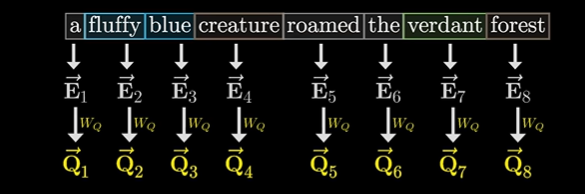
1. Unembedding Matrix

50257×12288，恰好与 Embedding Matrix 对调

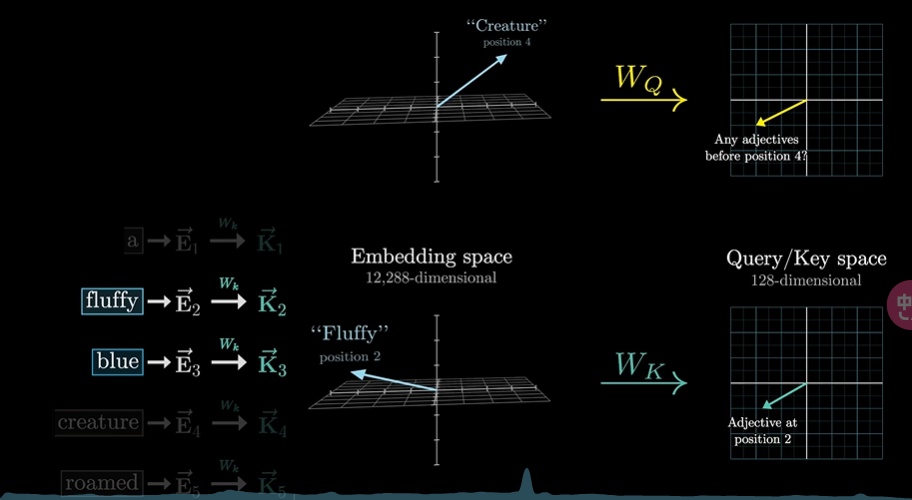
1. 关于模型的 temperature

大概就是对于 logits 做 softmax 时，e^x 改为 e^{x/T}，这样随着 T 增大，概率分布会更平滑，也就是模型对自己的输出没那么自信了，会导致模型的选择更多。

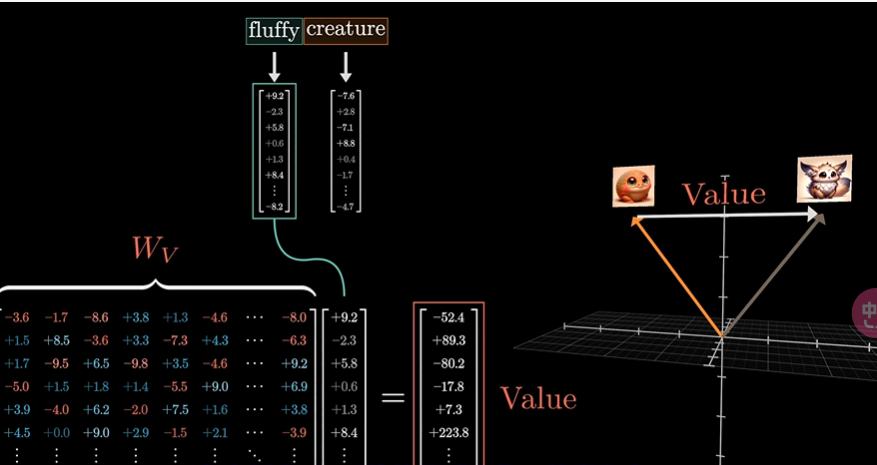
1. Query

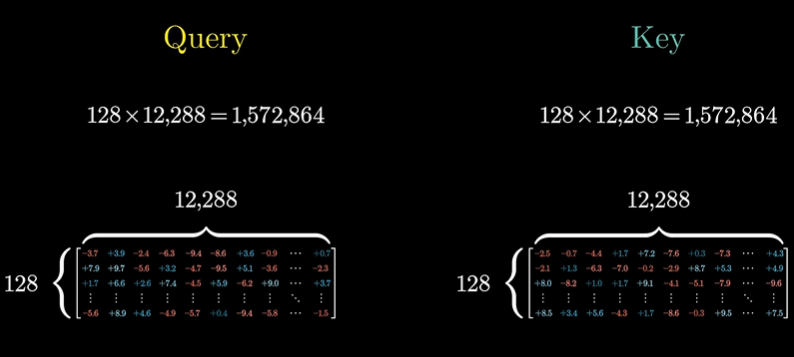


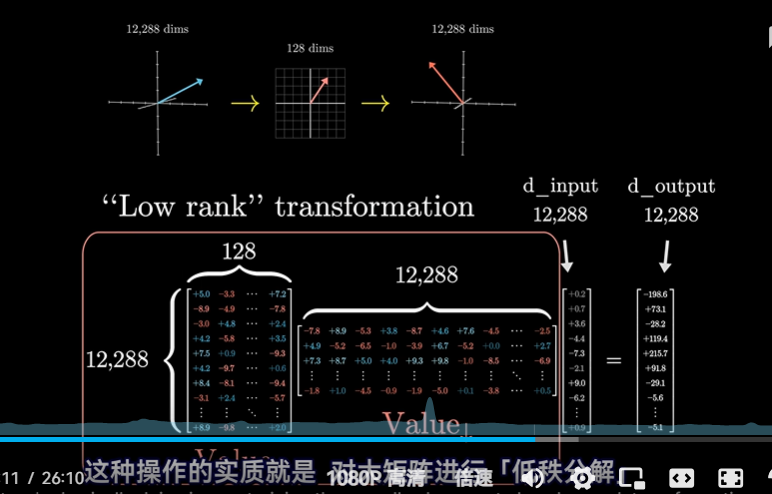
1. Key

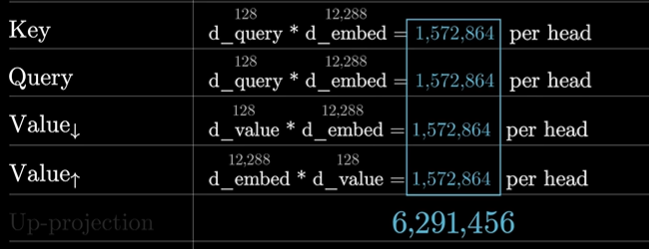


1. Value

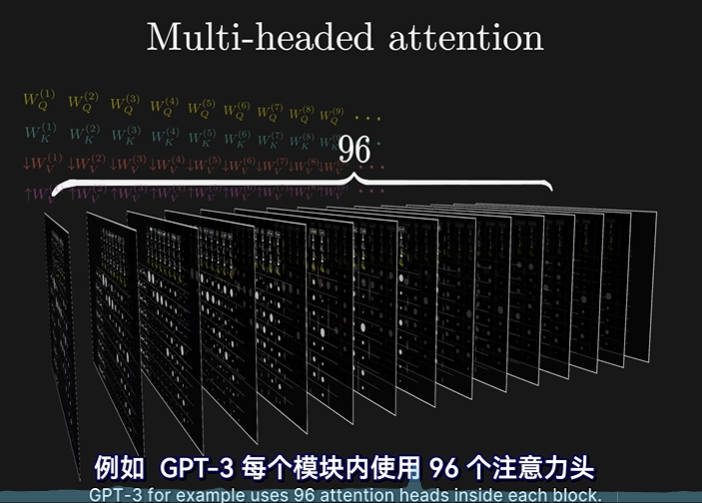




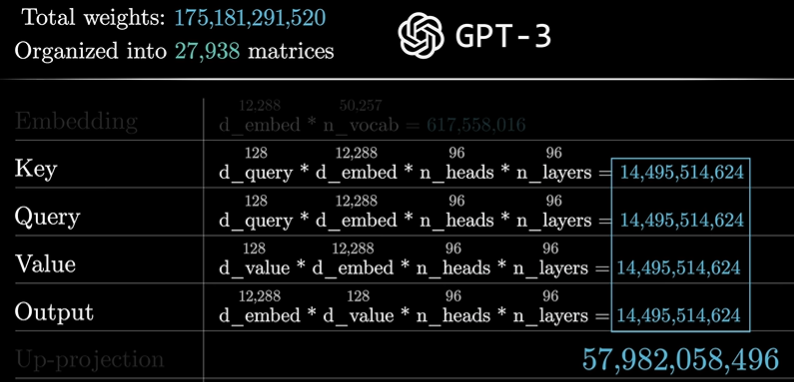




1. 96头注意力机制



1. Attention is about 1/3 of what you need >\_<



03 transformer 结构

