# MoE: Mixture of Experts

对于传统的 Embedding 前向使用 FFN：

A diagram of a cat and a cat

AI-generated content may be incorrect.A diagram of a router

AI-generated content may be incorrect.

如果 Expert 是 ，那么我们可以认为是一个概率加权

A diagram of a network

AI-generated content may be incorrect.

Router Collapse: 向专家模型分配令牌(tokens)不均衡的情况。缓解方法包括：

1. Token Dropping - 如果专家 超载，则不计算
2. Auxiliary Loss: 通过损失函数惩罚不平衡的专家利用率
3. Bias Term: 手动向亲和度分数(affinity scores)添加/减去偏置项以平衡专家

# LoRA

# Quantisation

Absolute Maximum Quantization

对于特别大的数据数据会立群，导致小数据偏移更大，因此将数据分块，每一块使用单独的 。

4bit NF Q 是把 FP32 压缩到了 NF4

我们假设每个块有64个数，每数 4bit。而每一个块需要一个FP32 的，因此 。这占比有点高，我们把256个量化常数作为一组，进行一个 8bit 量化，也就是会多一个 32bit。

因此我们可以看作第一部分的量化常数从原来的 FP32 变成 NP8。

也就是先把所有的数据压缩到 64个NF4组成的Chunk, 每个chunk有一个 FP8的常数()。每256个chunk有一个 FP 32的常数（）。

# QLoRA

QLoRA = 4bit Normal Float Quantisation + Double Quantisation + Page Optimisation