

**大模型的主要参数**

**1. 模型输入参数**

* **vocab\_size**（词表大小）：
  + 表示模型可以生成的字符或单词的种类数量，即词表的大小。
  + 在字符级别的模型中，vocab\_size 通常是所有可能的字符（包括汉字、标点符号等）的数量。
* **seq\_length**（序列长度）：
  + 表示输入序列的最大长度。
  + 例如，输入的诗句是 20 个字符长，则 seq\_length=20。

**2. 嵌入层参数**

* **embedding\_dim**（嵌入维度）：
  + 表示嵌入层将每个字符或单词映射到的向量空间的维度。
  + 例如，embedding\_dim=128 表示每个字符会被映射到一个 128 维的向量。

**3. Transformer 模型参数**

* **num\_heads**（注意力头的数量）：
  + 表示多头注意力机制中并行注意力头的数量。
  + 多头注意力可以捕获输入序列中不同位置之间的关系。
  + 例如，num\_heads=4 表示每层 Transformer 有 4 个注意力头。
* **num\_layers**（Transformer 层的数量）：
  + 表示 Transformer 编码器和解码器堆叠的层数。
  + 层数越多，模型的表达能力越强，但训练时间也会增加。
* **ff\_dim**（前馈神经网络的维度）：
  + 表示 Transformer 中每一层中前馈神经网络的隐藏层维度。
  + 通常是 embedding\_dim 的 2 倍或 4 倍。
* **dropout**（丢弃率）：
  + 表示在训练过程中随机丢弃神经元的概率，用于防止过拟合。
  + 例如，dropout=0.1 表示以 10% 的概率丢弃神经元。

**4. 训练相关参数**

* **learning\_rate**（学习率）：
  + 表示优化器更新模型参数的步长。
  + 较小的学习率（如 1e-4）可以提高训练稳定性，但可能需要更长的训练时间。
* **batch\_size**（批量大小）：
  + 表示每次训练中输入模型的数据样本数量。
  + 批量大小越大，训练越稳定，但对显存的需求也越高。
* **epochs**（训练轮数）：
  + 表示模型遍历整个训练数据集的次数。
  + 通常需要多次遍历数据集以使模型收敛。

**5. 推理相关参数**

* **temperature**（温度参数）：
  + 控制生成的随机性。
  + 较低的值（如 0.7）会使生成的内容更确定，较高的值（如 1.0）会增加随机性。
* **top\_k**（Top-K 采样）：
  + 表示从模型输出的概率分布中选择前 K 个概率最高的字符进行采样。
  + 限制采样范围，避免生成低概率的字符。
* **top\_p**（Top-P 采样，Nucleus Sampling）：
  + 表示从模型输出的概率分布中选择累积概率大于或等于 P 的字符进行采样。
  + 动态调整采样范围，既能限制范围，又能增加多样性。

**6. 其他参数**

* **positional\_encoding**（位置编码）：
  + 表示输入序列中每个位置的位置信息。
  + 由于 Transformer 模型没有循环结构，需要通过位置编码显式地告诉模型字符的位置信息。
* **tgt\_mask**（目标序列掩码）：
  + 防止模型在生成下一个字符时看到未来的字符。
  + 这是 Transformer 模型中自回归生成的关键部分。