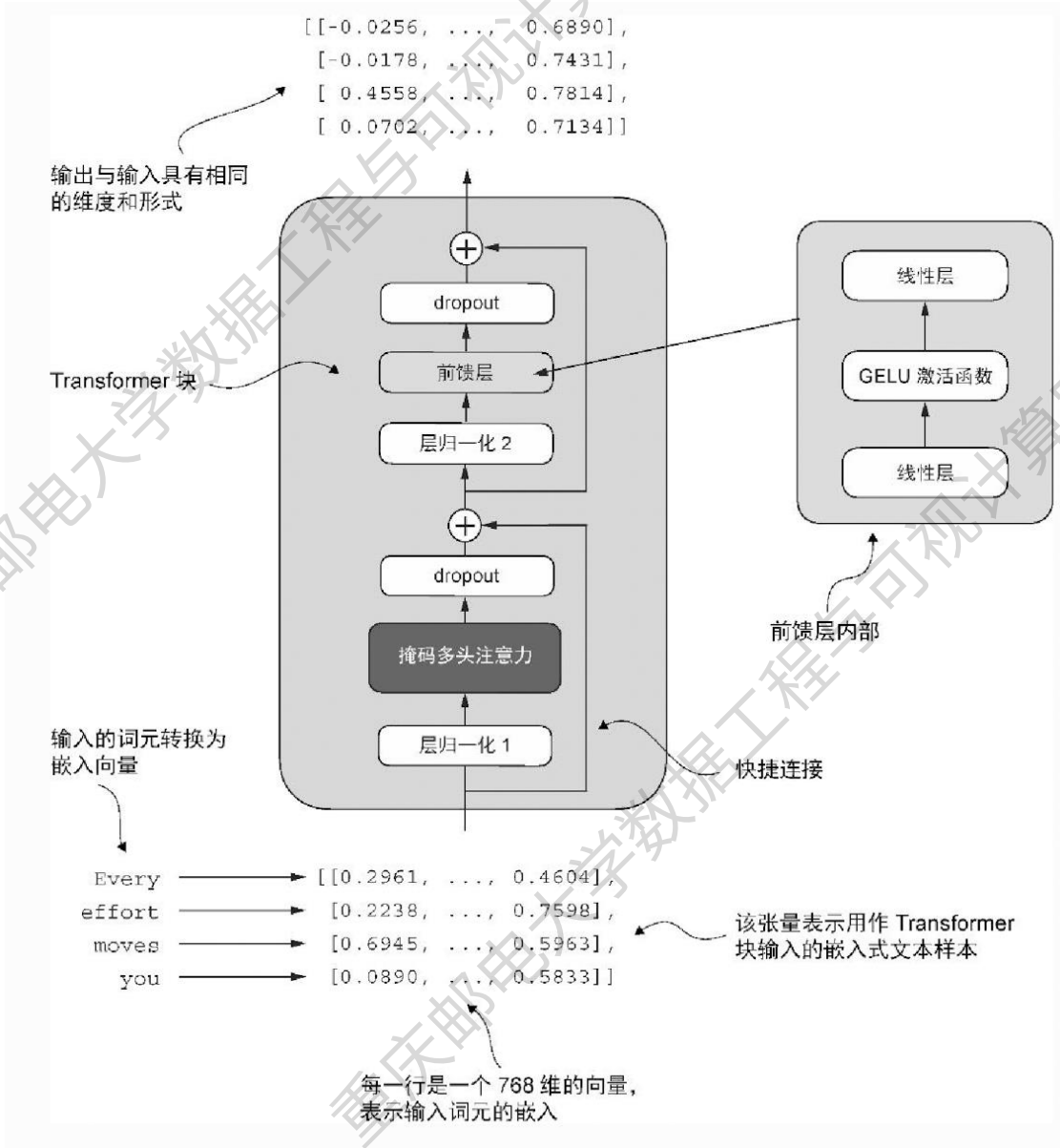
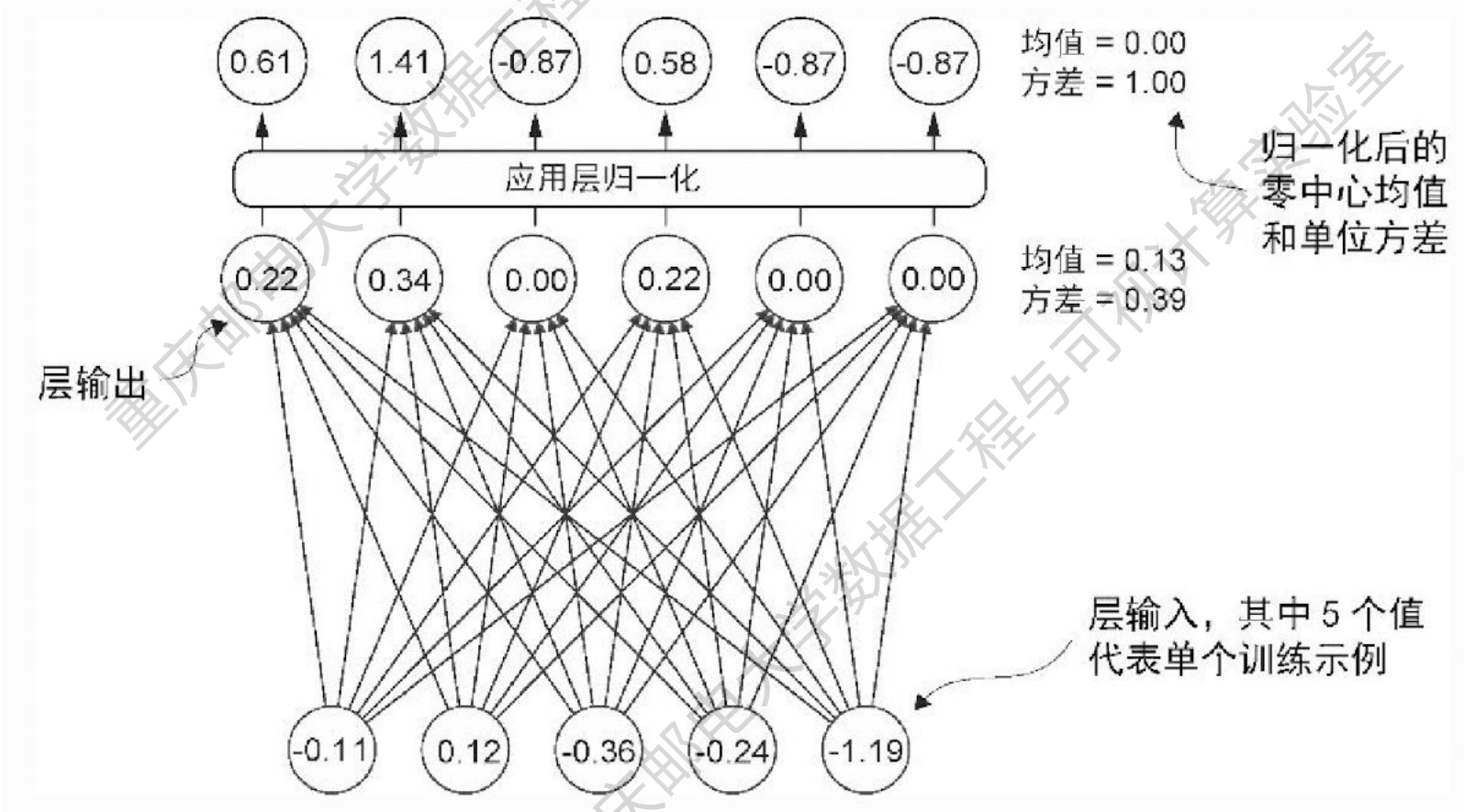
大语言模型预训练（二）
——GPT模型预训练

Transformer模块的结构



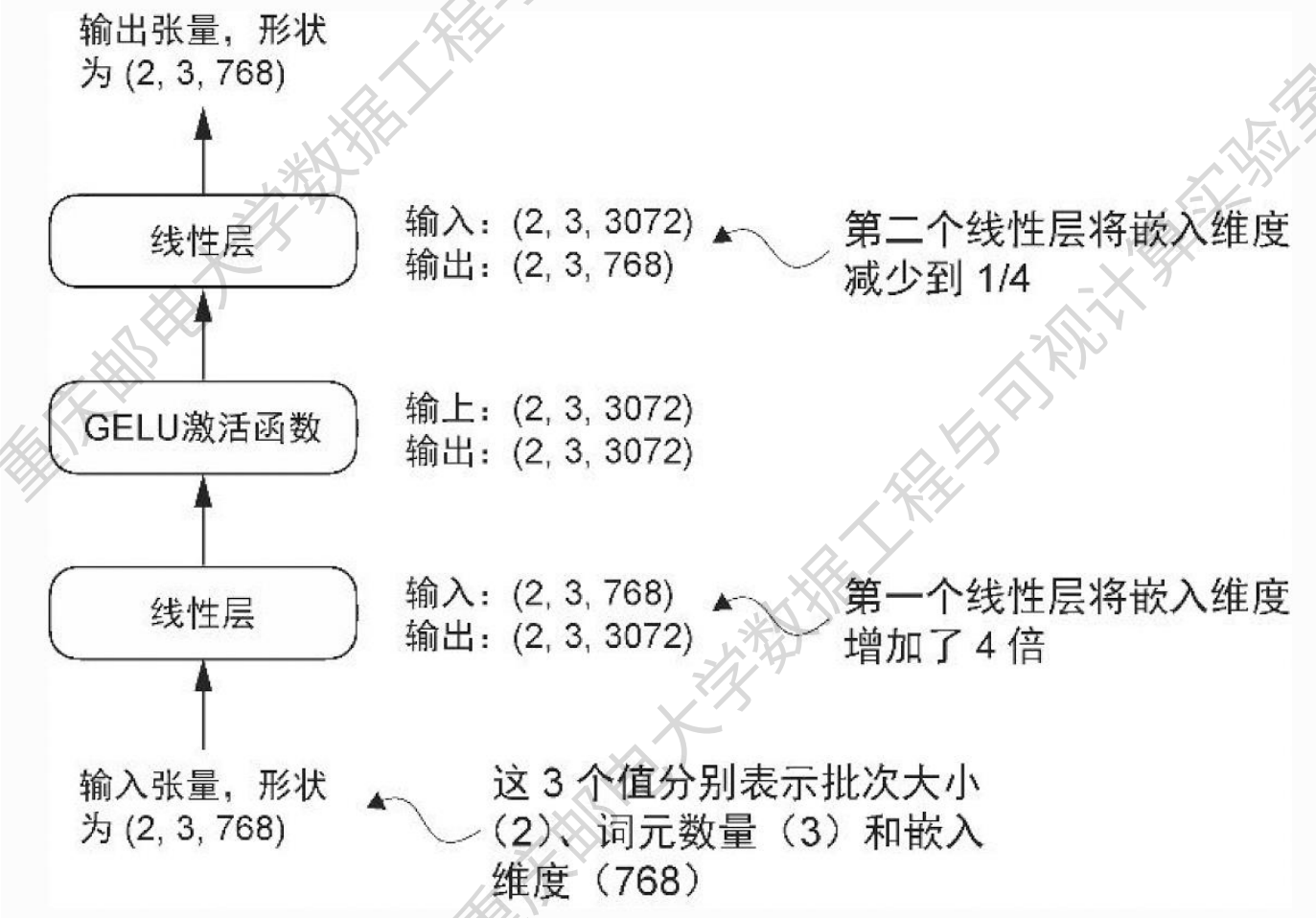
层归一化

层归一化是大语言模型中的重要组件。其目的是调整神经网络层的输出，使其均值为0，方差为1，提高模型训练过程的可靠性和稳定性。完成实验任务19-20。



前馈神经网络层

前馈神经网络层在提升模型学习和泛化能力方面非常关键。该模块保持输入和输出维度一致，通过第一个线性层将嵌入维度扩展到更高维度，应用GELU激活函数，通过第二个线性层将维度缩回原始大小。

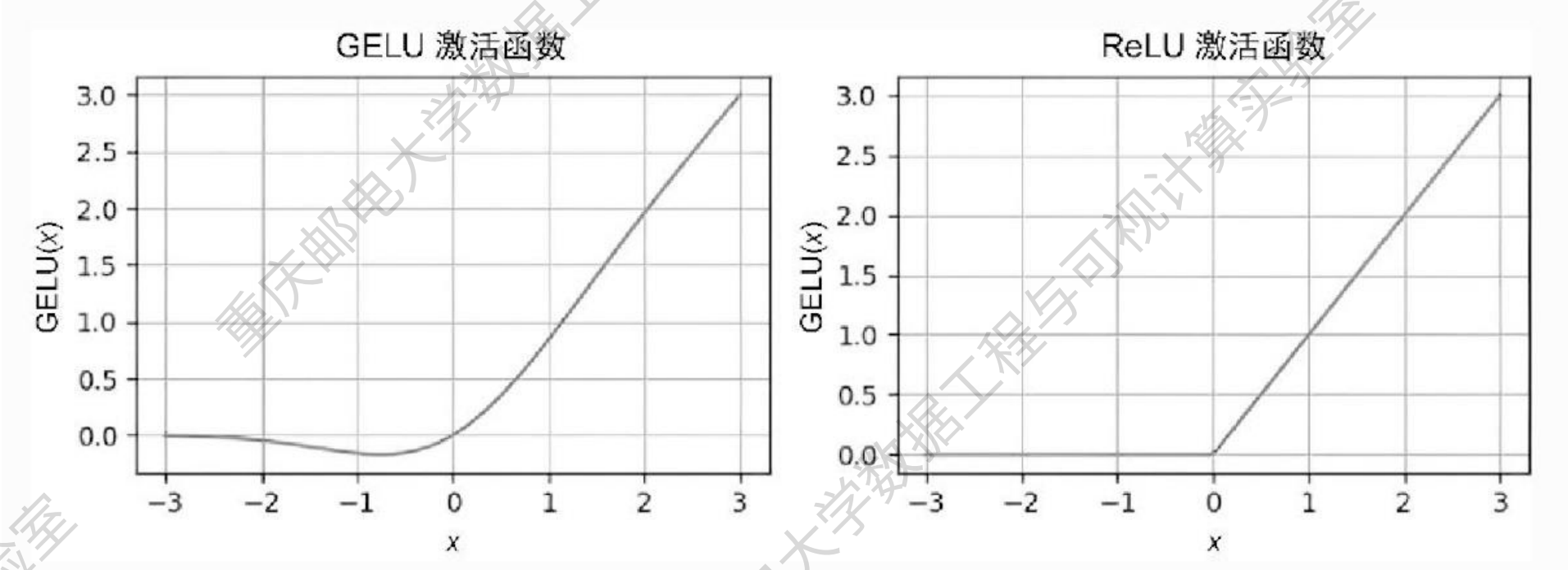


前馈神经网络层

与ReLU激活函数相比，GELU是更为复杂且平滑的激活函数，能够提升深度学习模型的性能。

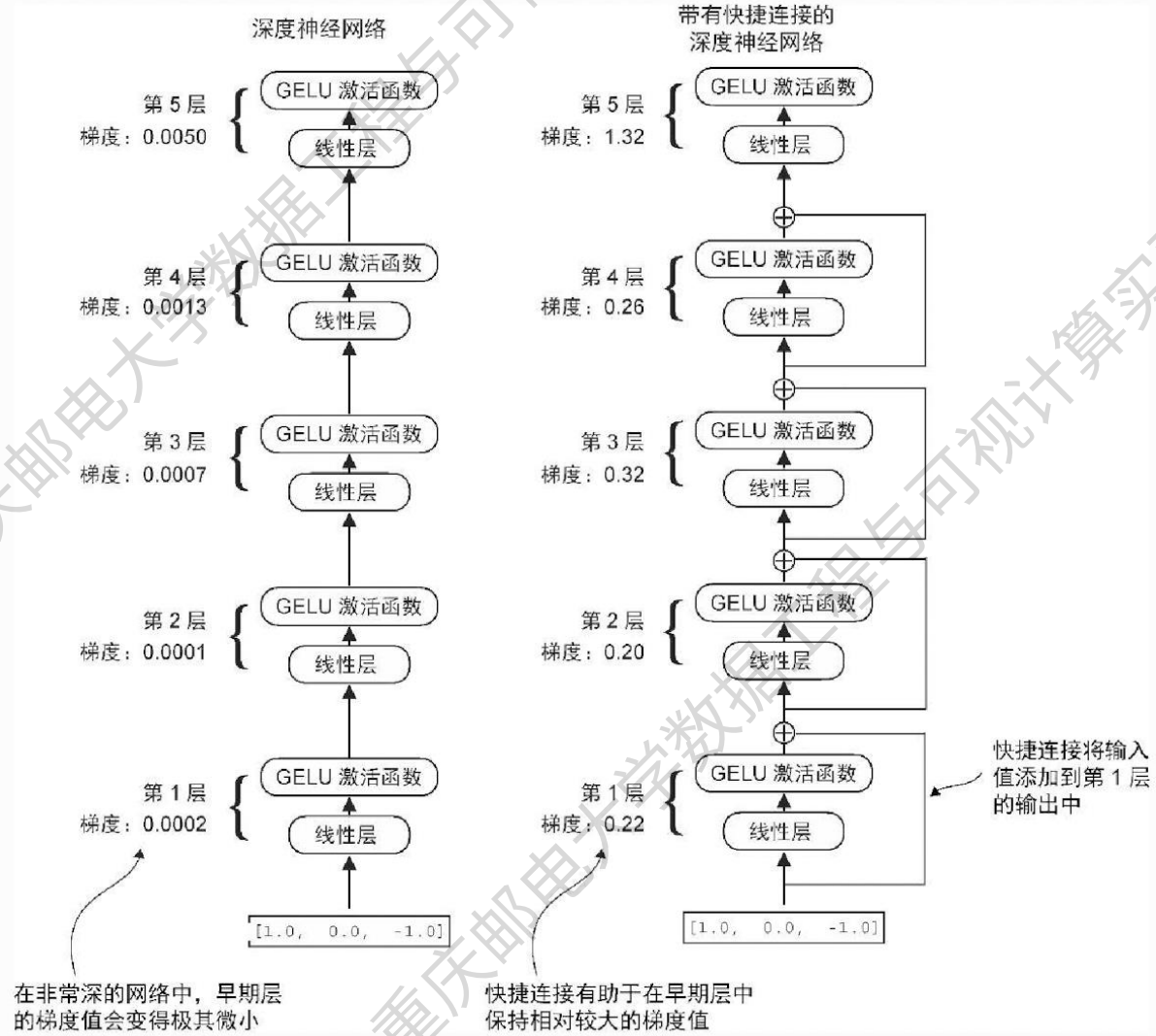
完成实验任务21。

$$\mathrm{GELU}(x) \approx 0.5 \cdot x \cdot \left(1 + \tanh\left[\sqrt{\frac{2}{\pi}} \cdot \left(x + 0.044\,715 \cdot x^3\right)\right]\right)$$



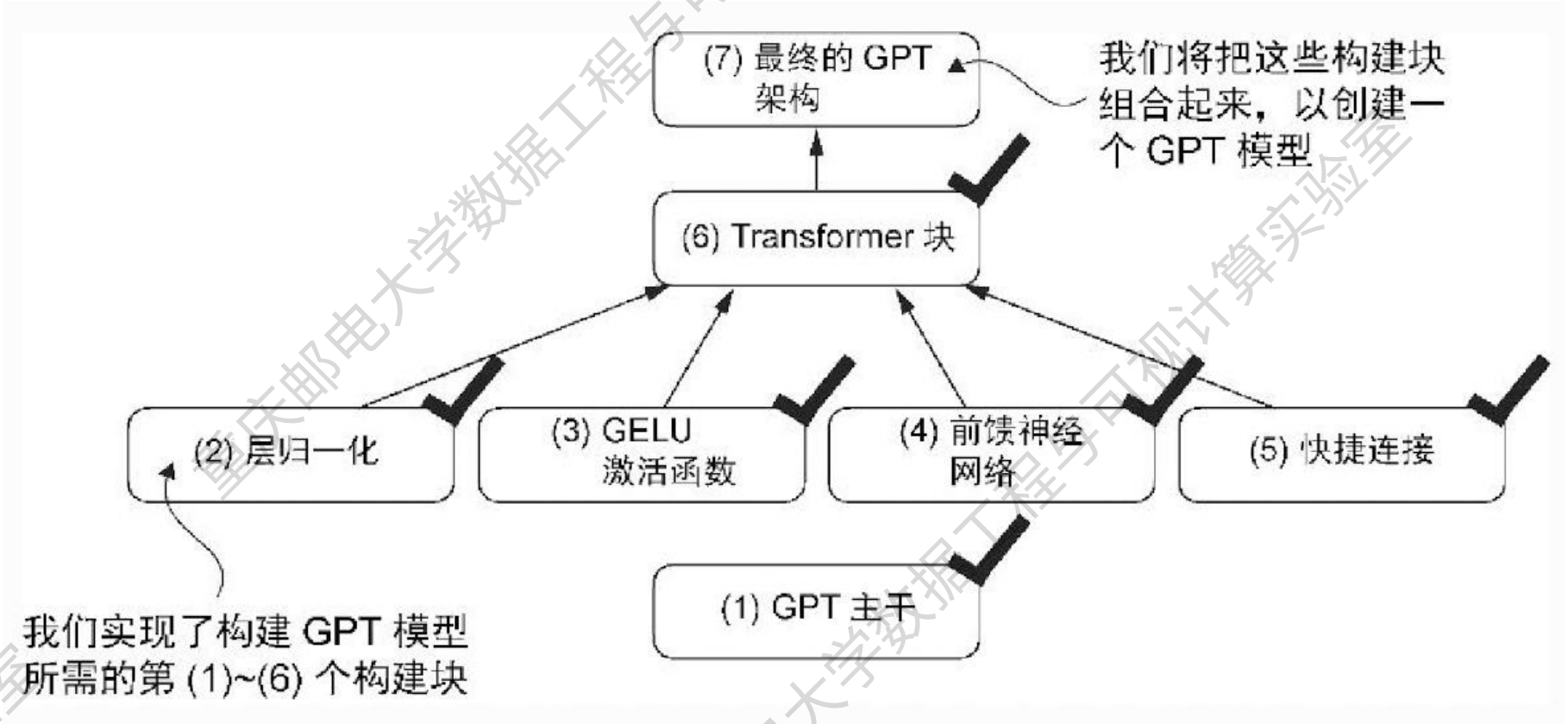
实现Transformer模块

快捷连接可以跳过一个或多个层，能够缓解梯度消失问题。



实现Transformer模块

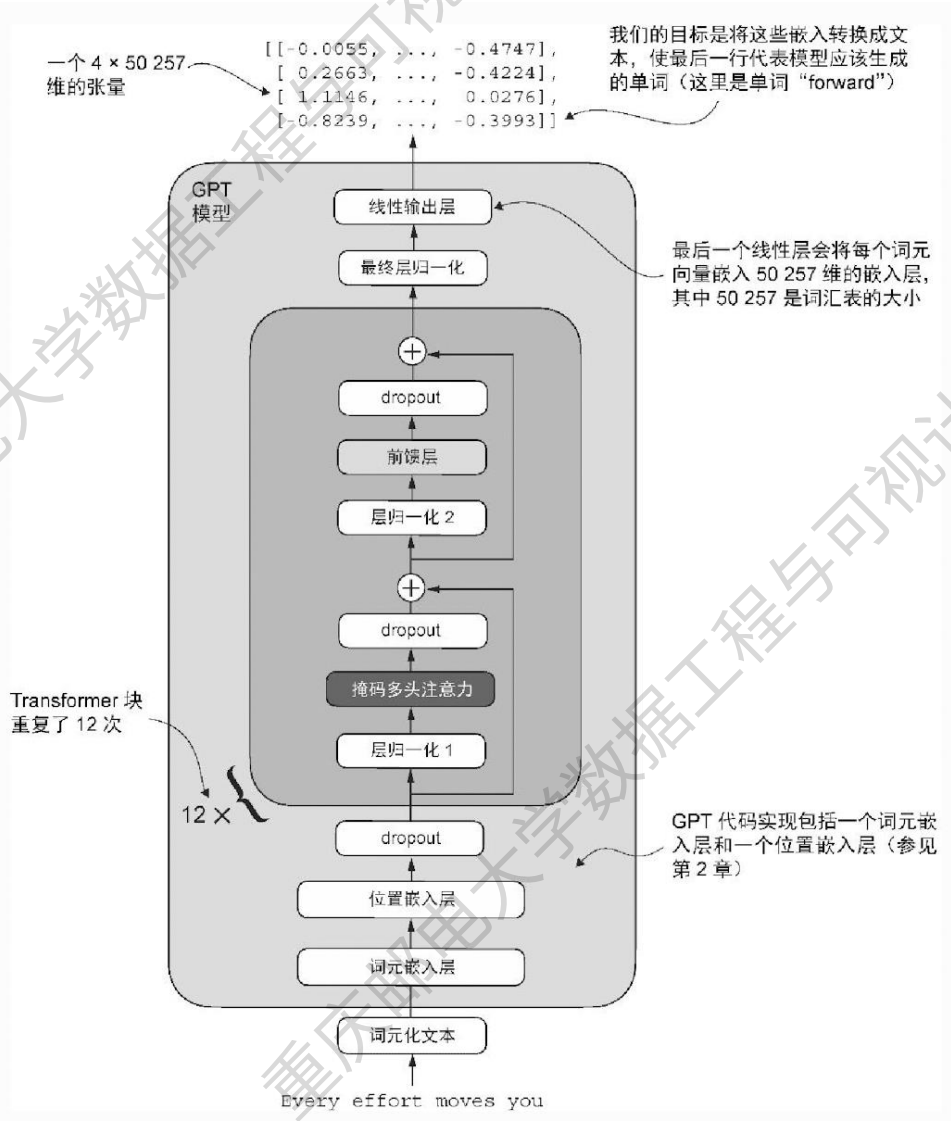
实现Transformer模块，包括：层归一化、前馈神经网络和快捷连接。完成实验任务22。



实现GPT模型

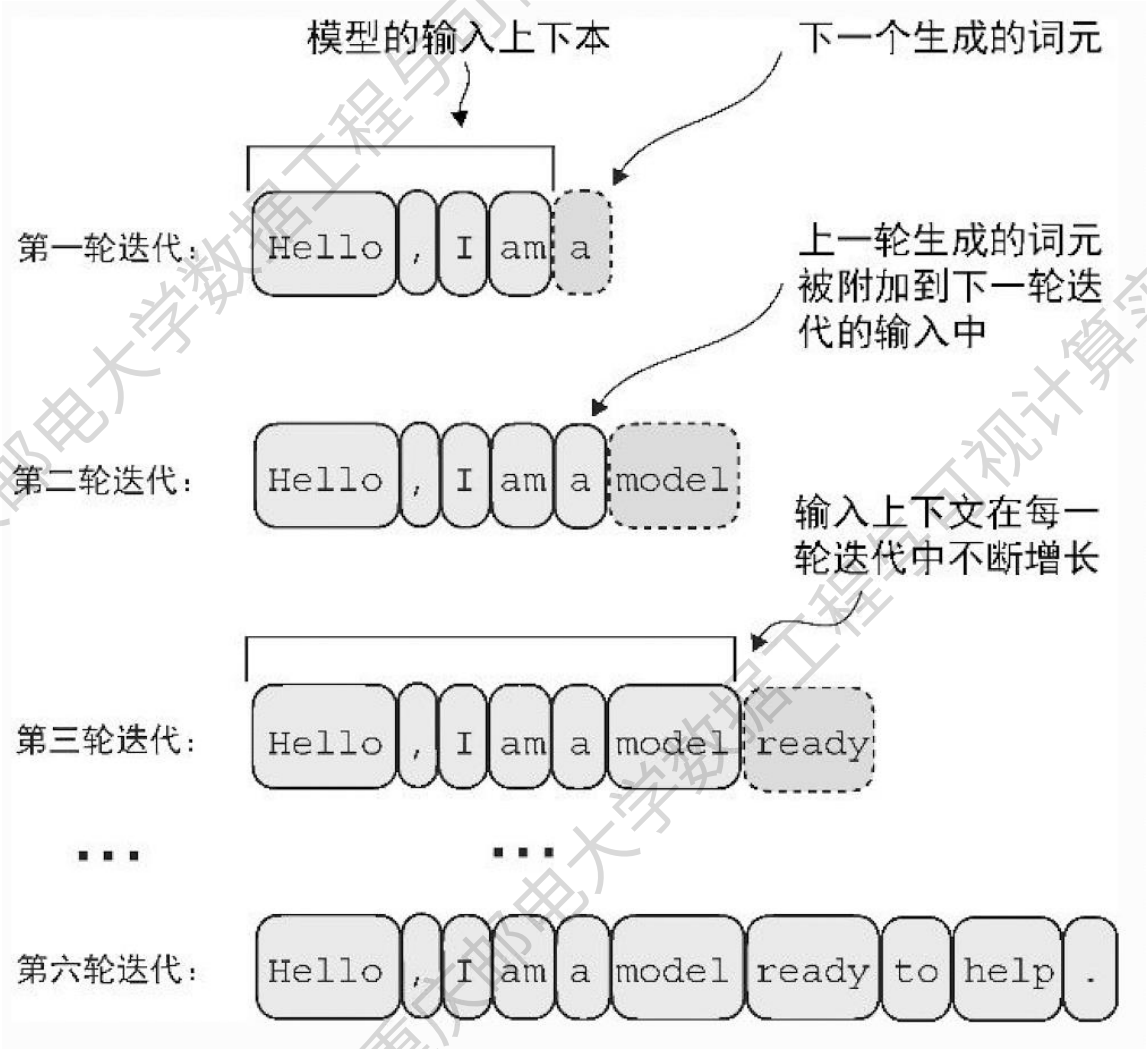
实现参数量为1.24亿的GPT-2模型，模型参数通过GPT_CONFIG_124M字典指定。完成实验任务23。

为什么实现GPT-2？GPT-2是开源模型，规模较小，可以在笔记本电脑上运行。相比之下，在单个V100GPU上训练GPT-3需要355年。



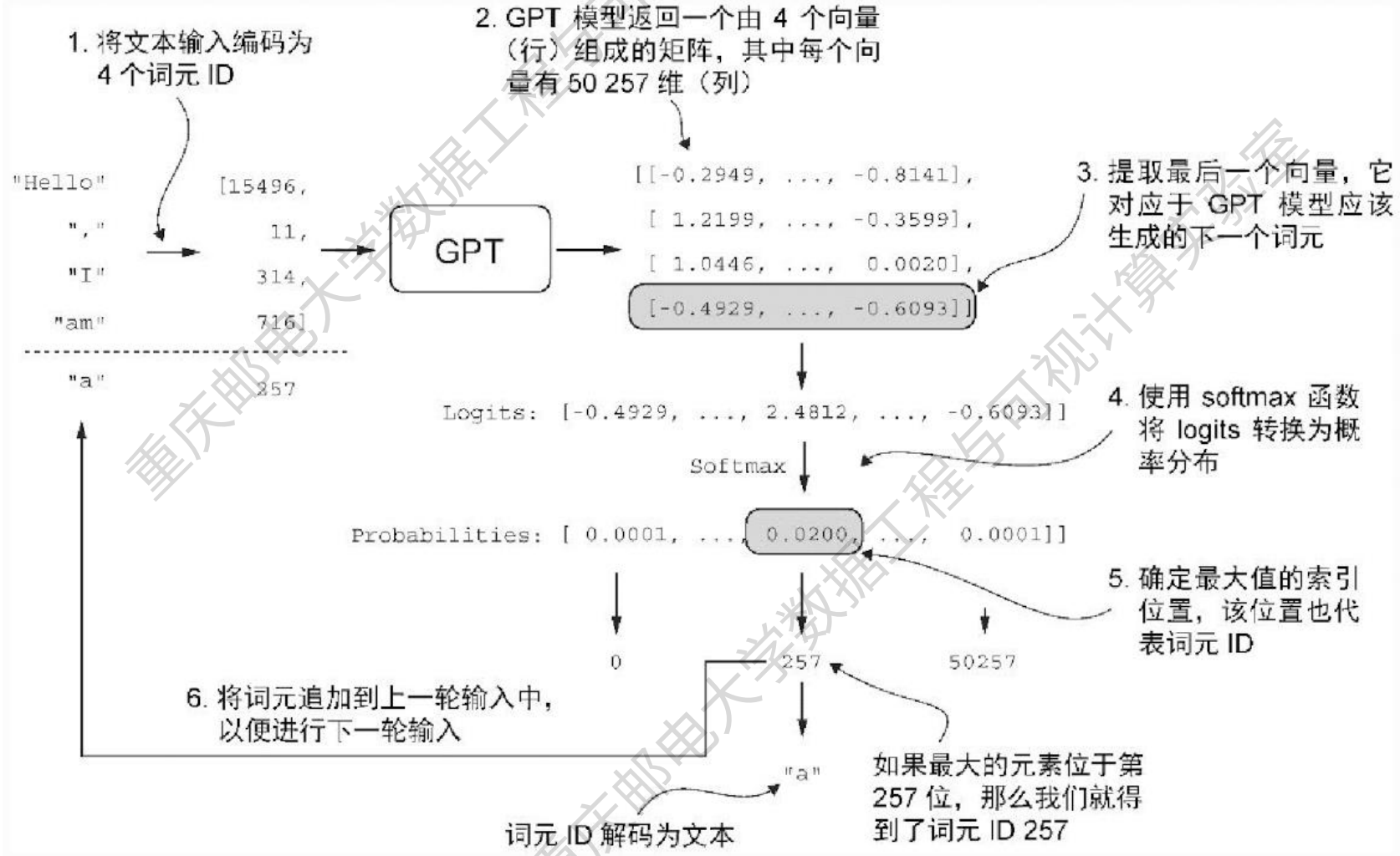
使用GPT模型生成文本

大语言模型逐步生成文本的过程



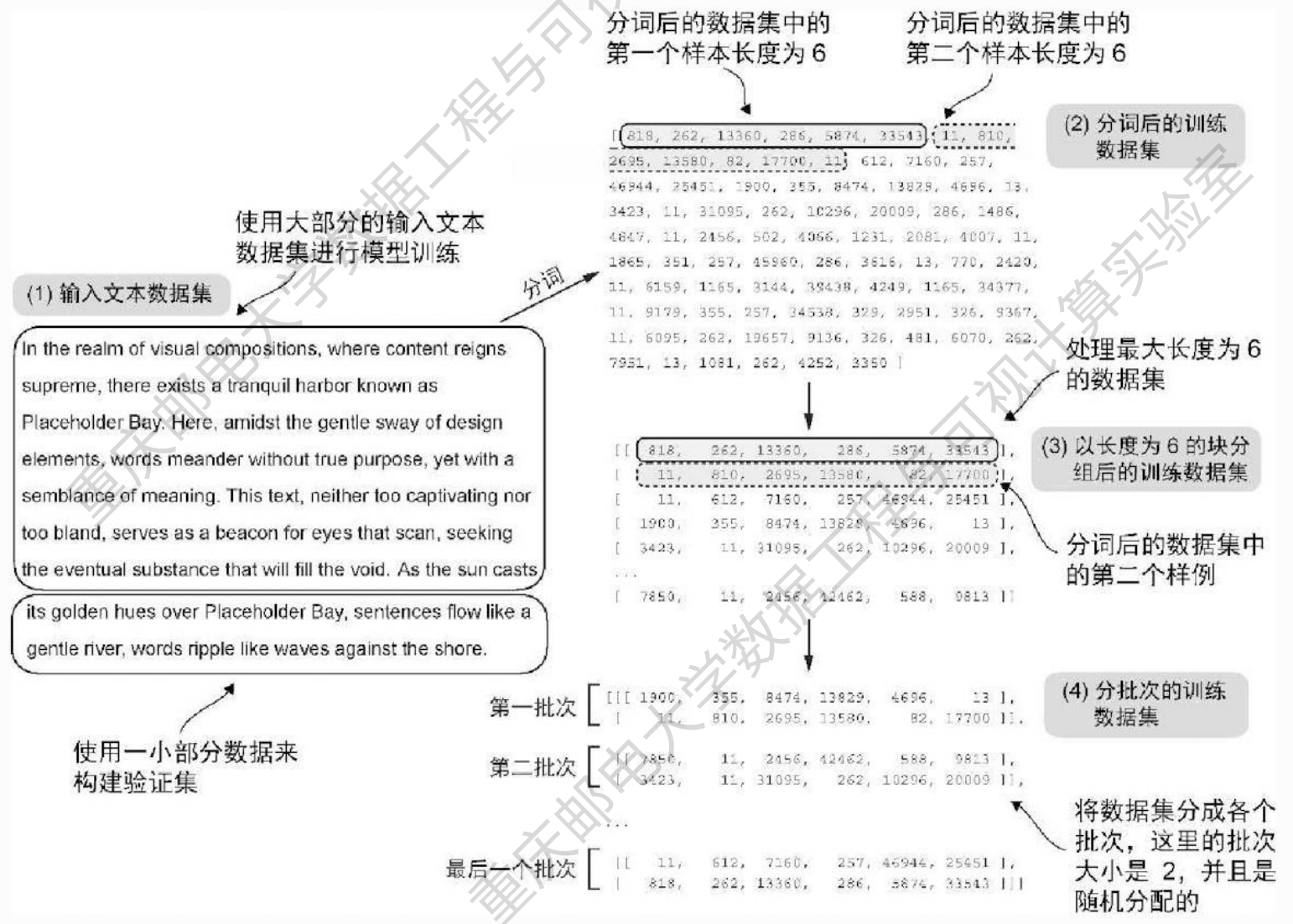
使用GPT模型生成文本

大语言模型在一次迭代中生成词元的过程。完成实验任务24-25。



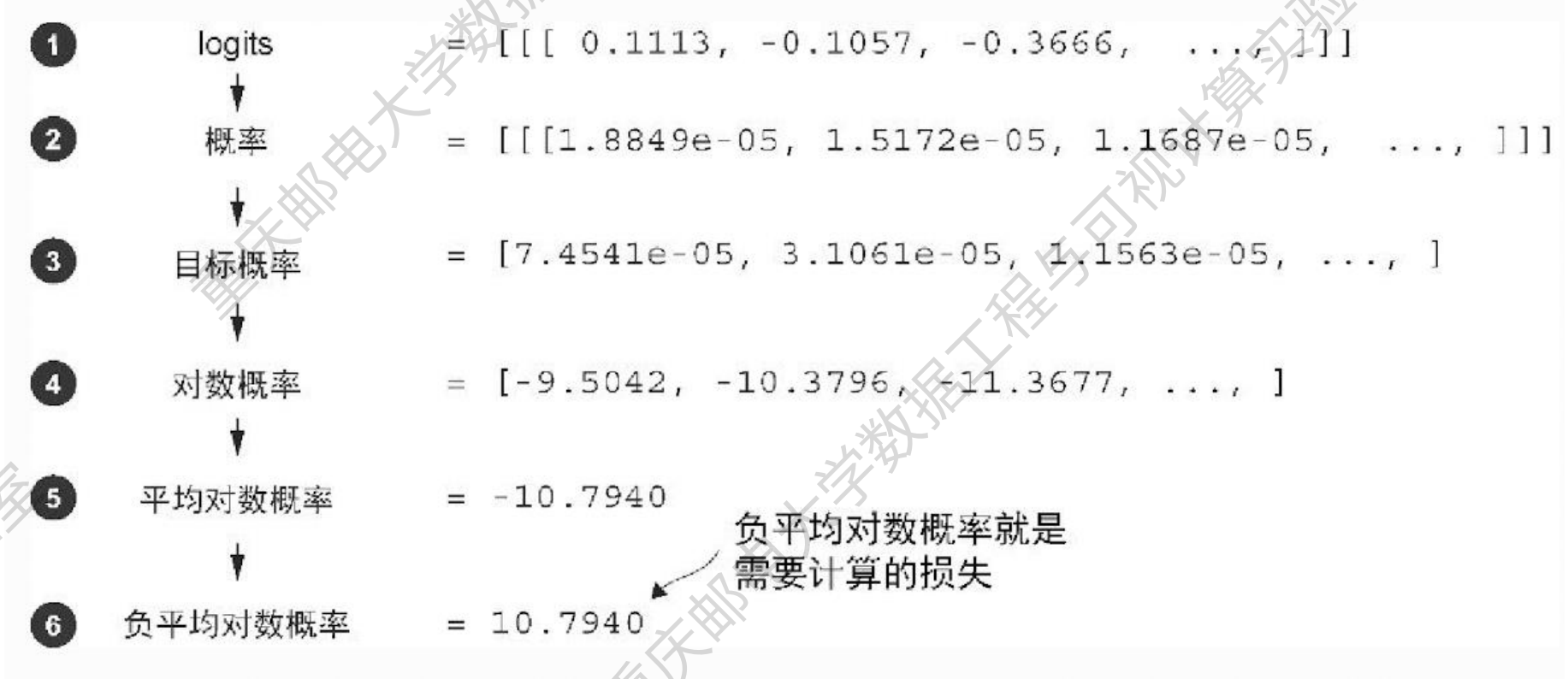
实现GPT模型的预训练

使用短篇小说作为数据集，准备预训练数据。完成实验任务26。



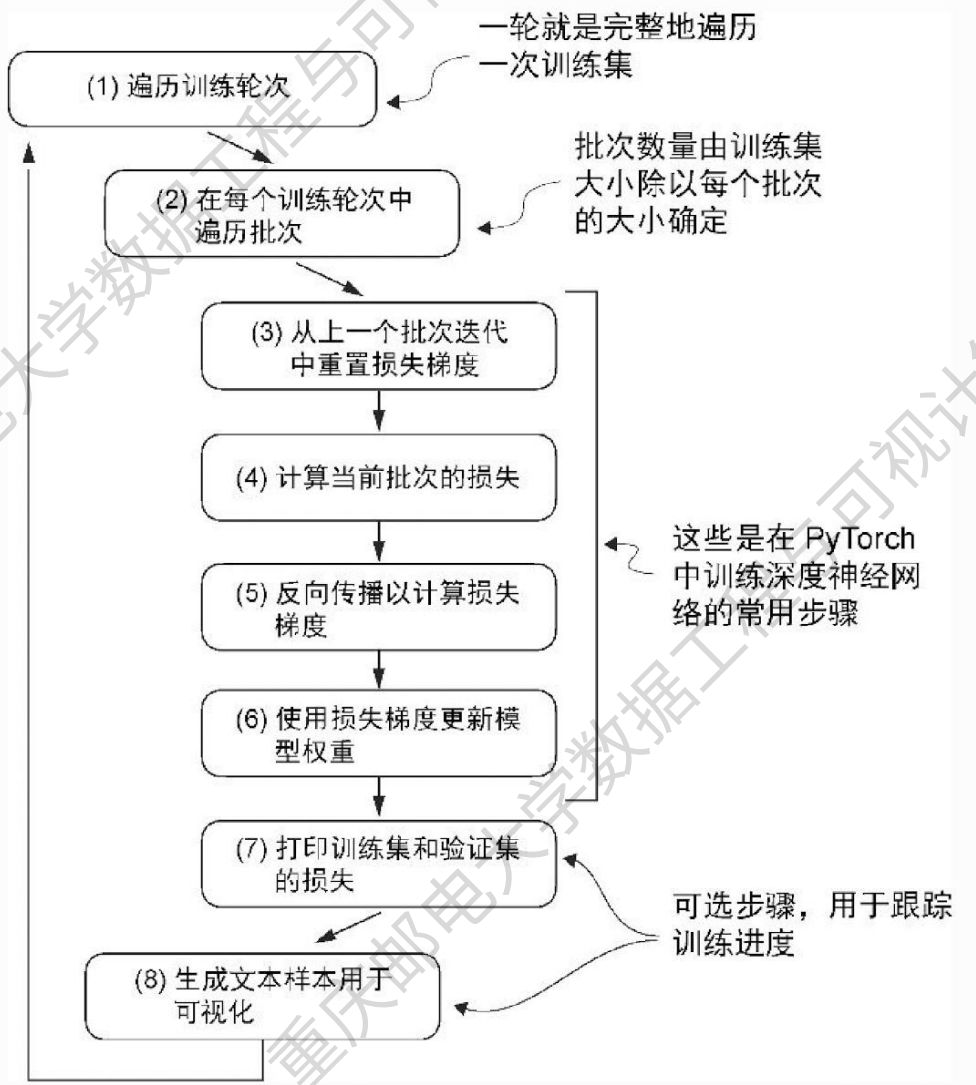
实现GPT模型的预训练

大模型预训练的损失函数：交叉熵（负平均对数概率），交叉熵度量预测结果概率分布和真实标签概率分布之间的差距。交叉熵越小说明模型预测结果越准确。模型的训练过程就是采用反向传播最小化损失函数。



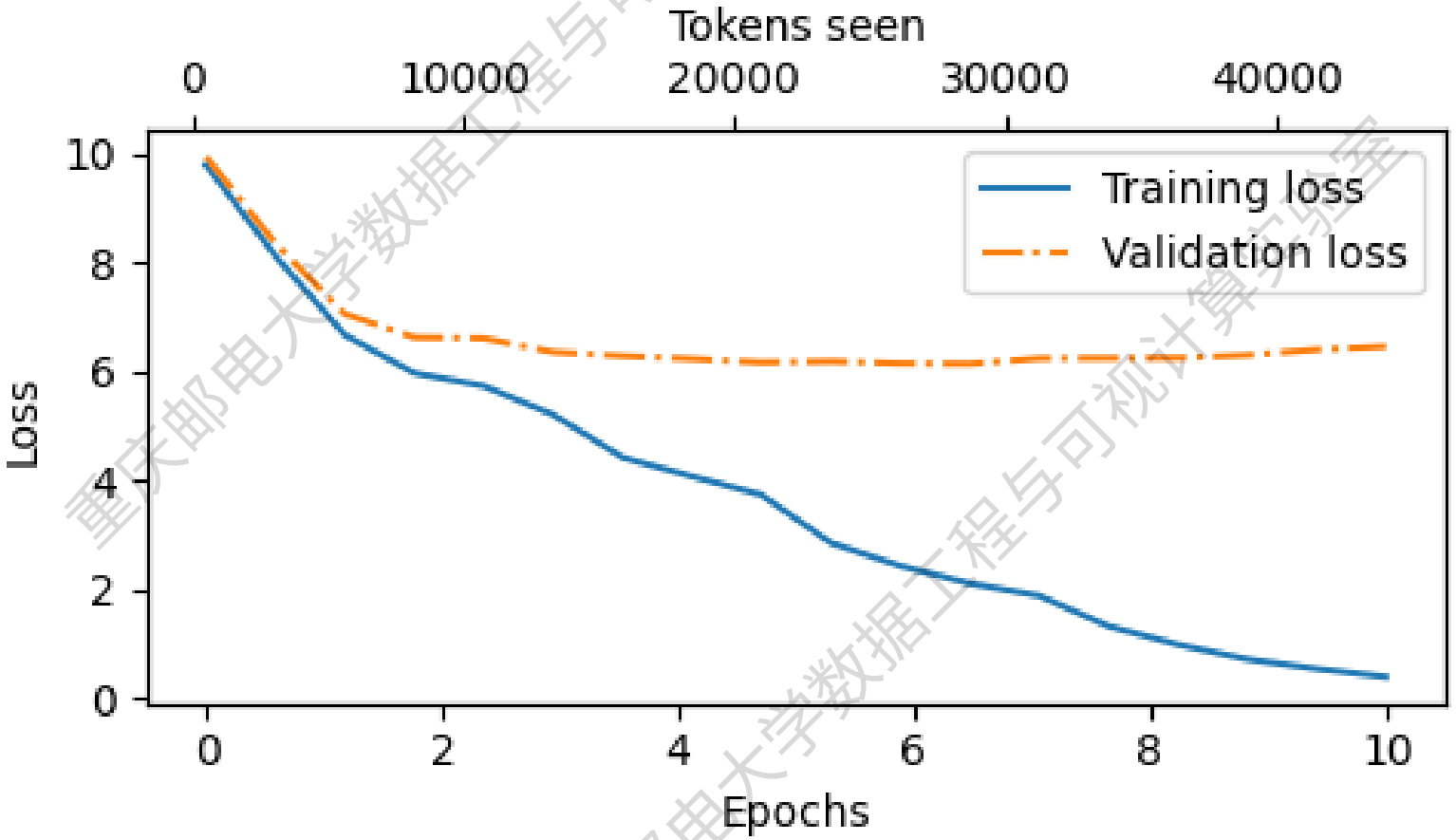
实现GPT模型的预训练

使用Pytorch训练大语言模型的步骤。完成实验任务27。



实现GPT模型的预训练

画图查看模型的训练集损失和验证集损失。完成实验任务28。



课后实验任务：在一个中文数据集上实现GPT模型预训练，并使用它生成文本