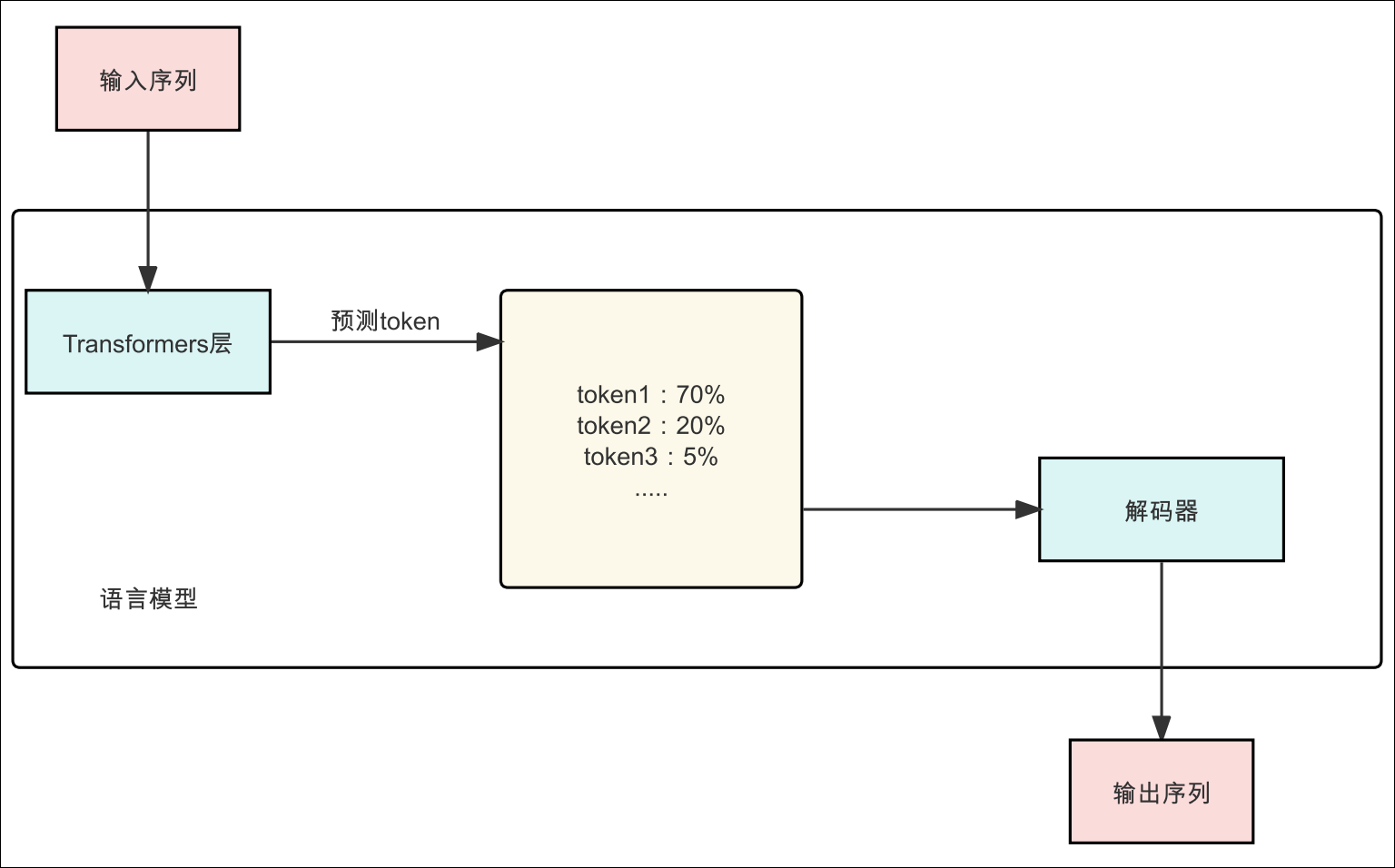
输入序列
Transformers层
预测token
token1：70%
token2：20%
token3：5%
.....
解码器
语言模型
输出序列