### ****1. 任务背景与目标****

系统需在给定光标位置的情况下，预测并生成后续文本。

### ****2. 方案调研****

本项目可使用以下技术构建：

**语言模型：**GPT 系列或开源模型（Llama 2、Mistral、Gemma）。微调模型进行文本补全任务。

**推理策略：**采样策略（Greedy、Top-k、Top-p 采样）。

**深度学习框架**：PyTorch。

**数据集**：BBC新闻语料（如https://huggingface.co/datasets/RealTimeData/bbc\_news\_alltime）。

**评估指标**：

1.生成文本的流畅性与合理性（BLEU、ROUGE）。

2.上下文相关性（BERTScore、Perplexity）。

### ****3. 方案设计****

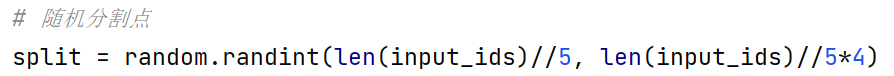
### 给定一段文本和光标位置，把光标前的文本进行预处理（判断长度是否超标，如果是则截断）

之后送入微调后的模型，解码输出得到补全文本。

### ****4. 详细实现****

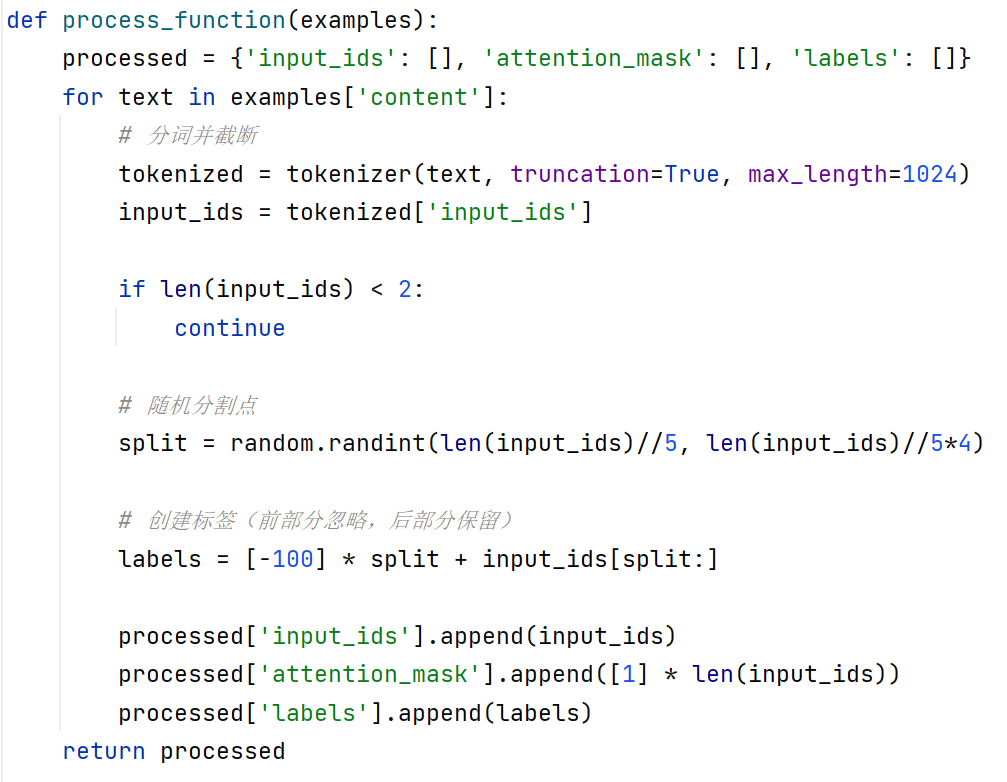
下载预训练的gpt2模型和bbc\_news\_2023\_may作为数据集，按9：1划分为训练集和验证集

按照token分割数据：随机选取数据中的某个点



分割点之前作为模型输入，之后作为label训练

流程如下：



加载数据集和gpt2的预训练模型开始微调。

微调结束后在验证集进行BERTScore评估。

代码链接如下：https://github.com/Chen-Zhen-Hui/Text-Auto-Completer

先运行train.py在数据集训练微调。

运行eval.py进行评估（BERTScore=0.826）。

在main.py修改example\_text的内容作为想要的输入送入模型得到预测结果。