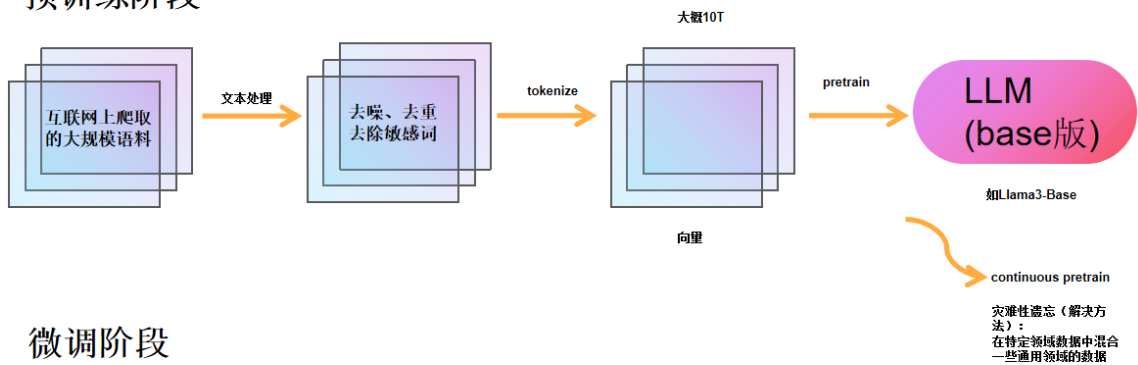
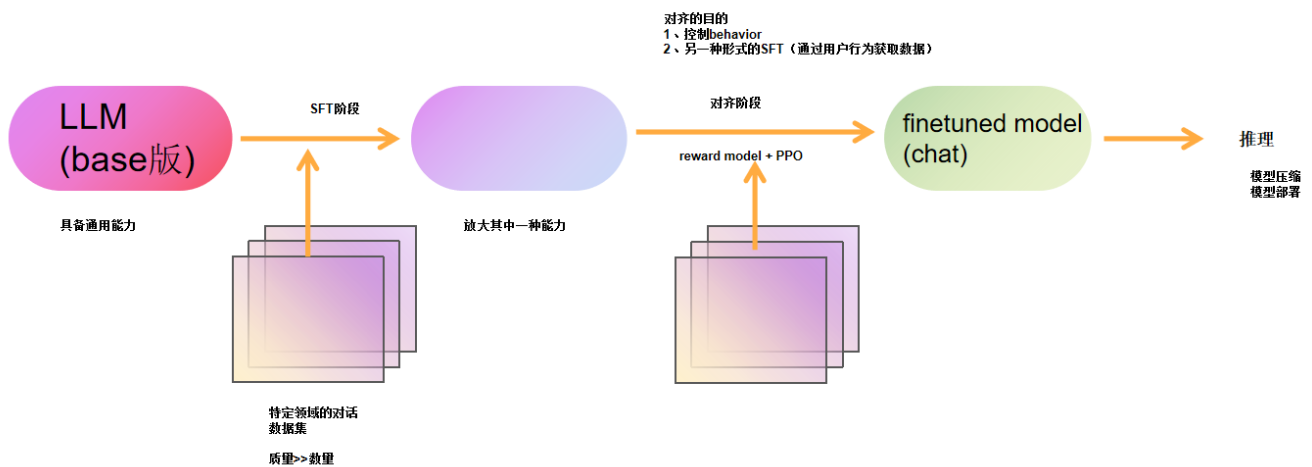


LLM训练流程

LLM 预训练阶段



微调阶段



【大语言模型LLM基础之Tokenizer完全介绍-哔哩哔哩】 <https://b23.tv/2kdTKxf>

LLM中的tokenizers

三种不同分词粒度的Tokenizers

- word-based
- character-based
- subword-based
 - WordPiece: BERT、DistilBERT
 - Unigram: XLNet、ALBERT
 - BPE (Byte-Pair Encoding) : GPT-2、RoBERTa
 - SentencePiece

BPE

词频统计->词表合并

设置：BPE的合并次数

Byte-Pair Encoding (BPE) Tokenization

■ BPE 算法包含两部分：“词频统计”与“词表合并”

(("h" "ug", 10), ("p" "ug", 5), ("p" "un", 12), ("b" "un", 4), ("h" "ug" "s", 5))



使用刚刚添加到词表的“hug”替换

(("hug", 10), ("p" "ug", 5), ("p" "un", 12), ("b" "un", 4), ("hug" "s", 5))



停止合并，得到最终词表（BPE的合并次数是超参数）



使用词表进行 tokenization

bug => ["b", "ug"]
mug => ["<unk>", "ug"]

每合并一次
词汇表大小+1

GPT的词汇表大小为40478，因为它有478个基本字符，并且在40000次合并后停止。

["b", "g", "h",
"n", "p", "s",
"u", "ug", "un",
"hug"]

词表

- Sennrich, Rico, Barry Haddow, and Alexandra Birch. "Neural machine translation of rare words with subword units." *arXiv preprint arXiv:1508.07909* (2015).
- https://huggingface.co/docs/transformers/tokenizer_summary

22

改进：BBPE

Byte-level BPE (BBPE)

■ BPE 的缺点

- 包含所有可能的基本字符 (token) 的基本词汇表可能相当大
- 例如，将所有 Unicode 字符都被视为基本字符（如中文）

■ 改进：Byte-level BPE

- 将字节 (byte) 视为基本 token
- 两个字节合并即可以表示 Unicode
 - ▣ 比如中文、日文、阿拉伯文、表情符号等等

GPT-2的词汇表大小为50257，对应于256字节的基本token、一个特殊的文本结束token和通过50000个合并得到的token。

WordPiece

23

WordPiece Tokenization

<https://static.googleusercontent.com/media/research.google.com/ja//pubs/archive/37842.pdf>

■ 大体和BPE类似

- 构建基本词表时，除第一个字母，会添加##作为前缀（BERT）

word => [w, ##o, ##r, ##d]

- 使用类似联合概率的大小而不是次数对 token 进行合并，公式如下：

$$\text{pair得分} = \frac{\text{pair出现的次数}}{\text{token1出现的次数} \times \text{token2出现的次数}}$$

- 通过将 pair 的频率除以其单个 token 的频率的乘积，该算法优先考虑单个 token 在词表中不太频繁的 pair 进行合并

pair得分越大，表明分子越大、分母越小

<https://huggingface.co/learn/nlp-course/chapter6/6?fw=pt>

24

Unigram

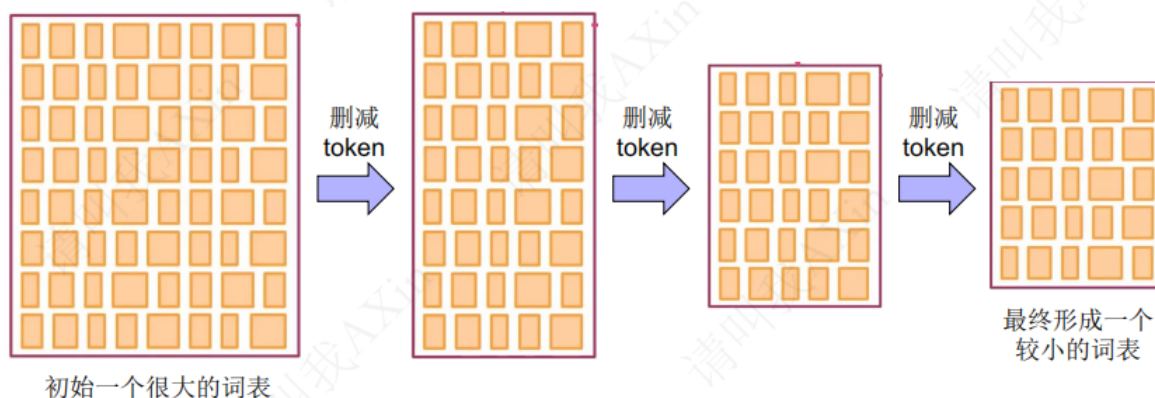
先初始化一个很大的词表（字母、单词、subword都包括）

设置：删减的次数

Unigram Tokenization

<https://arxiv.org/pdf/1804.10959.pdf>

- Unigram 算法经常在 SentencePiece 中使用，是 ALBERT、T5、mBART、Big Bird 和 XLNet 等模型使用的 tokenization 算法



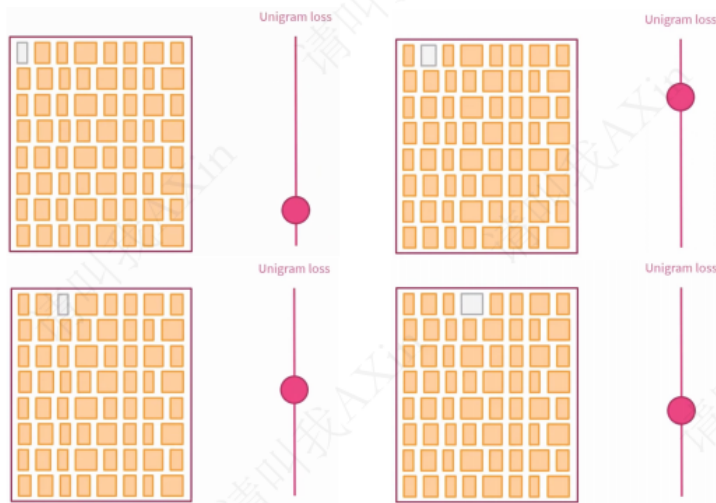
<https://huggingface.co/learn/nlp-course/chapter6/7>
<https://www.youtube.com/watch?v=TGZfZVuF9Yc>

27

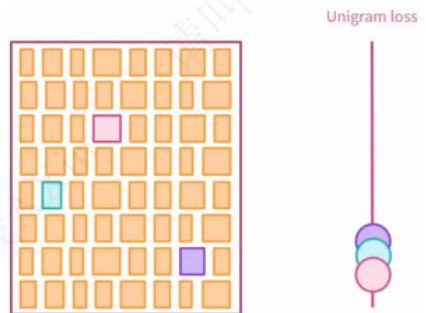
删去对词表的表达能力影响不大的token

Unigram Tokenization

■ 如何删减 token?



尝试删去一个 token，并计算对应的 unigram loss，删除 p% 使得 loss 增加最少的 token



<https://www.youtube.com/watch?v=TGZfZVuF9Yc>

28

基于统计的划分

loss: 负对数似然

SentencePiece (使用BBPE或Unigram)

解决多国语言的分词问题，输入都当做字节流（含空格）