LLM Notes

Y R

Updated on 2025-09-24

一些专业名词:

中文分词	Chinese Word Segmentation, CWS	处理中文文本时，由于词与词之间灭有明显分隔（空格），所以无法直接通过空格来确定词的边界。其目的是将连续的中文文本切分成有意义的词汇序列。
子词切分	Subword Segmentation	特别适合处理词汇稀疏的问题，即当遇到罕见词或者未见过的新词时，能够通过已知的子词单位来理解或生成这些词汇。在处理拼写复杂，合成词多的语言（德语）或预训练语言模型（BERT，GPT）中尤为重要。常用的方法有Byte Pair Encoding (BPE), WordPiece, Unigram, SentencePiece。
词性标注	Part of speech Tagging, POS Tagging	为文本中的每一个单词分配一个词性标签，如名词动词形容词。 POS tagging对理解句子结构，进行句法分析，语义角色标注等高级NLP任务至关重要。计算机可以更好地理解文本的含义，进行信息提取，情感飞，机器翻译。。通常依赖于机器学习模型，如隐马尔科夫模型(Hidden Markov Model HMM), 条件随机场（COnditional Random Field CRF),或RNN，LSTM等。通过学习大量的标注数据来预测新句子中每个单词的词性。
文本分类	Text Classfication	将给定的文本自动分配到一个或多个预定义的类别中。应用包括但不限于情感分析，垃圾邮件检测，新闻分类，主题识别等。文本分类的关键在于理解文本的含义和上下文，并基于此将文本映射到特定的类别。文本分类的关键在于选择合适的特征表示和分类算法，以及拥有高质量的训练数据。

实体识别 (又名，命名实体识别)	Named Entity Recognition, NER	自动识别文本中具有特定意义的实体，并将它们分类为预定的类别，如人名，地点，组织，日期，时间等。实体识别任务对于信息提取，知识图谱构建，问答系统，内容推荐等应用很重要，它能够帮助系统理解文本中的关键元素及其属性。

表 1: LLM