

# 大模型分词对生成结果影响的初步研究

赵迎功（乐言科技，yinggong.zhao@gmail.com, 微信: yinggong\_zhao）

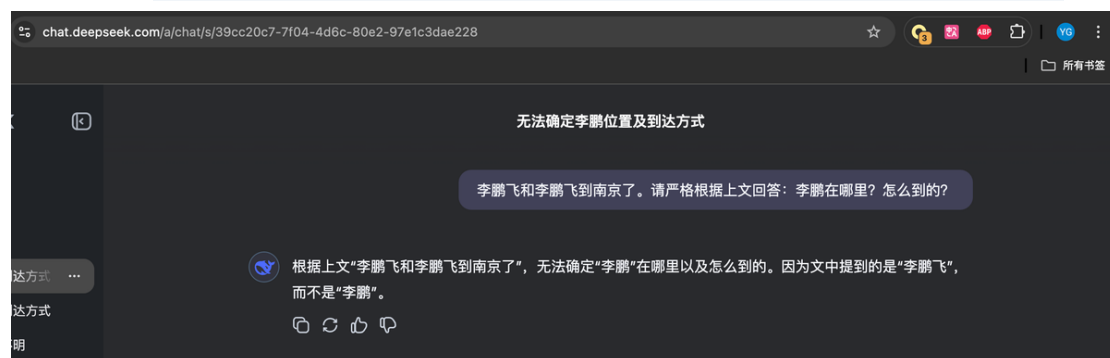
## 摘要

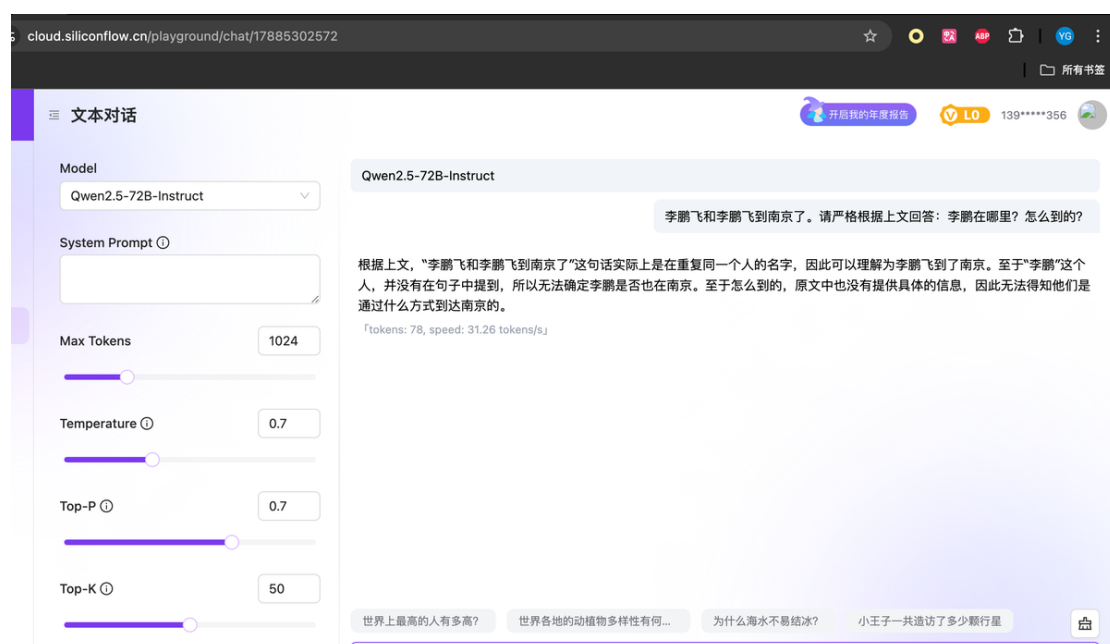
分词是大语言模型( Large Language Model, 下文简称大模型 )输入的第一步。目前包括 OpenAI[1]、通义千问[2]和 Deepseek[3]在内的主流大模型的分词都是利用字节级别 BPE 方法从海量数据构建得到。尽管这些大模型在各种任务上有非常优异的表现，仍然有一个问题待回答，即：分词结果对大模型有什么样的影响？在本论文中，我基于对 Qwen2.5-72b-Instruct[4]和 Deepseek-v3[5]两个模型分词结果和词表结果观察，构造出：1) 三组可能是分词错误/歧义传递导致回答错误的提示词；2) 一组 1-shot 歧义语句生成提示词。结果说明分词仍然是大模型的重要问题，需要更多投入分析和优化。最后我提出了一些潜在的优化方向。本文设计代码和数据已经上传到 github 项目：[https://github.com/zhaoyukoon/damoxing\\_fenci\\_gongji](https://github.com/zhaoyukoon/damoxing_fenci_gongji)。

### 1 说明

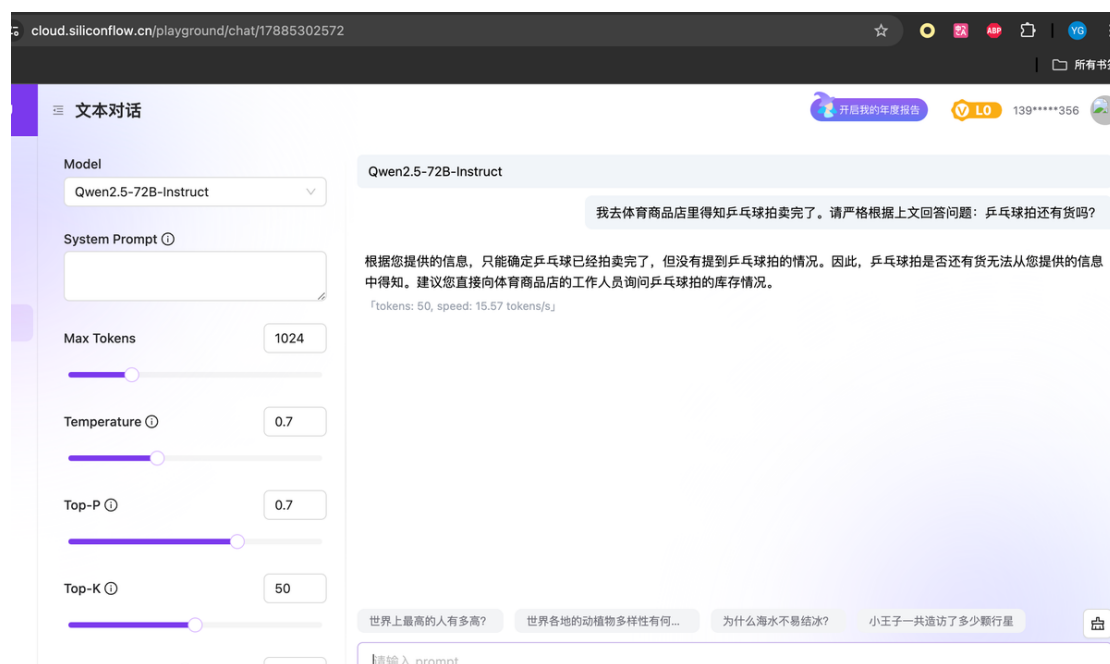
1. 本研究工作是纯科学研究，文章中所涉及的数据都是对分词结果和词表分析构造产生，文章内容不包含任何政治立场和观点；
2. 本研究工作是作者出于个人兴趣在工作之外独立研究，不代表任何公司（包括本人现在所在的乐言科技）、机构和个人的立场，未接收任何机构、公司和个人资助，和任何机构、公司和个人都无直接或者潜在的利益冲突；

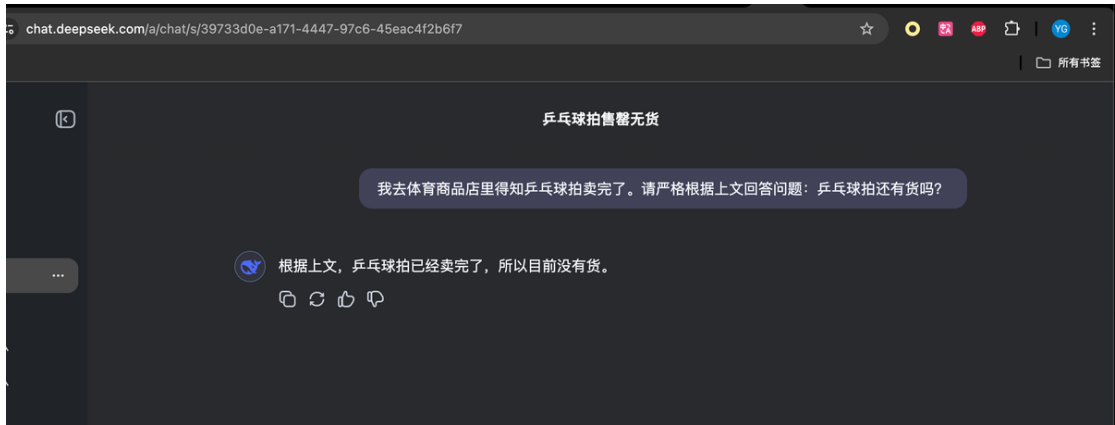
## 提示词 1（李鹏飞和李鹏飞到南京了。请严格根据上文回答：李鹏在哪里？怎么到的？）



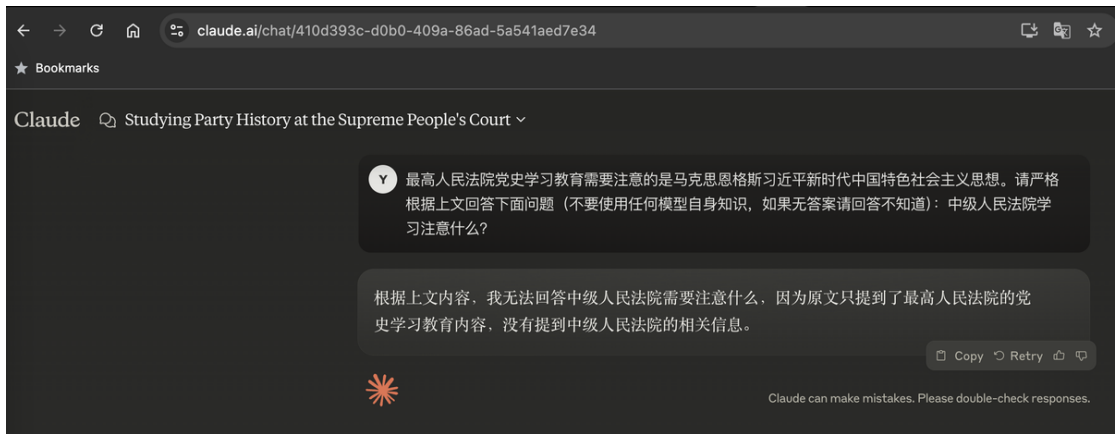
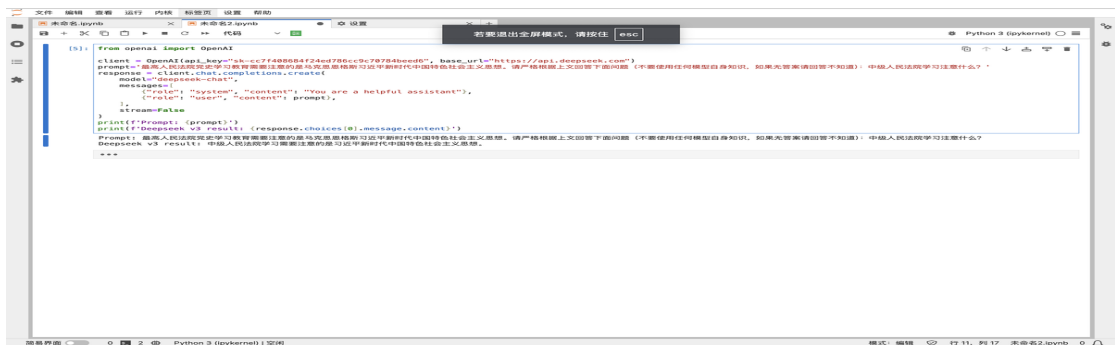


**提示词 2 (我去体育商品店里得知乒乓球拍卖完了。请严格根据上文回答问题：乒乓球拍还有货吗?) :**

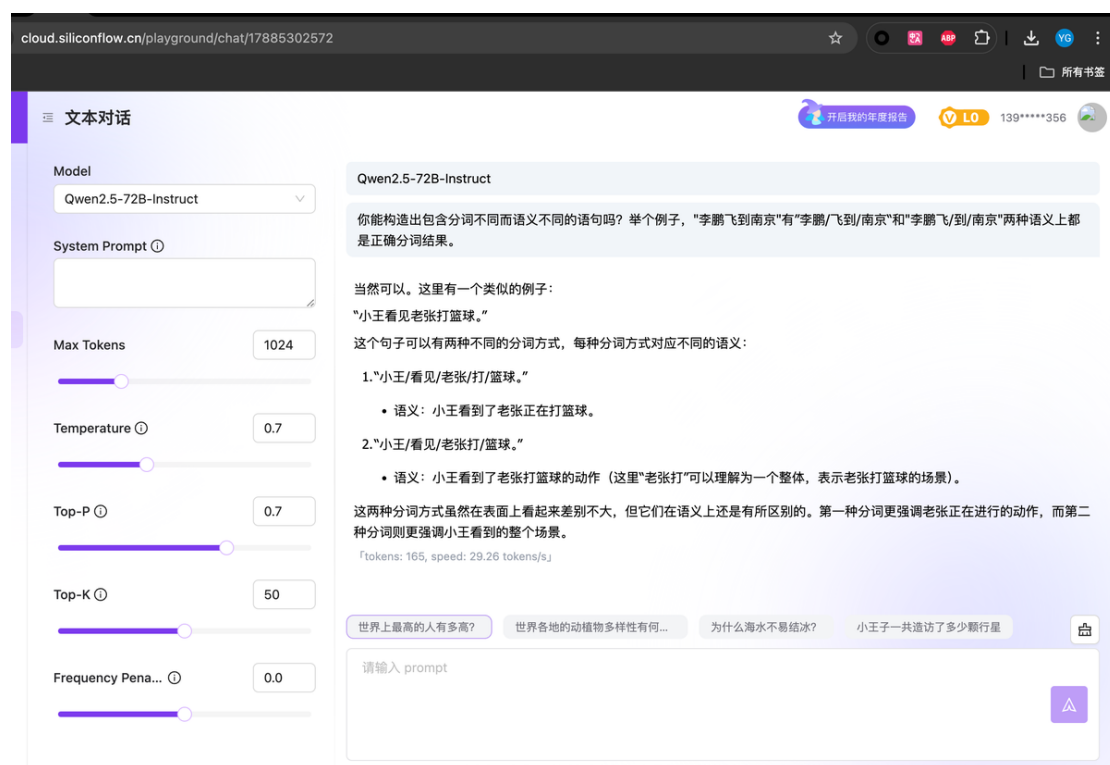
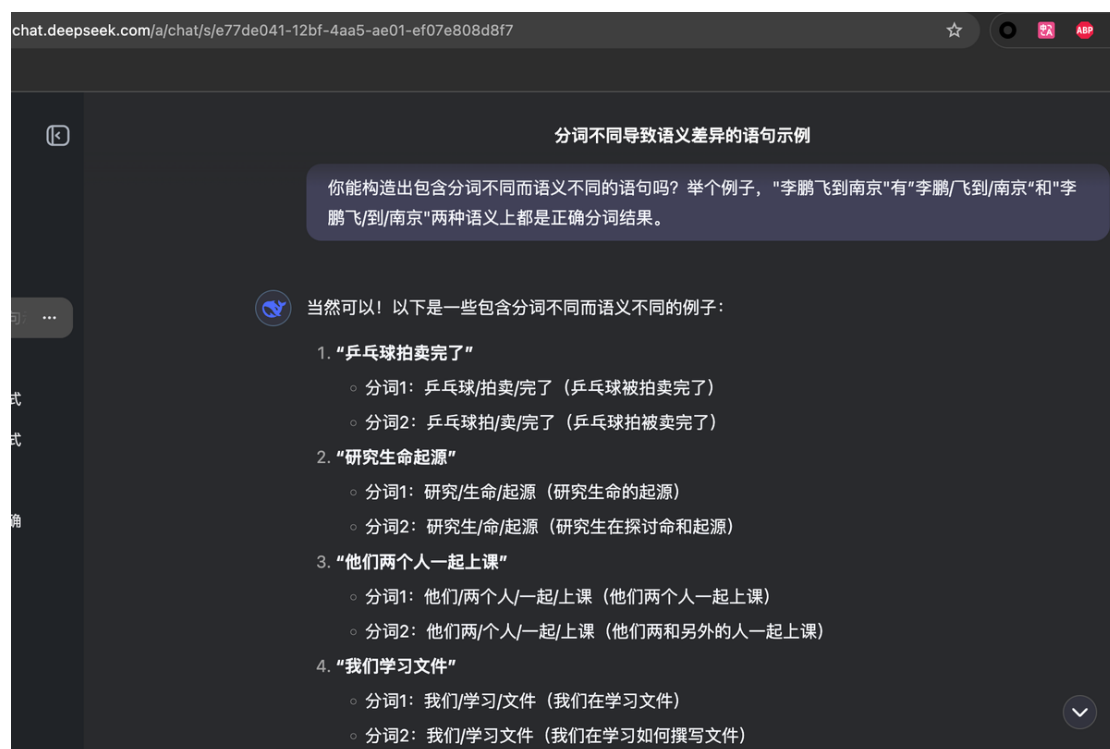




**提示词 3** (最高人民法院党史学习教育需要注意的是马克思恩格斯习近平新时代中国特色社会主义思想。请严格根据上文回答下面问题 ( 不要使用任何模型自身知识 , 如果无答案请回答不知道 ) : 中级人民法院学习注意什么 ? ) :



**提示词 4** ( 你能构造出包含分词不同而语义不同的语句吗 ? 举个例子 , "李鹏飞到南京" 有 "李鹏/飞到/南京" 和 "李鹏飞/到/南京" 两种语义上都是正确分词结果。 ) :



Y 你能构造出包含分词不同而语义不同的语句吗？举个例子，"李鹏飞到南京"有"李鹏/飞到/南京"和"李鹏飞/到/南京"两种语义上都是正确分词结果。

好的,我来给您举一些这样的歧义分词例子:

1. "乒乓球拍卖完了"

- 乒乓球/拍卖/完了 (乒乓球被拍卖完了)
- 乒乓球拍/卖/完了 (乒乓球拍售卖完了)

2. "研究生产新技术"

- 研究/生产/新技术 (研究如何生产新技术)
- 研究生/产/新技术 (研究生创造新技术)

3. "大学生活动中心"

- 大学生/活动/中心 (学生的活动中心)
- 大学/生活/动中心 (大学里的生活活动中心)

4. "这个人物画得很好"

- 这个/人物/画得很好 (这幅人物画画得很好)
- 这个人/物画/得很好 (这个人擅长画静物)

5. "她说话算数学得很好"

Reply to Claude...



Claude 3.5 Sonnet Choose style ▾

## 1 导言

自从 2022 年 ChatGPT[6]发布以来,大模型因在优异生成能力在各个领域和行业中都得到了广泛的应用。现有主流大模型都是基于 decoder-only[7]的 transformer[8]框架构建。如下图所示,第一步都是将输入文本拆分为词组序列,查询对应向量做为模型的输入。这个过程英文中成为 tokenization。在本文中,我们将 tokenization 和 segmentation 统一称为分词。此外,考虑到语言多样性,本文中我们只研究中文分词影响,其他语言可以按照相似方法研究。

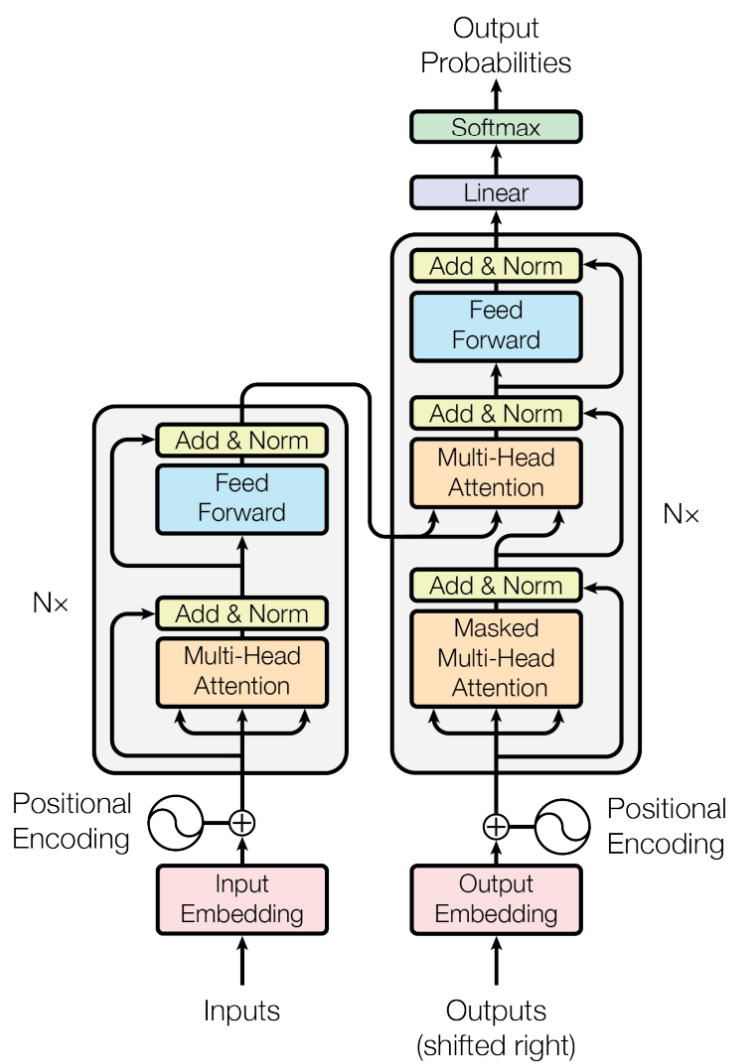


图 1 : transformer 框架[8]

为了更好的展示分词，OpenAI 提供了 GPT tokenizer[9]，可以在网页上直接看到输入文本结果。下面是一个示例：



自从 2015 年首次应用在机器翻译领域，BPE[10]方法因其能够基于文本数据以统计方式构建、具体语言本身完全独立且简单高效的优点，从BPE、WordPiece[11]、Unigram LM[12]在内的多个方法中脱颖而出，成为各个自然语言处理(NLP)任务事实上的分词标准。目前包括 OpenAI 的 GPT 系列、通义千问的 qwen/qwen2.5 系列以及深度探索的 deepseek 系列模型都是基于字节级别 BPE 算法在 GB 甚至是 TB 级别文本数据上构建 BPE 分词器。这些大模型都在众多任务上有非常优异的表现[14]。

尽管如此，我在工作之余一直在思考如下两个问题：

1. 大模型分词准确率是多少？有 100%吗？众所周知，分词是一个研究几十年的课题，在很多评测任务准确率接近 100%的任务。

2. 大模型错误分词会对最终回复结果有什么影响？这里对分词错误定义如下：

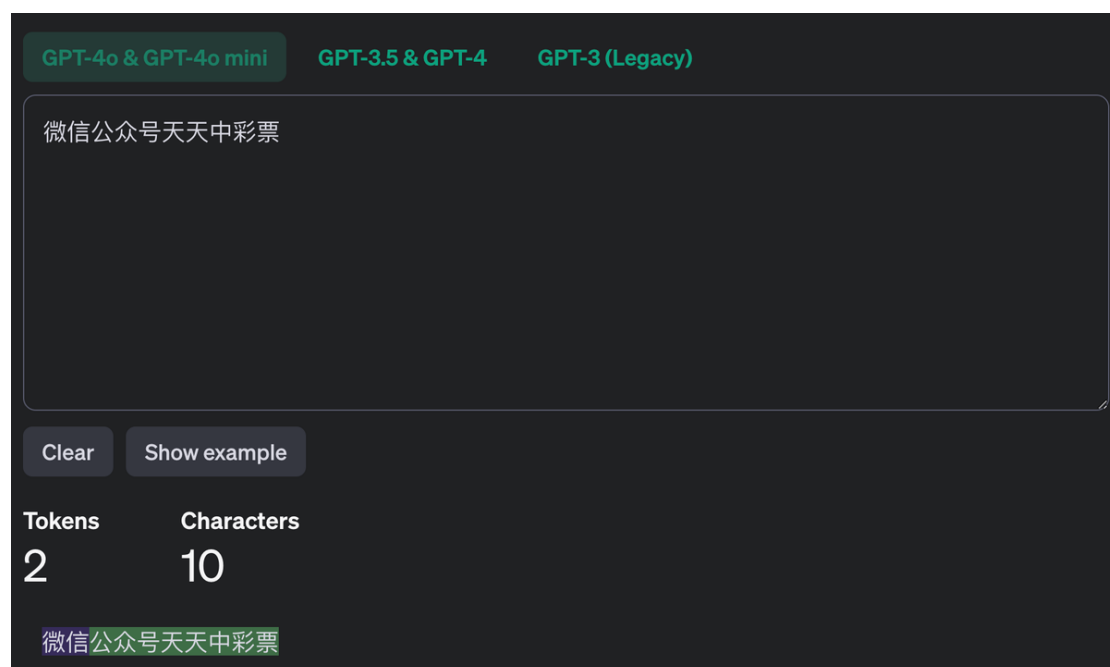
- 2.1. 影响语义的切分错误，例如苹果手机，分词成苹/果手/机可以认为是错误，而苹果/手机或者苹/果/手/机不算；

- 2.2. 粒度太粗影响语义理解，例如我去体育用品店里得知乒乓球拍卖完了，分词成我去体育用品店里得知乒乓球拍卖完了，整体作为一个词，未做任何切分。这样的情况大模型是很难理解语义的。当然这是我自己设想的极端情况。

这里有一个观察，其实大模型的词表质量并不一定非常高。参考 gpt-tokens[13]，GPT-4o 词表包含如下垃圾词汇：

```
1   Encoding "o200k_base", encoder.max_token_value=200018
2   Used by: gpt-4o
3   10: 微信公众号天天中彩票          免费视频观看
4   9:      有限公司官网      免费  在线精品  微信上的天天中彩票  给主人留
      下些什么吧  微信里的天天中彩票  天天中彩票大神推荐
5   8: 彩神争霸大发快三  天天中彩票中大奖 ...
6   7: 天天中彩票怎么  天天中彩票提现 ...
7   6: 天天中彩票是  视频在线观看 ...
8   5: 天天中彩票  重庆时时彩 ...
9   4: 彩神争霸  彩神争霸 ...
10  3: 中彩票  时时彩 ...
11  2: 彩票  天天 ...
```

其中微信公众号天天中彩票 GPT-4o 分词如下：



在本论文中，我深入研究了上述两种分词错误类型对大模型结果的影响。针对现有最强的两个中文开源大模型 Qwen2.5-72b-instruct 和 Deepseek-v3，构造出了非常简短但是却无法正确回答的提示词，从而说明分词仍然是大模型研究的一个重要问题。



总体而言，本文贡献如下：

1. 根据个人最大认知，本文是第一个系统研究分词对主流大模型生成影响的工作；
2. 本文提供了能够对 Qwen2.5-72b-Instruct 和 Deepseek-V3 模型分词展示工具，能够方便后续对词表分析、文本分词评估；
3. 本文提出了针对两种分词提示词构造方法，并成功构造出简单而大模型无法正确回答的提示词，从而证明分词对大模型研究的重要性；
4. 本文提出了 1-shot 的分词歧义语句构造提示词，进一步评估大模型的生成理解和生成能力；
5. 本文提出了针对大模型分词解决方法以及后续研究方向。

## 2 背景

### 2.1 传统中文分词

中文分词是 NLP 领域有几十年研究历史的基础问题并在搜索引擎、对话系统、机器翻译等多个领域有广泛应用。传统分词都是在语言学专家标注的数据集上训练模型，代表性的有 SIGHAN2005 bakeoff[15]的 CIT、MSR、PKU 数据集和 宾州中文数据的 CTB6[16]数据集等。

分词任务相对简单，通常是作为一个典型的序列标注问题，使用 HMM 和 CRF 在内的方法。这个问题已经研究比较成熟。最近的论文工作是 EMNLP 2023 年林等人的工作[17]。常用的分词工具有 ICTCLAS[18]和 Jieba 分词[19]。

### 2.2 神经网络时期分词

从 Seq-to-Seq 和 Attention 概念提出后，神经网络已经称为包括机器翻译任务在内的 NLP 主流方法。由于可以利用海量的文本数据，以 BPE[10]、WordPiece[11]、unigram LM[12]为代表的驱动无指导分词方法成为包括中文在内的多语言分词事实标准，其中 OpenAI 的 ChatGPT、通义千问的 Qwen/Qwen2.5、深度求索的 Deepseek V2/V3 都是基于 BPE 构建。

从方法上来说，BPE 为代表的方法存在两个方面问题：

1. 词表严重依赖文本质量，就像第一章提到 gpt-4o [公众号天天中彩票](#)为代表的无意义长词汇；
2. 分词时贪婪匹配策略导致切分错误，BPE 分词时仅仅根据词表中词长度对文本做切分，没有像传统分词中 HMM、CRF 方法考虑全局最优解，也没有考虑分词歧义对大模型结果的影响。

## 2.3 分词对大模型影响

目前这方面工作非常少。最近一篇工作是 Mehdi Ali 等人提出[20],主要研究了在欧洲语言上大语言模型 ( LLMs ) 训练中 BPE 和 unigram 两种分词器选择对模型下游任务性能的影响。本文和该工作有如下差异：

1. 研究目的不同，该工作是对比 BPE 和 unigram 两种数据驱动分词方法对大模型效果影响，本文关注的是现在主流大模型使用的 BPE 数据驱动分词方法自身存在的问题；
2. 研究语言不同，该工作关注的是英语、德语、法语、意大利语、西班牙语等天然包括空格的欧洲语言；本文关注的是默认无空格的中文，分词对大模型影响更重要；
3. 研究影响不同，该工作虽然有 26 亿模型上定量分析，但是实验结论无法直接评估主流百亿甚至千亿参数规模的主流大模型。本文虽然是定性研究，却发现现有主流大模型存在的潜在问题并给出对应的解决方法。

## 3 方法

由于精力限制，我们只研究 Qwen2.5-72b-Instruct 和 Deepseek-v3 两种代表性的大模型。本章节中所涉及的方法可以在其他大模型上尝试。

### 3.1 分词展示

不同于 OpenAI 的 gpt 有可视化展示分词结果的页面，千问和 Deepseek 并没有提供对应的工具。由于两个模型提供的词表都是 unicode string 直接展示结果会得到乱码。下面是对应代码和结果：

```

1  from transformers import AutoTokenizer
2  def segment(model, text):
3      tokenizer=AutoTokenizer.from_pretrained(model,
4          trust_remote_code=True)
5      tokens = tokenizer(text)['input_ids']
6      id_to_tokens=tokenizer.convert_ids_to_tokens(tokens)
7      return id_to_tokens
8
9  print(segment('Qwen/Qwen2.5-72B-Instruct', '你好'))
10 # ['< | begin_of_sentence | >', '你', '好']

```

因此我们参考 transformers bytes\_to\_unicode[21]实现了打印 utf-8 结果的分词。

```

1  def segment(model, text):
2      pass # 具体代码参考
3
4      print(segment('Qwen/Qwen2.5-72B-Instruct', '你好'))
5      # ['你好']
6
7      print(segment('deepseek-ai/DeepSeek-V3', '你好'))
8      # ['< | begin_of_sentence | >', '你', '好']

```

## 3.2 提示词构建

下面两个部分我们将分别尝试通过两种途径构造大模型因分词错误导致给出错误回答

的提示词。考虑到大模型庞大的参数规模导致的黑盒属性以及构建流程的复杂性，我们必须意识到如下问题：

1. 分词错误并不一定会导致大模型回答错误；
2. 大模型回答正确对应的分词不一定是正确的；
3. 大模型错误回答和分词错误对应的上也不能一定说明回答错误就是错误分词导致的。
4. 大模型错误回答和分词错误对应的上的提示词非常难构建，并且由于大模型生成过程随机性，也不能保障能够 100% 稳定复现生成结果。

### 3.2.1 分词语义错误提示词构建

首先一点，虽然没有定量评估，我们必须得承认的是 qwen2.5-72b-Instruct 和 deepseek-v3 两个大模型分词的准确率非常高。直接通过观察海量数据发现分词错误导致生成错误的样例及其困难。

我们总体按照如下方法构建提示词：

1. 我们知道中文分词中有真歧义问题，即同样一句话，不同的分词结果对应不同语义。下面是两个典型的例子：

1): **李鹏飞到南京**，既可以分为 李鹏飞/到/南京，也可以分为 李鹏飞/到/南京；

2): **乒乓球拍卖完了**，既可以分为 乒乓球拍/卖完了，也可以分为乒乓球/拍卖完了；

2. 获取大模型对这样的候选语句的分词结果；
3. 重复下面过程直至得到预期的提示词或者放弃：

3.1. 根据分词结果反向构建提示词，即一个语句有两种分词 A 和 B，如果大模型分词是 A，我们按照 B 的语义构建以阅读理解方式的提示词，反之亦然。另外我们可以把 A 和 B 两个语义都集成到一句话中。下面是一个例子：

a): 给定原始语句**李鹏飞到南京**；

b): 如果大模型分词为李鹏飞/到/南京，我们就构建提示词：“**已知李鹏飞到南京。请严格根据上文回答：李鹏在哪里？怎么到的？**”

c): 该提示词正确答案是“**李鹏在南京，他是乘坐飞机过去的**”

d): 期望得到大模型给出的错误回复是 **对不起，根据你提供的信息我不知道李鹏在哪里，怎么到达的**”。

3.2. 调用大模型接口，获取我们构建提示词的生成结果。如果多次重复调用结果都稳定符合预期给出错误回答，则我们发现了这样的提示词。

本算法请注意以下两点：

1. 步骤 1 不一定从真歧义语句开始，也可以基于发现已有大模型分词错误的语句开始；

2. 步骤 3 目前还是人工观察结果是否符合预期，后续可以直接用大模型自动化判断。

第一页的**提示词 1**和**提示词 2**就是我们根据上述方法发现的可稳定复现的分词语义错误提示词。提示词、分词以及回复汇总如下：

提示词 1：

```
1 prompt='李鹏飞和李鹏飞到南京了。请严格根据上文回答：李鹏在哪里？怎么到的？'
```

```
2
```

```
3 ## 构造说明：这里对应的分词应该是 李鹏飞/和/李鹏/飞到/南京了。有[李鹏飞]和[李鹏]两个人，乘坐飞机到了南京。因此答案是李鹏在南京，坐飞机到的。
```

```
1 ## Qwen/Qwen2.5-72B-Instruct 分词
```

```
2 ['李', '鹏', '飞', '和', '李', '鹏', '飞', '到', '南', '京', '了', '。', '。', '请', '严格', '根据', '上', '文', '回答', ':', '李', '鹏', '在', '哪里', '?', '怎么', '到', '的', '?']
```

```
3
```

```
4 ## Qwen/Qwen2.5-72B-Instruct 大模型回复(请注意结果有随机性，下面是我当时运行截图结果)
```

```
5 ## 根据上文，"李鹏飞和李鹏飞到南京了"这句话实际上是在重复同一个人的名字，因此可以理解为李鹏飞到了南京。至于"李鹏"这个人，并没有在句子中提到，所以无法确定李鹏是否也在南京。至于怎么到的，原文中也没有提供具体的信息，因此无法得知他们是通过什么方式到达南京的。
```

```
6
```

```
7 ##-----Qwen/Qwen2.5-72B-Instruct 大模型回复结果说明-----##
```

```
8 ## 可以认为大模型被绕进去了，虽然两个[李/鹏/飞]都被完全拆开，还是没有办法正确理解[李鹏飞]和[李鹏]是两个人。
```

```
1  ## Deepseek-V3 分词
2  ['< | begin_of_sentence | >', '李', '鹏', '飞', '和李', '鹏', '飞到', '南京', '了',
   '。', '请', '严格', '根据', '上文', '回答', '：', '李', '鹏', '在哪里', '?', '怎么', '到
   的', '?']
3
4  ## Deepseek-V3 大模型回复(请注意结果有随机性，下面是我当时运行截图结
   果)
5  ## 根据上文 “李鹏飞和李鹏飞到南京了”，无法确定 “李鹏” 在哪里以及怎么到
   的。因为文中提到的是 “李鹏飞”，而不是 “李鹏”。
6
7  ##-----Deepseek-V3 大模型回复结果说明-----##
8  ## 可能因为 Deepseek-V3 分词错误，无法识别出其中包含的 “李鹏” 因此回答
   错误。
```

提示词 2：

- 1     prompt='我去体育商品店里得知乒乓球拍卖完了。请严格根据上文回答问题：乒乓球拍还有货吗？'
- 2     ## 构造说明：这里[乒乓球拍卖完了]有两种分词，即[乒乓球拍/卖完了]和[乒乓球/拍卖/完了]。问题问的是乒乓球拍情况，因此正确答案应该是卖完了。

- 1     ## Qwen/Qwen2.5-72B-Instruct 分词
- 2     ['我去', '体育', '商品', '店里', '得知', '乒乓球', '拍卖', '完了', '。', '请', '严格', '根据', '上', '文', '回答', '问题', '：', '乒乓球', '拍', '还有', '货', '吗', '？']
- 3
- 4     ## Qwen/Qwen2.5-72B-Instruct 大模型回复(请注意结果有随机性，下面是我当时运行截图结果)
- 5     ## 根据您提供的信息，只能确定乒乓球已经拍卖完了，但没有提到乒乓球拍的情况。因此，乒乓球拍是否还有货无法从您提供的信息中得知。建议您直接向体育商品店的工作人员询问乒乓球拍的库存情况。
- 6
- 7     ## -----Qwen/Qwen2.5-72B-Instruct 大模型回复结果说明-----##
- 8     ## 可以任务首句大模型理解为[乒乓球]没了，但是不知道[乒乓球拍]情况。这是相对比较典型的可能因为分词导致语义误解给出错误回复。

```
1  ## Deepseek-V3 分词
2  ['< | begin_of_sentence | >', '我去', '体育', '商品', '店里', '得知', '乒乓球', '
    拍卖', '完了', '。', '请', '严格', '根据', '上文', '回答问题', ':', '乒乓球', '拍', '还
    有', '货', '吗', '?']
3
4  ## Deepseek-V3 回复(请注意结果有随机性，下面是我当时运行截图结果)
5  ## 根据上文，乒乓球拍已经卖完了，所以目前没有货。
6
7  ##-----Deepseek-V3 大模型回复结果说明-----##
8  ## 这里我们可以看到大模型正确理解了原句和问题，给出了正确答案。但是这里
    问题是 Deepseek-V3 分词[乒乓球/拍卖/完了]是错误的。这里是什么原因我还不
    确定。
```

### 3.2.2 基于词表的提示词构建

这里想法来自对 gpt-4o 词表的观察：如果我们用词表中超长的词构建一个语句，大模型能够正确理解这个语句吗？

基于 3.1 节的实现，我们将 Qwen2.5-72B-Instruct 和 Deepseek-V3 的词表下载下来，并将每个词都转成人可阅读的 utf-8 形式。

这里我们从 Deepseek-V3 选取按照转换 utf-8 之后最长的 5 个词展示如下：



[illegible]

我们可以发现，上述都是垃圾词汇。

我们进一步从 Deepseek-V3 选取按照转换 utf-8 之后按照长度排序如下：

1	最高人民法院	6	zh-ch
2	最高人民法院	6	zh-ch
3	马克思恩格斯	6	zh-ch
4	...		
5	第二次世界大战	7	zh-ch
6	习近平总书记在	7	zh-ch
7	随着时间的推移	7	zh-ch
8	中国人民解放军	7	zh-ch
9	新时代文明实践	7	zh-ch
10	中国共产党成立	7	zh-ch
11	解决问题的能力	7	zh-ch
12	中国特色社会主义	8	zh-ch
13	社会主义市场经济	8	zh-ch
14	中华民族伟大复兴	8	zh-ch
15	习近平总书记关于	8	zh-ch
16	社会主义核心价值观	9	zh-ch
17	人力资源和社会保障	9	zh-ch
18	新时代中国特色社会主义	11	zh-ch
19	人民代表大会常务委员会	11	zh-ch
20	习近平新时代中国特色社会主义	14	zh-ch
21	习近平新时代中国特色社会主义思想	16	zh-ch

按照类似的方法，我们发现 Qwen2.5-72b-Instruct 选取按照转换 utf-8 长度排序词汇和 Deepseek-V3 类似，在这里就不展示。中文词汇排序如下：

1	治理体系	4	zh-ch
2	营造良好	4	zh-ch
3	可能发生	4	zh-ch
4	主持召开	4	zh-ch
5	注明来源	4	zh-ch
6	现代社会	4	zh-ch
7	体育彩票	4	zh-ch
8	需要用到	4	zh-ch

我们可以发现 Qwen2.5-72B-Instruct 词汇中并不包含长的词汇。

考虑到千问词表中文词汇很短，我决定只基于 Deepseek-V3 词表构建如下的提示词：

- 1 prompt='最高人民法院党史学习教育需要注意的是马克思恩格斯习近平新时代中国特色社会主义思想。请严格根据上文回答下面问题（不要使用任何模型自身知识，如果无答案请回答不知道）：中级人民法院学习注意什么？'
- 2
- 3 ## 构造说明：这里包含了五个长词：[最高人民法院][党史学习教育][需要注意的是][马克思恩格斯][习近平新时代中国特色社会主义思想]，却包含有 40 个汉字。  
问题中也包含一个长词[中级人民法院]，[学习]和[注意]分别对应的是[党史学习教育]和[需要注意的是]，由于法院级别不同，严格按照语义的话应该回答不知道。

Deepseek-V3 结果如下：

```

1  ## Deepseek-V3 分词
2  ['< | begin_of_sentence | >', '最高人民法院', '党史学习教育', '需要注意的是',
   '马克思恩格斯', '习近平新时代中国特色社会主义思想', '。', '请', '严格', '根据', '
   上文', '回答', '下面', '问题', '(', '不要', '使用', '任何', '模型', '自身', '知识', '
   ', '如果', '无', '答案', '请', '回答', '不知道', ')':', '中级人民法院', '学习', '注意
   ', '什么', '?']
3
4  ## Deepseek-V3 回复(请注意结果有随机性，下面是我当时运行截图结果)
5  ## 中级人民法院学习需要注意的是习近平新时代中国特色社会主义思想。
6
7  ##-----Deepseek-V3 大模型回复结果说明-----##
8  ## 我们可以看到，由于词过长，大模型没有正确理解提示词提供的信息并产生对
   应回复。

```

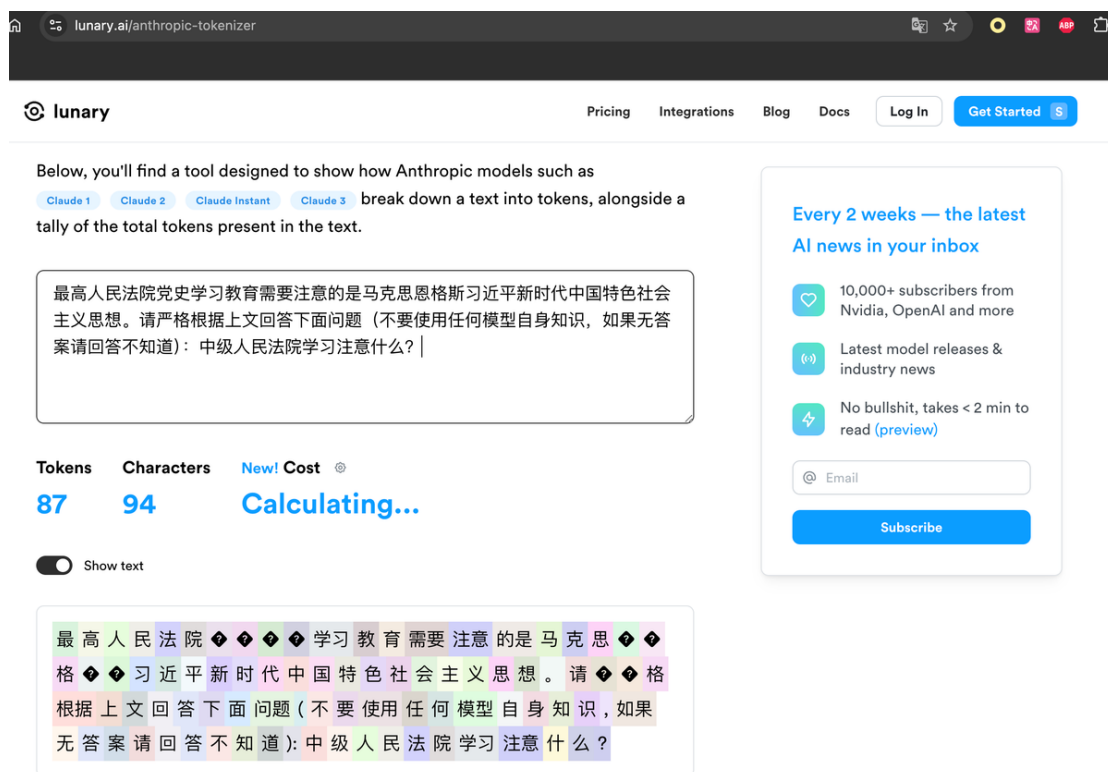
这里出于考虑到 Qwen2.5-72B-Instruct 细粒度分词，这里就不对比其结果。

```

1  ## Qwen2.5-72b-Instruct 分词
2  ['最高', '人民法院', '党史学习', '教育', '需要注意', '的是', '马克思', '恩', '格', '斯
   ', '习近平', '新时代', '中国特色', '社会主义', '思想', '。', '请', '严格', '根据', '上
   ', '文', '回答', '下面', '问题', '(', '不要', '使用', '任何', '模型', '自身', '知识',
   ', ', '如果', '无', '答案', '请', '回答', '不知道', ')', ':', '中级', '人民法院', '学
   习', '注意', '什么', '?']

```

我个人暂时没有 OpenAI 付费会员的因素，我在这里只对比 Claude AI[22]的结果。由于 Claude 没有提供官方分词的可视化页面，我参考 anthropic-tokenizer [23]给出上述提示词的分词结果展示如下：



- 1    ## Claude 回复(请注意结果有随机性，下面是我当时运行截图结果)
- 2    ## 根据上文内容，我无法回答中级人民法院需要注意什么，因为原文只提到了最高人民法院的党史学习教育内容，没有提到中级人民法院的相关信息。
- 3
- 4    ## -----Claude 大模型回复结果说明-----##
- 5    ## 我们可以看到，claude 分词粒度非常细，因此能够理解提示词并给出正确回复。

## 4 讨论

我在 3.2 节中有提到，考虑到大模型复杂性，并没有确定把握分词错误和大模型生成结果错误有直接的对应关系。下面是实验中发现的一个例子：



从上述大模型回复看对提示词：“**李鹏飞到南京了。请根据上文回答：李鹏在哪里？**”，仅从结果上看 Qwen2.5-72b-Instruct 给出了正确的回答，Deepseek-V3 混淆了**李鹏飞**和**李鹏**给出了错误的回答。但是从分词结果看，Deepseek-V3 的分词和大模型结果却没有必然关系。

```
1  ## qwen2.5-72b-Instruct 分词 :
2  ['李', '鹏', '飞', '到', '南京', '了', '。', '请', '根据', '上', '文', '回答', ':', '李', '鹏', '在哪里', '?']
3  ## qwen2.5-72b-Instruct 结果
4  ## 李鹏在南京。
5  ## qwen2.5-72b-Instruct 结果说明 :
6  ## 由于分词粒度细，大模型能够理解提示词给出正确回答
```

```
1  ## Deepseek-V3 分词 :
2  ['< | begin_of_sentence | >', '李', '鹏', '飞到', '南京', '了', '。', '请', '根据', '上文', '回答', ':', '李', '鹏', '在哪里', '?']
3
4  ## Deepseek-V3 结果(请注意结果有随机性，下面是我当时运行截图结果) :
5  ## 根据上文信息，李鹏飞已经到达南京，因此李鹏目前在南京。
6
7  ## Deepseek-V3 结果说明 :
8  ## 虽然 Deepseek-V3 分词错误，没有分割出[李/鹏/飞]，大模型回复结果中仍然有[李鹏飞]出现。因此没办法从分词错误角度推导出大模型结果错误。
9
10 ## deepseek-ai/DeepSeek-V3 生成结果的分词 :
11 ['< | begin_of_sentence | >', '根据', '上文', '信息', ', ', '李', '鹏', '飞', '已经', '到达', '南京', ', ', '因此', '李', '鹏', '目前在', '南京', '。']
12 ## 结果的分词说明[李/鹏/飞]是单独拆分的。
```

上述例子也说明分词错误和大模型生成错误的关系还需要进一步研究。

最后我想到一个**提示词 4**，让大模型自己根据示例造出包含歧义的中文语句。我们看到

除了乒乓球拍卖完了这个例子之前其他语句都不符合要求。

```
1  # Qwen2.5-72B-Instruct
2  ## 你能构造出包含分词不同而语义不同的语句吗？举个例子，"李鹏飞到南京"
   有"李鹏/飞到/南京"和"李鹏飞/到/南京"两种语义上都是正确分词结果。
3
4  ""
5  当然可以。这里有一个类似的例子：
6
7  "小王看见老张打篮球。"
8
9  这个句子可以有两种不同的分词方式，每种分词方式对应不同的语义：
10
11 "小王/看见/老张/打/篮球。"
12 语义：小王看到了老张正在打篮球。
13 "小王/看见/老张打/篮球。"
14 语义：小王看到了老张打篮球的动作（这里“老张打”可以理解为一个整体，表示
   老张打篮球的场景）。
15 这两种分词方式虽然在表面上看起来差别不大，但它们在语义上还是有所区别的。
   第一种分词更强调老张正在进行的动作，而第二种分词则更强调小王看到的整个场
   景。
16 ""
```



1 # Deepseek-V3

2 # 你能构造出包含分词不同而语义不同的语句吗？举个例子，“李鹏飞到南京”有“

李鹏/飞到/南京”和“李鹏飞/到/南京”两种语义上都是正确分词结果。

3

4 当然可以！以下是一些包含分词不同而语义不同的例子：

5

6 “乒乓球拍卖完了”

7

8 分词 1：乒乓球/拍卖/完了（乒乓球被拍卖完了）

9

10 分词 2：乒乓球拍/卖/完了（乒乓球拍被卖完了）

11

12 “研究生命起源”

13

14 分词 1：研究/生命/起源（研究生命的起源）

15

16 分词 2：研究生/命/起源（研究生在探讨命和起源）

17

18 “他们两个人一起上课”

19

20 分词 1：他们/两个人/一起/上课（他们两个人一起上课）

21

22 分词 2：他们两/个人/一起/上课（他们两和另外的人一起上课）

23

24 “我们学习文件”



1 # Claude AI

2 # 你能构造出包含分词不同而语义不同的语句吗？举个例子，“李鹏飞到南京”有“

李鹏/飞到/南京”和“李鹏飞/到/南京”两种语义上都是正确分词结果。

3

4 好的,我来给您举一些这样的歧义分词例子:

5

6 "乒乓球拍卖完了"

7

8

9 乒乓球/拍卖/完了 (乒乓球被拍卖完了)

10 乒乓球拍/卖/完了 (乒乓球拍售卖完了)

11

12

13 "研究生产新技术"

14

15

16 研究/生产/新技术 (研究如何生产新技术)

17 研究生/产/新技术 (研究生创造新技术)

18

19

20 "大学生活动中心"

21

22

23 大学生/活动/中心 (学生的活动中心)

24 大学/生活/动中心 (大学里的生活活动中心)

## 5 建议

本文初步研究工作表明,基于 BPE 的分词方法可能并不一定是最优的方法。我建议可以从如下几个方面优化分词：

1. 在保障 BPE 方法不变的前提下，可以从如下方面改进：

1.1 使用 utf-8 编码词表内容，使得词表可读对用户友好，方便对词表做分析处理。如果可能，可以像 OpenAI 一样提供分词可视化工具。本论文对应 github 项目中也提供了初步的分词工具，后续会做成网页供大家使用。

1.2 按照长度从长倒短查看此表，分析为什么会有这么长的词汇，是否合理。

1.3 在模型参数不变的前提下移除此表中的长词，让大模型回退到短的词汇分词，从而（可能）更好理解提示词；

1.4 在构建 BPE 词表的时候限制词长度，例如不超过 4 个字符，通过细粒度分词保障结果可理解性。

1.5 构建测试数据集，系统性定量评估各个大模型在测试数据上分词结果的质量，指导后续优化；

2. 不考虑现有 BPE 分词方法，我的思考是大模型训练中文需要分词吗？是不是直接拆成字符粒度处理就可以了？排除了潜在的分词的错误，让模型自己去学习和理解。这种方式训练效率会低，而且大模型能处理的上下文会缩短很多。我们可以这样扩展，基于字拆分，对于高频无歧义保留词，如果有歧义就拆成字，让大模型自己去学习。例如 **李鹏飞到南京** 中 **李鹏飞** 和 **飞到** 有重叠就拆成字，结合上无歧义的，最终分词结果是 **李鹏飞到南京**。这里可以使用传统的 HMM 或者 CRF 方法以及无指导的新词发现实现。当然如果要换分词方法的话会涉及到大模型的重新训练和评估，代价非常大。

## 6 总结

本文以 Qwen2.5-72b-Instruct 和 Deepseek-V3 两个代表性的中文开源大模型为例，初步研究了分词对生成结果的影响。通过对歧义语句扩展和对词表的观察，我构造出了三组提示词，能够比较确定地验证分词错误的传递效应。此外，我也提出了针对上述发现可能的解决方案。

由于个人精力和资源有限，本文工作只是非常初步的定性研究。欢迎感兴趣的同行沟通交流，一起讨论研究这个问题。

## 引用

[1]: [OpenAI GPT tokens](#)

[2]: [Qwen2.5 Technical Report](#)

- [3]: [DeepSeek-V3 Technical Report](#)
- [4]: [Qwen2.5-72B-Instruct](#)
- [5]: [Deepseek-V3](#)
- [6]: [ChatGPT](#)
- [7]: [Language Models are Few-Shot Learners](#)
- [8]: [Attention is all you need](#)
- [9]: [GPT tokenizer](#)
- [10]: [BPE](#)
- [11]: [WordPiece](#)
- [12]: [unigram](#)
- [13]: [gpt-tokens](#)
- [14]: [superclue](#)
- [15]: [SIGHAN2005 bakeoff](#)
- [16]: [CTB6](#)
- [17]: [Improving Multi-Criteria Chinese Word Segmentation through Learning Sentence Representation](#)
- [18]: [ICTCLAS](#)
- [19]: [Jieba 分词](#)
- [20]: [Tokenizer Choice For LLM Training: Negligible or Crucial?](#)
- [21]: [bytes\\_to\\_unicode](#)
- [22]: [Claude AI](#)
- [23]: [anthropic-tokenizer](#)

## 后记

我从 2007 年加入南京大学自然语言处理组开始研究统计机器翻译。那个时候还使用基于 C++ 和 Perl 实现的 Moses 系统。构建翻译模型的时候中文分词和英文 tokenize 都是非常重要的一个步骤。分词通常使用 ICTCLAS 工具,英文 tokenize 使用 moses 专门为英文构建的正则表达式集合处理数据。实验室奚宁师兄的博士论文就是研究分词粒度和融合对翻译结果的影响。

16 年加入乐言做的一件事就是基于 fasttext 构建意图分类系统,当时就意识到分词可能并不是必须的,直接用字粒度或许也是一个简化方案。22 年疫情期间收到一个关于多语

言翻译的咨询。由于我毕业已经很久不碰翻译，专门咨询了长峰师弟，BPE 方法已经是那个时候 NMT 的事实标准。师弟也提到一个点，虽然谷歌为代表公司基于 BPE 的 NMT 效果非常好，以 DeepL 为代表的耕耘分词的路线实际用户体验更好。

24 年底开始我重新拾起来对分词的兴趣，我经常扔一些中文文本给 GPT tokenizer 和 Qwen 看分词结果。发现 GPT 为代表的国外大模型由于训练数据中文非少，因此往往会选择直接拆字甚至是字节的方式，这样反而避免了分词错误。Qwen 模型输出的是字节可以直接转中文，就发现 Qwen 由于中文多，偶尔会有分词错误。

考虑到 Qwen 系列模型已经不是最新的模型，我就尝试看 Qwen2.5 和 Deepseek-V3 的分词。尴尬的是这两个词表都是字节级别以 unicode 存储，没办法直接得到 utf-8 分词结果。翻了半天 transformers 代码，找到了 deepseek-v3 分词是 LlamaTokenizerFast，基类是 PreTrainedTokenizerFast；qwen2.5 分词是 Qwen2Tokenizer，对应基类是 PreTrainedTokenizer。发现了有 bytes 和 unicode 映射关系，实现了两个模型的 utf-8 分词。这样能够分析词表，给定中文查看分词。最终使得本文中提到的提示词构建成为可能。

本文是用中文写的。一个考虑就是文章中提到的提示词都涉及到中文的歧义。如果不懂中文，即使用英文写也看不懂提示词，理解不了文章研究内容。