

Example: 医生：你好，有哪儿里不舒服的吗？患者：肚子上面那里有点痛。就很闷。医生：这个有多少年了呢患者：有七年了。最近三个月痛的更厉害了，次数还多了。医生：一般是什么时候痛呢，吃饭前还是吃饭后患者：吃饭前。医生：吃完饭以后有没有好一点患者：好很多了。但这几个月吃完东西还是感觉很痛。医生：最近皮肤颜色有变化吗，会不会有想吐的情况患者：没有。医生：大便正常吗，有没有黑便什么的患者：有时候有一点黑。医生：以前有没有吃过什么药患者：没有。医生：之前有做过胃镜什么的吗患者：没有，没做过什么检查。医生：最近吃的怎么样，体重有没有轻了患者：吃的挺好的，没有轻。医生：以前还有其他的病吗，糖尿病高血压这些有吗，冠心病有吗患者：有高血脂 3 年了，别的都没有。

Generate 函数

需要你协助完成病历信息抽取任务，当我给你一个病人自述时，帮我抽取出句子中的主诉(即患者自我表达自身症状及其持续时间、性质或程度等，可为医生提供诊断疾病的方向。不超过 20 字)，现病史(现病史是主要部分，包括询问患者的症状发作日期以及症状、原因等，并且要询问患者的一般情况，如食欲、体重、睡眠等)，既往病史(即询问患者过去曾患有过什么疾病，以及既往有无外伤、手术等各种情况，尤其是与患者就诊时相关的病史要仔细询问)，并按照 JSON 的格式输出，如果缺少信息用‘无’来表示，多个值之间用‘;’分隔。请根据以上对话，生成相应的 JSON 格式病历信息。

```
```json
{
 "patient_age": 45,
 "main_complaint": "肚子上方疼痛",
 "duration": "7 years",
 "severity": "加重，频率增加",
 "location": "上腹部",
 "frequency": "饭前",
 "relief": "饭后好转",
 "color_change": "无",
 "vomiting_sensation": "无",
 "black_stools": "有时",
 "previous_medications": "无",
 "previous_examinations": "无",
 "recent_饮食": "良好，未减轻体重",
 "previous_conditions": "高血脂 3 年",
 "other_systemic_conditions": "无",
 "work_pressure": "大"
}
```
```

出现问题：太多的种类

修改后的 prompt:

"需要你协助完成病历信息抽取任务，当我给你一个病人自述时，帮我抽取出句子中的主诉(即患者自我表达自身症状及其持续时间、性质或程度等，可为医生提供诊断疾病的方向。

不超过 20 字), 现病史(现病史是主要部分, 包括询问患者的症状发作日期以及症状、原因等, 并且要询问患者的一般情况, 如食欲、体重、睡眠等), 既往病史(即询问患者过去曾患有过什么疾病, 以及既往有无外伤、手术等各种情况, 尤其是与患者就诊时相关的病史要仔细询问), 并按照 JSON 的格式输出, 一共有三个 key, 注意不能使用原句, 如果缺少信息用‘无’来表示, 多个值之间用','分隔。"



```
Run qwen x
D:\python爬虫\pythonProject\.venv\Scripts\python.exe D:\python爬虫\pythonProject\.venv\qwen.py
Loading checkpoint shards: 100%|██████████| 8/8 [00:00<00:00, 13.56it/s]
The attention mask is not set and cannot be inferred from input because pad token is same as eos token.
Both 'max_new_tokens' (=512) and 'max_length' (=512) seem to have been set. 'max_new_tokens' will take p
原始句子: 医生: 你好, 有哪里不舒服的吗? 患者: 肚子上面那里有点痛。就很闷。医生: 这个有多少年了呢患者: 有七年了。最近三个月痛的
格式化结果: 需要你协助完成病历信息抽取任务, 当我给你一个病人自述时, 帮我抽取出句子中的主诉(即患者自我表达自身症状及其持续时间、
```json
{
 "主诉": "肚子上面疼痛",
 "现病史": "有七年了。最近三个月痛的更厉害了, 次数还多了。饭前痛, 吃完饭后好很多。这几个月吃完东西还是感觉很痛。",
 "既往病史": "无"
}
```

### Chat 函数

"需要你协助完成病历信息抽取任务, 当我给你一个病人自述时, 帮我抽取出句子中的主诉(即患者自我表达自身症状及其持续时间、性质或程度等, 可为医生提供诊断疾病的方向。不超过 20 字), 现病史(现病史是主要部分, 包括询问患者的症状发作日期以及症状、原因等, 并且要询问患者的一般情况, 如食欲、体重、睡眠等), 既往病史(即询问患者过去曾患有过什么疾病, 以及既往有无外伤、手术等各种情况, 尤其是与患者就诊时相关的病史要仔细询问), 并按照 JSON 的格式输出, 如果缺少信息用‘无’来表示, 多个值之间用','分隔。"



```
D:\python爬虫\pythonProject\.venv\Scripts\python.exe D:\python爬虫\pythonProject\.venv\qwen.py
Loading checkpoint shards: 100%|██████████| 8/8 [00:00<00:00, 12.82it/s]
[The attention mask is not set and cannot be inferred from input because pad token is same as eos token.As a consequence, you may observe unexpected behavior. Please pass v
2024-08-14 13:33:15.570998: I tensorflow/core/util/port.cc:153] oneDNN custom operations are on. You may see slightly different numerical results due to floating-point rou
2024-08-14 13:33:20.045645: I tensorflow/core/util/port.cc:153] oneDNN custom operations are on. You may see slightly different numerical results due to floating-point rou
医生: 你好, 有哪里不舒服的吗? 患者: 肚子上面那里有点痛。就很闷。医生: 这个有多少年了呢患者: 有七年了。最近三个月痛的更厉害了, 次数还多了。医生: 一般是什么时候痛呢, 吃饭前还是吃饭后患者, 吃饭前。医生, 吃完饭后有
{
 "主诉": "肚子上面那里有点痛。就很闷",
 "现病史": "有七年了。最近三个月痛的更厉害了, 次数还多了。一般是什么时候痛呢, 吃饭前还是吃饭后, 吃完饭后有没有好一点。但这几个月吃完东西还是感觉很痛。",
 "既往病史": "有高血压3年了, 别的都没有。"
}
Process finished with exit code -1
```

### 出现问题: 使用原话

#### 修改后的 prompt:

需要你协助完成病历信息抽取任务, 当我给你一个病人自述时, 帮我总结出句子中的主诉(即患者自我表达自身症状及其持续时间、性质或程度等, 可为医生提供诊断疾病的方向。不超过 20 字), 现病史(现病史是主要部分, 包括询问患者的症状发作日期以及症状、原因等, 并且要询问患者的一般情况, 如食欲、体重、睡眠等), 既往病史(即询问患者过去曾患有过什么疾病, 以及既往有无外伤、手术等各种情况, 尤其是与患者就诊时相关的病史要仔细询问), 注意不能使用原句, 并按照 JSON 的格式输出, 如果缺少信息用‘无’来表示, 多个值之间用','分隔。

```
D:\python爬虫\pythonProject\.venv\Scripts\python.exe D:\python爬虫\pythonProject\.venv\qwen.py
Loading checkpoint shards: 100%|██████████| 8/8 [00:00<00:00, 17.32it/s]
The attention mask is not set and cannot be inferred from input because pad token is same as eos token.As a consequence, you may ob
2024-08-15 10:29:35.237547: I tensorflow/core/util/port.cc:153] oneDNN custom operations are on. You may see slightly different num
2024-08-15 10:29:39.943271: I tensorflow/core/util/port.cc:153] oneDNN custom operations are on. You may see slightly different num
医生：你好，有哪儿里不舒服的吗？患者：肚子上面那里有点痛。就很闷。医生：这个有多少年了呢患者：有七年了。最近三个月痛的更厉害了，次数还多了。医生：一般是什么时候痛呢
```json
[
  {
    "类型": "主诉",
    "内容": "肚子上面那里有点痛。就很闷。"
  },
  {
    "类型": "现病史",
    "内容": "有七年了。最近三个月痛的更厉害了，次数还多了。一般是什么时候痛呢，吃饭前还是吃饭后？吃完饭以后有没有好一点？但这几个月吃完东西还是感觉很痛。"
  },
  {
    "类型": "既往病史",
    "内容": "有高血脂3年了。别的都没有。"
  }
]
```

结论：generate 函数的输出效果更好（总结效果好+理解 prompt 能力强）

参考资料: <https://kimi.moonshot.cn/share/cquqb81mqu0a74b98avg>
https://blog.csdn.net/m0_51300331/article/details/139305726



 Κυνοδοντα

大模型的generate函数和chat函数有什么区别

 Kimi

在大型语言模型（LLM）中，generate 函数和 chat 函数是两种不同的文本生成方法，它们各自有不同的特点和使用场景。

generate 函数是模型的原生方法，主要用于生成文本。它通常用于批量生成文本数据，可以根据特定的输入和条件生成一组文本。使用 generate 时，可以传递多种参数，例如 max_length（生成文本的最大长度）、num_beams（束搜索的数量，用于增强生成的多样性）等，来控制生成过程。generate 方法的输出结果是生成的 token ID 序列，这些 token ID 可以用相应的 tokenizer 解码回文本。

chat 函数是一个更高级的便捷方法，通常用于模拟对话。它提供了一个更简单的用户交互方式，以模拟对话流程，尤其在聊天应用中更为方便。chat 函数内部调用了 generate 方法，但提供了一个更简化的输入输出接口。使用 chat 时，主要参数包括 tokenizer、query（用户输入的待查询内容）以及 history（对话历史）。

总的来说，generate 方法更加灵活，适用于更多的生成任务，而 chat 方法则提供了更高级别、更易于使用的接口，特别适用于聊天式应用中。开发者可以根据具体的需求和使用场景选择使用哪个方法。

 Κυνοδοντα

在使用大模型根据问诊文本生成门诊档案时，你推荐generate还是chat

 Kimi

在大模型生成门诊档案的应用中，选择generate函数还是chat函数取决于具体的应用场景和需求。generate函数提供了更多的灵活性和生成策略，适用于批量生成文本数据和多样化的生成任务，可以根据特定的输入和条件生成一组文本，适合需要对生成文本有更多的控制和定制化的场景。而chat函数则提供了更简单易用的接口，专门用于模拟对话流程，适用于聊天式应用中，特别是当需要模拟问诊对话并生成门诊档案时。

门诊病历生成系统，例如云知声的系统，通过结合大模型技术、智能语音识别和医疗知识图谱，能够实现门诊病历的自动生成。这种系统可以在医患对话过程中自动进行语音采集、语音识别、角色分离、信息摘要和文书撰写，从而显著提升医生的工作效率和患者的就诊体验。

综合来看，如果目标是生成结构化、标准化的门诊档案，且需要在问诊过程中实时生成，chat函数可能更加合适，因为它可以更好地模拟真实的医患对话并生成连贯的文本。如果需要对生成的文本进行更多的定制化处理，或者应用在非对话式的文本生成任务中，generate函数将提供更大的灵活性。