

通用大模型原理及训练实践

实验课④：性能测试



性能测试

- 模型解码策略
- 通用任务性能测试
- 特定任务性能测试
- 自我认知任务性能测试
- 解码脚本

模型解码策略

- 自回归，从左到右逐步解码
- 每一步解码的概率分布为 $P(y_i|y_{1:i-1}, \boxed{X})$ instruction+input
- 常用的解码策略
 - 束搜索（beam search）
 - 每一步仅考虑概率最大的几个候选词，目标是搜索概率最大的句子
 - 采样（sampling）
 - 考虑所有可能的候选词，从概率分布中随机采样
 - 两个关键参数：temperature、top-p
- 根据不同任务特点选择不同的解码策略及参数

模型解码参数：temperature与top-p

■ temperature：温度系数

- 温度越小，采样分布越尖锐，模型输出越确定
- 温度越大，采样分布越平滑，模型输出越随机
- 温度趋近于0时，近似于贪婪搜索
- 温度趋近于无穷时，近似于从均匀分布中随机采样

■ top-p (与top-k相区分)

- 根据概率从高到低在词表上选择累积概率恰好超过p的候选词作为采样集合
- 从中进行随机采样

HuggingFace中设置不同的解码策略

- https://huggingface.co/docs/transformers/main_classes/text_generation#transformers.GenerationConfig
- 束搜索 (beam search)
 - num_beams=4
- 采样 (sampling)
 - do_sample=True
 - temperature=1.0
 - top_p=0.7
- 最大生成长度: max_new_tokens=1024, 防止输出太长被截断

通用任务性能测试

- 通用任务的测试集通常只包含instruction，无明确的参考答案
- 将instruction输入模型，得到模型输出
- Vicuna项目提供的80条测试指令，中英文2个版本
 - vicuna_80_en.jsonl和vicuna_80_zh.jsonl稍后发布在微信群中

```
{
  "instruction": "How can I improve my time management skills?",
  "input": "",
  "output": ""
},
{
  "instruction": "What are the most effective ways to deal with stress?",
  "input": "",
  "output": ""
},
```

```
{
  "instruction": "我该如何提高我的时间管理技能？",
  "input": "",
  "output": ""
},
{
  "instruction": "如何有效地应对压力？",
  "input": "",
  "output": ""
},
```

通用任务性能测试

■ 得到模型的输出后，如何进行评价？

□ 人工评价

- 比较两个模型输出的好坏，最后统计胜率

□ GPT-4评价

- 输入问题(Q)、两个模型的回答(R1, R2)
- 让GPT-4综合考虑多方面打分，给出解释
- 根据每个问题的打分，统计总分/胜率

[Question]

{Q}

[The Start of Assistant 1's Answer]

{R1}

[The End of Assistant 1's Answer]

[The Start of Assistant 2's Answer]

{R2}

[The End of Assistant 2's Answer]

[System]

We would like to request your feedback on the performance of two AI assistants in response to the user question displayed above.

Please rate the helpfulness, relevance, accuracy, level of details of their responses. Each assistant receives an overall score on a scale of 1 to 10, where a higher score indicates better overall performance.

Please first output a single line containing only two values indicating the scores for Assistant 1 and 2, respectively.

The two scores are separated by a space. In the subsequent line, please provide a comprehensive explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment.

通用任务性能测试-举例

Q

[Question]

植物蛋白和动物蛋白来源之间有什么区别？

[The Start of Assistant 1's Answer]

植物蛋白和动物蛋白在原料来源、成分组成、营养价值等方面有明显的区别。植物蛋白主要来自于农产品，如谷物、豆类、豆腐等，它们是天然蛋白质的来源。而动物蛋白则主要来自于动物肉类和蛋类，它们是人类直接消耗的蛋白质来源。植物蛋白的营养价值较高，富含维生素、矿物质和膳食纤维，对人体健康有益处。相比之下，动物蛋白的营养价值较低，但是它们富含蛋白质和某些维生素，对人体健康也有益处。在膳食中，应充分考虑营养价值，根据个人需要选择合适的蛋白质来源。

[The End of Assistant 1's Answer]

[The Start of Assistant 2's Answer]

植物蛋白和动物蛋白之间的主要区别在于其来源和构成。

植物蛋白是由植物组织，如豆类，谷物，坚果和种子中的蛋白质组成的。它们富含氨基酸，对人体健康有益，但在摄入时应量控制。

动物蛋白是由动物组织中的蛋白质组成的，如肉类，鸡蛋，奶制品和海鲜。动物蛋白一般富含铁和磷，对人体健康也有益，但应适量摄入，避免过量摄入会导致炎症等问题。

两者的构成不同，植物蛋白的氨基酸分布不同，动物蛋白的氨基酸分布也不同，因此对人体对氨基酸的利用也不同，对人体健康的影响也不同。

[The End of Assistant 2's Answer]

[System]

We would like to request your feedback on the performance of two AI assistants in response to the user question displayed above.

Please rate the helpfulness, relevance, accuracy, level of details of their responses. Each assistant receives an overall score on a scale of 1 to 10, where a higher score indicates better overall performance.

Please first output a single line containing only two values indicating the scores for Assistant 1 and 2, respectively.

The two scores are separated by a space. In the subsequent line, please provide a comprehensive explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment.



98

Assistant 1: 9

Assistant 2: 8

Assistant 1 provided a highly informative and comprehensive response. It explained the differences between plant protein and animal protein in terms of their sources, composition, and nutritional value. The answer highlighted the higher nutritional value of plant protein, including its rich content of vitamins, minerals, and dietary fiber. It also mentioned the lower nutritional value of animal protein but acknowledged its protein content and certain vitamins. The response was well-structured and covered the main aspects of the topic.

Assistant 2 also provided a good response but lacked some of the details provided by Assistant 1. It briefly mentioned the sources of plant and animal protein and their respective benefits to human health. The response noted that plant protein is rich in amino acids and beneficial to health, but should be consumed in moderation. It also mentioned that animal protein is generally rich in iron and phosphorus and can be beneficial but should be consumed in moderation to avoid potential issues. While the answer was accurate, it lacked the same level of detail and comprehensive coverage as Assistant 1.

Overall, both assistants provided satisfactory answers, but Assistant 1 demonstrated a higher level of detail, clarity, and comprehensiveness in its response, leading to a slightly higher score.

特定任务性能测试

- 常见NLP任务的测试集形式通常为<input, output>
- 根据任务特点，将其组织成<instruction, input, output>的形式
- 根据模型输出与参考答案计算指标（Accuracy、BLEU、ROUGE等）

```
{
  "instruction": "Translate the following sentences from Chinese to English.",
  "input": "有骨和辣泡菜不适合小朋友吃",
  "output": "There are bones and kimchi in it and it's not suitable for children."
},
{
  "instruction": "Translate the following sentences from Chinese to English.",
  "input": "如果是，请尽快 order",
  "output": "If so, please place an order soon."
},
{
  "instruction": "Translate the following sentences from Chinese to English.",
  "input": "暂时没有，谢谢",
  "output": "We don't have it at the moment, thank you."
},
{
  "instruction": "Translate the following sentences from Chinese to English.",
  "input": "他叫我跟你们说他所在的位置#PRS_ORG#",
  "output": "He asked me to tell you his location #PRS_ORG#."
},
}
```

自我认知任务性能测试

- 构造若干条不同的instruction用于自我认知测试。
- 自我认知任务有标准答案（模型的名字），对于模型给出的回答，检测是否包含正确的名字，计算准确率。
- 自我认知任务同样可以评估整体回复的内容，人工评估或者用模型评估（参考通用任务评估）。

解码脚本

- generate_file.py, 供参考, 稍后发布在微信群
- 指定基座模型路径、LoRA路径、prompt模板、测试集路径、输出路径
- `python generate_file.py \`
 - `--base_model ~/bayling-2-7b/ \`
 - `--lora_weights /path/to/lora/weights \`
 - `--prompt_template bayling \`
 - `--test_data vicuna-80-en.json \`
 - `--output_path /path/to/output`

课后任务（7月10日23:59前提交）

■ 通用任务性能测试

- 选取20条通用指令（可从Vicuna-80中选取，或自行选择/构造其他指令），分别获得百聆2-7B的输出，以及自己小组模型的输出，采取人工打分的方式进行评价，统计胜率。（注：胜率高低不影响本课程得分）
- 将结果组织成json格式，每个条目包含<instruction, output1, output2, win>，分别为指令、模型1的输出、模型2的输出、win=1/2，表示模型1表现更好或模型2表现更好。

课后任务（7月10日23:59前提交）

■ 通用任务性能测试

```
[
{
  "instruction": "植物蛋白和动物蛋白来源之间有什么区别？",
  "output1": "植物蛋白和动物蛋白在原料来源、成分组成、营养价值等方面有明显的区别。植物蛋白主要来自于农产品，如谷物、豆类、豆腐等，它们是天然蛋白质的来源。而动物蛋白则主要来自于动物肉类和蛋类，它们是人类直接消耗的蛋白质来源。植物蛋白的营养价值较高，富含维生素、矿物质和膳食纤维，对人体健康有益处。相比之下，动物蛋白的营养价值较低，但是它们富含蛋白质和某些维生素，对人体健康也有益处。在膳食中，应充分考虑营养价值，根据个人需要选择合适的蛋白质来源。",
  "output2": "植物蛋白和动物蛋白之间的主要区别在于其来源和构成。植物蛋白是由植物组织，如豆类，谷物，坚果和种子中的蛋白质组成的。它们富含氨基酸，对人体健康有益，但在摄入时应量控制。动物蛋白是由动物组织中的蛋白质组成的，如肉类，鸡蛋，奶制品和海鲜。动物蛋白一般富含铁和磷，对人体健康也有益，但应适量摄入，避免过量摄入会导致炎症等问题。两者的构成不同，植物蛋白的氨基酸分布不同，动物蛋白的氨基酸分布也不同，因此对人体对氨基酸的利用也不同，对人体健康的影响也不同。",
  "win": 1
},
{
  "instruction": "描述一种场景，人工智能可以用来提高医疗服务的质量和效率。",
  "output1": "在未来的医疗场景中，人工智能（AI）被广泛应用，以提高医疗服务的质量和效率。以下是一个描述这种场景的例子：\n\n在一家现代化的医疗中心，AI技术被应用于各个环节，从患者接待到诊断和治疗。当患者进入医疗中心时，他们被引导到一个智能自助注册系统。这个系统使用自然语言处理和人脸识别技术，能够理解患者的需求，并准确获取他们的个人和医疗信息。接下来，AI可以分析患者的病历、症状和生理数据，以辅助医生进行初步诊断。AI系统能够比人类更快速地从大量的医疗数据库中提取相关知识，并根据患者的情况生成可能的诊断结果。这样，医生可以更准确地评估患者的状况，缩短诊断时间，并制定更有效的治疗计划。在治疗过程中，AI还可以提供实时监测和建议。通过与患者佩戴的传感器设备进行连接，AI能够收集和分析患者的生理参数，如心率、血压和血氧饱和度。如果出现异常情况，AI会立即向医生发送警报，并提供相应的建议和指导。此外，AI还在药物研发和个性化治疗方面发挥重要作用。基于大数据和机器学习算法，AI可以分析大量的生物医学数据，识别潜在的药物靶点和疾病相关基因。这种精确的预测和筛选有助于加速新药研发的过程，为患者提供更好的治疗选择。总的来说，AI的应用使医疗服务更加智能化和高效化。它可以帮助医生做出更准确的诊断，提供实时监测和个性化治疗，加速药物研发进程，并改善患者的医疗体验。这种场景下，医疗服务变得更加精细和人性化，为人们的健康提供更好的保障。",
  "output2": "在不久的将来，人工智能将在医疗服务中发挥越来越重要的作用，以提高质量和效率。下面是一种场景，展示了人工智能在医疗服务中的应用：\n\n假设某地区的医院引入了一个先进的医疗人工智能系统。这个系统使用了自然语言处理和机器学习技术，可以分析海量的医学文献、临床试验数据、病历记录和患者反馈等信息。一天，一名50岁的男性患者来到医院寻求诊断和治疗。他感到胸闷、气短，并且有轻度胸痛。他被导诊到医院的急诊科。在过去，医生可能会依靠自己的经验和知识来进行诊断，但现在有了这个医疗人工智能系统，医生可以更好地利用它来提高诊断的准确性和速度。医生与患者进行了初步询问和体检，然后将相关信息输入到人工智能系统中。系统立即开始分析该患者的症状、病史和实验室结果，并与庞大的医学数据库进行比对。人工智能系统的深度学习算法能够识别潜在的疾病模式和风险因素，并生成有关可能诊断的建议。在这种情况下，系统可以迅速排除一些常见的疾病，如胃酸倒流，而把注意力集中在潜在的严重疾病上。医生得到系统的建议后，结合自己的临床判断，可以更准确地诊断患者的病情。医生还可以利用系统提供的最新研究和治疗指南，制定最佳的治疗方案。同时，这个人工智能系统还具有监测和预测的能力。它可以根据大量的实时数据，监测患者的健康状况和治疗效果，并及时提醒医生进行调整。此外，系统还可以预测患者的住院风险和并发症发展趋势，帮助医生采取相应的预防措施。总的来说，这个医疗人工智能系统可以提高医疗服务的质量和效率。它能够辅助医生进行准确的诊断，提供最新的治疗指南，并监测患者的健康状况。这种系统的引入有助于减少错误诊断和治疗延误，提高患者的治疗效果和满意度。同时，它还可以帮助医院更好地利用医疗资源，提高整体效率和生产力。",
  "win": 2
}
]
```

课后任务（7月10日23:59前提交）

■ 特定任务性能测试

- 根据自己选择的NLP任务，找一个测试集，最好是公开的基准测试集
 - 如没有公开测试集，可自行构建测试集
- 提交若干测试用例，用于说明任务内容
- 测试百聆2-7B和小组模型在整个测试集上的自动评价指标，进行比较
- 如没有合适的自动评价指标，可在整个测试集上采用人工评价或GPT打分（此时测试集可以不用过大，几十条到几百条为宜）

课后任务（7月10日23:59前提交）

■ 自我认知任务性能测试

- 选取10条自我认知指令，统计模型回复的准确率
- 将结果组织成json格式，每个条目包含<instruction, output, T/F>，其中T/F为是否包含正确的名字

谢谢！