第十章 评估（下）——不存在简单的正确答案

在上一章中，我们探索了如何评估 LLM 模型在 **有明确正确答案** 的情况下的性能，并且我们学会了编写一个函数来验证 LLM 是否正确地进行了分类列出产品。

然而，如果我们想要使用 LLM 来生成文本，而不仅仅是用于解决分类问题，我们又应该如何评估其回答准确率呢？在本章，我们将讨论如何评估LLM在这种应用场景中的输出的质量。

一、运行问答系统获得一个复杂回答

我们首先运行在之前章节搭建的问答系统来获得一个复杂的、不存在一个简单正确答案的回答：

```
import utils_zh

'''
注意：限于模型对中文理解能力较弱，中文 Prompt 可能会随机出现不成功，可以多次运行；也非常欢迎同学探究更稳定的中文 Prompt
'''
# 用户消息
customer_msg = f"""
告诉我有关 the smartx pro phone 和 the fotosnap camera, the dslr one 的信息。
另外，你们这有什么 TVs ？"""

# 从问题中抽取商品名
products_by_category = utils_zh.get_products_from_query(customer_msg)
# 将商品名转化为列表
category_and_product_list = utils_zh.read_string_to_list(products_by_category)
# 查找商品对应的信息
product_info = utils_zh.get_mentioned_product_info(category_and_product_list)
# 由信息生成回答
assistant_answer = utils_zh.answer_user_msg(user_msg=customer_msg,
product_info=product_info)

print(assistant_answer)
```

```
关于SmartX Pro手机和FotoSnap DSLR相机的信息：

1. SmartX Pro手机（型号：SX-PP10）是一款功能强大的智能手机，拥有6.1英寸显示屏、128GB存储空间、12MP双摄像头和5G网络支持。价格为899.99美元，保修期为1年。

2. FotoSnap DSLR相机（型号：FS-DSLR200）是一款多功能的单反相机，拥有24.2MP传感器、1080p视频拍摄、3英寸液晶屏和可更换镜头。价格为599.99美元，保修期为1年。

关于电视的信息：

我们有以下电视可供选择：
1. CineView 4K电视（型号：CV-4K55）- 55英寸显示屏，4K分辨率，支持HDR和智能电视功能。价格为599.99美元，保修期为2年。
2. CineView 8K电视（型号：CV-8K65）- 65英寸显示屏，8K分辨率，支持HDR和智能电视功能。价格为2999.99美元，保修期为2年。
3. CineView OLED电视（型号：CV-OLED55）- 55英寸OLED显示屏，4K分辨率，支持HDR和智能电视功能。价格为1499.99美元，保修期为2年。
```

```
请问您对以上产品有任何特别的要求或其他问题吗？
```

二、使用 GPT 评估回答是否正确

我们希望您能从中学到一个设计模式，即当您可以指定一个评估 LLM 输出的标准列表时，您实际上可以使用另一个 API 调用来评估您的第一个 LLM 输出。

```
from tool import get_completion_from_messages

# 问题、上下文
cust_prod_info = {
    'customer_msg': customer_msg,
    'context': product_info
}

def eval_with_rubric(test_set, assistant_answer):
    """
    使用 GPT API 评估生成的回答

    参数:
    test_set: 测试集
    assistant_answer: 助手的回复
    """

    cust_msg = test_set['customer_msg']
    context = test_set['context']
    completion = assistant_answer

    # 人设
    system_message = """\
    你是一位助理，通过查看客户服务代理使用的上下文来评估客户服务代理回答用户问题的情况。
    """

    # 具体指令
    user_message = f"""\
    你正在根据代理使用的上下文评估对问题的提交答案。以下是数据：
    [开始]
    ************
    [用户问题]: {cust_msg}
    ************
    [使用的上下文]: {context}
    ************
    [客户代理的回答]: {completion}
    ************
    [结束]

    请将提交的答案的事实内容与上下文进行比较，忽略样式、语法或标点符号上的差异。
    回答以下问题：
    助手的回应是否只基于所提供的上下文？（是或否）
    回答中是否包含上下文中未提供的信息？（是或否）
    回应与上下文之间是否存在任何不一致之处？（是或否）
    计算用户提出了多少个问题。（输出一个数字）
    对于用户提出的每个问题，是否有相应的回答？
    问题1：（是或否）
    问题2：（是或否）
```

```
    ...
    问题N：（是或否）
    在提出的问题数量中，有多少个问题在回答中得到了回应？（输出一个数字）
"""

    messages = [
        {'role': 'system', 'content': system_message},
        {'role': 'user', 'content': user_message}
    ]

    response = get_completion_from_messages(messages)
    return response

evaluation_output = eval_with_rubric(cust_prod_info, assistant_answer)
print(evaluation_output)
```

```
助手的回应只基于所提供的上下文。是
回答中不包含上下文中未提供的信息。是
回应与上下文之间不存在任何不一致之处。是
用户提出了2个问题。
对于用户提出的每个问题，都有相应的回答。
问题1：是
问题2：是
在提出的问题数量中，有2个问题在回答中得到了回应。
```

三、评估生成回答与标准回答的差距

在经典的自然语言处理技术中，有一些传统的度量标准用于衡量 LLM 输出与人类专家编写的输出的相似度。例如，BLUE 分数可用于衡量两段文本的相似程度。

实际上有一种更好的方法，即使用 Prompt。您可以指定 Prompt，使用 Prompt 来比较由 LLM 自动生成的客户服务代理响应与人工理想响应的匹配程度。

```
'''基于中文Prompt的验证集'''
test_set_ideal = {
    'customer_msg': """\
告诉我有关 the Smartx Pro 手机 和 FotoSnap DSLR相机, the dslr one 的信息。\n另外，你们这有什么电视 ？""",
    'ideal_answer':"""\
SmartX Pro手机是一款功能强大的智能手机，拥有6.1英寸显示屏、128GB存储空间、12MP双摄像头和5G网络支持。价格为899.99美元，保修期为1年。
FotoSnap DSLR相机是一款多功能的单反相机，拥有24.2MP传感器、1080p视频拍摄、3英寸液晶屏和可更换镜头。价格为599.99美元，保修期为1年。

我们有以下电视可供选择：
1. CineView 4K电视（型号：CV-4K55）- 55英寸显示屏，4K分辨率，支持HDR和智能电视功能。价格为599.99美元，保修期为2年。
2. CineView 8K电视（型号：CV-8K65）- 65英寸显示屏，8K分辨率，支持HDR和智能电视功能。价格为2999.99美元，保修期为2年。
3. CineView OLED电视（型号：CV-OLED55）- 55英寸OLED显示屏，4K分辨率，支持HDR和智能电视功能。价格为1499.99美元，保修期为2年。
    """
}
```

我们首先在上文中定义了一个验证集，其包括一个用户指令与一个标准回答。

接着我们可以实现一个评估函数，该函数利用 LLM 的理解能力，要求 LLM 评估生成回答与标准回答是否一致。

```
def eval_vs_ideal(test_set, assistant_answer):
    """
    评估回复是否与理想答案匹配

    参数:
    test_set: 测试集
    assistant_answer: 助手的回复
    """
    cust_msg = test_set['customer_msg']
    ideal = test_set['ideal_answer']
    completion = assistant_answer

    system_message = """\
    您是一位助理，通过将客户服务代理的回答与理想（专家）回答进行比较，评估客户服务代理对用户问题的回答质量。
    请输出一个单独的字母（A 、B、C、D、E），不要包含其他内容。
    """

    user_message = f"""\
    您正在比较一个给定问题的提交答案和专家答案。数据如下:
    [开始]
    ************
    [问题]: {cust_msg}
    ************
    [专家答案]: {ideal}
    ************
    [提交答案]: {completion}
    ************
    [结束]

    比较提交答案的事实内容与专家答案，关注在内容上，忽略样式、语法或标点符号上的差异。
    你的关注核心应该是答案的内容是否正确，内容的细微差异是可以接受的。
    提交的答案可能是专家答案的子集、超集，或者与之冲突。确定适用的情况，并通过选择以下选项之一回答问题:
    （A）提交的答案是专家答案的子集，并且与之完全一致。
    （B）提交的答案是专家答案的超集，并且与之完全一致。
    （C）提交的答案包含与专家答案完全相同的细节。
    （D）提交的答案与专家答案存在分歧。
    （E）答案存在差异，但从事实的角度来看这些差异并不重要。
    选项：ABCDE
"""

    messages = [
        {'role': 'system', 'content': system_message},
        {'role': 'user', 'content': user_message}
    ]

    response = get_completion_from_messages(messages)
    return response
```

这个评分标准来自于 OpenAI 开源评估框架，这是一个非常棒的框架，其中包含了许多评估方法，既有 OpenAI 开发人员的贡献，也有更广泛的开源社区的贡献。

在这个评分标准中，我们要求 LLM 针对提交答案与专家答案进行信息内容的比较，并忽略其风格、语法和标点符号等方面的差异，但关键是我们要求它进行比较，并输出从A到E的分数，具体取决于提交的答案是否是专家答案的子集、超集或完全一致，这可能意味着它虚构或编造了一些额外的事实。

LLM 将选择其中最合适的描述。

LLM 生成的回答为：

```
print(assistant_answer)
```

```
关于SmartX Pro手机和FotoSnap DSLR相机的信息：

1. SmartX Pro手机（型号：SX-PP10）是一款功能强大的智能手机，拥有6.1英寸显示屏、128GB存储空间、12MP双摄像头和5G网络支持。价格为899.99美元，保修期为1年。

2. FotoSnap DSLR相机（型号：FS-DSLR200）是一款多功能的单反相机，拥有24.2MP传感器、1080p视频拍摄、3英寸液晶屏和可更换镜头。价格为599.99美元，保修期为1年。

关于电视的信息：

我们有以下电视可供选择：
1. CineView 4K电视（型号：CV-4K55）- 55英寸显示屏，4K分辨率，支持HDR和智能电视功能。价格为599.99美元，保修期为2年。
2. CineView 8K电视（型号：CV-8K65）- 65英寸显示屏，8K分辨率，支持HDR和智能电视功能。价格为2999.99美元，保修期为2年。
3. CineView OLED电视（型号：CV-OLED55）- 55英寸OLED显示屏，4K分辨率，支持HDR和智能电视功能。价格为1499.99美元，保修期为2年。

请问您对以上产品有任何进一步的问题或者需要了解其他产品吗？
```

```
eval_vs_ideal(test_set_ideal, assistant_answer)
```

```
'C'
```

对于该生成回答，GPT 判断生成内容与标准答案一致

```
assistant_answer_2 = "life is like a box of chocolates"

eval_vs_ideal(test_set_ideal, assistant_answer_2)
```

```
'D'
```

对于明显异常答案，GPT 判断为不一致

希望您从本章中学到两个设计模式。

1. 即使没有专家提供的理想答案，只要能制定一个评估标准，就可以使用一个 LLM 来评估另一个 LLM 的输出。

2. 如果您可以提供一个专家提供的理想答案，那么可以帮助您的 LLM 更好地比较特定助手输出是否与专家提供的理想答案相似。

希望这可以帮助您评估 LLM 系统的输出，以便在开发期间持续监测系统的性能，并使用这些工具不断评估和改进系统的性能。

四、英文版

1. 对问答系统提问

```
import utils_en

# 用户消息
customer_msg = f"""
tell me about the smartx pro phone and the fotosnap camera, the dslr one.
Also, what TVs or TV related products do you have?"""

# 从问题中抽取商品名
products_by_category = utils_en.get_products_from_query(customer_msg)
# 将商品名转化为列表
category_and_product_list = utils_en.read_string_to_list(products_by_category)
# 查找商品对应的信息
product_info = utils_en.get_mentioned_product_info(category_and_product_list)
# 由信息生成回答
assistant_answer = utils_en.answer_user_msg(user_msg=customer_msg,
product_info=product_info)
```

```
print(assistant_answer)
```

```
Sure! Let me provide you with some information about the SmartX ProPhone and the
FotoSnap DSLR Camera.

The SmartX ProPhone is a powerful smartphone with advanced camera features. It
has a 6.1-inch display, 128GB storage, a 12MP dual camera, and supports 5G
connectivity. The SmartX ProPhone is priced at $899.99 and comes with a 1-year
warranty.

The FotoSnap DSLR Camera is a versatile camera that allows you to capture
stunning photos and videos. It features a 24.2MP sensor, 1080p video recording, a
3-inch LCD screen, and supports interchangeable lenses. The FotoSnap DSLR Camera
is priced at $599.99 and also comes with a 1-year warranty.

As for TVs and TV-related products, we have a range of options available. Some of
our popular TV models include the CineView 4K TV, CineView 8K TV, and CineView
OLED TV. We also have home theater systems like the SoundMax Home Theater and
SoundMax Soundbar. Could you please let me know your specific requirements or
preferences so that I can assist you better?
```

2. 使用GPT评估

```
# 问题、上下文
cust_prod_info = {
    'customer_msg': customer_msg,
    'context': product_info
}
```

```
def eval_with_rubric(test_set, assistant_answer):
    """
    使用 GPT API 评估生成的回答

    参数:
    test_set: 测试集
    assistant_answer: 助手的回复
    """

    cust_msg = test_set['customer_msg']
    context = test_set['context']
    completion = assistant_answer

    # 要求 GPT 作为一个助手评估回答正确性
    system_message = """\
    You are an assistant that evaluates how well the customer service agent \
    answers a user question by looking at the context that the customer service \
    agent is using to generate its response.
    """

    # 具体指令
    user_message = f"""\
You are evaluating a submitted answer to a question based on the context \
that the agent uses to answer the question.
Here is the data:
    [BEGIN DATA]
    ************
    [Question]: {cust_msg}
    ************
    [Context]: {context}
    ************
    [Submission]: {completion}
    ************
    [END DATA]

Compare the factual content of the submitted answer with the context. \
Ignore any differences in style, grammar, or punctuation.
Answer the following questions:
    - Is the Assistant response based only on the context provided? (Y or N)
    - Does the answer include information that is not provided in the context? (Y
or N)
    - Is there any disagreement between the response and the context? (Y or N)
    - Count how many questions the user asked. (output a number)
    - For each question that the user asked, is there a corresponding answer to
it?
      Question 1: (Y or N)
      Question 2: (Y or N)
      ...
```

```
      Question N: (Y or N)
    - Of the number of questions asked, how many of these questions were
addressed by the answer? (output a number)
"""

    messages = [
        {'role': 'system', 'content': system_message},
        {'role': 'user', 'content': user_message}
    ]

    response = get_completion_from_messages(messages)
    return response
```

```
evaluation_output = eval_with_rubric(cust_prod_info, assistant_answer)
print(evaluation_output)
```

```
- Is the Assistant response based only on the context provided? (Y or N)
Y

- Does the answer include information that is not provided in the context? (Y or
N)
N

- Is there any disagreement between the response and the context? (Y or N)
N

- Count how many questions the user asked. (output a number)
2

- For each question that the user asked, is there a corresponding answer to it?
Question 1: Y
Question 2: Y

- Of the number of questions asked, how many of these questions were addressed by
the answer? (output a number)
2
```

3. 评估生成回答与标准回答的差距

```
test_set_ideal = {
    'customer_msg': """\
tell me about the smartx pro phone and the fotosnap camera, the dslr one.
Also, what TVs or TV related products do you have?""",
    'ideal_answer':"""\
Of course!  The SmartX ProPhone is a powerful \
smartphone with advanced camera features. \
For instance, it has a 12MP dual camera. \
Other features include 5G wireless and 128GB storage. \
It also has a 6.1-inch display.  The price is $899.99.

The FotoSnap DSLR Camera is great for \
capturing stunning photos and videos. \
Some features include 1080p video, \
3-inch LCD, a 24.2MP sensor, \
```

```
and interchangeable lenses. \
The price is 599.99.

For TVs and TV related products, we offer 3 TVs \


All TVs offer HDR and Smart TV.

The CineView 4K TV has vibrant colors and smart features. \
Some of these features include a 55-inch display, \
'4K resolution. It's priced at 599.

The CineView 8K TV is a stunning 8K TV. \
Some features include a 65-inch display and \
8K resolution.  It's priced at 2999.99

The CineView OLED TV lets you experience vibrant colors. \
Some features include a 55-inch display and 4K resolution. \
It's priced at 1499.99.

We also offer 2 home theater products, both which include bluetooth.\
The SoundMax Home Theater is a powerful home theater system for \
an immmersive audio experience.
Its features include 5.1 channel, 1000W output, and wireless subwoofer.
It's priced at 399.99.

The SoundMax Soundbar is a sleek and powerful soundbar.
It's features include 2.1 channel, 300W output, and wireless subwoofer.
It's priced at 199.99

Are there any questions additional you may have about these products \
that you mentioned here?
Or may do you have other questions I can help you with?
    """
}
```

```
def eval_vs_ideal(test_set, assistant_answer):
    """
    评估回复是否与理想答案匹配

    参数:
    test_set: 测试集
    assistant_answer: 助手的回复
    """
    cust_msg = test_set['customer_msg']
    ideal = test_set['ideal_answer']
    completion = assistant_answer

    system_message = """\
    You are an assistant that evaluates how well the customer service agent \
    answers a user question by comparing the response to the ideal (expert)
response
    Output a single letter and nothing else.
    """
```

```
    user_message = f"""\
You are comparing a submitted answer to an expert answer on a given question.
Here is the data:
    [BEGIN DATA]
    ************
    [Question]: {cust_msg}
    ************
    [Expert]: {ideal}
    ************
    [Submission]: {completion}
    ************
    [END DATA]

Compare the factual content of the submitted answer with the expert answer.
Ignore any differences in style, grammar, or punctuation.
    The submitted answer may either be a subset or superset of the expert answer,
or it may conflict with it. Determine which case applies.
    Answer the question by selecting one of the following options:
    (A) The submitted answer is a subset of the expert answer and is fully
consistent with it.
    (B) The submitted answer is a superset of the expert answer and is fully
consistent with it.
    (C) The submitted answer contains all the same details as the expert answer.
    (D) There is a disagreement between the submitted answer and the expert
answer.
    (E) The answers differ, but these differences don't matter from the
perspective of factuality.
  choice_strings: ABCDE
"""

    messages = [
        {'role': 'system', 'content': system_message},
        {'role': 'user', 'content': user_message}
    ]

    response = get_completion_from_messages(messages)
    return response
```

```
print(assistant_answer)
```

```
Sure! Let me provide you with some information about the SmartX ProPhone and the FotoSnap DSLR Camera.

The SmartX ProPhone is a powerful smartphone with advanced camera features. It has a 6.1-inch display, 128GB storage, a 12MP dual camera, and supports 5G connectivity. The SmartX ProPhone is priced at $899.99 and comes with a 1-year warranty.

The FotoSnap DSLR Camera is a versatile camera that allows you to capture stunning photos and videos. It features a 24.2MP sensor, 1080p video recording, a 3-inch LCD screen, and supports interchangeable lenses. The FotoSnap DSLR Camera is priced at $599.99 and also comes with a 1-year warranty.

As for TVs and TV-related products, we have a range of options available. Some of our popular TV models include the CineView 4K TV, CineView 8K TV, and CineView OLED TV. We also have home theater systems like the SoundMax Home Theater and SoundMax Soundbar. Could you please let me know your specific requirements or preferences so that I can assist you better?
```

```
# 由于模型的更新，目前在原有 Prompt 上不再能够正确判断
eval_vs_ideal(test_set_ideal, assistant_answer)
```

```
'D'
```

```
assistant_answer_2 = "life is like a box of chocolates"
```

```
eval_vs_ideal(test_set_ideal, assistant_answer_2)
# 对于明显异常答案，GPT 判断为不一致
```

```
'D'
```