第七章 检查结果

随着我们深入本书的学习，本章将引领你了解如何评估系统生成的输出。在任何场景中，无论是自动化流程还是其他环境，我们都必须确保在向用户展示输出之前，对其质量、相关性和安全性进行严格的检查，以保证我们提供的反馈是准确和适用的。我们将学习如何运用审查(Moderation) API 来对输出进行评估，并深入探讨如何通过额外的 Prompt 提升模型在展示输出之前的质量评估。

一、检查有害内容

我们主要通过 OpenAI 提供的 Moderation API 来实现对有害内容的检查。

```
import openai
from tool import get_completion_from_messages

final_response_to_customer = f"""
SmartX ProPhone 有一个 6.1 英寸的显示屏，128GB 存储、\
1200 万像素的双摄像头，以及 5G。FotoSnap 单反相机\
有一个 2420 万像素的传感器，1080p 视频，3 英寸 LCD 和\
可更换的镜头。我们有各种电视，包括 CineView 4K 电视，\
55 英寸显示屏，4K 分辨率、HDR，以及智能电视功能。\
我们也有 SoundMax 家庭影院系统，具有 5.1 声道，\
1000W 输出，无线重低音扬声器和蓝牙。关于这些产品或\
我们提供的任何其他产品您是否有任何具体问题？
"""
# Moderation 是 OpenAI 的内容审核函数，旨在评估并检测文本内容中的潜在风险。
response = openai.Moderation.create(
    input=final_response_to_customer
)
moderation_output = response["results"][0]
print(moderation_output)
```

```
{
  "categories": {
    "harassment": false,
    "harassment/threatening": false,
    "hate": false,
    "hate/threatening": false,
    "self-harm": false,
    "self-harm/instructions": false,
    "self-harm/intent": false,
    "sexual": false,
    "sexual/minors": false,
    "violence": false,
    "violence/graphic": false
  },
  "category_scores": {
    "harassment": 4.2861907e-07,
    "harassment/threatening": 5.9538485e-09,
    "hate": 2.079682e-07,
    "hate/threatening": 5.6982725e-09,
    "self-harm": 2.3966843e-08,
    "self-harm/instructions": 1.5763412e-08,
    "self-harm/intent": 5.042827e-09,
```

```
    "sexual": 2.6989035e-06,
    "sexual/minors": 1.1349888e-06,
    "violence": 1.2788286e-06,
    "violence/graphic": 2.6259923e-07
  },
  "flagged": false
}
```


如你所见，这个输出没有被标记为任何特定类别，并且在所有类别中都获得了非常低的得分，说明给出的结果评判是合理的。

总体来说，检查输出的质量同样是十分重要的。例如，如果你正在为一个对内容有特定敏感度的受众构建一个聊天机器人，你可以设定更低的阈值来标记可能存在问题的输出。通常情况下，如果审查结果显示某些内容被标记，你可以采取适当的措施，比如提供一个替代答案或生成一个新的响应。

值得注意的是，随着我们对模型的持续改进，它们越来越不太可能产生有害的输出。

检查输出质量的另一种方法是向模型询问其自身生成的结果是否满意，是否达到了你所设定的标准。这可以通过将生成的输出作为输入的一部分再次提供给模型，并要求它对输出的质量进行评估。这种操作可以通过多种方式完成。接下来，我们将通过一个例子来展示这种方法。

二、检查是否符合产品信息

在下列示例中，我们要求 LLM 作为一个助理检查回复是否充分回答了客户问题，并验证助理引用的事实是否正确。

```
# 这是一段电子产品相关的信息
system_message = f"""
您是一个助理，用于评估客服代理的回复是否充分回答了客户问题，\
并验证助理从产品信息中引用的所有事实是否正确。
产品信息、用户和客服代理的信息将使用三个反引号（即 ```）\
进行分隔。
请以 Y 或 N 的字符形式进行回复，不要包含标点符号：\
Y - 如果输出充分回答了问题并且回复正确地使用了产品信息\
N - 其他情况。

仅输出单个字母。
"""

#这是顾客的提问
customer_message = f"""
告诉我有关 smartx pro 手机\
和 fotosnap 相机（单反相机）的信息。\
还有您电视的信息。
"""
```

```
product_information = """{ "name": "SmartX ProPhone", "category": "Smartphones and Accessories", "brand": "SmartX", "model_number": "SX-PP10", "warranty": "1 year", "rating": 4.6, "features": [ "6.1-inch display", "128GB storage", "12MP dual camera", "5G" ], "description": "A powerful smartphone with advanced camera features.", "price": 899.99 } { "name": "FotoSnap DSLR Camera", "category": "Cameras and Camcorders", "brand": "FotoSnap", "model_number": "FS-DSLR200", "warranty": "1 year", "rating": 4.7, "features": [ "24.2MP sensor", "1080p video", "3-inch LCD", "Interchangeable lenses" ], "description": "Capture stunning photos and videos with this versatile DSLR camera.", "price": 599.99 } { "name": "CineView 4K TV", "category": "Televisions and Home Theater Systems", "brand": "CineView", "model_number": "CV-4K55", "warranty": "2 years", "rating": 4.8, "features": [ "55-inch display", "4K resolution", "HDR", "Smart TV" ], "description": "A stunning 4K TV with vibrant colors and smart features.", "price": 599.99 } { "name": "SoundMax Home Theater", "category": "Televisions and Home Theater Systems", "brand": "SoundMax", "model_number": "SM-HT100", "warranty": "1 year", "rating": 4.4, "features": [ "5.1 channel", "1000W output", "Wireless subwoofer", "Bluetooth" ], "description": "A powerful home theater system for an immersive audio experience.", "price": 399.99 } { "name": "CineView 8K TV", "category": "Televisions and Home Theater Systems", "brand": "CineView", "model_number": "CV-8K65", "warranty": "2 years", "rating": 4.9, "features": [ "65-inch display", "8K resolution", "HDR", "Smart TV" ], "description": "Experience the future of television with this stunning 8K TV.", "price": 2999.99 } { "name": "SoundMax Soundbar", "category": "Televisions and Home Theater Systems", "brand": "SoundMax", "model_number": "SM-SB50", "warranty": "1 year", "rating": 4.3, "features": [ "2.1 channel", "300W output", "Wireless subwoofer", "Bluetooth" ], "description": "Upgrade your TV's audio with this sleek and powerful soundbar.", "price": 199.99 } { "name": "CineView OLED TV", "category": "Televisions and Home Theater Systems", "brand": "CineView", "model_number": "CV-OLED55", "warranty": "2 years", "rating": 4.7, "features": [ "55-inch display", "4K resolution", "HDR", "Smart TV" ], "description": "Experience true blacks and vibrant colors with this OLED TV.", "price": 1499.99 }"""

q_a_pair = f"""
顾客的信息: ```{customer_message}```
产品信息: ```{product_information}```
代理的回复: ```{final_response_to_customer}```

回复是否正确使用了检索的信息?
回复是否充分地回答了问题?

输出 Y 或 N
"""
#判断相关性
messages = [
    {'role': 'system', 'content': system_message},
    {'role': 'user', 'content': q_a_pair}
]

response = get_completion_from_messages(messages, max_tokens=1)
print(response)
```

```
Y
```

在上一个示例中，我们给了一个正例，LLM 很好地做出了正确的检查。而在下一个示例中，我们将提供一个负例，LLM 同样能够正确判断。

```
another_response = "生活就像一盒巧克力"
q_a_pair = f"""
顾客的信息: ```{customer_message}```
产品信息: ```{product_information}```
代理的回复: ```{another_response}```

回复是否正确使用了检索的信息?
回复是否充分地回答了问题?

输出 Y 或 N
"""
messages = [
    {'role': 'system', 'content': system_message},
    {'role': 'user', 'content': q_a_pair}
]

response = get_completion_from_messages(messages)
print(response)
```

```
N
```

因此，你可以看到，模型具有提供生成输出质量反馈的能力。你可以使用这种反馈来决定是否将输出展示给用户，或是生成新的回应。你甚至可以尝试为每个用户查询生成多个模型回应，然后从中挑选出最佳的回应呈现给用户。所以，你有多种可能的尝试方式。

总的来说，借助审查 API 来检查输出是一个可取的策略。但在我看来，这在大多数情况下可能是不必要的，特别是当你使用更先进的模型，比如 GPT-4 。实际上，在真实生产环境中，我们并未看到很多人采取这种方式。这种做法也会增加系统的延迟和成本，因为你需要等待额外的 API 调用，并且需要额外的 token 。如果你的应用或产品的错误率仅为0.0000001%，那么你可能可以尝试这种策略。但总的来说，我们并不建议在实际应用中使用这种方式。在接下来的章节中，我们将把我们在评估输入、处理输出以及审查生成内容所学到的知识整合起来，构建一个端到端的系统。

三、英文版

1.1 检查有害信息

```
final_response_to_customer = f"""
The SmartX ProPhone has a 6.1-inch display, 128GB storage, \
12MP dual camera, and 5G. The FotoSnap DSLR Camera \
has a 24.2MP sensor, 1080p video, 3-inch LCD, and \
interchangeable lenses. We have a variety of TVs, including \
the CineView 4K TV with a 55-inch display, 4K resolution, \
HDR, and smart TV features. We also have the SoundMax \
Home Theater system with 5.1 channel, 1000W output, wireless \
subwoofer, and Bluetooth. Do you have any specific questions \
about these products or any other products we offer?
"""


response = openai.Moderation.create(
```

```
    input=final_response_to_customer
)
moderation_output = response["results"][0]
print(moderation_output)
```


```
{
  "categories": {
    "harassment": false,
    "harassment/threatening": false,
    "hate": false,
    "hate/threatening": false,
    "self-harm": false,
    "self-harm/instructions": false,
    "self-harm/intent": false,
    "sexual": false,
    "sexual/minors": false,
    "violence": false,
    "violence/graphic": false
  },
  "category_scores": {
    "harassment": 3.4429521e-09,
    "harassment/threatening": 9.538529e-10,
    "hate": 6.0008998e-09,
    "hate/threatening": 3.5339007e-10,
    "self-harm": 5.6997046e-10,
    "self-harm/instructions": 3.864466e-08,
    "self-harm/intent": 9.3394e-10,
    "sexual": 2.2777907e-07,
    "sexual/minors": 2.6869095e-08,
    "violence": 3.5471032e-07,
    "violence/graphic": 7.8637696e-10
  },
  "flagged": false
}
```


2.1 检查是否符合产品信息

```
# 这是一段电子产品相关的信息
system_message = f"""
You are an assistant that evaluates whether \
customer service agent responses sufficiently \
answer customer questions, and also validates that \
all the facts the assistant cites from the product \
information are correct.
The product information and user and customer \
service agent messages will be delimited by \
3 backticks, i.e. ```.
Respond with a Y or N character, with no punctuation:
Y - if the output sufficiently answers the question \
AND the response correctly uses product information
N - otherwise

Output a single letter only.
"""
```

```
#这是顾客的提问
customer_message = f"""
tell me about the smartx pro phone and \
the fotosnap camera, the dslr one. \
Also tell me about your tvs"""
product_information = """{ "name": "SmartX ProPhone", "category": "Smartphones
and Accessories", "brand": "SmartX", "model_number": "SX-PP10", "warranty": "1
year", "rating": 4.6, "features": [ "6.1-inch display", "128GB storage", "12MP
dual camera", "5G" ], "description": "A powerful smartphone with advanced camera
features.", "price": 899.99 } { "name": "FotoSnap DSLR Camera", "category":
"Cameras and Camcorders", "brand": "FotoSnap", "model_number": "FS-DSLR200",
"warranty": "1 year", "rating": 4.7, "features": [ "24.2MP sensor", "1080p
video", "3-inch LCD", "Interchangeable lenses" ], "description": "Capture
stunning photos and videos with this versatile DSLR camera.", "price": 599.99 } {
"name": "CineView 4K TV", "category": "Televisions and Home Theater Systems",
"brand": "CineView", "model_number": "CV-4K55", "warranty": "2 years", "rating":
4.8, "features": [ "55-inch display", "4K resolution", "HDR", "Smart TV" ],
"description": "A stunning 4K TV with vibrant colors and smart features.",
"price": 599.99 } { "name": "SoundMax Home Theater", "category": "Televisions and
Home Theater Systems", "brand": "SoundMax", "model_number": "SM-HT100",
"warranty": "1 year", "rating": 4.4, "features": [ "5.1 channel", "1000W output",
"Wireless subwoofer", "Bluetooth" ], "description": "A powerful home theater
system for an immersive audio experience.", "price": 399.99 } { "name": "CineView
8K TV", "category": "Televisions and Home Theater Systems", "brand": "CineView",
"model_number": "CV-8K65", "warranty": "2 years", "rating": 4.9, "features": [
"65-inch display", "8K resolution", "HDR", "Smart TV" ], "description":
"Experience the future of television with this stunning 8K TV.", "price": 2999.99
} { "name": "SoundMax Soundbar", "category": "Televisions and Home Theater
Systems", "brand": "SoundMax", "model_number": "SM-SB50", "warranty": "1 year",
"rating": 4.3, "features": [ "2.1 channel", "300W output", "Wireless subwoofer",
"Bluetooth" ], "description": "Upgrade your TV's audio with this sleek and
powerful soundbar.", "price": 199.99 } { "name": "CineView OLED TV", "category":
"Televisions and Home Theater Systems", "brand": "CineView", "model_number": "CV-
OLED55", "warranty": "2 years", "rating": 4.7, "features": [ "55-inch display",
"4K resolution", "HDR", "Smart TV" ], "description": "Experience true blacks and
vibrant colors with this OLED TV.", "price": 1499.99 }"""

q_a_pair = f"""
Customer message: ```{customer_message}```
Product information: ```{product_information}```
Agent response: ```{final_response_to_customer}```

Does the response use the retrieved information correctly?
Does the response sufficiently answer the question?

Output Y or N
"""
#判断相关性
messages = [
    {'role': 'system', 'content': system_message},
    {'role': 'user', 'content': q_a_pair}
]

response = get_completion_from_messages(messages, max_tokens=1)
print(response)
```

```
Y
```

```
another_response = "life is like a box of chocolates"
q_a_pair = f"""
Customer message: ```{customer_message}```
Product information: ```{product_information}```
Agent response: ```{another_response}```

Does the response use the retrieved information correctly?
Does the response sufficiently answer the question?

Output Y or N
"""
messages = [
    {'role': 'system', 'content': system_message},
    {'role': 'user', 'content': q_a_pair}
]

response = get_completion_from_messages(messages)
print(response)
```

```
N
```