第六章 评估

评估是检验语言模型问答质量的关键环节。评估可以检验语言模型在不同文档上的问答效果，找出其弱点。还可以通过比较不同模型，选择最佳系统。此外，定期评估也可以检查模型质量的衰减。评估通常有两个目的：

- 检验LLM应用是否达到了验收标准
- 分析改动对于LLM应用性能的影响

基本的思路就是利用语言模型本身和链本身，来辅助评估其他的语言模型、链和应用程序。我们还是以上一章节的文档问答应用为例，在本章节中讨论如何在 LangChain 中处理和考虑评估的内容。

一、 创建LLM应用

首先，按照 langchain 链的方式构建一个 LLM 的文档问答应用

```
from langchain.chains import RetrievalQA #检索QA链，在文档上进行检索
from langchain.chat_models import ChatOpenAI #openai模型
from langchain.document_loaders import CSVLoader #文档加载器，采用csv格式存储
from langchain.indexes import VectorstoreIndexCreator #导入向量存储索引创建器
from langchain.vectorstores import DocArrayInMemorySearch #向量存储
#加载中文数据
file = '../data/product_data.csv'
loader = CSVLoader(file_path=file)
data = loader.load()

#查看数据
import pandas as pd
test_data = pd.read_csv(file,skiprows=0)
display(test_data.head())
```

	product_name	description
0	全自动咖啡机	规格:\n大型 - 尺寸：13.8'' x 17.3''。\n中型 - 尺寸：11.5'' ...
1	电动牙刷	规格:\n一般大小 - 高度：9.5''，宽度：1''。\n\n为什么我们热爱它:\n我们的...
2	橙味维生素C泡腾片	规格:\n每盒含有20片。\n\n为什么我们热爱它:\n我们的橙味维生素C泡腾片是快速补充维...
3	无线蓝牙耳机	规格:\n单个耳机尺寸：1.5'' x 1.3''。\n\n为什么我们热爱它:\n这款无线蓝...
4	瑜伽垫	规格:\n尺寸：24'' x 68''。\n\n为什么我们热爱它:\n我们的瑜伽垫拥有出色的...

```
# 将指定向量存储类,创建完成后，我们将从加载器中调用,通过文档记载器列表加载

index = VectorstoreIndexCreator(
    vectorstore_cls=DocArrayInMemorySearch
).from_loaders([loader])
```

```
#通过指定语言模型、链类型、检索器和我们要打印的详细程度来创建检索QA链
llm = ChatOpenAI(temperature = 0.0)
qa = RetrievalQA.from_chain_type(
    llm=llm,
    chain_type="stuff",
    retriever=index.vectorstore.as_retriever(),
    verbose=True,
    chain_type_kwargs = {
        "document_separator": "<<<<>>>>>"
    }
)
```

上述代码的主要功能及作用在上一章节中都已说明，这里不再赘述

1.1 设置测试的数据

我们查看一下经过档加载器 CSVLoad 加载后生成的 data 内的信息，这里我们抽取 data 中的第九条和第十条数据，看看它们的主要内容：

第九条数据：

```
data[10]
```

```
Document(page_content="product_name: 高清电视机\ndescription: 规格:\n尺寸：50''。\n\n为什么我们热爱它:\n我们的高清电视机拥有出色的画质和强大的音效，带来沉浸式的观看体验。\n\n材质与护理:\n使用干布清洁。\n\n构造:\n由塑料、金属和电子元件制成。\n\n其他特性:\n支持网络连接，可以在线观看视频。\n配备遥控器。\n在韩国制造。\n\n有问题？请随时联系我们的客户服务团队，他们会解答您的所有问题。", metadata={'source': '../data/product_data.csv', 'row': 10})
```

第十条数据：

```
data[11]
```

```
Document(page_content="product_name: 旅行背包\ndescription: 规格:\n尺寸：18'' x 12'' x 6''。\n\n为什么我们热爱它:\n我们的旅行背包拥有多个实用的内外袋，轻松装下您的必需品，是短途旅行的理想选择。\n\n材质与护理:\n可以手洗，自然晾干。\n\n构造:\n由防水尼龙制成。\n\n其他特性:\n附带可调节背带和安全锁。\n在中国制造。\n\n有问题？请随时联系我们的客户服务团队，他们会解答您的所有问题。", metadata={'source': '../data/product_data.csv', 'row': 11})
```

看上面的第一个文档中有高清电视机，第二个文档中有旅行背包，从这些细节中，我们可以创建一些例子查询和答案

1.2 手动创建测试数据

需要说明的是这里我们的文档是 csv 文件，所以我们使用的是文档加载器是 CSVLoader ， CSVLoader 会对 csv 文件中的每一行数据进行分割，所以这里看到的 data[10], data[11]的内容则是 csv 文件中的第10条，第11条数据的内容。下面我们根据这两条数据手动设置两条“问答对”，每一个“问答对”中包含一个 query ， 一个 answer ：

```
examples = [
    {
        "query": "高清电视机怎么进行护理？",
        "answer": "使用干布清洁。"
    },
    {
        "query": "旅行背包有内外袋吗？",
        "answer": "有。"
    }
]
```

1.3 通过LLM生成测试用例

在前面的内容中，我们使用的方法都是通过手动的方法来构建测试数据集，比如说我们手动创建10个问题和10个答案，然后让 LLM 回答这10个问题，再将 LLM 给出的答案与我们准备好的答案做比较，最后再给 LLM 打分，评估的流程大概就是这样。但是这里有一个问题，就是我们需要手动去创建所有的问题集和答案集，那会是一个非常耗费时间和人力的成本。那有没有一种可以自动创建大量问答测试集的方法呢？那当然是有的，今天我们就来介绍 Langchain 提供的方法：`QAGenerateChain`，我们可以通过`QAGenerateChain`来为我们的文档自动创建问答集：

由于`QAGenerateChain`类中使用的`PROMPT`是英文，故我们继承`QAGenerateChain`类，将`PROMPT`加上“请使用中文输出”。下面是`generate_chain.py`文件中的`QAGenerateChain`类的源码

```
from langchain.evaluation.qa import QAGenerateChain #导入QA生成链，它将接收文档，并从每个文档中创建一个问题答案对

# 下面是langchain.evaluation.qa.generate_prompt中的源码，在template的最后加上“请使用中文输出”
from langchain.output_parsers.regex import RegexParser
from langchain.prompts import PromptTemplate
from langchain.base_language import BaseLanguageModel
from typing import Any

template = """You are a teacher coming up with questions to ask on a quiz.
Given the following document, please generate a question and answer based on that document.

Example Format:
<Begin Document>
...
<End Document>
QUESTION: question here
ANSWER: answer here

These questions should be detailed and be based explicitly on information in the document. Begin!

<Begin Document>
{doc}
<End Document>
请使用中文输出
"""
output_parser = RegexParser(
    regex=r"QUESTION: (.*?)\nANSWER: (.*)", output_keys=["query", "answer"]
```

```
)
PROMPT = PromptTemplate(
    input_variables=["doc"], template=template, output_parser=output_parser
)

# 继承QAGenerateChain
class ChineseQAGenerateChain(QAGenerateChain):
    """LLM Chain specifically for generating examples for question answering."""

    @classmethod
    def from_llm(cls, llm: BaseLanguageModel, **kwargs: Any) -> QAGenerateChain:
        """Load QA Generate Chain from LLM."""
        return cls(llm=llm, prompt=PROMPT, **kwargs)



example_gen_chain = ChineseQAGenerateChain.from_llm(ChatOpenAI())#通过传递chat
open AI语言模型来创建这个链
new_examples = example_gen_chain.apply([{"doc": t} for t in data[:5]])

#查看用例数据
new_examples
```

```
[{'qa_pairs': {'query': '这款全自动咖啡机的尺寸是多少？',
   'answer': "大型尺寸为13.8'' x 17.3''，中型尺寸为11.5'' x 15.2''。"}},
 {'qa_pairs': {'query': '这款电动牙刷的规格是什么？', 'answer': "一般大小 - 高度：
9.5''，宽度：1''。"}},
 {'qa_pairs': {'query': '这种产品的名称是什么？', 'answer': '这种产品的名称是橙味维生素C
泡腾片。'}},
 {'qa_pairs': {'query': '这款无线蓝牙耳机的尺寸是多少？',
   'answer': "该无线蓝牙耳机的尺寸为1.5'' x 1.3''。"}},
 {'qa_pairs': {'query': '这款瑜伽垫的尺寸是多少？', 'answer': "这款瑜伽垫的尺寸是24'' x
68''。"}}]
```

在上面的代码中，我们创建了一个`QAGenerateChain`，然后我们应用了`QAGenerateChain`的 apply 方法对 data 中的前5条数据创建了5个“问答对”，由于创建问答集是由 LLM 来自动完成的，因此会涉及到 token 成本的问题，所以我们这里出于演示的目的，只对 data 中的前5条数据创建问答集。

```
new_examples[0]
```

```
{'qa_pairs': {'query': '这款全自动咖啡机的尺寸是多少？',
  'answer': "大型尺寸为13.8'' x 17.3''，中型尺寸为11.5'' x 15.2''。"}}
```

源数据：

```
data[0]
```

```
Document(page_content="product_name: 全自动咖啡机\ndescription: 规格:\n大型 - 尺寸: 13.8'' x 17.3''。\n中型 - 尺寸: 11.5'' x 15.2''。\n\n为什么我们热爱它:\n这款全自动咖啡机是爱好者的理想选择。 一键操作，即可研磨豆子并沏制出您喜爱的咖啡。它的耐用性和一致性使它成为家庭和办公室的理想选择。\n\n材质与护理:\n清洁时只需轻擦。\n\n构造:\n由高品质不锈钢制成。\n\n其他特性:\n内置研磨器和滤网。\n预设多种咖啡模式。\n在中国制造。\n\n有问题？ 请随时联系我们的客户服务团队，他们会解答您的所有问题。", metadata={'source': '../data/product_data.csv', 'row': 0})
```

1.4 整合测试集

还记得我们前面手动创建的两个问答集吗？现在我们需要将之前手动创建的问答集合并到 `QAGenerateChain` 创建的问答集中，这样在答集中既有手动创建的例子又有 llm 自动创建的例子，这会使我们的测试集更加完善。

接下来我们就需要让之前创建的文档问答链 `qa` 来回答这个测试集里的问题，来看看 LLM 是怎么回答的吧：

```
examples += [ v for item in new_examples for k,v in item.items()]
qa.run(examples[0]["query"])
```

```
> Entering new RetrievalQA chain...

> Finished chain.
```

```
'高清电视机的护理非常简单。您只需要使用干布清洁即可。避免使用湿布或化学清洁剂，以免损坏电视机的表面。'
```

这里我们看到 `qa` 回答了第0个问题："高清电视机怎么进行护理？"，这里的第0个问题就是先前我们手动创建的第一个问题，并且我们手动创建的 answer 是："使用干布清洁。" 这里我们发现问答链 `qa` 回答的也是“您只需要使用干布清洁即可”，只是它比我们的答案还多了一段说明：“高清电视机的护理非常简单。您只需要使用干布清洁即可。避免使用湿布或化学清洁剂，以免损坏电视机的表面。”。

二、 人工评估

你想知道 `qa` 是怎么找到问题的答案的吗？下面让我们打开 `debug`，看看 `qa` 是如何找到问题的答案！

```
import langchain
langchain.debug = True

#重新运行与上面相同的示例，可以看到它开始打印出更多的信息
qa.run(examples[0]["query"])
```

```
[chain/start] [1:chain:RetrievalQA] Entering Chain run with input:
{
  "query": "高清电视机怎么进行护理？"
}
[chain/start] [1:chain:RetrievalQA > 3:chain:StuffDocumentsChain] Entering Chain run with input:
[inputs]
[chain/start] [1:chain:RetrievalQA > 3:chain:StuffDocumentsChain > 4:chain:LLMChain] Entering Chain run with input:
```

```
{
  "question": "高清电视机怎么进行护理？",
  "context": "product_name: 高清电视机\ndescription: 规格:\n尺寸：50''。\n\n为什么我们热爱它:\n我们的高清电视机拥有出色的画质和强大的音效，带来沉浸式的观看体验。\n\n材质与护理:\n使用干布清洁。\n\n构造:\n由塑料、金属和电子元件制成。\n\n其他特性:\n支持网络连接，可以在线观看视频。\n配备遥控器。\n在韩国制造。\n\n有问题？请随时联系我们的客户服务团队，他们会解答您的所有问题。<<<<>>>>>product_name: 空气净化器\ndescription: 规格:\n尺寸：15'' x 15'' x 20''。\n\n为什么我们热爱它:\n我们的空气净化器采用了先进的HEPA过滤技术，能有效去除空气中的微粒和异味，为您提供清新的室内环境。\n\n材质与护理:\n清洁时使用干布擦拭。\n\n构造:\n由塑料和电子元件制成。\n\n其他特性:\n三档风速，附带定时功能。\n在德国制造。\n\n有问题？请随时联系我们的客户服务团队，他们会解答您的所有问题。<<<<>>>>>product_name: 宠物自动喂食器\ndescription: 规格:\n尺寸：14'' x 9'' x 15''。\n\n为什么我们热爱它:\n我们的宠物自动喂食器可以定时定量投放食物，让您无论在家或外出都能确保宠物的饮食。\n\n材质与护理:\n可用湿布清洁。\n\n构造:\n由塑料和电子元件制成。\n\n其他特性:\n配备LCD屏幕，操作简单。\n可以设置多次投食。\n在美国制造。\n\n有问题？请随时联系我们的客户服务团队，他们会解答您的所有问题。<<<<>>>>>product_name: 玻璃保护膜\ndescription: 规格:\n适用于各种尺寸的手机屏幕。\n\n为什么我们热爱它:\n我们的玻璃保护膜可以有效防止手机屏幕刮伤和破裂，而且不影响触控的灵敏度。\n\n材质与护理:\n使用干布擦拭。\n\n构造:\n由高强度的玻璃材料制成。\n\n其他特性:\n安装简单，适合自行安装。\n在日本制造。\n\n有问题？请随时联系我们的客户服务团队，他们会解答您的所有问题。"
}
[llm/start] [1:chain:RetrievalQA > 3:chain:StuffDocumentsChain > 4:chain:LLMChain > 5:llm:ChatOpenAI] Entering LLM run with input:
{
  "prompts": [
    "System: Use the following pieces of context to answer the users question.\nIf you don't know the answer, just say that you don't know, don't try to make up an answer.\n----------------\nproduct_name: 高清电视机\ndescription: 规格:\n尺寸：50''。\n\n为什么我们热爱它:\n我们的高清电视机拥有出色的画质和强大的音效，带来沉浸式的观看体验。\n\n材质与护理:\n使用干布清洁。\n\n构造:\n由塑料、金属和电子元件制成。\n\n其他特性:\n支持网络连接，可以在线观看视频。\n配备遥控器。\n在韩国制造。\n\n有问题？请随时联系我们的客户服务团队，他们会解答您的所有问题。<<<<>>>>>product_name: 空气净化器\ndescription: 规格:\n尺寸：15'' x 15'' x 20''。\n\n为什么我们热爱它:\n我们的空气净化器采用了先进的HEPA过滤技术，能有效去除空气中的微粒和异味，为您提供清新的室内环境。\n\n材质与护理:\n清洁时使用干布擦拭。\n\n构造:\n由塑料和电子元件制成。\n\n其他特性:\n三档风速，附带定时功能。\n在德国制造。\n\n有问题？请随时联系我们的客户服务团队，他们会解答您的所有问题。<<<<>>>>>product_name: 宠物自动喂食器\ndescription: 规格:\n尺寸：14'' x 9'' x 15''。\n\n为什么我们热爱它:\n我们的宠物自动喂食器可以定时定量投放食物，让您无论在家或外出都能确保宠物的饮食。\n\n材质与护理:\n可用湿布清洁。\n\n构造:\n由塑料和电子元件制成。\n\n其他特性:\n配备LCD屏幕，操作简单。\n可以设置多次投食。\n在美国制造。\n\n有问题？请随时联系我们的客户服务团队，他们会解答您的所有问题。\n<<<<>>>>>product_name: 玻璃保护膜\ndescription: 规格:\n适用于各种尺寸的手机屏幕。\n\n为什么我们热爱它:\n我们的玻璃保护膜可以有效防止手机屏幕刮伤和破裂，而且不影响触控的灵敏度。\n\n材质与护理:\n使用干布擦拭。\n\n构造:\n由高强度的玻璃材料制成。\n\n其他特性:\n安装简单，适合自行安装。\n在日本制造。\n\n有问题？请随时联系我们的客户服务团队，他们会解答您的所有问题。\nHuman: 高清电视机怎么进行护理？"
  ]
}
[llm/end] [1:chain:RetrievalQA > 3:chain:StuffDocumentsChain > 4:chain:LLMChain > 5:llm:ChatOpenAI] [2.86s] Exiting LLM run with output:
{
  "generations": [
    [
      {
        "text": "高清电视机的护理非常简单。您只需要使用干布清洁即可。避免使用湿布或化学清洁剂，以免损坏电视机的表面。",
        "generation_info": {
          "finish_reason": "stop"
```

```
        },
        "message": {
          "lc": 1,
          "type": "constructor",
          "id": [
            "langchain",
            "schema",
            "messages",
            "AIMessage"
          ],
          "kwargs": {
            "content": "高清电视机的护理非常简单。您只需要使用干布清洁即可。避免使用湿布或化学清洁剂，以免损坏电视机的表面。",
            "additional_kwargs": {}
          }
        }
      }
    ]
  ],
  "llm_output": {
    "token_usage": {
      "prompt_tokens": 823,
      "completion_tokens": 58,
      "total_tokens": 881
    },
    "model_name": "gpt-3.5-turbo"
  },
  "run": null
}
[chain/end] [1:chain:RetrievalQA > 3:chain:StuffDocumentsChain >
4:chain:LLMChain] [2.86s] Exiting Chain run with output:
{
  "text": "高清电视机的护理非常简单。您只需要使用干布清洁即可。避免使用湿布或化学清洁剂，以免损坏电视机的表面。"
}
[chain/end] [1:chain:RetrievalQA > 3:chain:StuffDocumentsChain] [2.87s] Exiting
Chain run with output:
{
  "output_text": "高清电视机的护理非常简单。您只需要使用干布清洁即可。避免使用湿布或化学清洁剂，以免损坏电视机的表面。"
}
[chain/end] [1:chain:RetrievalQA] [3.26s] Exiting Chain run with output:
{
  "result": "高清电视机的护理非常简单。您只需要使用干布清洁即可。避免使用湿布或化学清洁剂，以免损坏电视机的表面。"
}
```


```
'高清电视机的护理非常简单。您只需要使用干布清洁即可。避免使用湿布或化学清洁剂，以免损坏电视机的表面。'
```

我们可以看到它首先深入到检索 QA 链中，然后它进入了一些文档链。如上所述，我们正在使用 stuff 方法，现在我们正在传递这个上下文，可以看到，这个上下文是由我们检索到的不同文档创建的。因此，在进行问答时，当返回错误结果时，通常不是语言模型本身出错了，实际上是检索步骤出错了，仔细查看问题的确切内容和上下文可以帮助调试出错的原因。

然后，我们可以再向下一级，看看进入语言模型的确切内容，以及 OpenAI 自身，在这里，我们可以看到传递的完整提示，我们有一个系统消息，有所使用的提示的描述，这是问题回答链使用的提示，我们可以看到提示打印出来，使用以下上下文片段回答用户的问题。

如果您不知道答案，只需说您不知道即可，不要试图编造答案。然后我们看到一堆之前插入的上下文，我们还可以看到有关实际返回类型的更多信息。我们不仅仅返回一个答案，还有 token 的使用情况，可以了解到 token 数的使用情况

由于这是一个相对简单的链，我们现在可以看到最终的响应，通过链返回给用户。这部分我们主要讲解了如何查看和调试单个输入到该链的情况。

三、 通过LLM进行评估实例

来简要梳理一下问答评估的流程:

- 首先，我们使用 LLM 自动构建了问答测试集，包含问题及标准答案。
- 然后，同一 LLM 试图回答测试集中的所有问题，得到响应。
- 下一步，需要评估语言模型的回答是否正确。这里奇妙的是，我们再使用另一个 LLM 链进行判断，所以 LLM 既是“球员”，又是“裁判”。

具体来说，第一个语言模型负责回答问题。第二个语言模型链用来进行答案判定。最后我们可以收集判断结果，得到语言模型在这一任务上的效果分数。需要注意的是，回答问题的语言模型链和答案判断链是分开的，职责不同。这避免了同一个模型对自己结果的主观判断。

总之,语言模型可以自动完成构建测试集、回答问题和判定答案等全流程，使评估过程更加智能化和自动化。我们只需要提供文档并解析最终结果即可。

```
langchain.debug = False

#为所有不同的示例创建预测
predictions = qa.apply(examples)

# 对预测的结果进行评估，导入QA问题回答，评估链，通过语言模型创建此链
from langchain.evaluation.qa import QAEvalChain #导入QA问题回答，评估链

#通过调用chatGPT进行评估
llm = ChatOpenAI(temperature=0)
eval_chain = QAEvalChain.from_llm(llm)

#在此链上调用evaluate，进行评估
graded_outputs = eval_chain.evaluate(examples, predictions)
```

```
> Entering new RetrievalQA chain...

> Finished chain.
```

```python
#我们将传入示例和预测，得到一堆分级输出，循环遍历它们打印答案
for i, eg in enumerate(examples):
    print(f"Example {i}:")
    print("Question: " + predictions[i]['query'])
    print("Real Answer: " + predictions[i]['answer'])
    print("Predicted Answer: " + predictions[i]['result'])
    print("Predicted Grade: " + graded_outputs[i]['results'])
    print()
```

```
Example 0:
Question: 高清电视机怎么进行护理？
Real Answer: 使用干布清洁。
Predicted Answer: 高清电视机的护理非常简单。您只需要使用干布清洁即可。避免使用湿布或化学清洁剂，以免损坏电视机的表面。
Predicted Grade: CORRECT

Example 1:
Question: 旅行背包有内外袋吗？
Real Answer: 有。
Predicted Answer: 是的，旅行背包有多个实用的内外袋，可以轻松装下您的必需品。
Predicted Grade: CORRECT

Example 2:
Question: 这款全自动咖啡机的尺寸是多少？
Real Answer: 大型尺寸为13.8'' x 17.3''，中型尺寸为11.5'' x 15.2''。
Predicted Answer: 这款全自动咖啡机有两种尺寸可选：
- 大型尺寸为13.8'' x 17.3''。
- 中型尺寸为11.5'' x 15.2''。
Predicted Grade: CORRECT

Example 3:
Question: 这款电动牙刷的规格是什么？
Real Answer: 一般大小 - 高度：9.5''，宽度：1''。
Predicted Answer: 这款电动牙刷的规格是：高度为9.5英寸，宽度为1英寸。
Predicted Grade: CORRECT

Example 4:
Question: 这种产品的名称是什么？
Real Answer: 这种产品的名称是橙味维生素C泡腾片。
Predicted Answer: 这种产品的名称是儿童益智玩具。
Predicted Grade: INCORRECT

Example 5:
Question: 这款无线蓝牙耳机的尺寸是多少？
Real Answer: 该无线蓝牙耳机的尺寸为1.5'' x 1.3''。
Predicted Answer: 这款无线蓝牙耳机的尺寸是1.5'' x 1.3''。
Predicted Grade: CORRECT

Example 6:
Question: 这款瑜伽垫的尺寸是多少？
Real Answer: 这款瑜伽垫的尺寸是24'' x 68''。
Predicted Answer: 这款瑜伽垫的尺寸是24'' x 68''。
Predicted Grade: CORRECT
```

从上面的返回结果中我们可以看到，在评估结果中每一个问题中都包含了`Question`，`Real Answer`，`Predicted Answer`和`Predicted Grade`四组内容，其中`Real Answer`是有先前的`QAGenerateChain`创建的问答测试集中的答案，而`Predicted Answer`则是由我们的`qa`链给出的答案，最后的`Predicted Grade`则是由上面代码中的`QAEvalChain`回答的。

在本章中，我们学习了使用 LangChain 框架实现 LLM 问答效果自动化评估的方法。与传统手工准备评估集、逐题判断等方式不同，LangChain 使整个评估流程自动化。它可以自动构建包含问答样本的测试集，然后使用语言模型对测试集自动产生回复，最后通过另一个模型链自动判断每个回答的准确性。**这种全自动的评估方式极大地简化了问答系统的评估和优化过程，开发者无需手动准备测试用例，也无需逐一判断正确性，大大提升了工作效率。**

借助LangChain的自动评估功能，我们可以快速评估语言模型在不同文档集上的问答效果，并可以持续地进行模型调优，无需人工干预。这种自动化的评估方法解放了双手，使我们可以更高效地迭代优化问答系统的性能。

总之，自动评估是 LangChain 框架的一大优势，它将极大地降低问答系统开发的门槛，使任何人都可以轻松训练出性能强大的问答模型。

英文版提示

1. 创建LLM应用

```
from langchain.chains import RetrievalQA
from langchain.chat_models import ChatOpenAI
from langchain.document_loaders import CSVLoader
from langchain.indexes import VectorstoreIndexCreator
from langchain.vectorstores import DocArrayInMemorySearch
from langchain.evaluation.qa import QAGenerateChain
import pandas as pd

file = '../data/OutdoorClothingCatalog_1000.csv'
loader = CSVLoader(file_path=file)
data = loader.load()



test_data = pd.read_csv(file,skiprows=0,usecols=[1,2])
display(test_data.head())

llm = ChatOpenAI(temperature = 0.0)

index = VectorstoreIndexCreator(
    vectorstore_cls=DocArrayInMemorySearch
).from_loaders([loader])

qa = RetrievalQA.from_chain_type(
    llm=llm,
    chain_type="stuff",
    retriever=index.vectorstore.as_retriever(),
    verbose=True,
    chain_type_kwargs = {
        "document_separator": "<<<<>>>>>"
    }
)
```

```
print(data[10],"\n")
print(data[11],"\n")

examples = [
    {
        "query": "Do the Cozy Comfort Pullover Set have side pockets?",
        "answer": "Yes"
    },
    {
        "query": "What collection is the Ultra-Lofty 850 Stretch Down Hooded
Jacket from?",
        "answer": "The DownTek collection"
    }
]


example_gen_chain = QAGenerateChain.from_llm(ChatOpenAI())


from langchain.evaluation.qa import QAGenerateChain #导入QA生成链，它将接收文档，并从
每个文档中创建一个问题答案对
example_gen_chain = QAGenerateChain.from_llm(ChatOpenAI())#通过传递chat open AI语言
模型来创建这个链
new_examples = example_gen_chain.apply([{"doc": t} for t in data[:5]])

#查看用例数据
print(new_examples)

examples += [ v for item in new_examples for k,v in item.items()]
qa.run(examples[0]["query"])
```

	name	description
0	Women's Campside Oxfords	This ultracomfortable lace-to-toe Oxford boast...
1	Recycled Waterhog Dog Mat, Chevron Weave	Protect your floors from spills and splashing ...
2	Infant and Toddler Girls' Coastal Chill Swimsu...	She'll love the bright colors, ruffles and exc...
3	Refresh Swimwear, V-Neck Tankini Contrasts	Whether you're going for a swim or heading out...
4	EcoFlex 3L Storm Pants	Our new TEK O2 technology makes our four-seaso...

```
page_content=": 10\nname: Cozy Comfort Pullover Set, Stripe\ndescription: Perfect for lounging, this striped knit set lives up to its name. We used ultrasoft fabric and an easy design that's as comfortable at bedtime as it is when we have to make a quick run out.\n\nSize & Fit\n- Pants are Favorite Fit: Sits lower on the waist.\n- Relaxed Fit: Our most generous fit sits farthest from the body.\n\nFabric & Care\n- In the softest blend of 63% polyester, 35% rayon and 2% spandex.\n\nAdditional Features\n- Relaxed fit top with raglan sleeves and rounded hem.\n- Pull-on pants have a wide elastic waistband and drawstring, side pockets and a modern slim leg.\n\nImported." metadata={'source': '../data/OutdoorClothingCatalog_1000.csv', 'row': 10}

page_content=': 11\nname: Ultra-Lofty 850 Stretch Down Hooded Jacket\ndescription: This technical stretch down jacket from our DownTek collection is sure to keep you warm and comfortable with its full-stretch construction providing exceptional range of motion. With a slightly fitted style that falls at the hip and best with a midweight layer, this jacket is suitable for light activity up to 20° and moderate activity up to -30°. The soft and durable 100% polyester shell offers complete windproof protection and is insulated with warm, lofty goose down. Other features include welded baffles for a no-stitch construction and excellent stretch, an adjustable hood, an interior media port and mesh stash pocket and a hem drawcord. Machine wash and dry. Imported.' metadata={'source': '../data/OutdoorClothingCatalog_1000.csv', 'row': 11}

[{'qa_pairs': {'query': "What is the description of the Women's Campside Oxfords?", 'answer': "The description of the Women's Campside Oxfords is that they are an ultracomfortable lace-to-toe Oxford made of super-soft canvas. They have thick cushioning and quality construction, providing a broken-in feel from the first time they are worn."}}, {'qa_pairs': {'query': 'What are the dimensions of the small and medium sizes of the Recycled Waterhog Dog Mat, Chevron Weave?', 'answer': 'The dimensions of the small size of the Recycled Waterhog Dog Mat, Chevron Weave are 18" x 28". The dimensions of the medium size are 22.5" x 34.5".'}}, {'qa_pairs': {'query': "What are the features of the Infant and Toddler Girls' Coastal Chill Swimsuit, Two-Piece?", 'answer': "The swimsuit has bright colors, ruffles, and exclusive whimsical prints. It is made of four-way-stretch and chlorine-resistant fabric, which keeps its shape and resists snags. The fabric is UPF 50+ rated, providing the highest rated sun protection possible by blocking 98% of the sun's harmful rays. The swimsuit also has crossover no-slip straps and a fully lined bottom for a secure fit and maximum coverage."}}, {'qa_pairs': {'query': 'What is the fabric composition of the Refresh Swimwear, V-Neck Tankini Contrasts?', 'answer': 'The Refresh Swimwear, V-Neck Tankini Contrasts is made of 82% recycled nylon and 18% Lycra® spandex for the body, and 90% recycled nylon with 10% Lycra® spandex for the lining.'}}, {'qa_pairs': {'query': 'What is the fabric composition of the EcoFlex 3L Storm Pants?', 'answer': 'The EcoFlex 3L Storm Pants are made of 100% nylon, exclusive of trim.'}}]
```

```
> Entering new RetrievalQA chain...

> Finished chain.
```

```
'Yes, the Cozy Comfort Pullover Set does have side pockets.'
```

2. 人工评估

```
import langchain
langchain.debug = True

#重新运行与上面相同的示例，可以看到它开始打印出更多的信息
qa.run(examples[0]["query"])

langchain.debug = False
```

```
[chain/start] [1:chain:RetrievalQA] Entering Chain run with input:
{
  "query": "Do the Cozy Comfort Pullover Set have side pockets?"
}
[chain/start] [1:chain:RetrievalQA > 3:chain:StuffDocumentsChain] Entering Chain
run with input:
[inputs]
[chain/start] [1:chain:RetrievalQA > 3:chain:StuffDocumentsChain >
4:chain:LLMChain] Entering Chain run with input:
{
  "question": "Do the Cozy Comfort Pullover Set have side pockets?",
  "context": ": 10\nname: Cozy Comfort Pullover Set, Stripe\ndescription: Perfect
for lounging, this striped knit set lives up to its name. We used ultrasoft
fabric and an easy design that's as comfortable at bedtime as it is when we have
to make a quick run out.\n\nSize & Fit\n- Pants are Favorite Fit: Sits lower on
the waist.\n- Relaxed Fit: Our most generous fit sits farthest from the
body.\n\nFabric & Care\n- In the softest blend of 63% polyester, 35% rayon and 2%
spandex.\n\nAdditional Features\n- Relaxed fit top with raglan sleeves and
rounded hem.\n- Pull-on pants have a wide elastic waistband and drawstring, side
pockets and a modern slim leg.\n\nImported.<<<<>>>>>: 73\nname: Cozy Cuddles Knit
Pullover Set\ndescription: Perfect for lounging, this knit set lives up to its
name. We used ultrasoft fabric and an easy design that's as comfortable at
bedtime as it is when we have to make a quick run out. \n\nSize & Fit \nPants are
Favorite Fit: Sits lower on the waist. \nRelaxed Fit: Our most generous fit sits
farthest from the body. \n\nFabric & Care \nIn the softest blend of 63%
polyester, 35% rayon and 2% spandex.\n\nAdditional Features \nRelaxed fit top
with raglan sleeves and rounded hem. \nPull-on pants have a wide elastic
waistband and drawstring, side pockets and a modern slim leg. \nImported.
<<<<>>>>>: 151\nname: Cozy Quilted Sweatshirt\ndescription: Our sweatshirt is an
instant classic with its great quilted texture and versatile weight that easily
transitions between seasons. With a traditional fit that is relaxed through the
chest, sleeve, and waist, this pullover is lightweight enough to be worn most
months of the year. The cotton blend fabric is super soft and comfortable, making
it the perfect casual layer. To make dressing easy, this sweatshirt also features
a snap placket and a heritage-inspired Mt. Katahdin logo patch. For care, machine
wash and dry. Imported.<<<<>>>>>: 265\nname: Cozy Workout Vest\ndescription: For
serious warmth that won't weigh you down, reach for this fleece-lined vest, which
provides you with layering options whether you're inside or outdoors.\nSize &
Fit\nRelaxed Fit. Falls at hip.\nFabric & Care\nSoft, textured fleece lining.
Nylon shell. Machine wash and dry. \nAdditional Features \nTwo handwarmer
pockets. Knit side panels stretch for a more flattering fit. Shell fabric is
treated to resist water and stains. Imported."
}
```

```
[llm/start] [1:chain:RetrievalQA > 3:chain:StuffDocumentsChain > 4:chain:LLMChain
> 5:llm:ChatOpenAI] Entering LLM run with input:
{
  "prompts": [
    "System: Use the following pieces of context to answer the users question.
\nIf you don't know the answer, just say that you don't know, don't try to make
up an answer.\n----------------\n: 10\nname: Cozy Comfort Pullover Set,
Stripe\ndescription: Perfect for lounging, this striped knit set lives up to its
name. We used ultrasoft fabric and an easy design that's as comfortable at
bedtime as it is when we have to make a quick run out.\n\nSize & Fit\n- Pants are
Favorite Fit: Sits lower on the waist.\n- Relaxed Fit: Our most generous fit sits
farthest from the body.\n\nFabric & Care\n- In the softest blend of 63%
polyester, 35% rayon and 2% spandex.\n\nAdditional Features\n- Relaxed fit top
with raglan sleeves and rounded hem.\n- Pull-on pants have a wide elastic
waistband and drawstring, side pockets and a modern slim leg.\n\nImported.
<<<<>>>>>: 73\nname: Cozy Cuddles Knit Pullover Set\ndescription: Perfect for
lounging, this knit set lives up to its name. We used ultrasoft fabric and an
easy design that's as comfortable at bedtime as it is when we have to make a
quick run out. \n\nSize & Fit \nPants are Favorite Fit: Sits lower on the waist.
\nRelaxed Fit: Our most generous fit sits farthest from the body. \n\nFabric &
Care \nIn the softest blend of 63% polyester, 35% rayon and 2%
spandex.\n\nAdditional Features \nRelaxed fit top with raglan sleeves and rounded
hem. \nPull-on pants have a wide elastic waistband and drawstring, side pockets
and a modern slim leg. \nImported.<<<<>>>>>: 151\nname: Cozy Quilted
Sweatshirt\ndescription: Our sweatshirt is an instant classic with its great
quilted texture and versatile weight that easily transitions between seasons.
With a traditional fit that is relaxed through the chest, sleeve, and waist, this
pullover is lightweight enough to be worn most months of the year. The cotton
blend fabric is super soft and comfortable, making it the perfect casual layer.
To make dressing easy, this sweatshirt also features a snap placket and a
heritage-inspired Mt. Katahdin logo patch. For care, machine wash and dry.
Imported.<<<<>>>>>: 265\nname: Cozy Workout Vest\ndescription: For serious warmth
that won't weigh you down, reach for this fleece-lined vest, which provides you
with layering options whether you're inside or outdoors.\nSize & Fit\nRelaxed
Fit. Falls at hip.\nFabric & Care\nSoft, textured fleece lining. Nylon shell.
Machine wash and dry. \nAdditional Features \nTwo handwarmer pockets. Knit side
panels stretch for a more flattering fit. Shell fabric is treated to resist water
and stains. Imported.\nHuman: Do the Cozy Comfort Pullover Set have side
pockets?"
  ]
}
[llm/end] [1:chain:RetrievalQA > 3:chain:StuffDocumentsChain > 4:chain:LLMChain >
5:llm:ChatOpenAI] [879.746ms] Exiting LLM run with output:
{
  "generations": [
    [
      {
        "text": "Yes, the Cozy Comfort Pullover Set does have side pockets.",
        "generation_info": {
          "finish_reason": "stop"
        },
        "message": {
          "lc": 1,
          "type": "constructor",
          "id": [
            "langchain",
```

```
            "schema",
            "messages",
            "AIMessage"
          ],
          "kwargs": {
            "content": "Yes, the Cozy Comfort Pullover Set does have side
pockets.",
            "additional_kwargs": {}
          }
        }
      }
    ]
  ],
  "llm_output": {
    "token_usage": {
      "prompt_tokens": 626,
      "completion_tokens": 14,
      "total_tokens": 640
    },
    "model_name": "gpt-3.5-turbo"
  },
  "run": null
}
[chain/end] [1:chain:RetrievalQA > 3:chain:StuffDocumentsChain >
4:chain:LLMChain] [880.5269999999999ms] Exiting Chain run with output:
{
  "text": "Yes, the Cozy Comfort Pullover Set does have side pockets."
}
[chain/end] [1:chain:RetrievalQA > 3:chain:StuffDocumentsChain]
[881.4499999999999ms] Exiting Chain run with output:
{
  "output_text": "Yes, the Cozy Comfort Pullover Set does have side pockets."
}
[chain/end] [1:chain:RetrievalQA] [1.21s] Exiting Chain run with output:
{
  "result": "Yes, the Cozy Comfort Pullover Set does have side pockets."
}
```


3. 通过LLM进行评估实例

```
langchain.debug = False

#为所有不同的示例创建预测
predictions = qa.apply(examples)

# 对预测的结果进行评估，导入QA问题回答，评估链，通过语言模型创建此链
from langchain.evaluation.qa import QAEvalChain #导入QA问题回答，评估链

#通过调用chatGPT进行评估
llm = ChatOpenAI(temperature=0)
eval_chain = QAEvalChain.from_llm(llm)

#在此链上调用evaluate，进行评估
graded_outputs = eval_chain.evaluate(examples, predictions)
```

```python
#我们将传入示例和预测，得到一堆分级输出，循环遍历它们打印答案
for i, eg in enumerate(examples):
    print(f"Example {i}:")
    print("Question: " + predictions[i]['query'])
    print("Real Answer: " + predictions[i]['answer'])
    print("Predicted Answer: " + predictions[i]['result'])
    print("Predicted Grade: " + graded_outputs[i]['results'])
    print()
```

```
> Entering new RetrievalQA chain...

> Finished chain.
Example 0:
Question: Do the Cozy Comfort Pullover Set have side pockets?
Real Answer: Yes
Predicted Answer: Yes, the Cozy Comfort Pullover Set does have side pockets.
Predicted Grade: CORRECT

Example 1:
Question: What collection is the Ultra-Lofty 850 Stretch Down Hooded Jacket from?
Real Answer: The DownTek collection
Predicted Answer: The Ultra-Lofty 850 Stretch Down Hooded Jacket is from the
DownTek collection.
Predicted Grade: CORRECT

Example 2:
Question: What is the description of the Women's Campside Oxfords?
Real Answer: The description of the Women's Campside Oxfords is that they are an
ultracomfortable lace-to-toe Oxford made of super-soft canvas. They have thick
cushioning and quality construction, providing a broken-in feel from the first
time they are worn.
Predicted Answer: The description of the Women's Campside Oxfords is: "This
ultracomfortable lace-to-toe Oxford boasts a super-soft canvas, thick cushioning,
and quality construction for a broken-in feel from the first time you put them
on."
Predicted Grade: CORRECT

Example 3:
Question: What are the dimensions of the small and medium sizes of the Recycled
Waterhog Dog Mat, Chevron Weave?
Real Answer: The dimensions of the small size of the Recycled Waterhog Dog Mat,
Chevron Weave are 18" x 28". The dimensions of the medium size are 22.5" x 34.5".
Predicted Answer: The dimensions of the small size of the Recycled Waterhog Dog
Mat, Chevron Weave are 18" x 28". The dimensions of the medium size are 22.5" x
34.5".
Predicted Grade: CORRECT

Example 4:
Question: What are the features of the Infant and Toddler Girls' Coastal Chill
Swimsuit, Two-Piece?
```

Real Answer: The swimsuit has bright colors, ruffles, and exclusive whimsical prints. It is made of four-way-stretch and chlorine-resistant fabric, which keeps its shape and resists snags. The fabric is UPF 50+ rated, providing the highest rated sun protection possible by blocking 98% of the sun's harmful rays. The swimsuit also has crossover no-slip straps and a fully lined bottom for a secure fit and maximum coverage.
Predicted Answer: The features of the Infant and Toddler Girls' Coastal Chill Swimsuit, Two-Piece are:

- Bright colors and ruffles
- Exclusive whimsical prints
- Four-way-stretch and chlorine-resistant fabric
- UPF 50+ rated fabric for sun protection
- Crossover no-slip straps
- Fully lined bottom for a secure fit and maximum coverage
- Machine washable and line dry for best results
- Imported
Predicted Grade: CORRECT

Example 5:
Question: What is the fabric composition of the Refresh Swimwear, V-Neck Tankini Contrasts?
Real Answer: The Refresh Swimwear, V-Neck Tankini Contrasts is made of 82% recycled nylon and 18% Lycra® spandex for the body, and 90% recycled nylon with 10% Lycra® spandex for the lining.
Predicted Answer: The fabric composition of the Refresh Swimwear, V-Neck Tankini Contrasts is 82% recycled nylon with 18% Lycra® spandex for the body, and 90% recycled nylon with 10% Lycra® spandex for the lining.
Predicted Grade: CORRECT

Example 6:
Question: What is the fabric composition of the EcoFlex 3L Storm Pants?
Real Answer: The EcoFlex 3L Storm Pants are made of 100% nylon, exclusive of trim.
Predicted Answer: The fabric composition of the EcoFlex 3L Storm Pants is 100% nylon, exclusive of trim.
Predicted Grade: CORRECT