LLM大语言模型应用

田兴, 2023/7/1

摘要

(1)LLM 大模型最容易落地的应用是 document-QA，即根据产品说明书回答用户问题.
(2)接下来计划基于 LangChain 开发 document-QA 的演示项目.

目前对 LLM 应用最火的方案

三个方案, 第一个最重要.

(1)基于LLM的document-QA

智能客服机器人的QA问答场景包括: FAQ，Doc-QA，KBQA，TableQA，TaskQA.
其中，document-QA（即: Doc-QA）就是机器人根据产品说明书或操作手册回答用户的问题.

document-QA 以前的实现方法:
(1)从用户句子中提取:
* 产品名称: 即，核心词，锚定用户问题的主体.
* 用户意图: 即，场景词，识别用户的退货,换货,保修,等等大类别的意图.
(2)根据产品与用户意图可以从该产品文档中检索出相关信息. 然后执行 "阅读理解" 算法，根据文章段落和用户问题获取答案.

其中的问题:
(1) "阅读理解" 算法一般采用 "抽取" 或 "生成"，抽取的答案过于死板，生成的答案不可靠不安全. 其准确率不高，因此一般客服机器人主流的模块是 FAQ，TaskQA，KBQA.

document-QA 现在的实现方法:
即，将 "阅读理解" 算法部分替换成 ChatGPT 一类的 LLM.
(1)相关文章检测.
* 目前网络上提倡的是段落向量化，并通过向量检索. 这种方法不精准，不相关的上下文会降低答案的质量.
* 建议的方法是采用原来的传统方法. 根据实体，意图精准召回.
(2)LLM阅读.
* 将召回的段落填充到 prompt 中，让 LLM 生成答案.

优势:
(1)主要优势在于 LLM 的 document-QA 准确率更高，生成句子更流畅.
(2)产品说明书，操作手册，是一定会有的数据，降低了制作 FAQ 或对话流程的人工成本.

参考链接:
https://mp.weixin.qq.com/s/movaNCWjJGBaes6KxhpYpg

(2)CoT思维链,LLM会使用工具

大概是作一个机器人，LLM是它的大脑，通过固定流程设置 prompt，使其具备对问题的推理能力.

以下示例中，机器人可以访问浏览器，机器人通过 Thought，Action，Observation 三步，不断循环，最终得出答案.

问题:
What was the high temperature in SF yesterday in Fahrenheit? What is that number raised to the .023 power?

机器人思维过程:
Thought: I need to find the temperature first, then use the calculator to raise it to the .023 power.
Action: Search
Action Input: "High temperature in SF yesterday"
Observation: San Francisco Temperature Yesterday. Maximum temperature yesterday: 57 °F (at 1:56 pm) Minimum temperature yesterday: 49 °F (at 1:56 am) Average temperature ...

Thought: I now have the temperature, so I can use the calculator to raise it to the .023 power.
Action: Calculator
Action Input: 57^.023
Observation: Answer: 1.0974509573251117

Thought: I now know the final answer
Final Answer: The high temperature in SF yesterday in Fahrenheit raised to the .023 power is 1.0974509573251117.

最终答案:
The high temperature in SF yesterday in Fahrenheit raised to the .023 power is 1.0974509573251117.

备注:
这是 LangChain 官方的一个示例，还不清楚其具体实现过程.

参考链接:
https://arxiv.org/abs/2305.17390
https://python.langchain.com/docs/get_started/quickstart

(3)基于LLM和知识库,让大模型会自我学习

LLM大模型就相当于人的思考能力，知识库就相当于人的记忆.

步骤:
(1)LLM 与人互动的同时会得出一些答案，即会推理出一些事实.
(2)给与 LLM 能力，它可以决定是否将新产生的知识存储到知识库.
(3)LLM 在回答用户问题时，可以访问知识库，它也能访问到自己曾经存入的知识.

(4)LLM 根据从知识库中检索到的相关资料给用户输出答案.
以上过程中，LLM 在对话的过程中会不断修改知识库，即它的记忆改变了.

备注:
(1)LLM 与知识库的互交过程中，所产生的操作日志，可用于分析 LLM 为什么会有某些的行为. 这使大模型的可解释性增强.
(2)prompt 中的上下文相当于 LLM 的短期记忆，知识库中的资料，相当于LLM的长期记忆.
(3)他们认为知识在人的大脑中是向量形式存储的，所以向量数据库现在发展也很火.

参考链接:
https://new.qq.com/rain/a/20230426A08Y3T00