WebGPT: Browser-assisted question-answering with human feedback

基于 GPT-3，用SFT + RLHF训练 TOOL-AUGMENTED QA AGENT

Reiichiro Nakano* Jacob Hilton* Suchir Balaji* Jeff Wu Long Ouyang

Christina Kim Christopher Hesse Shantanu Jain Vineet Kosaraju

William Saunders Xu Jiang Karl Cobbe Tyna Eloundou Gretchen Krueger

Kevin Button Matthew Knight Benjamin Chess John Schulman

OpenAI

简介

WebGPT是基于GPT-3构建的早期Tool-Augmented QA Agent，通过“行为克隆（SFT）+ 人类偏好优化（RLHF）”两阶段训练，让模型能先浏览网页，再回答复杂问题。第一个创新点在于构建了文本版浏览器环境，即用语言描述网页状态和交互记录，然后将这些信息拼接进prompt 中，让模型“看到”它当前所处的浏览环境。第二个创新点模型在回答问题时必须提供参考资料（references, Answer with Citations），以帮助人类评估回答是否符合事实。

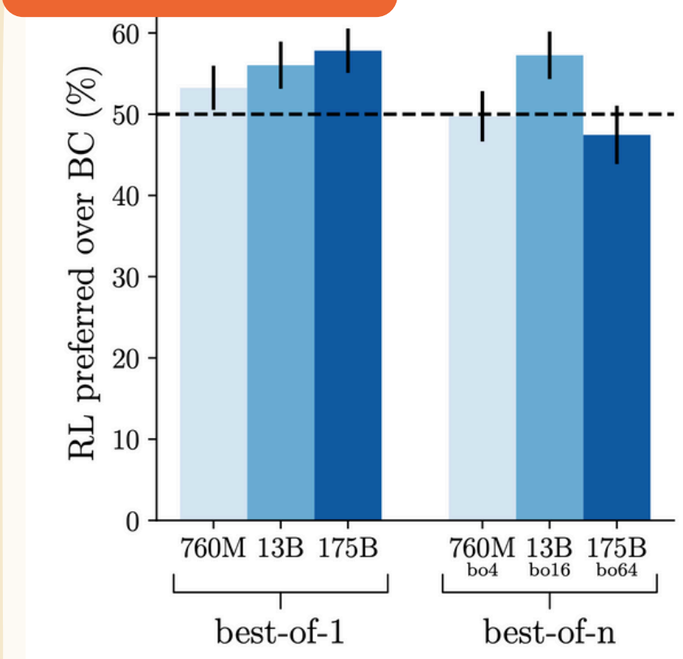
背景

在LLM刚刚展现出强大语言生成能力的2021年，GPT-3虽然能写出通顺的文本，但在面对长篇问答任务时幻觉问题还是比较严重的。OpenAI提出的WebGPT是一次大胆尝试：让GPT-3学会使用Bing搜索引擎，主动上网查资料，再来回答问题。它模拟了一个“文本版浏览器环境”，教模型像人一样搜索、点击网页、引用段落，并通过模仿人类行为（SFT） + 人类反馈优化（RLHF），成为早期具备“查→读→答”能力的工具增强型智能体。

实验设置

- 任务：long-form QA (LFQA, 长篇问答)
- 数据集：ELI5
- 模型：GPT-3 760M, 13B, 175B
- 两阶段训练：行为克隆（SFT）+ RLHF
- 不同训练阶段使用互不重叠的数据
- RL算法：PPO

SFT VS RLHF



Command	Effect
`Search <query>`	Send `<query>` to the Bing API and display a search results page
`Clicked on link <link ID>`	Follow the link with the given ID to a new page
`Find in page: <text>`	Find the next occurrence of `<text>` and scroll to it
`Quote: <text>`	If `<text>` is found in the current page, add it as a reference
`Scrolled down <1, 2, 3>`	Scroll down a number of times
`Scrolled up <1, 2, 3>`	Scroll up a number of times
`Top`	Scroll to the top of the page
`Back`	Go to the previous page
`End: Answer`	End browsing and move to answering phase
`End: <Nonsense, Controversial>`	End browsing and skip answering phase

虽然TOOL只有搜索引擎一个，但是ACTION真不少

RL模型在不使用拒绝采样的情况下，稍微提升了SFT模型的偏好评分。
但如果使用了拒绝采样，这个提升就不明显了。

思考

主要思考下WEBGPT和TIR的一些差异，首先GPT-3不是REASONING MODEL，通过SFT+RLHF的方式训练，，其次WEBGPT的GENERATION流程是预定义好的，只有两个阶段，先BROWSING再ANSWER，它不能像RLVR训练的TIR那样在TOOL CALLING之后可以再回到THINKING模式，换句话说，WEBGPT 的生成流程是“先浏览 ⇝ 后回答”，而不是“交替REASONING与TOOL调用”。