arxiv.org/abs/2410.03608

TICKING ALL THE BOXES: GENERATED CHECKLISTS IMPROVE LLM EVALUATION AND GENERATION

自动化checklist生成以及基于checklist进行response refine

Jonathan Cook*
FLAIR, University of Oxford
jonathan.cook2@hertford.ox.ac.uk

Tim Rocktäschel
Centre for AI, University College London

Jakob Foerster
FLAIR, University of Oxford

Dennis Aumiller† & Alex Wang†
Cohere

未开源

简介

本文提出TICK(Targeted Instruct-evaluation with Checklist)方法，不需要人工参与依赖llm就能评估llm的指令遵循能力，效果也不错还可解释。简单来说，就是让llm根据用户指令(instruction)自动提取一组checklist(是/否问题)，再让llm根据response对照checklist，逐项检查response是否满足这些要求，并统计通过率作为评分。这样不但给出response评分还清晰的说明了“哪些要求满足了、哪些没满足”。进一步作者提出了STICK(Self-TICK)方法，即将TICK用于llm response的refinement，简单说，让llm对自己的初始response用checklist打分然后将这些打分结果作为反馈，引导llm重新生成response，这个过程可以迭代多轮，类似于llm在“照着打分表查漏补缺”。此外，STICK还可用于Best-of-N场景，即对生成的多个候选response用 checklist打分，选出最优的response。

背景

如何高效评估llm指令跟随的能力? 传统的评估方式包括人工偏好排序(preference comparison)、打分以及Elo、LLM-as-a-Judge等机制，都或多或少存在缺陷：1) 这些方法将复杂多维度的指令目标压缩成一个数值，缺乏可解释性；2) 人工标注成本太高而使用llm自动打分又存在稳定性与一致性问题。

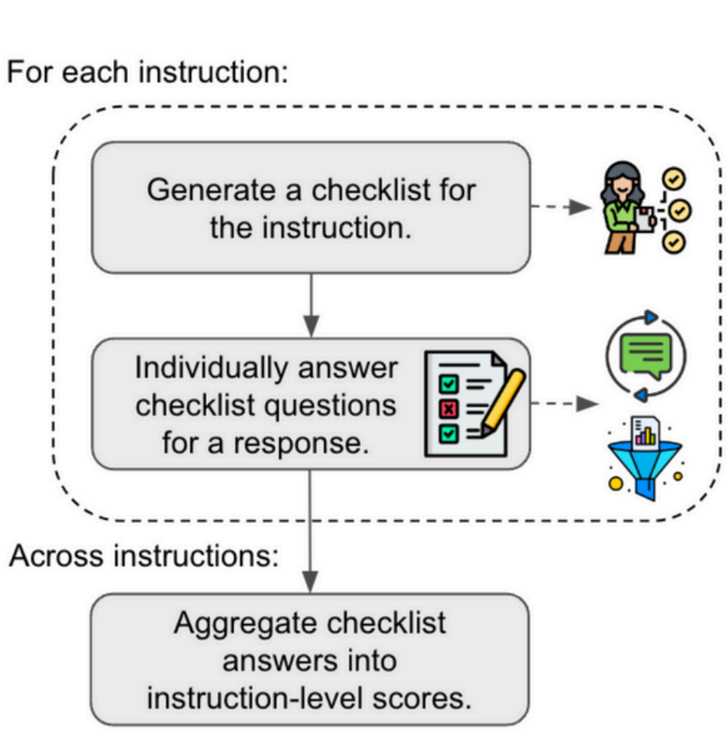
于是，有工作提出了checklist-based evaluation(基于清单的逐项评估)，本文在此基础上提出了基于llm的自动化的TICK/STICK框架，分别作用于自动指令跟随评估和llm自我改进。

实验设置

- 生成checklist以及打分的llm：GPT-4o、Command-R+、LLaMA 3.1-70B
- 实验数据集：InFoBench、WildBench、LiveBench
- 为了评估llm生成的checklist质量，作者找了一些instruction，将llm生成的checklist和人撰写的checklist作对比，计算BLUE、ROUGE指标

TICK

checklist就是一堆yes/no的问题，每个问题对应指令中的一个具体要求，用来判断llm的response是否满足这个要求，全部满足才能说明回答真正符合指令，从而使模型评估变得更可解释、更可靠。TICK方法的核心就是让llm自动生成这样的 checklist，并依此自动打分。



Instruction

Summarize the first season of the TV Series “Mr. Robot” using a similar tone to the show’s main character, Elliot Alderson. Use subheadings and bullet points for formatting, and stay under 500 words.

GPT-4o Generated Checklist

- Is the response a summary of the first season of the TV series “Mr. Robot”?
- Is the response written in a tone similar to the show’s main character, Elliot Alderson?
- Does the response use subheadings and bullet points for formatting?
- Is the response under 500 words?

思考

checklist方法不是本文提出来的，本文的创新是提出让llm自动生成checklist自己打分，然后验证llm生成的checklist和人写的差不多，进一步，作者想到既然checklist可以细致的评分，是否可以将这种反馈用于llm refine response呢？于是提出了STICK，实验效果也不错。