Large Language Models are Zero-Shot Reasoners

https://arxiv.org/abs/2205.11916

Problems

Few-shot CoT improves the reasoning ability, resulting in more accuracy of LLM resoning. But its difficult to create carefully-crafted and task-specific step-by-step examples.

The zero-shot fundamental capabilities of LLMs remains untapped and understudied in current research.

Contributions

A single prompt across diverse reasoning tasks to invoke the LLM reasoning ability.

Experimental results demonstrate that LLMs are zero-shot reasoners.

Methodology

提出了使用提示文本（不像CoT的例子那样，且与任务无关）引发LLM产生多跳推理。进而生成与CoT类似的推理过程和答案。

两阶段提示：1. 生成推理过程，2. 问题+步骤1生成的推理过程询问答案。

均使用贪婪解码来获取结果。

Experiment：

4类推理数据集：

算数推理（6）：

SingleEq

AddSub

MultiArith

AQUARAT

GSM8K

SVAMP

常识推理（2）：

CommonsenseQA

StrategyQA

符号推理（2）：

Last Letter Concatenation

Coin Flip

其他逻辑推理（2）：

Date Understanding

Tracking Shuffled Objects

17 个LLMs

Baseline: zero-shot CoT VS zero-shot

Also compared with few-shot, few-shot CoT (which are quoted from CoT paper).To keep fair, each seed will only run once.

Answer cleaning: 算术题：如果答案有两个，选择第一个数字。“probably 375 and 376”。选择题：第一个大写字母。匹配"The answer is "后的内容作为答案。如果没有"The answer is "，则从后匹配要求格式的文本作为答案。

Evaluation

More importantly, we observe that many generated chain of thought themselves are surprisingly logically correct or only contains human-understandable mistakes

zero-shot CoT对小模型无效，与CoT保持一致。

一些情况下，zero-shot CoT很难将问题缩小到一个可以推理的问题，而是输出多个答案对应多种情况。

不同模型size下，zero-shot CoT是否会随着模型增大逐渐接近few-shot CoT？

First, Zero-shot-CoT tends to output unnecessary steps of reasoning after getting the correct prediction, which results in changing the prediction to incorrect one. Zero-shot-CoT also sometimes does not start reasoning, just rephrasing the input question. In contrast, Few-shot-CoT tend to fail when generated chain of thought include ternary operation, e.g. (3 + 2) ∗ 4

鼓励推导，阻止误导，阻止无关推理。

可以借鉴的：

对于LLM的不规则输出，如何验证答案的正确性

对于few-shot 的数据集使用和实验设计，如何保证未来实验的公平。

给出了一些很好的思路，但是感觉并没有深入的解释为什么有不同的情况，和探索方法如何避免，例如不产生推理；很难将问题缩小的一个合理范围进行推理；给出正确的预判然后继续产生了误导的推理。。。