ASCII讨论班ChatGPT特辑 第二期

ChatGPT 相关原理概念介绍

张琬悦

2023/2/24

目录

CONTENTS

23

what are the differences between Instruct Learning and prompt learning?

In instructional design and education, prompt learning and instruct learning are two different teaching strategies that serve distinct purposes.

Prompt learning is a teaching method that involves presenting a cue or prompt to the learner, followed by a response from the learner. The prompt is usually a question or a problem that the learner must solve, and the response is the learner's attempt to answer or solve the prompt. Prompt learning is often used to test or assess the learner's knowledge or understanding of a topic.

Instruct learning, on the other hand, is a teaching method that involves presenting information or instruction to the learner, followed by the learner's attempt to apply or use that information. The instruction may take the form of a lecture, demonstration, or explanation, and the learner is expected to apply the knowledge or skills they have learned to complete a task or solve a problem.

In summary, prompt learning focuses on testing or assessing the learner's knowledge or understanding, while instruct learning focuses on teaching and application of knowledge or skills. Both strategies can be effective, depending on the learning goals and the context in which they are used.

23

No, i mean what are the differences between instruct tuning and prompt tuning?

I apologize for misunderstanding your question earlier. In the context of machine learning, "instruct tuning" and "prompt tuning" are not widely used terms, and it's possible that you may be referring to more specific concepts that use these terms. However, based on my understanding of machine learning techniques and terminology, I will try to provide an answer to your question.

I believe you may be referring to "supervised learning" and "reinforcement learning", which are two different approaches to machine learning that involve different types of feedback.

Supervised learning, also known as "instructional learning", is a type of machine learning where a model is trained using labeled data. The labeled data consists of input examples paired with their corresponding output or target values. The goal of supervised learning is for the model to learn the relationship between the input and output data, so that it can accurately predict the output for new, unseen inputs.

Reinforcement learning, also known as "prompt learning", is a type of machine learning where a model learns from feedback in the form of rewards or punishments. The model is trained to make a sequence of decisions in an environment, and receives rewards or punishment Regenerate response of its decisions. The goal of reinforcement learning is for the model to learn the optimal sequence of actions that

提示 Prompt

Prompt:是伴随着输入一起，给予模型的一种上下文。

cloze prompt：填充答案的位置在句中；
prefix promp： 填充答案的位置在句末。

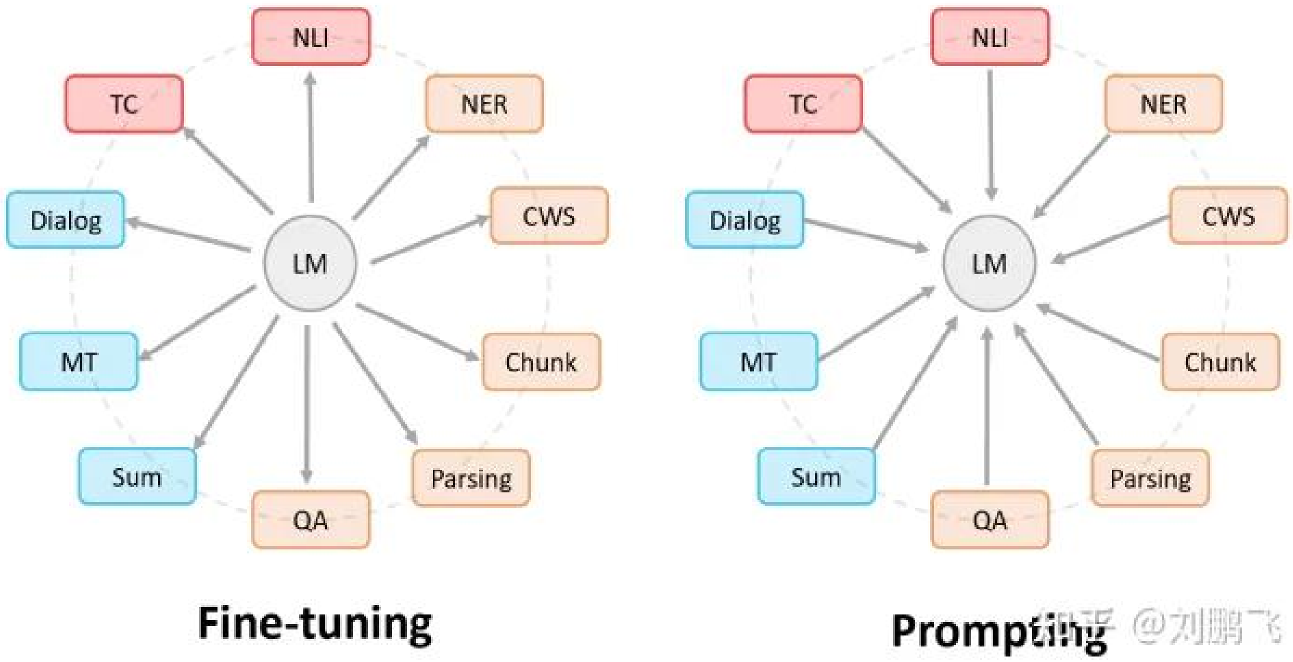
zero-shot learning：0个例子
one-shot learning：1个例子
few-shot learning：多个例子

```
Translate English to French: /* task description */
cheese => ___ /* prompt */
```

```
Translate English to French: /* task description */
sea otter => loutre de mer /* example */
cheese => ___ /* prompt */
```

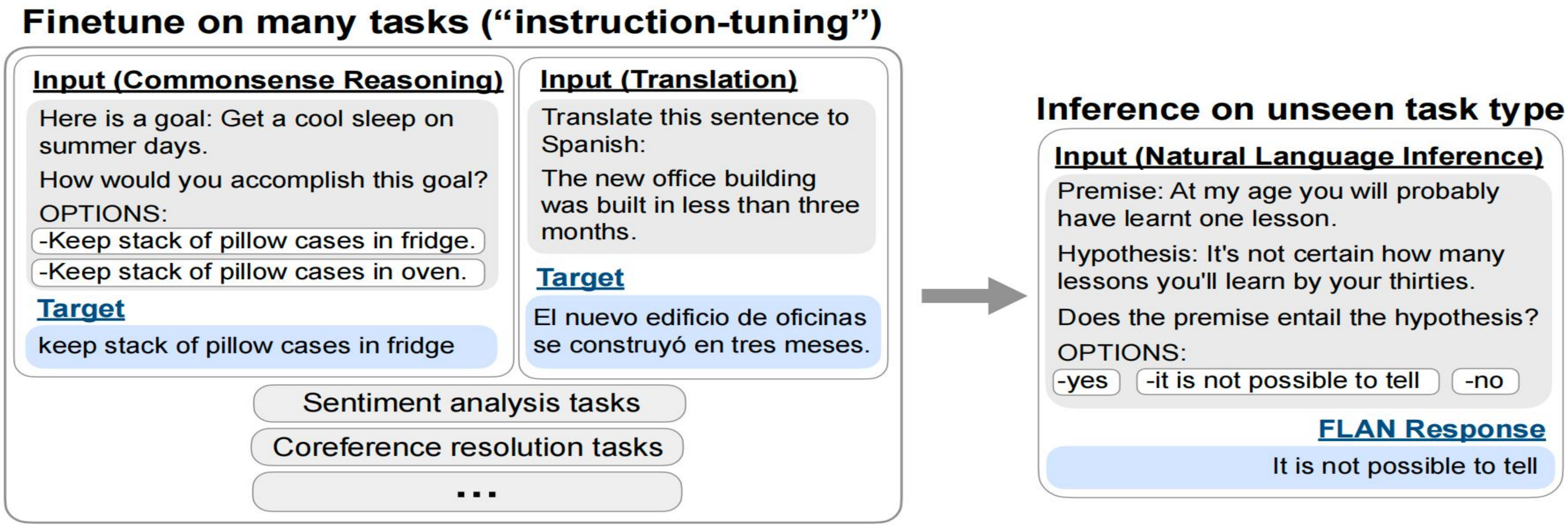
提示学习 Prompt tuning

提示学习(Prompt Learning)和Fine Tuning都是对预训练模型进行微调的方法。
提示学习的灵感源于GPT-3．所谓Prompt Tuning，就是在Prompt中插入一段task-specific的可以tune的prompt token。由于这个token对于每个任务都是不同的，所以可以帮助机器识别任务到底是什么。又因为机器自己学习（tune）这个prompt token，所以这个token对于机器会有非常好的效果。



指示学习 Instruct tuning

指示学习是谷歌Deepmind的Quoc V.Le团队在2021年的一篇名为《Finetuned Language Models Are Zero-Shot Learners》文章中提出的思想。提出一种基于instruction-tuning的方法叫做FLAN（Finetuned LAnguage Net）。对所有的NLP task，根据其任务类型和目标划分若干个簇，随机挑选一个簇内的所有task作为评估，其他所有簇的task用于instruction-tuning。

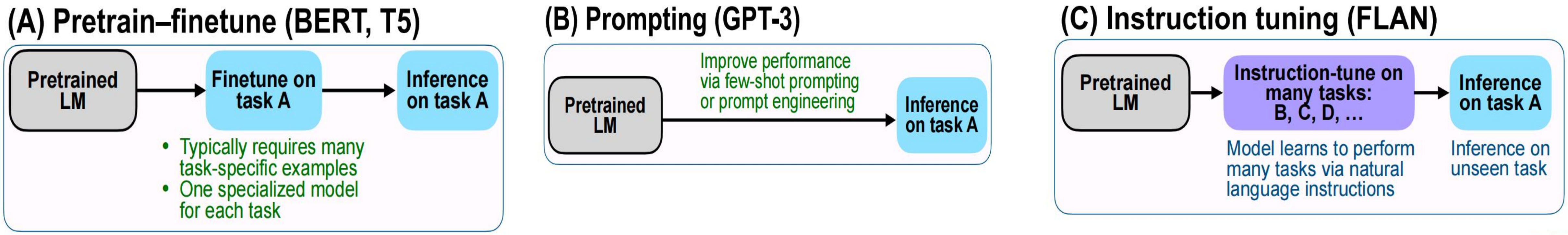


指示学习 Instruct tuning

Fine-tuning：先在大规模语料上进行预训练，然后再在某个下游任务上进行微调，如BERT、T5；

Prompt-tuning：先选择某个通用的大规模预训练模型，然后为具体的任务生成一个prompt模板以适应大模型进行微调，如GPT-3；

Instruction-tuning：仍然在预训练语言模型的基础上，先在多个已知任务上进行微调（通过自然语言的形式），然后再推理某个新任务上进行zero-shot。



指示学习 Instruct tuning

指示学习和提示学习的目的都是去挖掘语言模型本身具备的知识。不同的是Prompt是激发语言模型的补全能力，例如根据上半句生成下半句，或是完形填空等。Instruct是激发语言模型的理解能力，它通过给出更明显的指令，让模型去做出正确的行动。

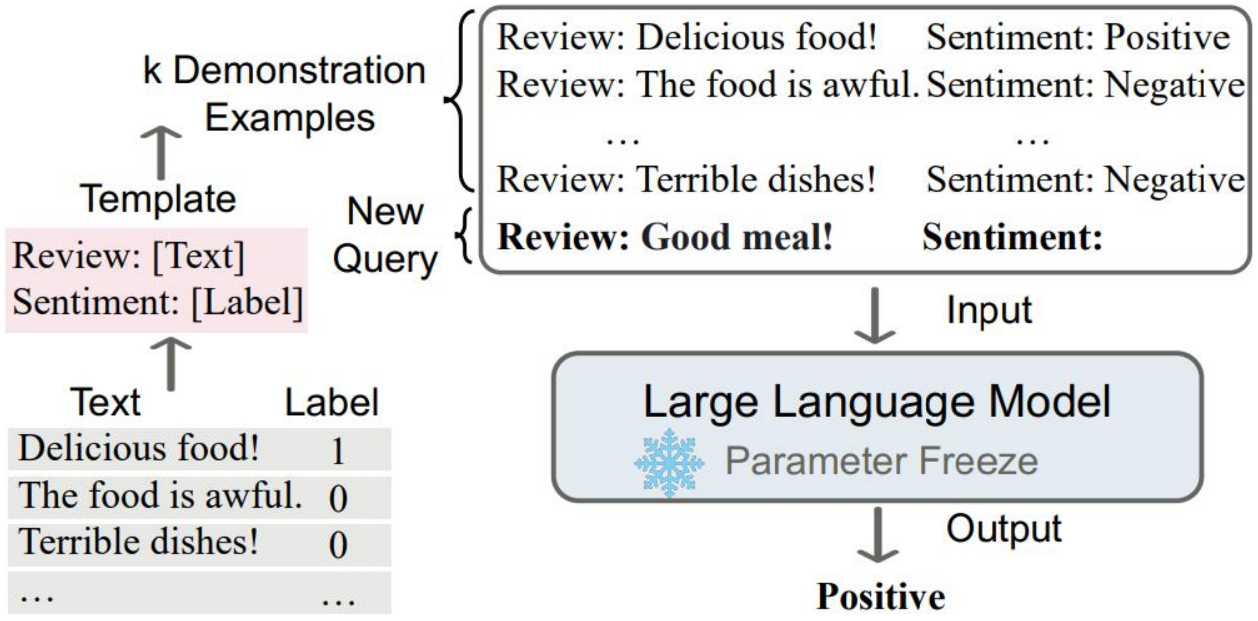
指示学习的优点是它经过多任务的微调后，也能够在其他任务上做zero-shot，而提示学习都是针对一个任务的。泛化能力不如指示学习。

上下文学习 In-context Learning

上下文学习In-context Learning又称情境学习，简称ICL，它避免了对大模型本身的修改，通过增加example（sentence_1，answer_1）的方式诱导模型生成更优的结果。

综述：A Survey for In-context Learning

值得注意的是，与需要使用反向梯度更新模型参数的训练阶段的监督学习不同，ICL不需要参数更新，并直接对预先训练好的语言模型进行预测。我们希望该模型学习隐藏在演示中的模式，并据此做出正确的预测。



上下文学习 In-context Learning

GPT3在GPT2已有的zero-shot的基础上，提出了In-Context learning的概念,有one-shot和few-shot两种方案，对应不同的增强Prompt的构建方式。随着模型参数量级的提升，few-shot，one-shot带来的效果提升更加显著.

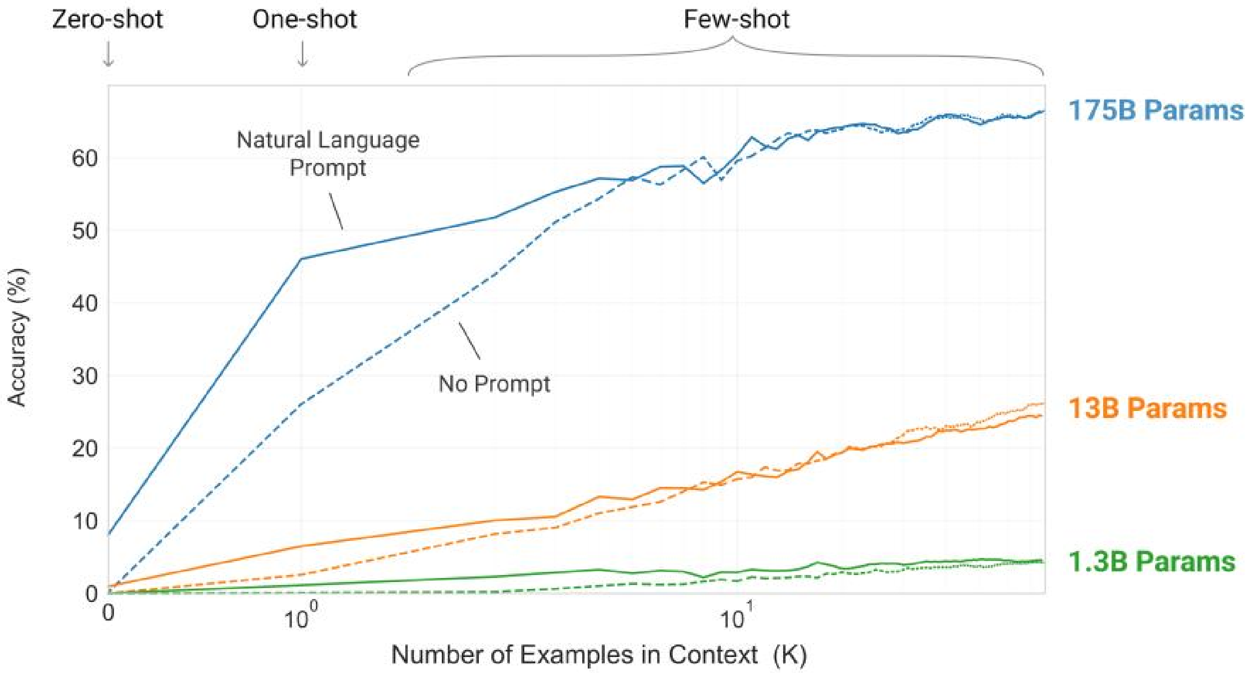


Figure 1.2: Larger models make increasingly efficient use of in-context information. We show in-context learning performance on a simple task requiring the model to remove random symbols from a word, both with and without a natural language task description (see Sec. 3.9.2). The steeper "in-context learning curves" for large models demonstrate improved ability to learn a task from contextual information. We see qualitatively similar behavior across a wide range of tasks.

思维链 Chain-of-thought(COT)

刚刚提到的两种tuning，虽然使得然大模型语言建模方面表现很好，但是在逻辑推理和数值计算领域（称其为第二类任务，system-2 task），即使是当时最大的模型也难以处理某些推理任务，如数学单词问题、符号操作和常识性推理。

C

有若干只鸡兔同在一个笼子里，从上面数，有35个头，从下面数，有94只脚。问笼中各有多少只鸡和兔?

设鸡的数量为x，兔子的数量为y。

从上面数，有35个头，则有35=x+y个头。

从下面数，有94只脚，则有94=2x+4y只脚。

将以上两式相减，得到35-94=3y-x，即y=12-x/3。

将结果带入第一式得到x+12-x/3=35，则x=36。

由于x=36，则y=12-x/3=12-36/3=4。

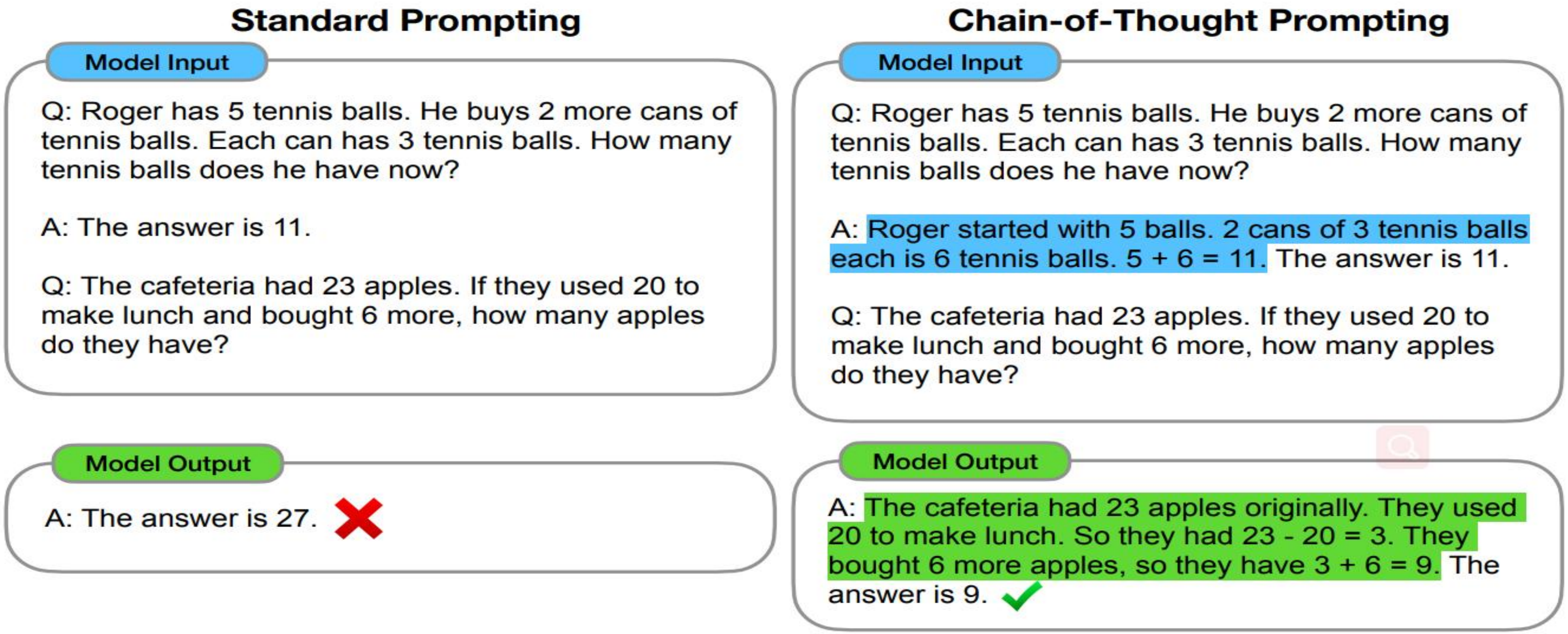
答案为：鸡有36只，兔子有4只。

思维链 Chain-of-thoughts(COT)

思维链概念的开山之作：Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

下图（左）为standard prompting，exemplary为<input, output>；

下图（右）为,chain-of-thought prompt的例子，exemplary为<input, chain-of-thought, output>，高亮部分为chain-of-thought。



思维链 Chain-of-thoughts(COT)

dataset(Arithmetic Reasoning): GSM8K, SVAMP, ASDiv, AQuA, MAWPS

dataset(Commonsense Reasoning): CSQA, StrategyQA, Date, Sports, SayCan

baseline: standard prompting

LLM：基于5个LLM（GPT-3, LaMDA, PaLM, UL2 20B, Codex）

RESUILT:

（1）chain-of-thought对大规模模型有帮助，但小规模模型会产生通顺但不合逻辑的推理链，从而使得效果差于standard prompting；

（2）问题越复杂，chain-of-thought提升效果越好；

（3）一些数据集的结果是state-of-the-art的。

（4）随机抽取最终答案正确的样本进行人工检验，推理链几乎都是正确的；对于最终答案错误的样本，推理链大多也只有一些小细节错误。

思维链 Chain-of-thoughts(COT)

消融实验ablation study：

这一部分对推理链进行了三种变式，检验chain-of-thought成功的原因。

（1）Equation only：prompt只包含数学公式，没有自然语言。

对于复杂问题效果差，简单问题（只有一步或两步）效果好。说明了自然语言表述的必要性。

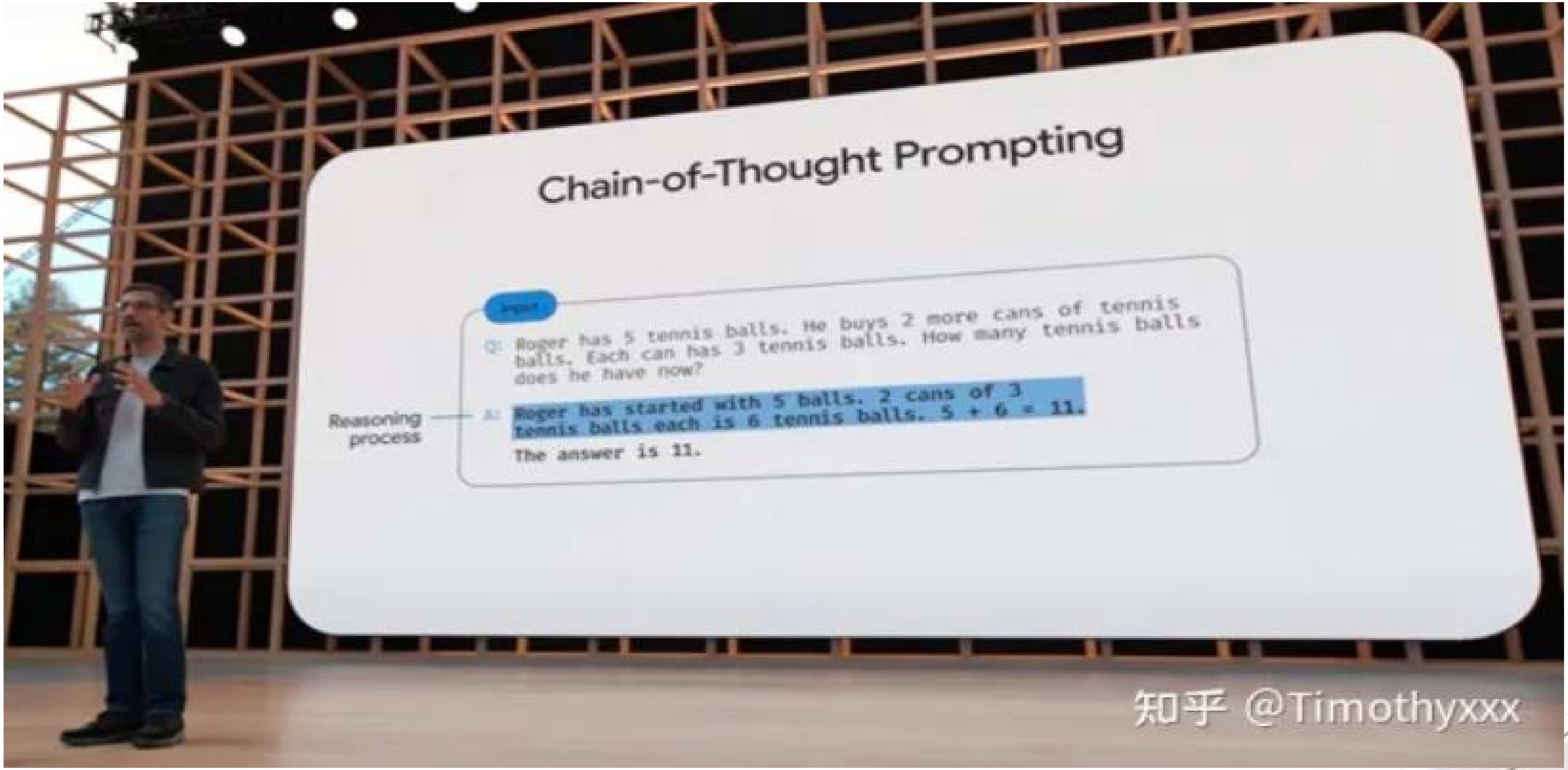
（2）Variable only：prompt只包含一个dot序列（…），与最终答案的计算所需的字符数相等。

表现与baseline（standard prompting）相同。同样说明了自然语言表述的必要性。

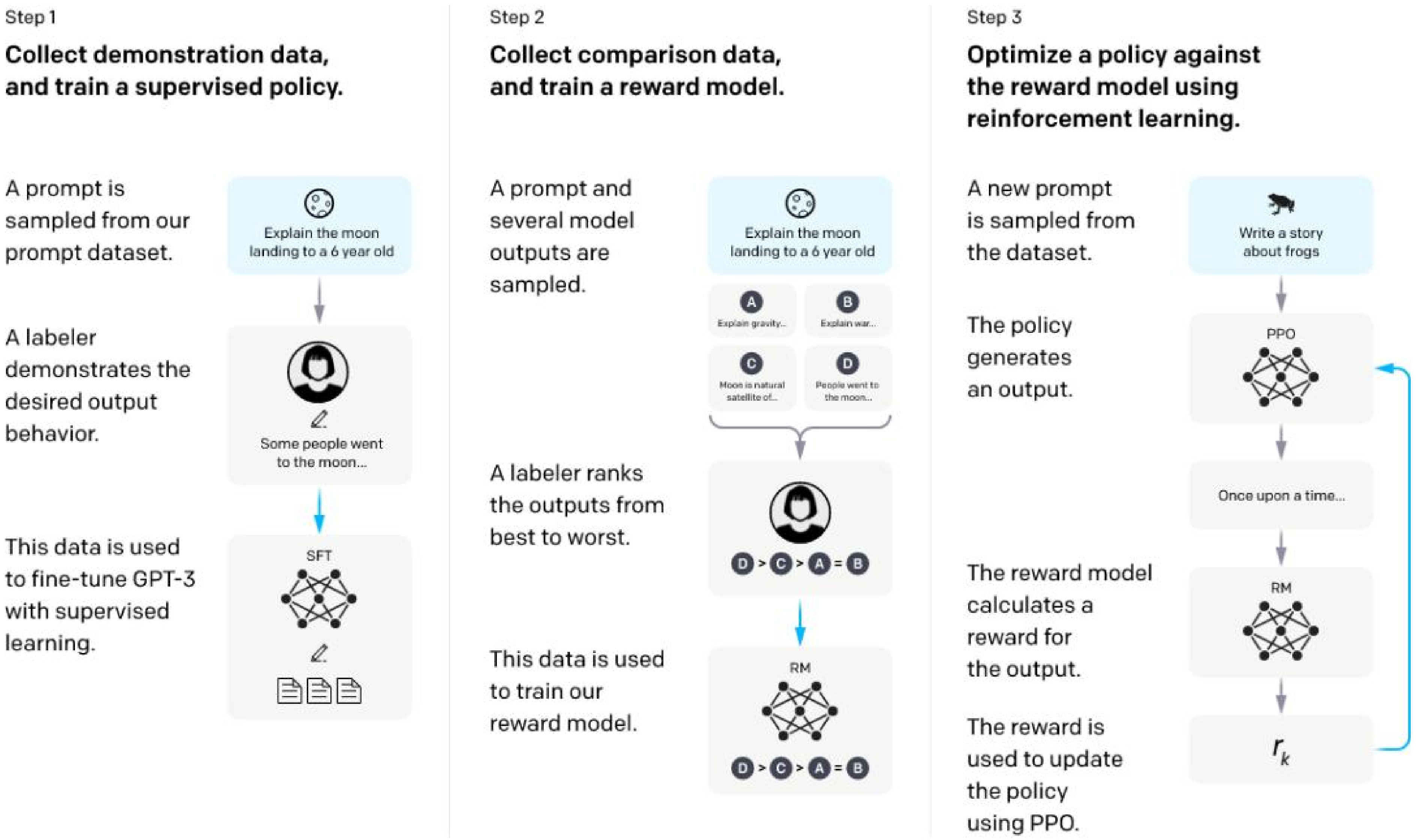
（3）Reasoning after answer：将推理过程放在answer之后，检验推理链是不是通过帮助模型在预训练中提取相关的信息得到最终答案。

表现与baseline（standard prompting）相同，说明推理链和答案的顺序很重要，答案是基于推理链得出的。

思维链 Chain-of-thoughts(COT)



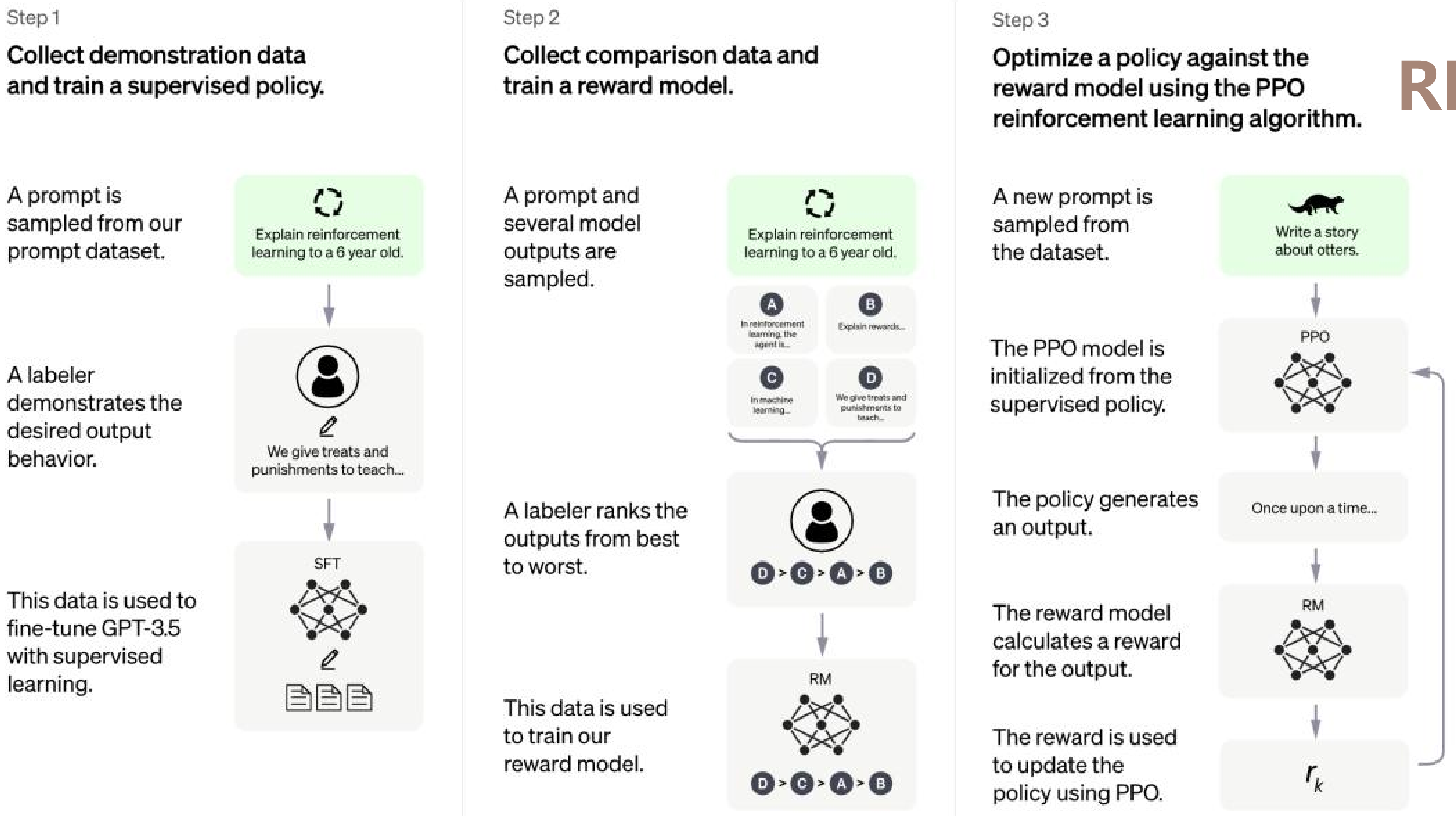
InstructGPT流程图-三步骤



[1] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[J]. arXiv preprint arXiv:2203.02155, 2022.

ChatGPT流程图-三步骤

RLHF



数据集

SFT Data			RM Data			PPO Data		
split	source	size	split	source	size	split	source	size
train	labeler	11,295	train	labeler	6,623	train	customer	31,144
train	customer	1,430	train	customer	26,584	valid	customer	16,185
valid	labeler	1,550	valid	labeler	3,488			
valid	customer	103	valid	customer	14,399			

InstructGPT/ChatGPT的训练分成3步，每一步需要的数据也有些许差异，下面我们分别介绍它们。
前两步的prompts，来自于OpenAI的在线API上的用户使用数据，以及雇佣的标注者手写的。最后一步则全都是从API数据中采样的，下表的具体数据：

1. SFT数据集

SFT数据集是用来训练第1步有监督的模型，即使用采集的新数据，按照GPT-3的训练方式对GPT-3进行微调。因为GPT-3是一个基于提示学习的生成模型，因此SFT数据集也是由提示-答复对组成的样本。SFT数据一部分来自使用OpenAI的PlayGround的用户，另一部分来自OpenAI雇佣的40名标注工（labeler）。并且他们对labeler进行了培训。在这个数据集中，标注工的工作是根据内容自己编写指示，并且要求编写的指示满足下面三点：

- Plain: We simply ask the labelers to come up with an arbitrary task, while ensuring the tasks had sufficient diversity.
- Few-shot: We ask the labelers to come up with an instruction, and multiple query/response pairs for that instruction.
- User-based: We had a number of use-cases stated in waitlist applications to the OpenAI API. We asked labelers to come up with prompts corresponding to these use cases.

数据集

Use-case	(%)
Generation	45.6%
Open QA	12.4%
Brainstorming	11.2%
Chat	8.4%
Rewrite	6.6%
Summarization	4.2%
Classification	3.5%
Other	3.5%
Closed QA	2.6%
Extract	1.9%

来自API提示数据集的用例类别分布

2. RM数据集

R M 数据集用来训练第 2 步的奖励模型， 我们也需要为InstructGPT/ChatGPT的训练设置一个奖励目标。这个奖励目标不必可导，但是一定要尽可能全面且真实的对齐我们需要模型生成的内容。很自然的，我们可以通过人工标注的方式来提供这个奖励，通过人工对可以给那些涉及偏见的生成内容更低的分从而鼓励模型不去生成这些人类不喜欢的内容。InstructGPT/ChatGPT的做法是先让模型生成一批候选文本，让后通过labeler根据生成数据的质量对这些生成内容进行排序。

3. PPO数据集

InstructGPT的PPO数据没有进行标注，它均来自GPT-3的API的用户。既又不同用户提供的不同种类的生成任务，其中占比最高的包括生成任务（45.6%），QA（12.4%），头脑风暴（11.2%），对话（8.4%）等。

能力	OpenAI模型	训练方法	OpenAI API	OpenAI论文	近似的开源模型
GPT-3系列					
语言生成 + 世界知识 + 上下文学习	GPT-3初始版本 **大部分的能力已经存在于模型中，尽管表面上看起来很弱。	语言建模	Davinci	GPT-3论文	Meta OPT
+ 遵循人类的指令 + 泛化到没有见过的任务	Instruct-GPT初始版本	指令微调	Davinci-Instruct-Beta	Instruct-GPT论文	T0论文 Google FLAN论文
+ 代码理解 + 代码生成	Codex初始版本	在代码上进行训练	Code-Cushman-001	Codex论文	Salesforce CodeGen
GPT-3.5系列					
++ 代码理解 ++ 代码生成 ++ 复杂推理 / 思维链 (为什么?) + 长距离的依赖（很可能)	现在的Codex **GPT3.5系列中最强大的模型	在代码+文本上进行训练 在指令上进行微调	Code-Davinci-002 (目前免费的版本 = 2022年12月)	Codex 论文	
++ 遵循人类指令 - 上下文学习 - 推理能力 ++ 零样本生成	有监督的Instruct-GPT **通过牺牲上下文学习换取零样本生成的能力	监督学习版的指令微调	Text-Davinci-002	Instruct-GPT论文, 有监督的部分	T0论文 Google FLAN论文
+ 遵循人类价值观 + 包含更多细节的生成 + 上下文学习 + 零样本生成	经过RLHF训练的Instruct-GPT **和002模型相比，和人类更加对齐，并且更少的性能损失	强化学习版的指令微调	Text-Davinci-003	Instruct-GPT论文, RLHF部分，从人类反馈中的学习摘要。	DeepMind Sparrow论文 AI2 RL4LMs
++ 遵循人类价值观 ++ 包含更多细节的生成 ++ 拒绝知识范围外的问题 (为什么?) ++ 建模对话历史的能力 -- 上下文学习	ChatGPT ** 通过牺牲上下文学习的能力换取建模对话历史的能力	使用对话数据进行强化学习指令微调			DeepMind Sparrow论文 AI2 RL4LMs

我们先关注GPT3.5系列，这是ChatGPT的基石。code-davinci-002和text-davinci-002这两兄弟是第一版的 GPT3.5 模型，一个用于代码，另一个用于文本。它们表现出了三种重要能力与初代 GPT-3 不同的能力：

响应人类指令：以前，GPT-3 的输出主要训练集中常见的句子。现在的模型会针对指令 / 提示词生成更合理的答案（而不是相关但无用的句子）。

泛化到没有见过的任务：当用于调整模型的指令数量超过一定的规模时，模型就可以自动在从没见过的新指令上也能生成有效的回答。 这种能力对于上线部署至关重要，因为用户总会提新的问题，模型得答得出来才行。

代码生成和代码理解：这个能力很显然，因为模型用代码训练过。

利用思维链 (chain-of-thought) 进行复杂推理：初代 GPT3 的模型思维链推理的能力很弱甚至没有。 code-davinci-002 和 text-davinci-002 是两个拥有足够强的思维链推理能力的模型。

总的来说：初代GPT-3模型通过预训练获得生成能力、世界知识和in-context learning。然后通过instruction tuning的模型分支获得了遵循指令和能泛化到没有见过的任务的能力。经过代码训练的分支模型则获得了代码理解的能力，作为代码训练的副产品，模型同时潜在地获得了复杂推理的能力。结合这两个分支，code-davinci-002似乎是具有所有强大能力的最强GPT-3.5模型。接下来通过有监督的instruction tuning和RLHF通过牺牲模型能力换取与人类对齐，即对齐税。RLHF 使模型能够生成更翔实和公正的答案，同时拒绝其知识范围之外的问题，使得ChatGPT以良好的体验名声大噪。

To be continued

02.17	张岚雪	GPTs & ChatGPT
02.24	张琬悦	Instruct Learning
03.03	**毕冠群**	**基于人类反馈的强化学习**
03.10	王青悦	ChatGPT的风潮和未来

参考资料 Reference

[1] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[J]. arXiv preprint arXiv:2203.02155, 2022.

[2] Qngxiu D, Lei L, Damai D, et al. A Survey for In-context Learningk[J]. arXiv preprint arXiv:2301.00234 , 2023.

[3] Or Honovich, Thomas Scialom, Omer Levy, Timo Schick. Unnatural Instructions: Tuning Language Models with (Almost) No Human Labor[J]. arXiv preprint arXiv:2212.09689 , 2022.

[4] Brown T B， Mann B， Ryder N , et al. Language Models are Few-Shot Learners[J]. arXiv preprint arXiv: 2005.14165, 2021.

[5] Jason W, Maarten B, Vincent Y, et al. Finetuned Language Models Are Zero-Shot Learners[J]. arXiv preprint arXiv: 2109.01652, 2021.

[6] OpenAI是如何“魔鬼调教” GPT的？——InstructGPT论文解读 https://cloud.tencent.com/developer/news/979148

[7] In-Context Learning（上下文学习）相关分享 https://zhuanlan.zhihu.com/p/603650082

[8] Instruction-Tuning论文列表：https://github.com/SinclairCoder/Instruction-Tuning-Papers

[9] Chain-of-Thought论文列表：https://github.com/Timothyxxx/Chain-of-ThoughtsPapers

[10] Prompt专栏-概述篇 https://zhuanlan.zhihu.com/p/464825153