



大型语言模型的能力分析与应用

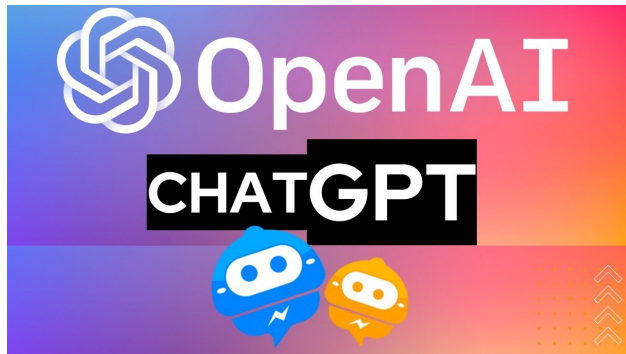
邱锡鹏

复旦大学

2023年2月18日 Saturday

将会改变世界的对话式通用人工智能模型ChatGPT

- 2022 年11 月30 日，OpenAI 发布了AI 对话模型，被认为是**人工智能里程碑式应用**



- 仅用两个月时间，月活跃用户已达1亿，是史上用户增速最快的消费应用。
- **ChatGPT不开源，其参数规模，技术细节都没有对外公布。**



微软联合创始人比尔·盖茨：像**ChatGPT**这样的AI聊天机器人将变得与**个人电脑**或**互联网**同样重要。

OpenAI发布AI对话模型ChatGPT，开启生成式AI商业化新机遇 ...

2022年12月19日 — 当地时间11月30日，美国人工智能公司OpenAI发布全新产品**ChatGPT**，一款基于GPT-3.5的免费对话模型。公司CEO Sam Altman 透露上线五天该模型的全球用户数量 ...

AI新物种：ChatGpt 不会止于写代码调Bug - 巴比特

2022年12月11日 — GPT-3发布于2020年，作为一个自监督模型，几乎可以完成自然语言处理的绝大部分任务，在参数上，GPT-1包含了1.17亿个参数，GPT-2包含了15亿个参数，而GPT-3 ...

ChatGPT 通过了美国MBA、法律和医学考试 - Showmetech

2023年1月26日 — 这次的新颖之处在于，根据在美国进行的一项研究，**OpenAI** 会轻松通过即使是学生也难以通过的复杂测试。重点是创建包含所有重要细节的法律文件和聊天GPT ...

微软100亿美元砸向OpenAI，ChatGPT要加入Office全家桶了？

2023年1月12日 — 如果100 亿美元的交易成真，**OpenAI** 将获得巨额资金，微软赢得广阔未来，双赢局面就此达成。过去一段时间，对话式AI 模型**ChatGPT** 火遍了整个社区，它 ...

GPT-4：人工智能的新语言方法被定义为“强大” - Showmetech

2023年1月2日 — GPT-2024 计划于4 年**发布**，应该会为**ChatGPT** 带来更好的理解和文本创建。... 不浪费时间，该公司已经在准备下一代AI 语言方法，预计将于2023 年推出， ...

美国大学89%的学生居然用ChatGPT写作业 - 国际竞赛

1天前 — **ChatGPT**的崛起并在高等教育领域的突然普及，让众多美国高校感觉措手不及！ ... **Nature**早就很有先见之明地发文，担心**ChatGPT**会成为学生写论文的工具。

ChatGPT发展路径

解码器部分成为GPT发展的基础

2017年6月,
Google提出
Transformer
模型

2019年2月,
OpenAI提出
GPT2模型

赋予GPT-3代码能力,
同时引入思维链能力

2021年12月,
OpenAI提出
CodeX模型

赋予GPT理解人类指令能力

2022年2月,
OpenAI提出
InstructGPT
模型

2018年6月,
OpenAI提出
GPT模型

预训练语言模型元年

2020年5月,
OpenAI提出
GPT3模型

首个千亿级模型

2021年12月,
OpenAI提出
WebGPT模
型

赋予GPT搜索能力

2022年11月,
OpenAI推出
ChatGPT模
型

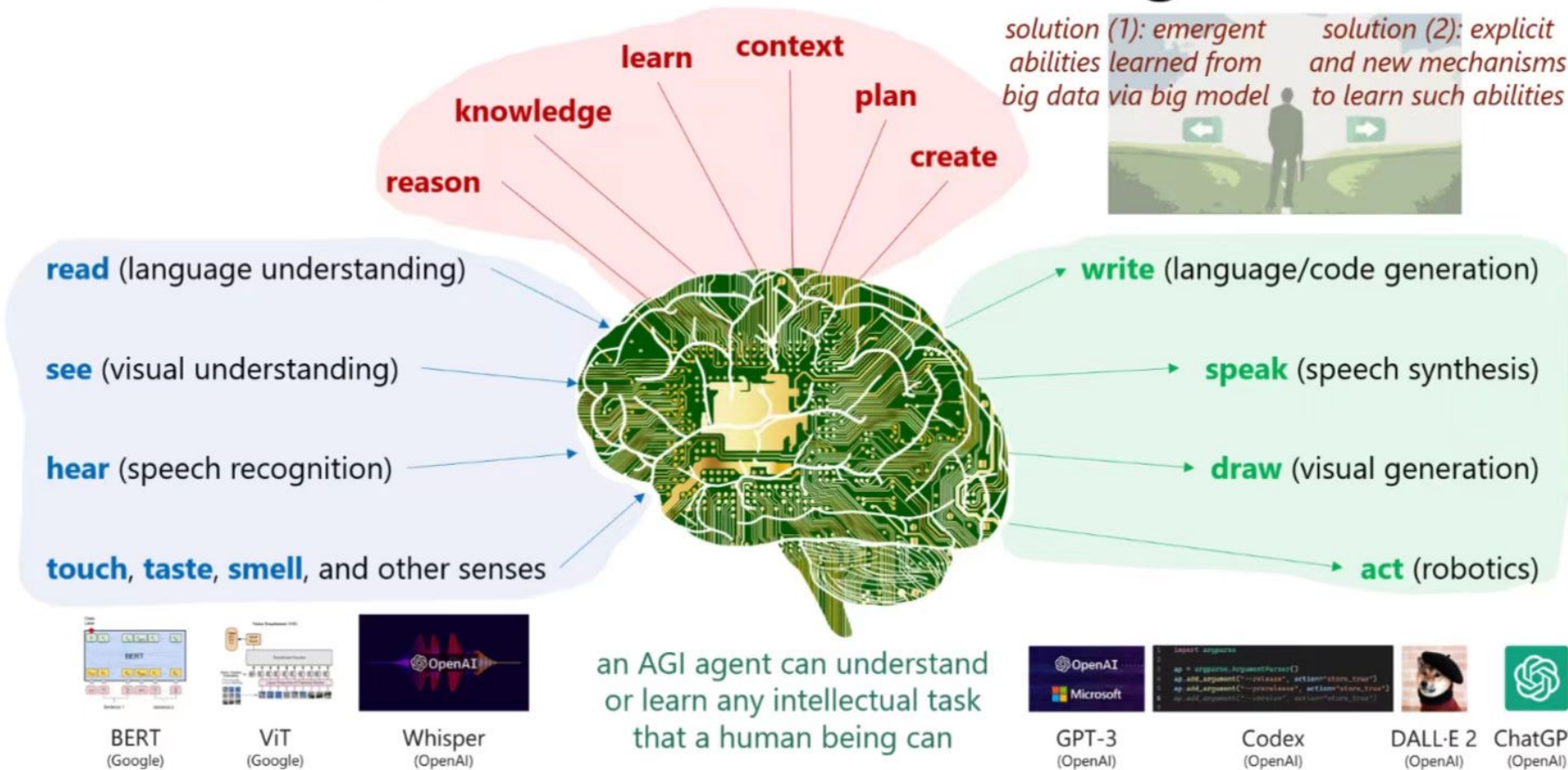
赋予GPT对话能力

大型语言模型：ChatGPT的基座

对人工智能技术的颠覆性影响

ChatGPT将加速通用人工智能的实现。

AGI (Artificial General Intelligence)





基础介绍

语言模型 (Language Model)

Example

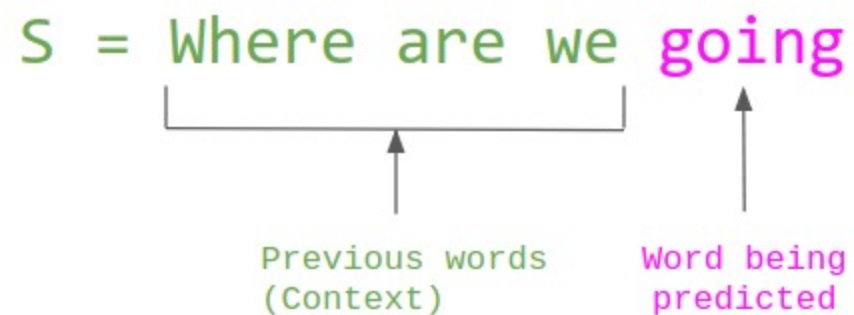
The cat **sat** on the mat

The cat **sad** on the mat

High wind tonight

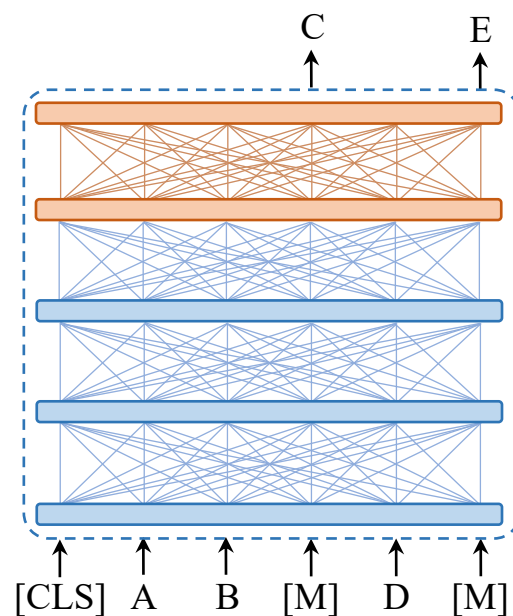
Large wind tonight

语言模型 (Language Model)

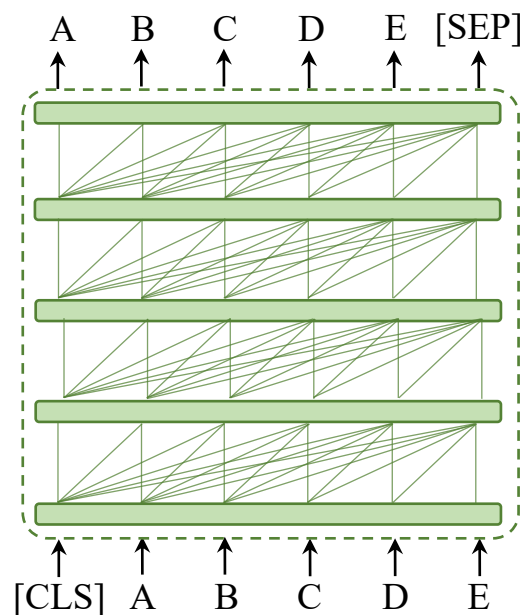


$$P(S) = P(\text{Where}) \times P(\text{are} \mid \text{Where}) \times P(\text{we} \mid \text{Where are}) \times P(\text{going} \mid \text{Where are we})$$

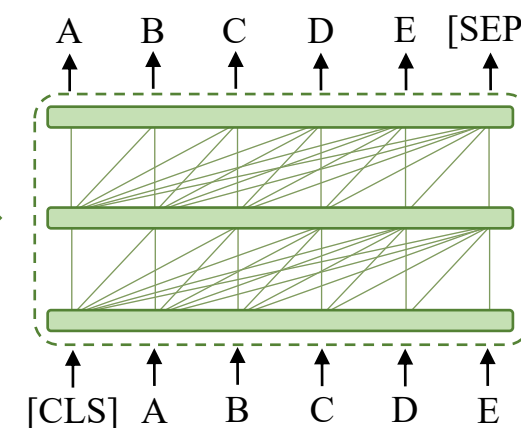
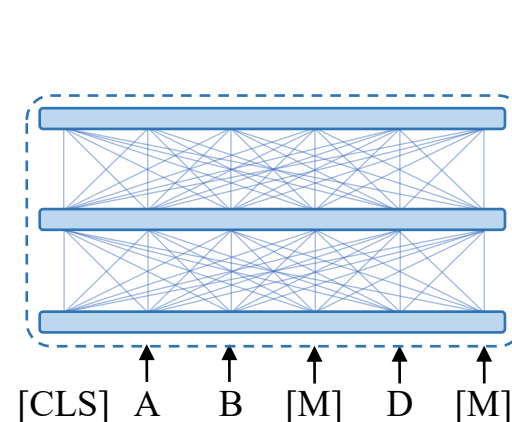
预训练语言模型 (Pretrained Language Model)



BERT



GPT



BART/T5

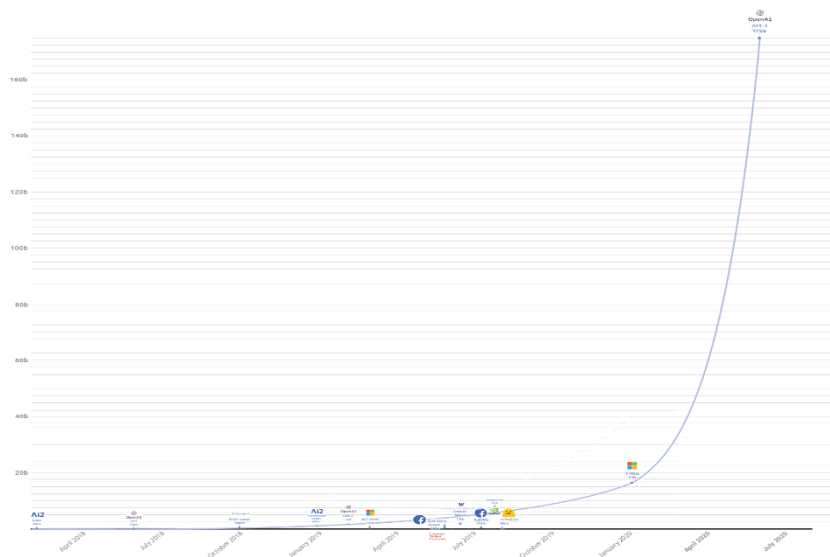
逐渐走向大模型时代

Pretraining + Adapting: New Paradigm for NLP

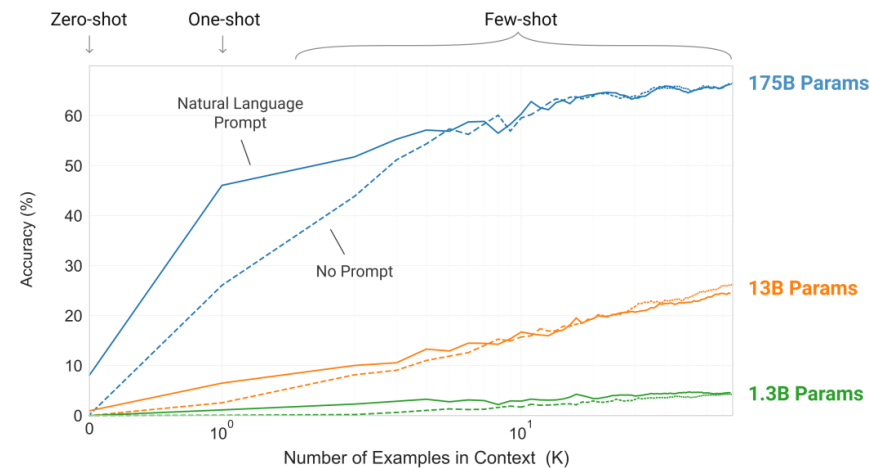
Upstream
Model Pretraining



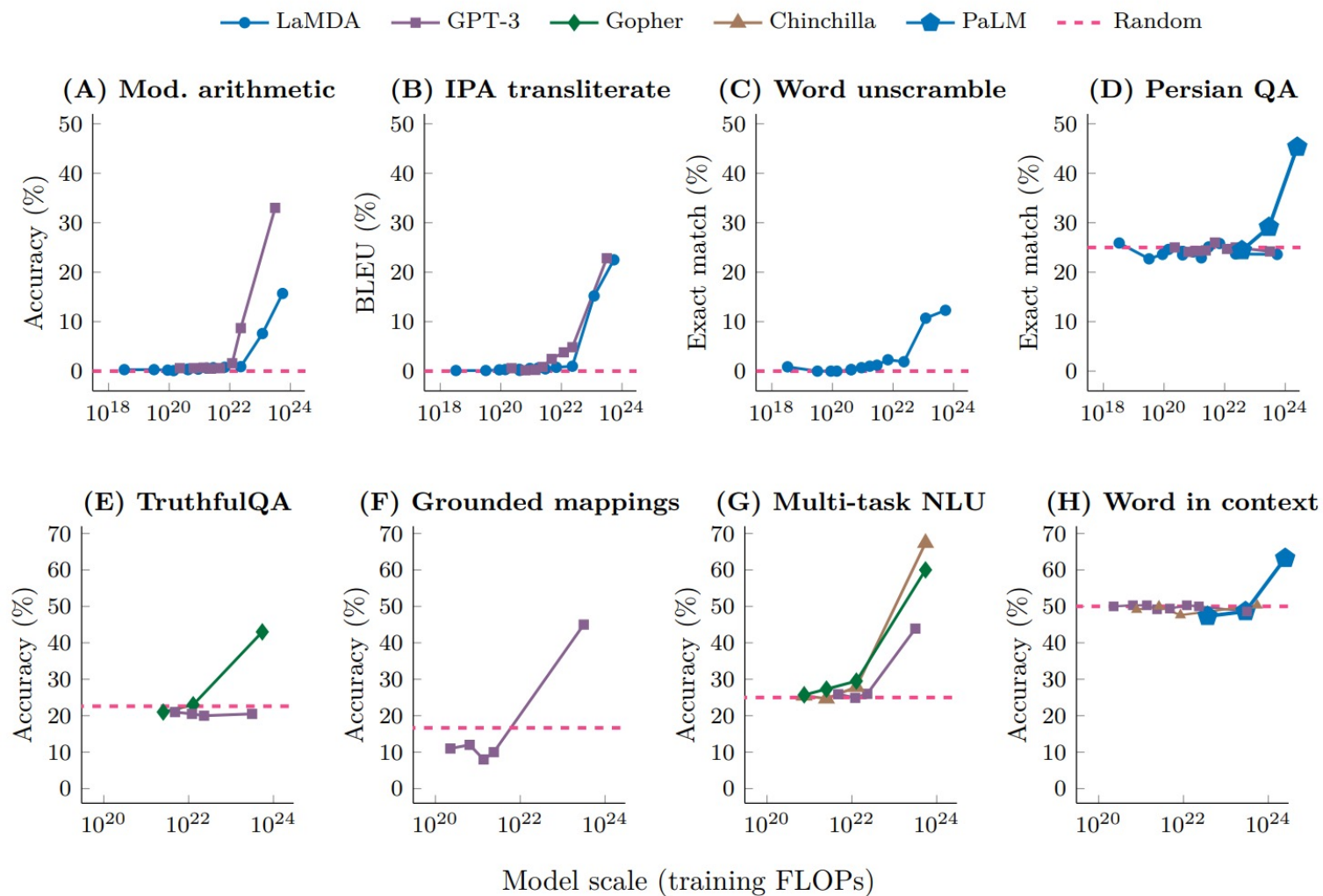
Downstream
Model Adapting



Increasing Size

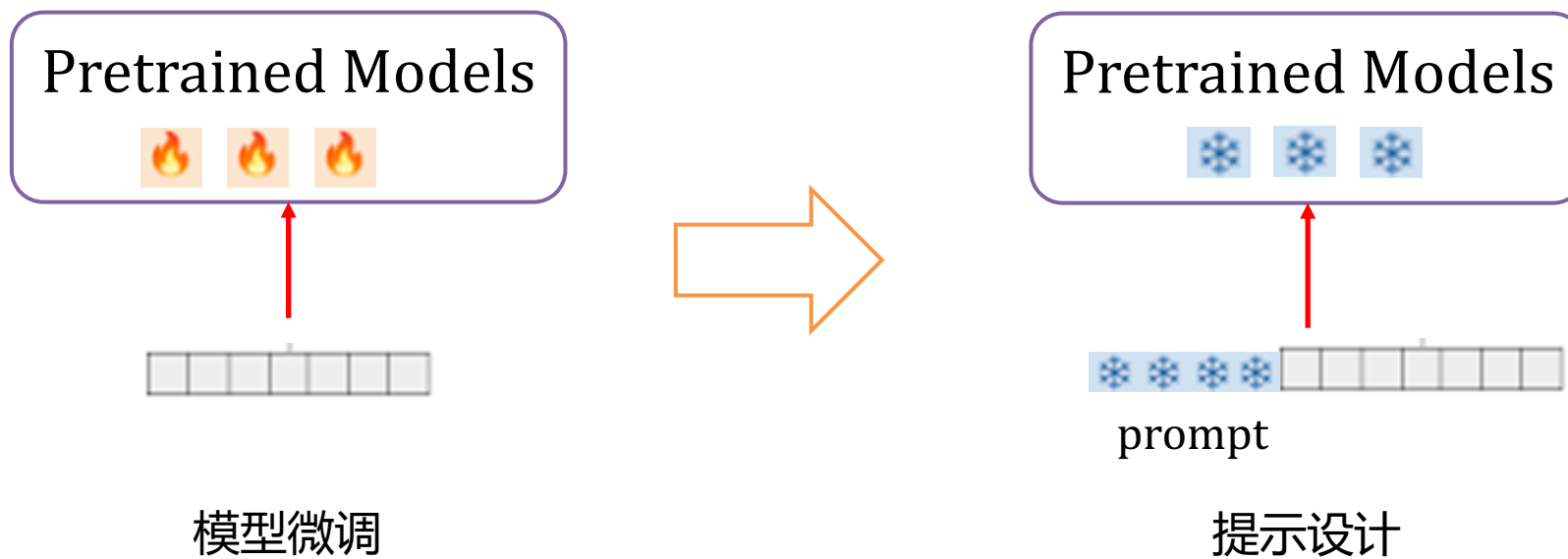


Increase Performance
for Few-shot Learning

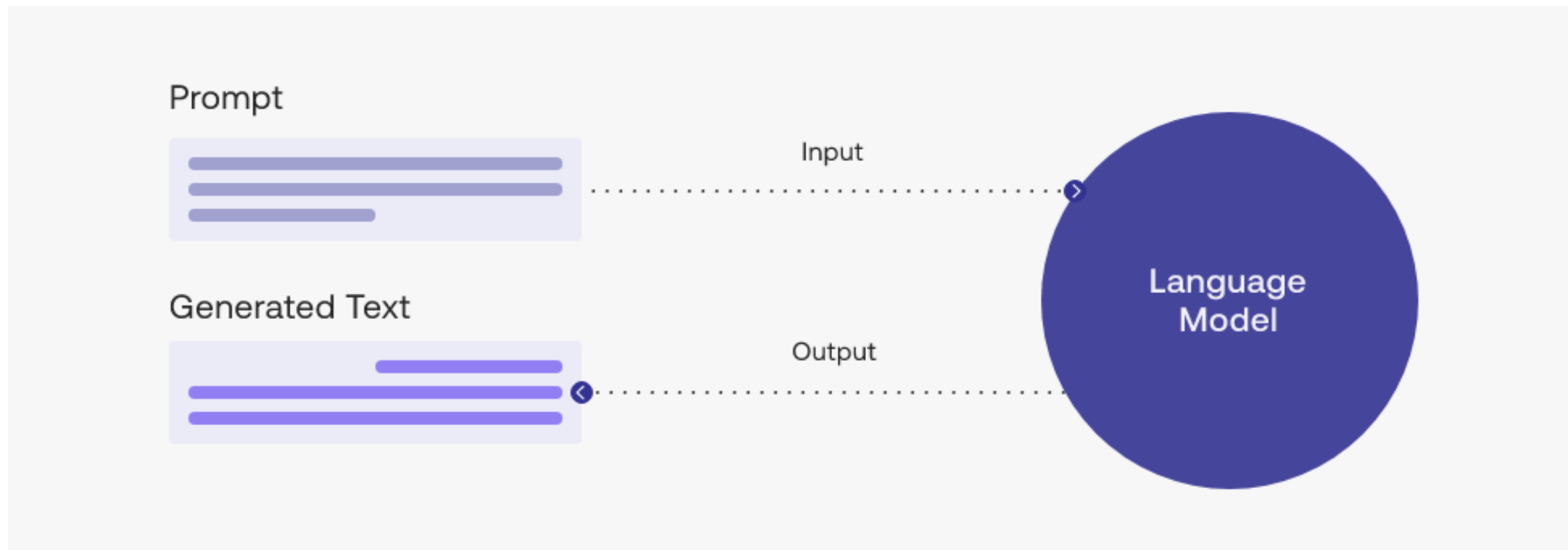


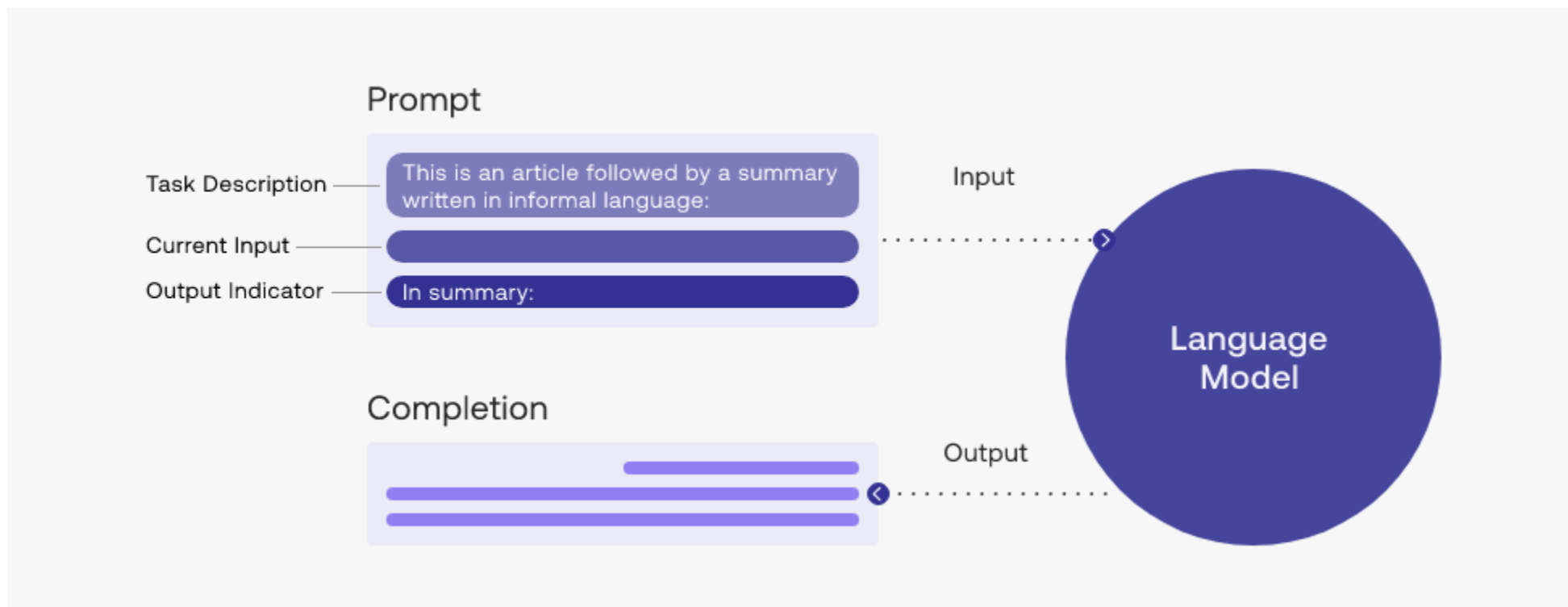
Emergent Abilities of Large Language Models, 2022

模型使用方式的变换

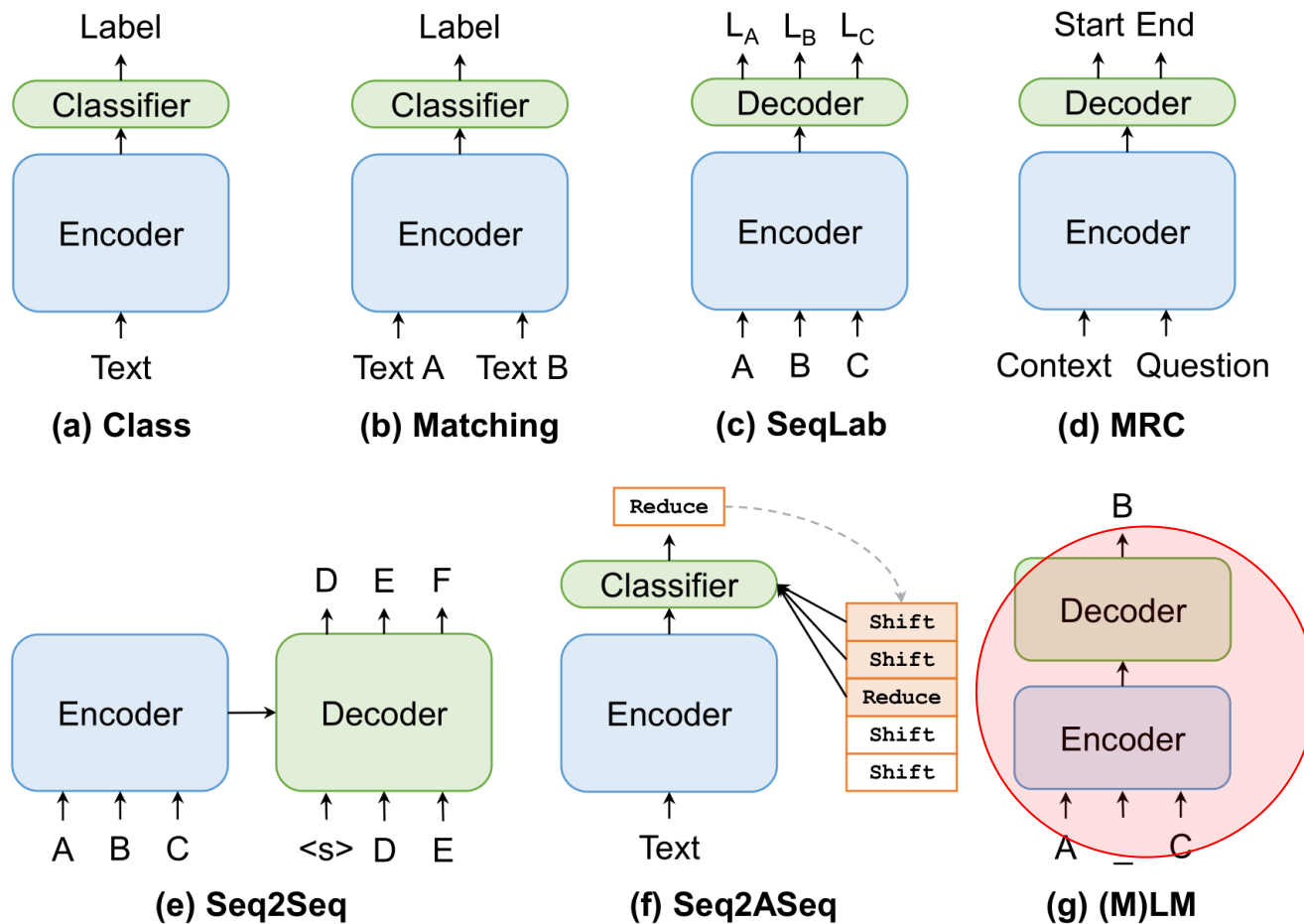


大型语言模型的使用方式



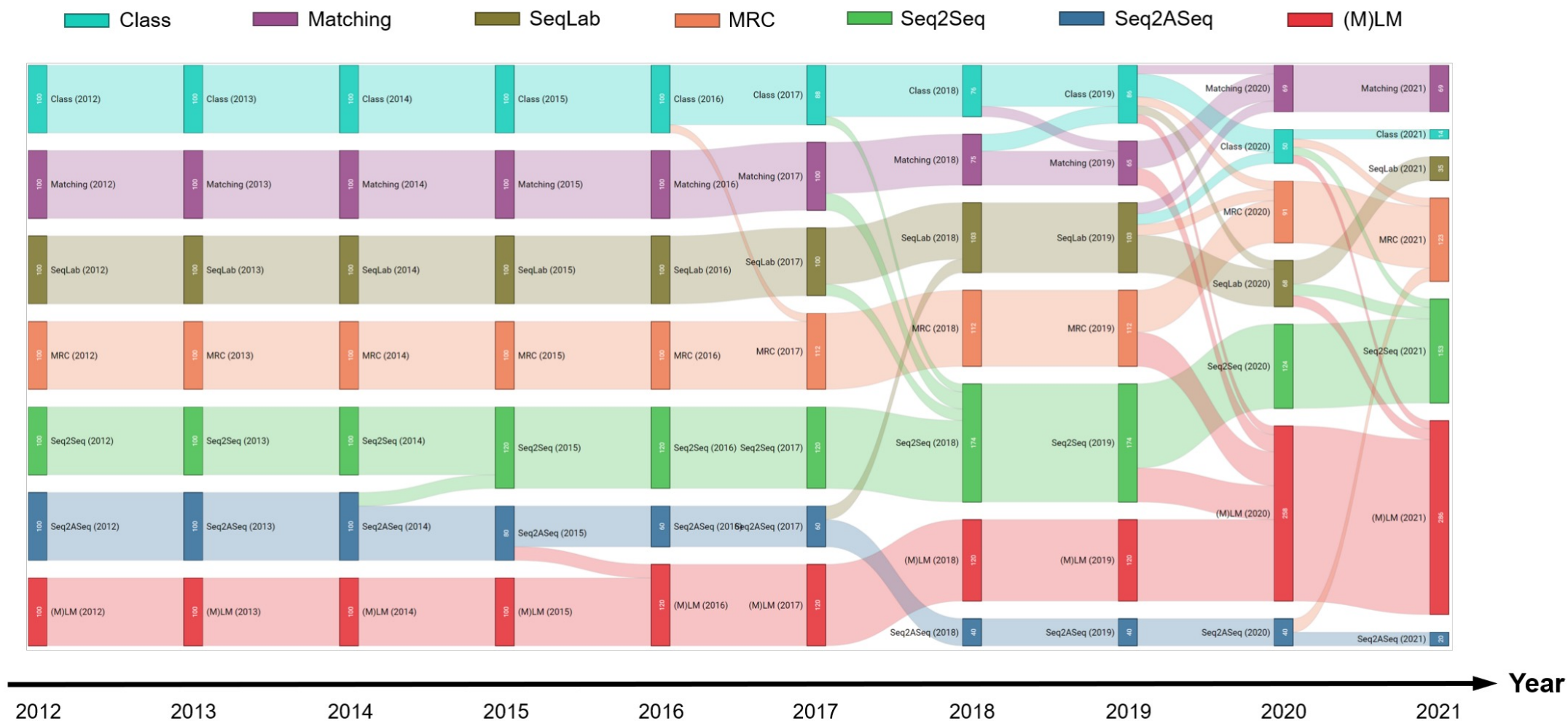


自然语言处理的范式迁移



T Sun, X Liu, X Qiu, X Huang, **Paradigm Shift in Natural Language Processing**, Machine Intelligence Research. 19, 169–183 (2022). <https://doi.org/10.1007/s11633-022-1331-6>

Towards Unified Paradigm



T Sun, X Liu, X Qiu, X Huang, **Paradigm Shift in Natural Language Processing**, Machine Intelligence Research. 19, 169–183 (2022). <https://doi.org/10.1007/s11633-022-1331-6>

ChatGPT的三个关键技术

shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

Translate English to French: ← task description
sea otter => loutre de mer ← examples
pepper
plush
cheese

情景学习

大模型的涌现能力
改变传统学习范式

Chain-of-Thought Prompting

Input
Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: Roger has started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.
Q: How
Mod
A: Th
Ma

思维链

大模型的涌现能力
打破模型参数约束

Natural Instructions

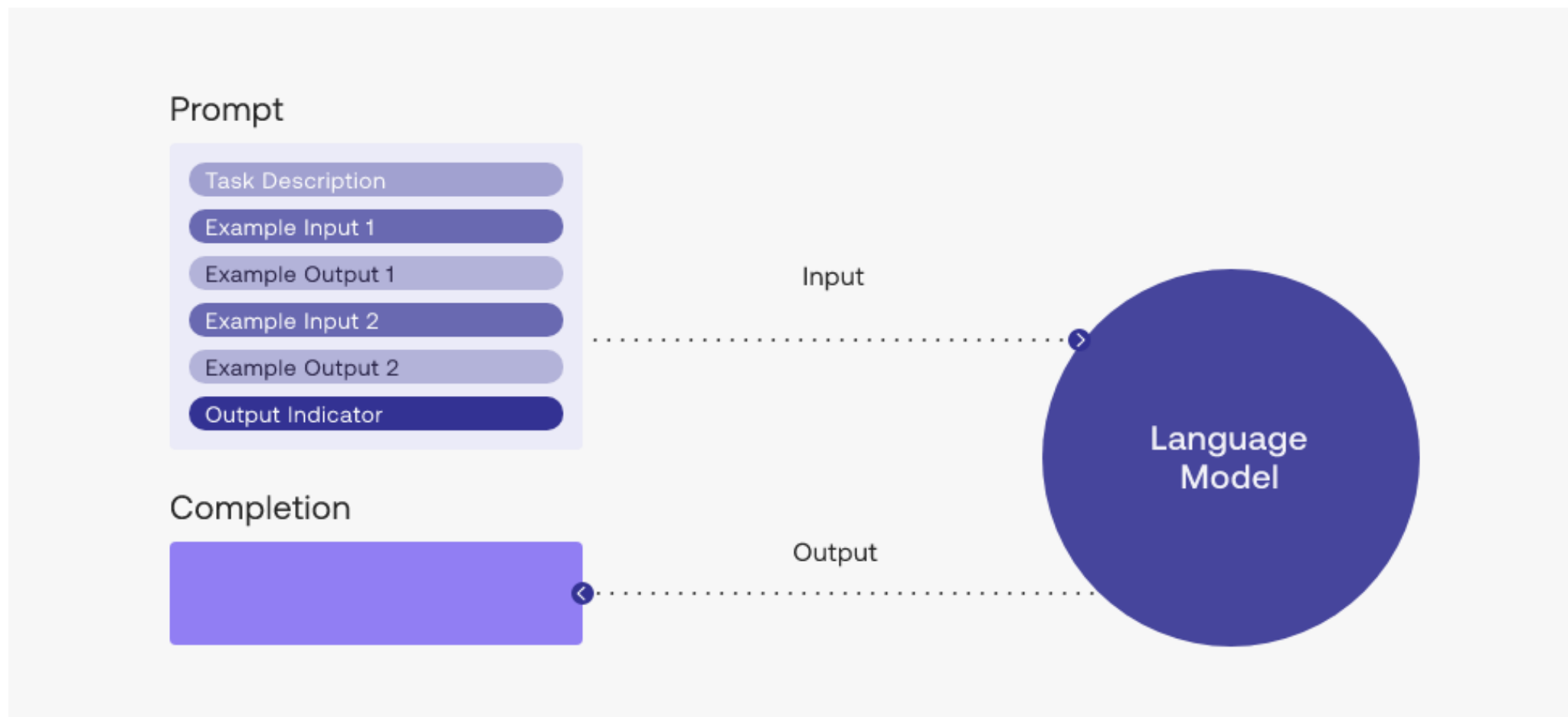
指令学习

人在环路增强
对齐人类意图

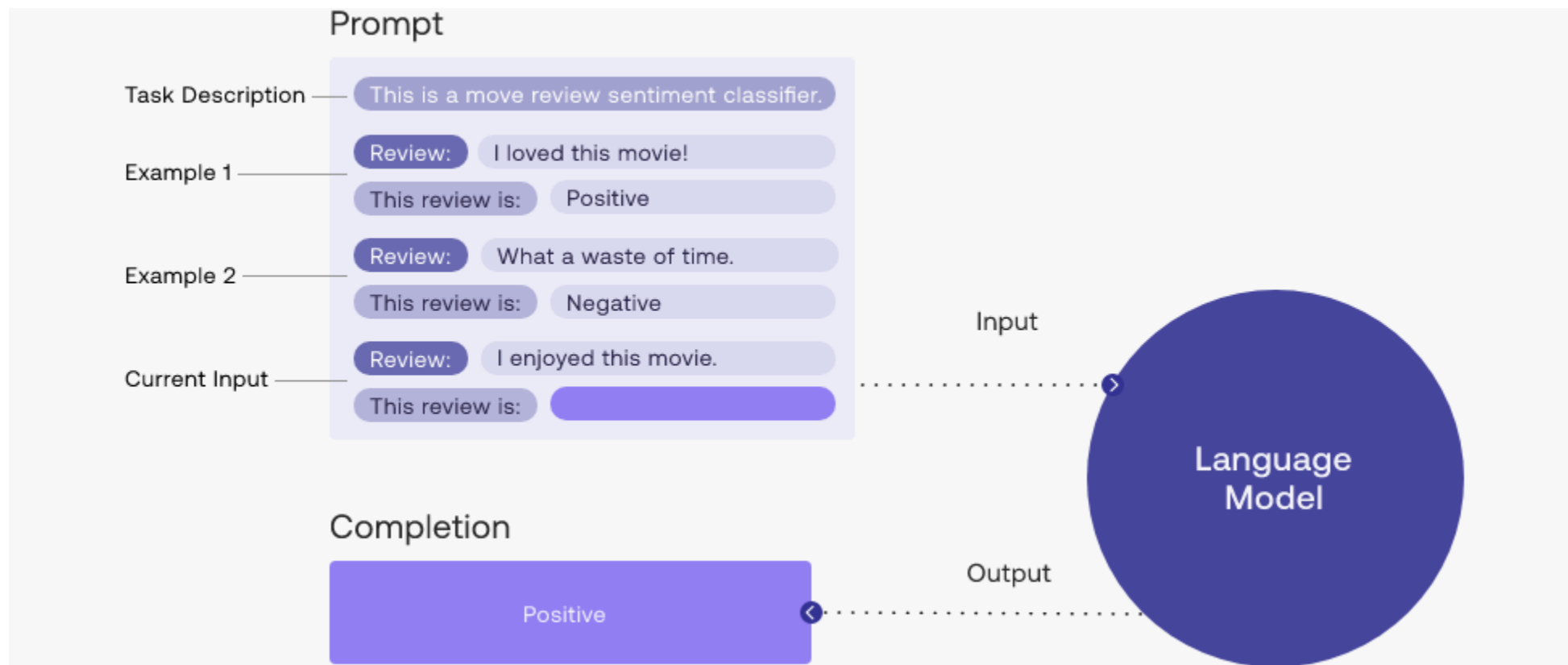


情景学习 (In-Context Learning)

情景学习 (In-Context Learning)



情景学习 (In-Context Learning)



This is a movie review sentiment classifier. Review: "I loved this movie!" This review is positive.
 Review: "I don't know, it was ok I guess.." This review is neutral. Review: "What a waste of time,
 would not recommend this movie." This review is negative. Review: "I really enjoyed this movie!"
 This review is

Apps built on GPT-3

情景学习可以大幅降低下游开发成本！

Describe a layout.

Just describe any layout you want, and it'll try to render below!

A div that contains 3 buttons each with a random color.

Generate

Equation description

x squared plus two times x

Translate

$$x^2 + 2x$$

Products

Select product

Collections

New

Popular

Upcoming

Requested

Categories

All	319
A/B Testing	2
Ad Generation	3
AI Copywriting	37
AI Writing Assistants	1
API Design	1
Avatars	1
Blog writing	2
Book Writing	1

New

Recently added GPT-3 apps



Customer Service
ActiveChat.ai



Chatbots
AI Buddy



Humor
AI Guru



LegalTech
aiLawDocs



Chatbots
AskBrian



Developer Tools
Azure OpenAI Service



Generative Art
Botto



Image captioning
ClipClap



Healthcare
Curai



Code Generation
DeepGenX



Summarization
Delv AI



API Design
Design an API with ...



Recruiting
Drafted

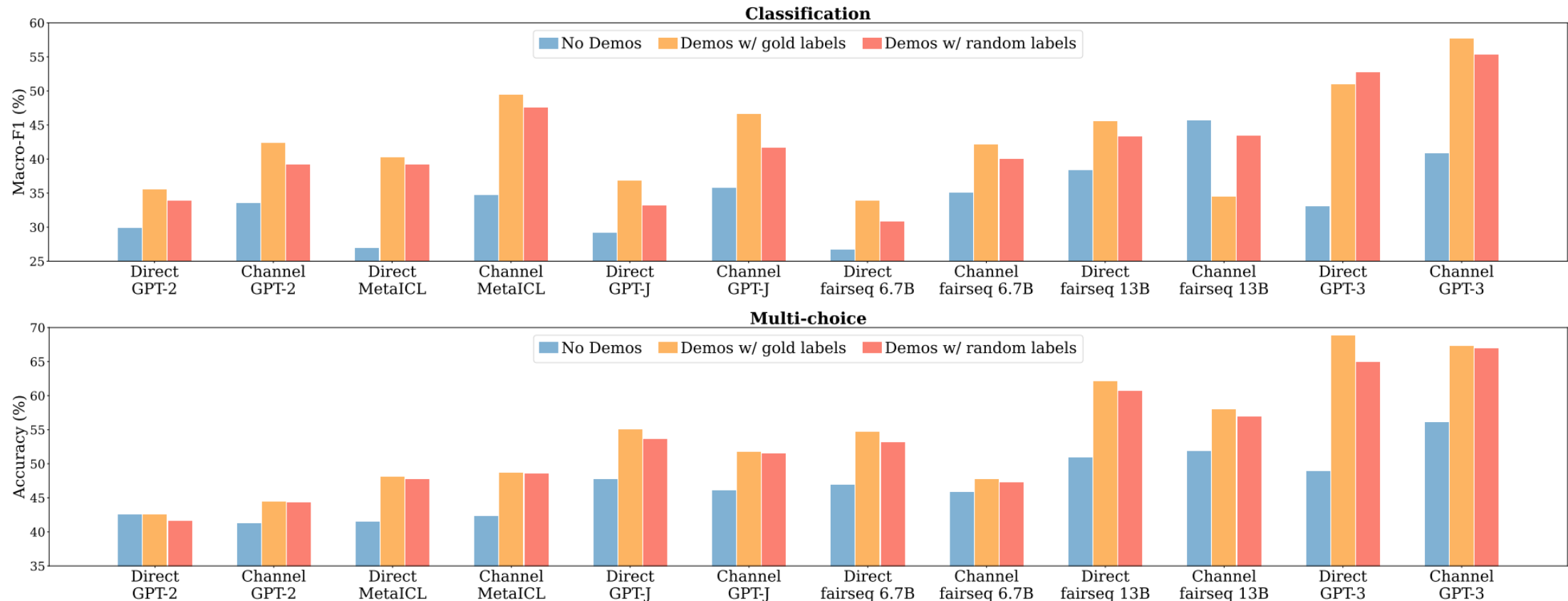


Language Learning
Duolingo



Research Assistants
Elicit

What Makes In-Context Learning Work?



Min et al, 2022, Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?



思维链 (Chain-of-Thought, CoT)

思维链 (Chain-of-Thought, CoT)

Standard Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain of Thought Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

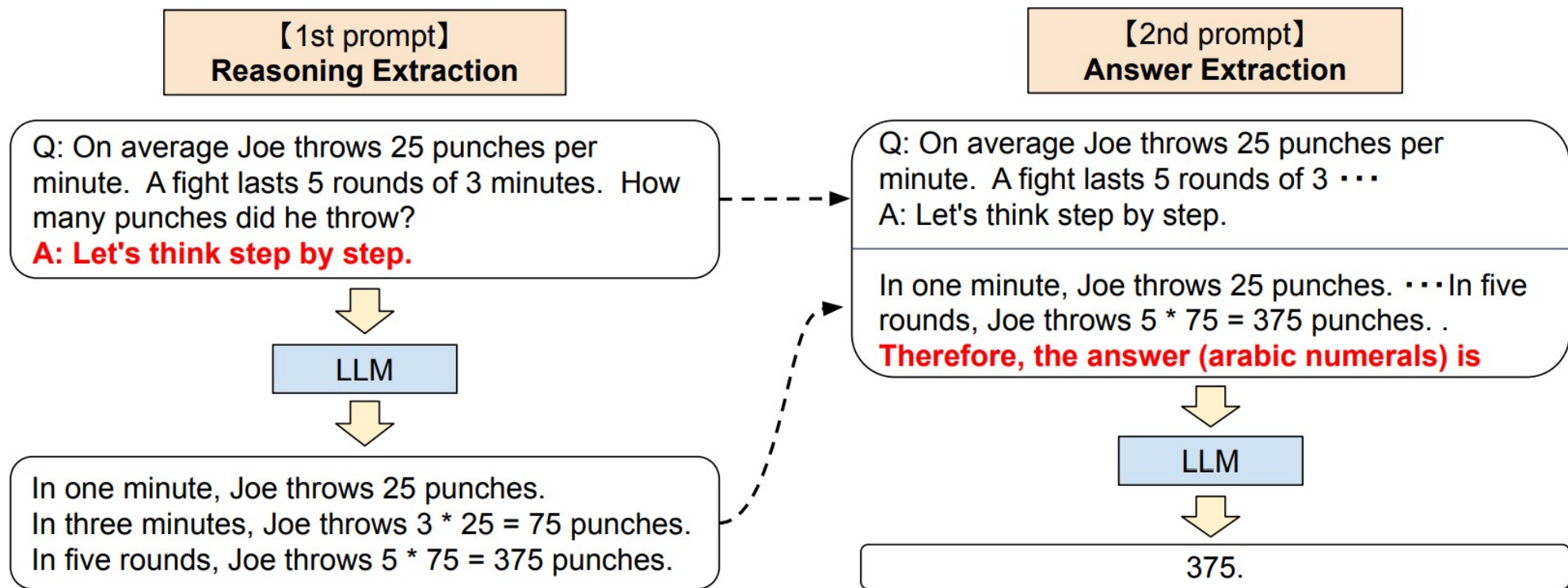
Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

Wei. et. al. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

思维链 (Chain-of-Thought, CoT)



Takeshi, Kojima, et al. **"Large Language Models are Zero-Shot Reasoners."** arXiv preprint arXiv:2205.11916 (2022).

Program-aided Reasoning

Chain-of-Thought (Wei et al., 2022)

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 tennis balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6$. The answer is 11.

Q: The bakers at the Beverly Hills Bakery baked 200 loaves of bread on Monday morning. They sold 93 loaves in the morning and 39 loaves in the afternoon. A grocery store returned 6 unsold loaves. How many loaves of bread did they have left?

Model Output

A: The bakers started with 200 loaves. They sold 93 in the morning and 39 in the afternoon. So they sold $93 + 39 = 132$ loaves. The grocery store returned 6 loaves. So they had $200 - 132 - 6 = 62$ loaves left. The answer is 62.



Program-aided Reasoning (this work)

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 tennis balls.

`tennis_balls = 5`

2 cans of 3 tennis balls each is

`bought_balls = 2 * 3`

tennis balls. The answer is

`answer = tennis_balls + bought_balls`

Q: The bakers at the Beverly Hills Bakery baked 200 loaves of bread on Monday morning. They sold 93 loaves in the morning and 39 loaves in the afternoon. A grocery store returned 6 unsold loaves. How many loaves of bread did they have left?

Model Output

A: The bakers started with 200 loaves

`loaves_baked = 200`

They sold 93 in the morning and 39 in the afternoon

`loaves_sold_morning = 93`

`loaves_sold_afternoon = 39`

The grocery store returned 6 loaves.

`loaves_returned = 6`

The answer is

`answer = loaves_baked - loaves_sold_morning
- loaves_sold_afternoon + loaves_returned`

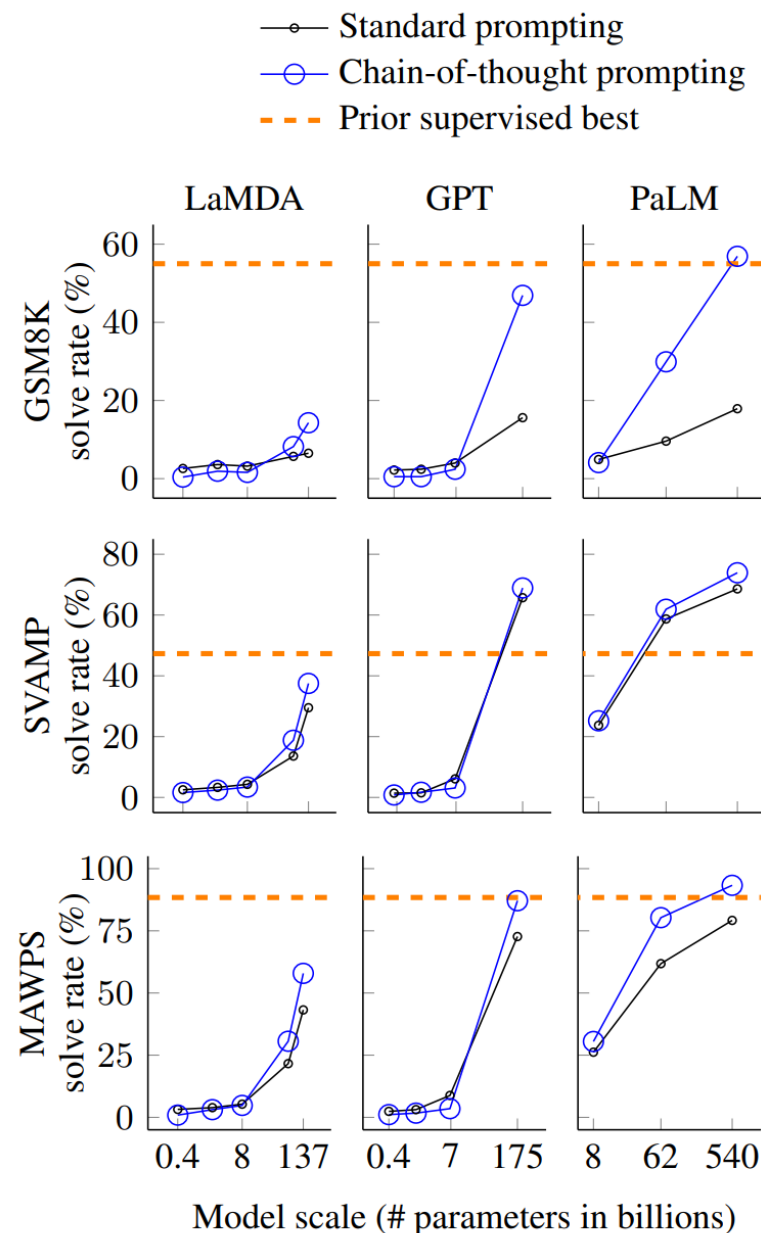
`>>> print(answer)`

74



思维链 (Chain-of-Thought, CoT)

打破Scaling Law的关键





自然指令学习

(Learning from Natural Instructions)

自然指令学习 (Learning from Natural Instructions)

Mach Learn (2014) 94:205–232
DOI 10.1007/s10994-013-5407-y

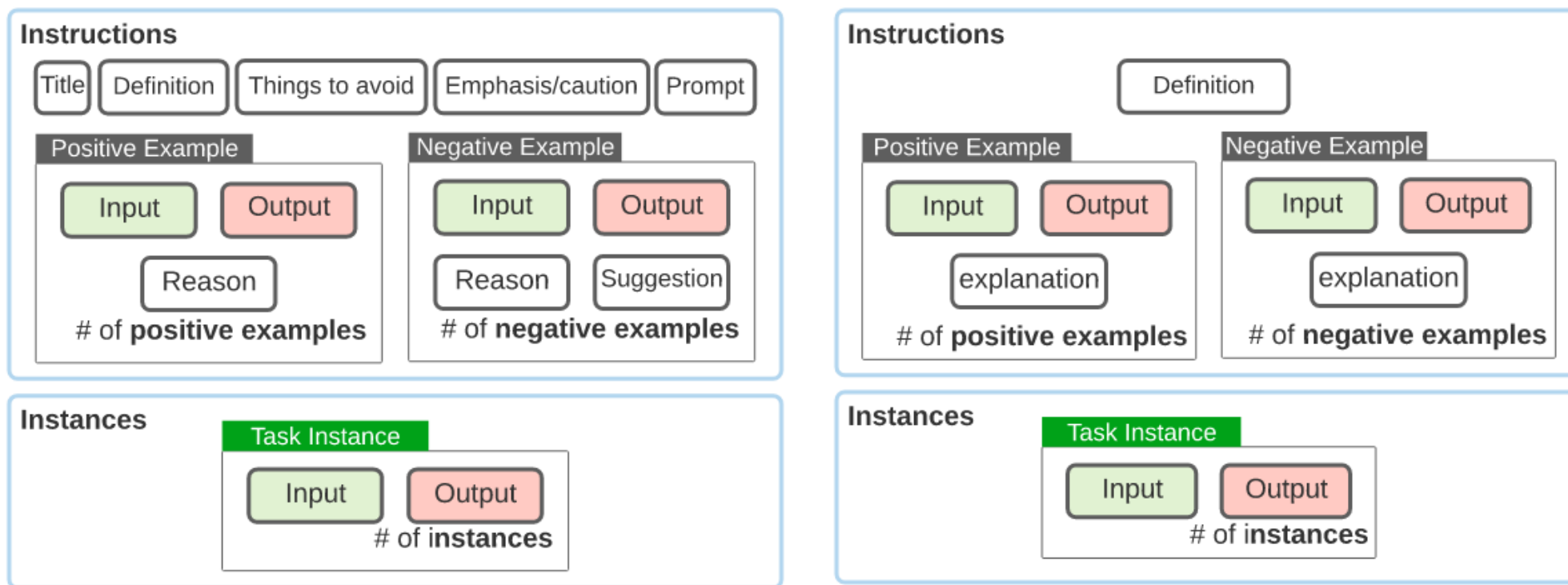
Learning from natural instructions

Dan Goldwasser · Dan Roth

Received: 15 May 2012 / Accepted: 8 August 2013 / Published online: 18 September 2013
© The Author(s) 2013

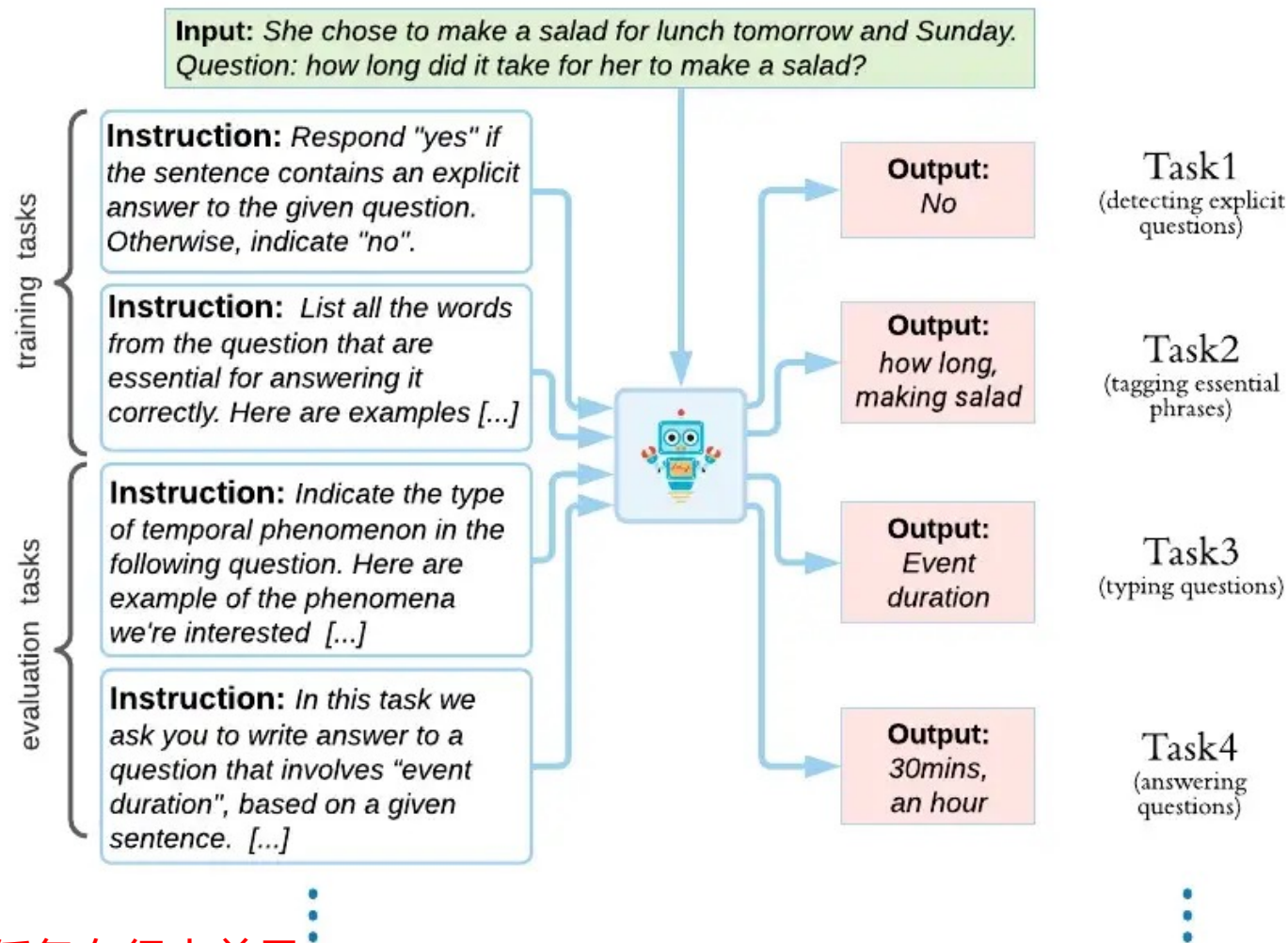
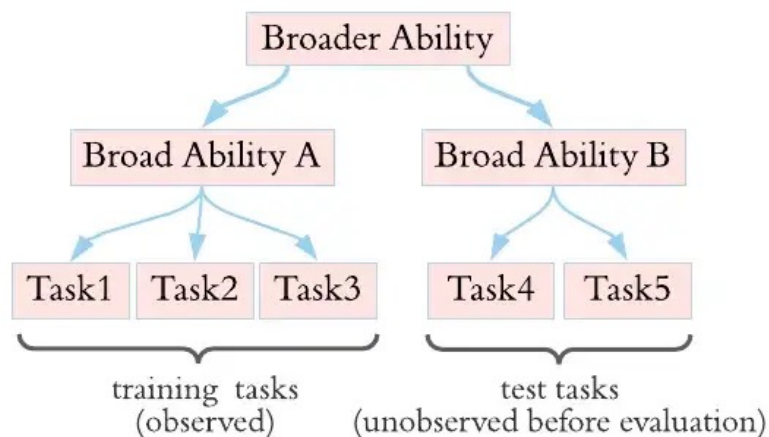
Schema of Instruction

Comparison of schema: v1.x (left) and v2.x (right)



<https://instructions.apps.allenai.org/>

自然指令学习 (Learning from Natural Instructions)



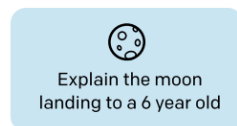
大幅提升了泛化能力，但是和人类的真实任务有很大差异。

通过人类反馈对齐人类指令

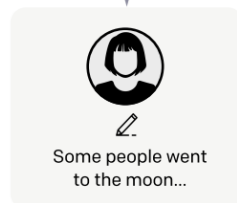
Step 1

Collect demonstration data, and train a supervised policy.

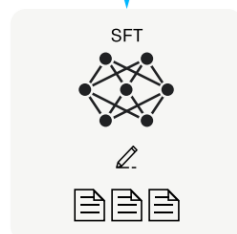
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



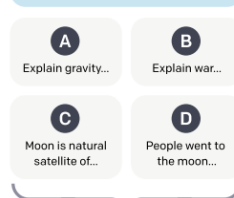
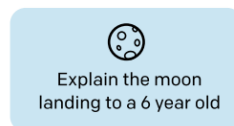
This data is used to fine-tune GPT-3 with supervised learning.



Step 2

Collect comparison data, and train a reward model.

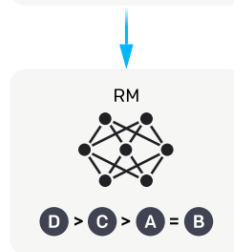
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



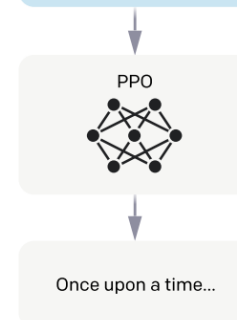
Step 3

Optimize a policy against the reward model using reinforcement learning.

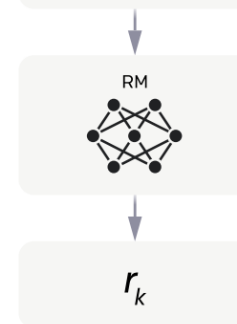
A new prompt is sampled from the dataset.



The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



ChatGPT: Optimizing Language Models for Dialogue

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.

Capability Development

GPT3



ChatGPT

Knows

Unknowns

Knows

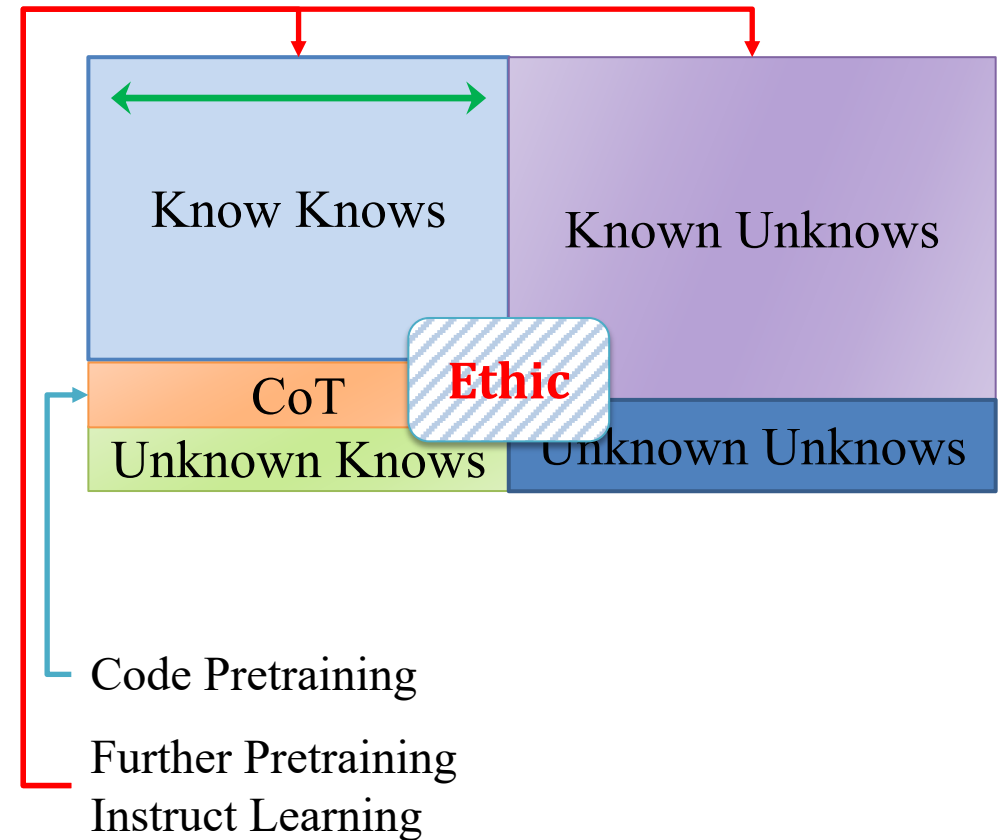
Know
Knows

Known Unknowns

Unknowns

Unknown
Knows

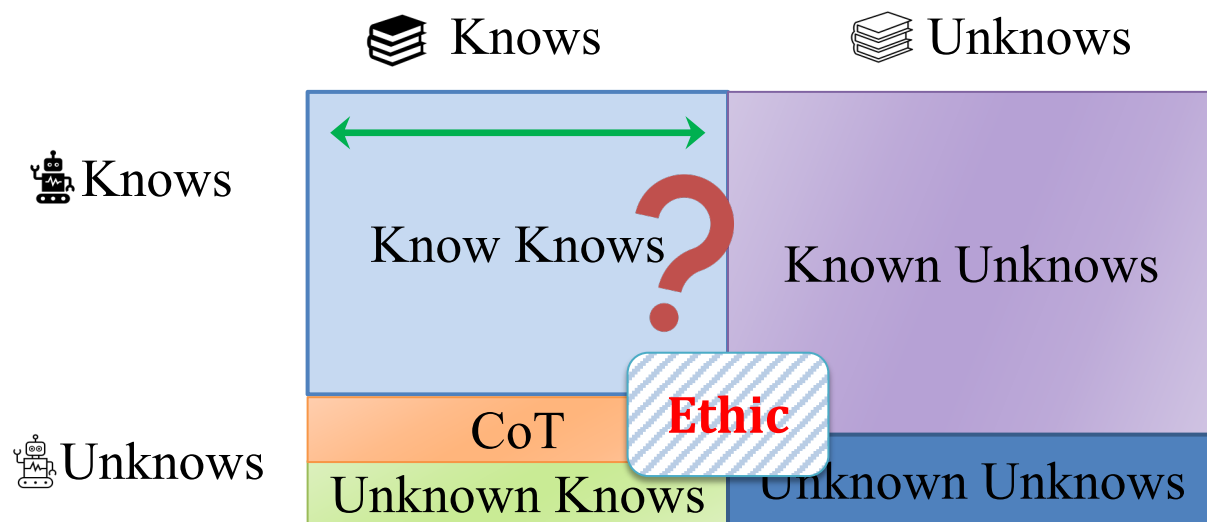
Unknown Unknowns





能力分析

Know-Unknown Quadrant



Holistic Evaluation of Language Models

Previous work

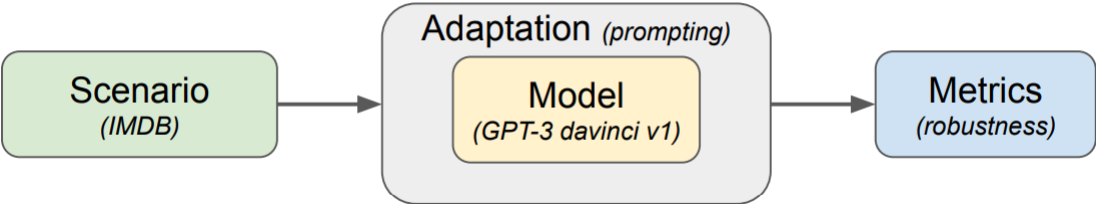
Scenarios	Metric	
	Natural Questions	✓ (Accuracy)
	XSUM	✓ (Accuracy)
	AdversarialQA	✓ (Robustness)
	RealToxicity Prompts	✓ (Toxicity)
	BBQ	✓ (Bias)

HELM

Scenarios	Metrics						
	Accuracy	Calibration	Robustness	Fairness	Bias	Toxicity	Efficiency
	RAFT	✓	✓	✓	✓	✓	✓
	IMDB	✓	✓	✓	✓	✓	✓
	Natural Questions	✓	✓	✓	✓	✓	✓
	QuAC	✓	✓	✓	✓	✓	✓
	XSUM	✓				✓	✓

Holistic Evaluation of Language Models

Holistic Evaluation of Language Models



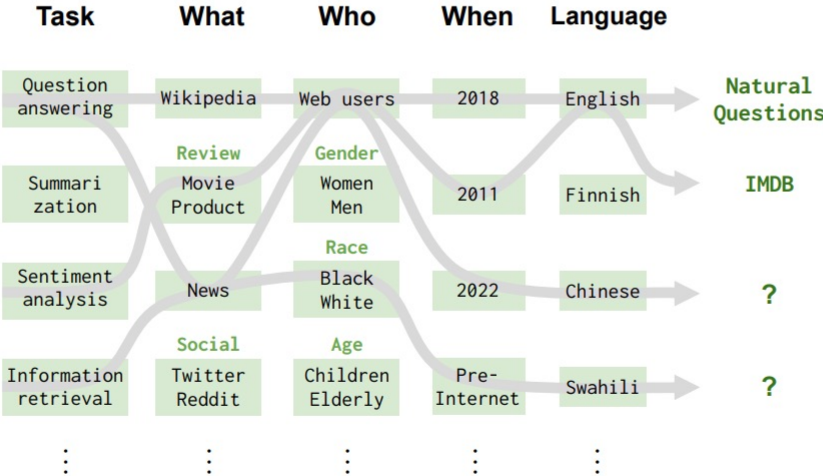
Previous work

Benchmark

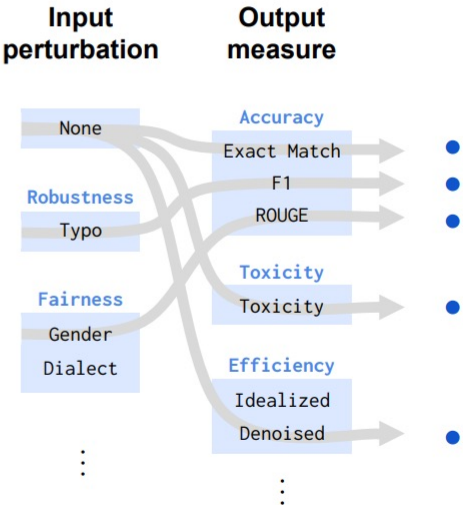
Natural Questions
 XSUM
 IMDB
 MS MARCO
 CivilComments
 WikiText-103
 WebNLG
 ANLI
 ⋮

HELM

Scenarios



Metrics



如何评价LLM？

- ▶ 当前LLM的Benchmark的局限性
 - ▶ 范围受限，多为单一领域
 - ▶ 单一的子任务，不够通用
 - ▶ 不够难
 - ▶ Benchmark难度的提升滞后于模型的迭代
 - ▶ 人类注释质量和噪声
 - ▶ 很难兼顾数量，多领域和数据标注质量
 - ▶ 不够直观或者过于复杂

GAOKAO-Bench:让Chat GPT做中国的高考题

▶ GAOKAO-Bench的优点:

- ▶ 同样具有挑战性但更灵活,
- ▶ 简单直观, 符合人类的直觉
- ▶ 同时满足数量多, 多领域和高质量
- ▶ 稳定, 有注释, 便于分析

▶ 测试方法

- ▶ 带有指示的直接提问
 - ▶ 实现方法: 建立题目数据库, 以API的形式请求测评
 - ▶ 规模: 10000+
- ▶ 不需要额外成本, 仅通过API调用, 全领域通用

GAOKAO-Bench: 让Chat GPT做中国的高考题

QV

Let's solve a Chinese college entrance examination mathematics question, please think carefully, think step by step, and choose the correct option: If $z=1+\mathrm{i}$. Then

$\mathrm{i} * z + 3 * \bar{z} = ()$

A. $4\sqrt{5}$

B. $4\sqrt{2}$

C. $2\sqrt{5}$

D. $2\sqrt{2}$



To solve this problem, we need to first expand $\mathrm{i} * z + 3 * \bar{z}$ using the rules of complex numbers. We have that:

$$\mathrm{i} * z = \mathrm{i} * (1 + \mathrm{i}) = \mathrm{i} - 1$$

$$3 * \bar{z} = 3 * (1 - \mathrm{i}) = 3 - 3\mathrm{i}$$

Therefore, $\mathrm{i} * z + 3 * \bar{z} = \mathrm{i} - 1 + (3 - 3\mathrm{i}) = -2 - 2\mathrm{i}$. To find the magnitude of a complex number, we use the formula $|a + b\mathrm{i}| = \sqrt{a^2 + b^2}$. Using this formula, we can calculate that $|-2 - 2\mathrm{i}| = \sqrt{2^2 + (-2)^2} = \sqrt{8} = 2\sqrt{2}$. Therefore, the correct answer is $\boxed{\textbf{(D)}} 2\sqrt{2}$.

Chat GPT求解一道高考数学题，
为了方便理解，题目中的公式预处理时被转为Latex格式

GAOKAO-Bench: 让Chat GPT做中国的高考题

我们以Chat GPT在2022高考全国卷上的结果为例：

GM

选择题，请运用你的中国哲学知识选择正确的答案，并说出你推理的过程。汉晋时期有多种文本记载，帝尧之时，“天下太和，百姓无事”。有老者“击壤”而戏，围观者称颂帝尧。老者歌云：“吾日出而作，日入而息，凿井而饮，耕地而食，帝何德于我哉！”上述记载所体现的政治理念最接近（ ）
A. 孔子 B. 老子 C. 韩非 D. 墨子



根据题干中给出的信息，上述记载所体现的政治理念最接近于老子的思想。老子认为，天下太平时人们不必干预政治，只需专注于自己的事业即可。因此，可以推断出本题的答案是B，老子。



引导Chat GPT做高考题的一个例子，Chat GPT展现了惊人地理解和推理能力

Chat GPT在2022高考全国卷客观题的表现

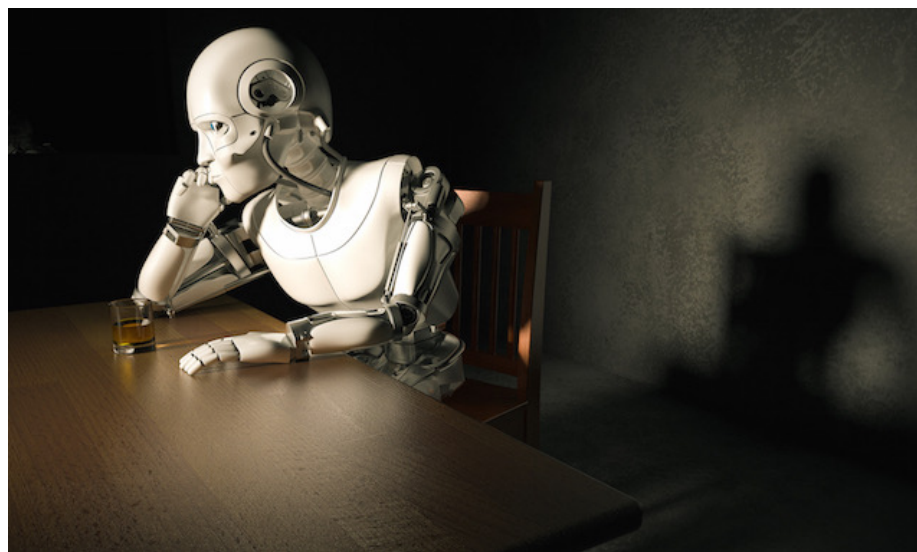
	得分率	准确率
语文	-	-
英语	93.0 / 95	56/60
地理	24/36	6/9
政治	44/48	11/12
历史	24/40	6/10
数学（理）	30/50	6/10
数学（文）	35/55	7/11
物理	0/12	0/2
化学	6/42	1/7
生物	18/30	3/5
合计	274/408= 0.67	96/126= 0.76

客观题：在排除了坏样本（带图题,听力题等）之后，在全部126个样本数可以达到**76%**的准确率和**67%**的得分率，其客观题能力与**两名**500分左右的高考生（文科和理科各一名）相当。

主观题：对Chat GPT结果的人工打分，在文科综合（历史，地理，政治）上取得了不错的成绩（得分率78%），在生物可以达到50%的得分率，而在数学，物理，化学，历史上则表现不佳（得分率低于30%）。

进一步改进

- ▶ LLM的高考分数能够说明什么?
 - ▶ 当LLM在高考上超越人类时会怎样?
- ▶ “听”，“看图”和“阅读”
 - ▶ 听: ASR -> Text -> LLM
 - ▶ 看图: MultiModal Model -> LLM
 - ▶ 阅读: 究竟是文本还是图片?



Limitations



Yann LeCun

@ylecun

Replying to [@bnjasim](#)

Limited in their current form.

Not easily controllable.

Very limited reasoning capabilities.

Superficial because disconnected the the underlying reality of the real world.

But potentially useful.

And progress will be made.

The big challenge for AI dialog systems over the next year or so is to make them factual, non-toxic, up to date, and capable of using tools like calculators, databases, search engines, simulators, or in this case, a simple calendar with today's date.

FA

list some research directions on large language models if i have not many GPUs



If you have limited access to GPUs, there are still many research directions you can explore in the field of large language models. Here are a few ideas:



1. Compression techniques: One direction to explore is to develop techniques to compress large language models without significantly sacrificing performance. This can involve methods like pruning or distillation, which can help reduce the number of parameters in a model.
2. Transfer learning: Another direction to explore is to investigate the effectiveness of transfer learning in language models. This involves fine-tuning a pre-trained model on a specific task or domain, rather than training a new model from scratch. Transfer learning can significantly reduce the amount of training required, and can be done with limited resources.
3. Multilingual models: Multilingual models can be trained on multiple languages, allowing for more efficient use of resources. By sharing parameters across languages, these models can be trained with less data and fewer GPUs.
4. Domain-specific models: Another direction is to train models for specific domains, such as biomedical or legal language. These models can be smaller in size than general language models and can be trained on smaller datasets.
5. Small-scale experiments: Finally, you can conduct small-scale experiments to explore different hyperparameters, architectures, or optimization algorithms. Even with limited GPUs, you can still conduct experiments on smaller datasets to gain insights into the behavior of different models.





谢谢