人工智能之自然语言处理

I.R. Lab
@shandong univ.

正在改变世界的对话式通用人工智能模型ChatGPT

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。**

SpaceX、特斯拉公司总裁 埃隆·马斯克：**ChatGPT好得吓人，我们离危险的强人工智能不远了。**

英伟达总裁黄仁勋：**ChatGPT是AI领域iPhone，是更伟大事物的开始。**

美国作家、Robust. AI公司创始人加里·马库斯：**生成式人工智能 将对社会结构产生切实的、迫在眉睫的威胁。**

Meta首席科学家、图灵奖得主杨立昆：就底层技术而言，**ChatGPT并不是多么了不得的创新。**虽然在公众眼中，它是革命性的，但是我们知道，**它就是一个组合得很好的产品**，仅此而已。

ChatGPT 的定义

- ChatGPT是2022年11月美国人工智能公司OpenAI所推出的**生成式对话预训练模型**。它通过对话的形式进行交互，对话的形式使得其**能够回答后续问题，承认自己的错误，质疑不正确的前提，并拒绝不适当的请求**。

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

自然语言处理发展历史

自然语言处理范式变迁

小规模专家知识
1950~1990

浅层机器学习算法
1990~2010

深度学习算法
2010~2017

大规模预训练模型
2018 ~

资源：规则库

1: If 源 =“我”, **then** 译 =“I”

2: If 源 =“你”, **then** 译 =“you”

3: If 源 =“感到满意”,
then 译 =“be satisfied with”

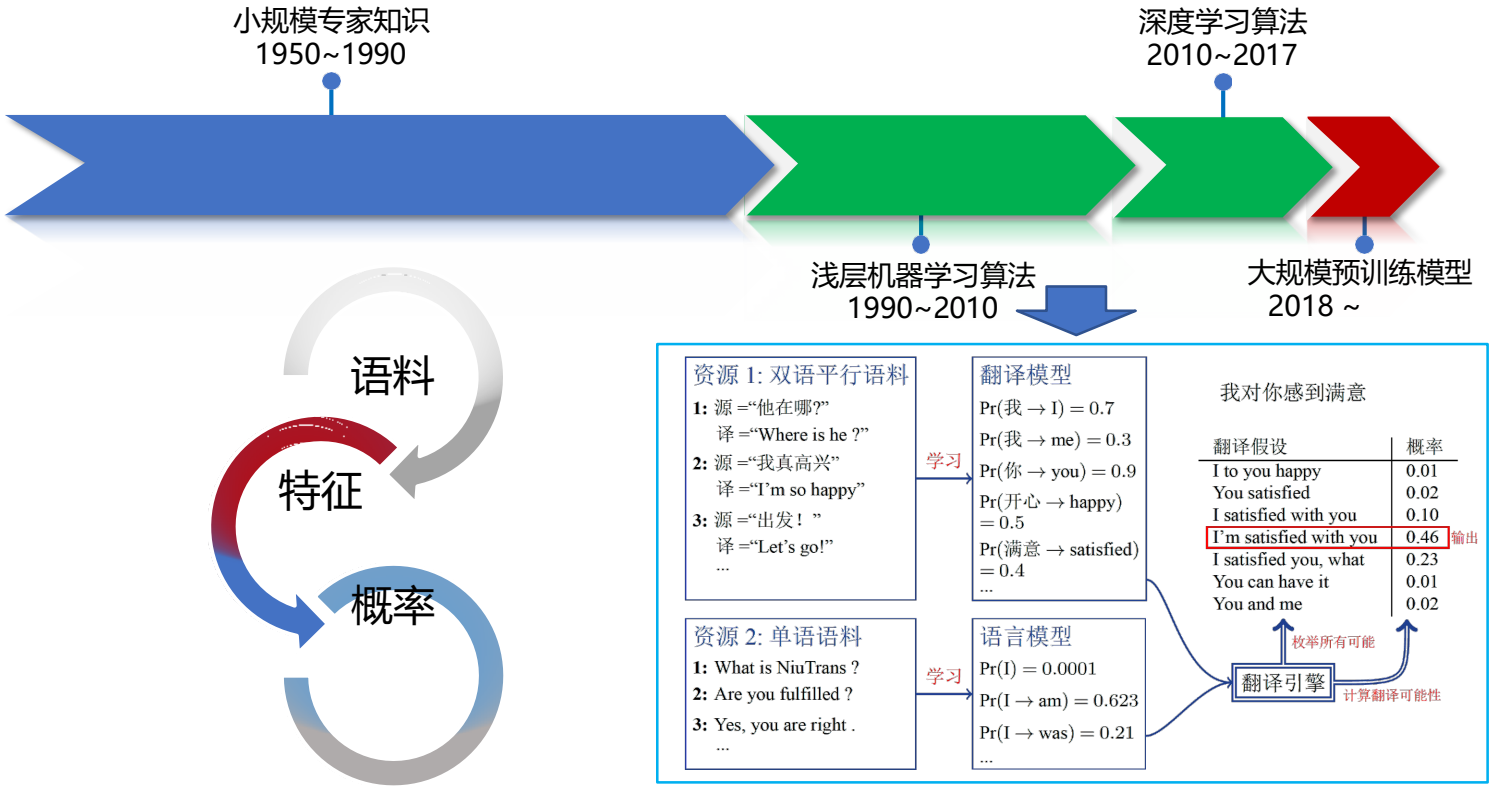
4: If 源 =“对... 动词 [表态度]”
then 调序 [动词 + 对象]

5: If 译文主语是 “I”
then be 动词为 “am/was”

6: If 源语是主谓结构
then 译文为主谓结构

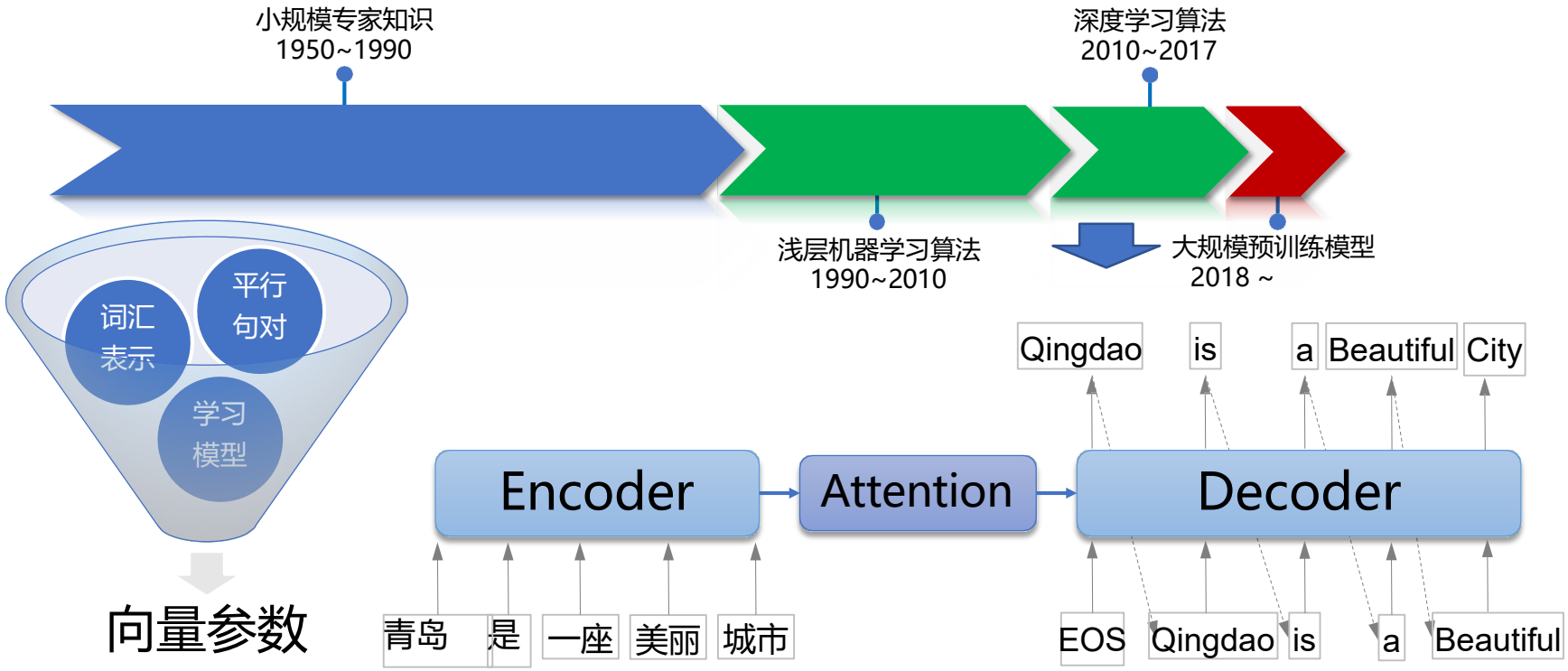
自然语言处理发展历史

自然语言处理范式变迁



自然语言处理发展历史

自然语言处理范式变迁



自然语言处理发展历史

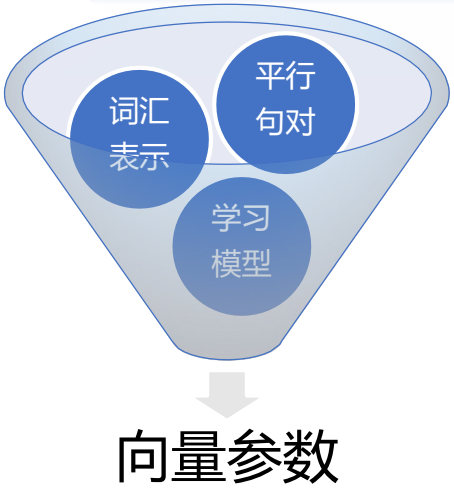
自然语言处理范式变迁

小规模专家知识
1950~1990

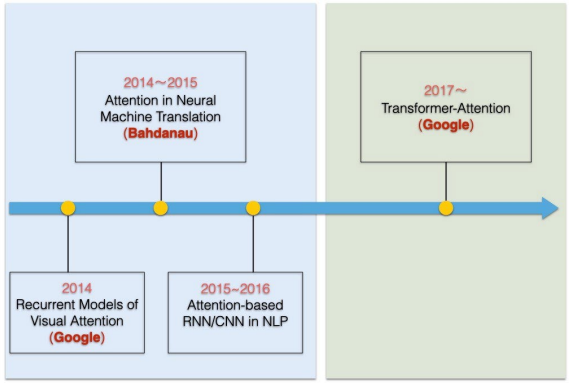
浅层机器学习算法
1990~2010

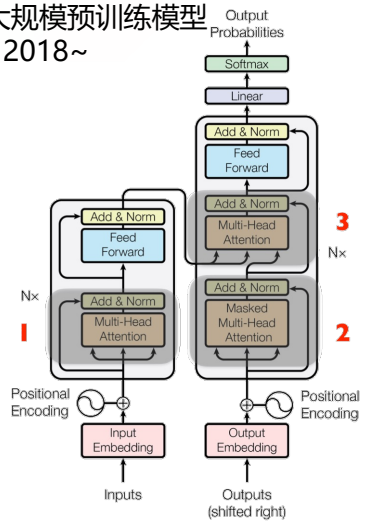
深度学习算法
2010~2017

大规模预训练模型
2018~



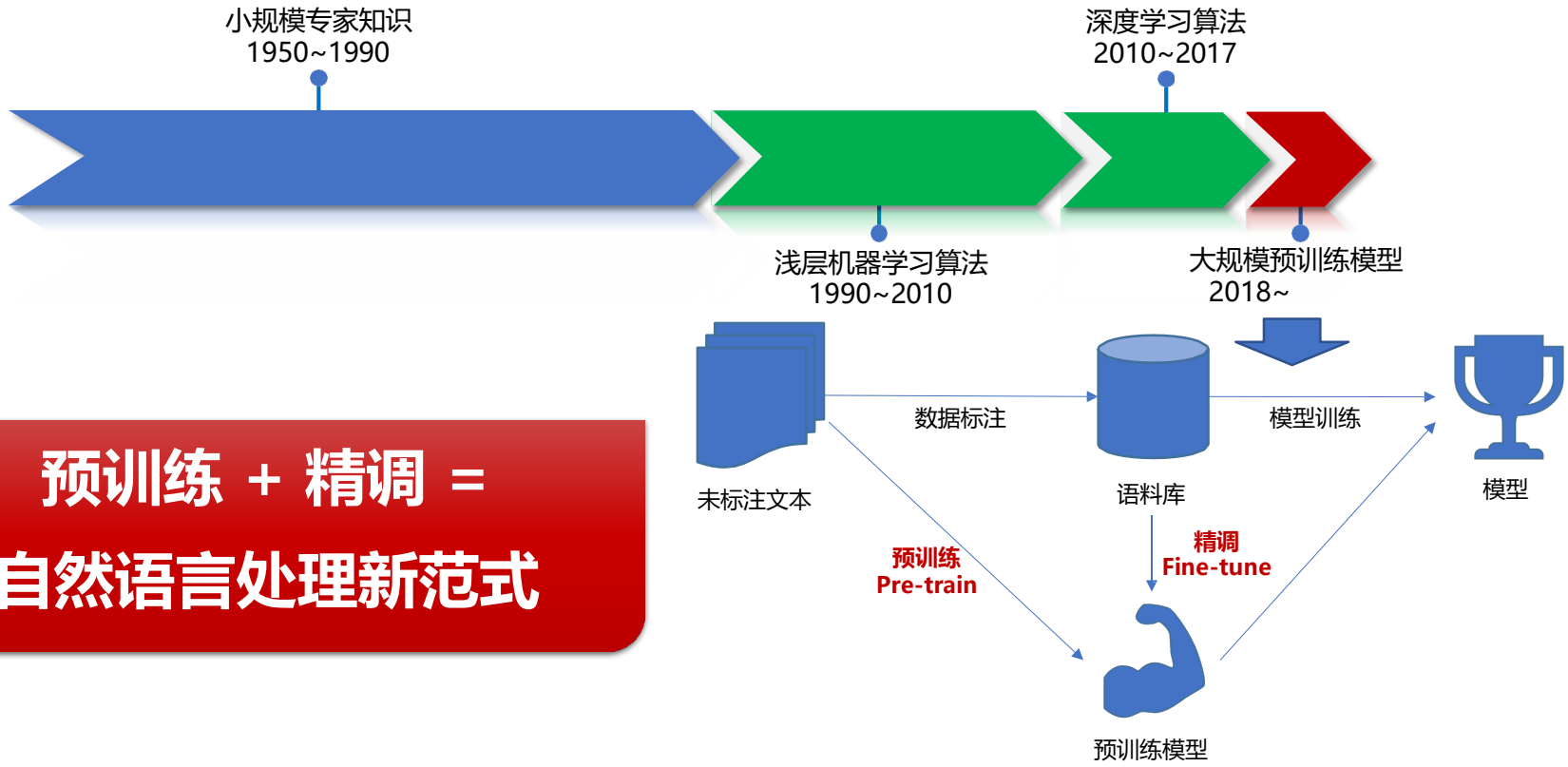
向量参数





自然语言处理发展历史

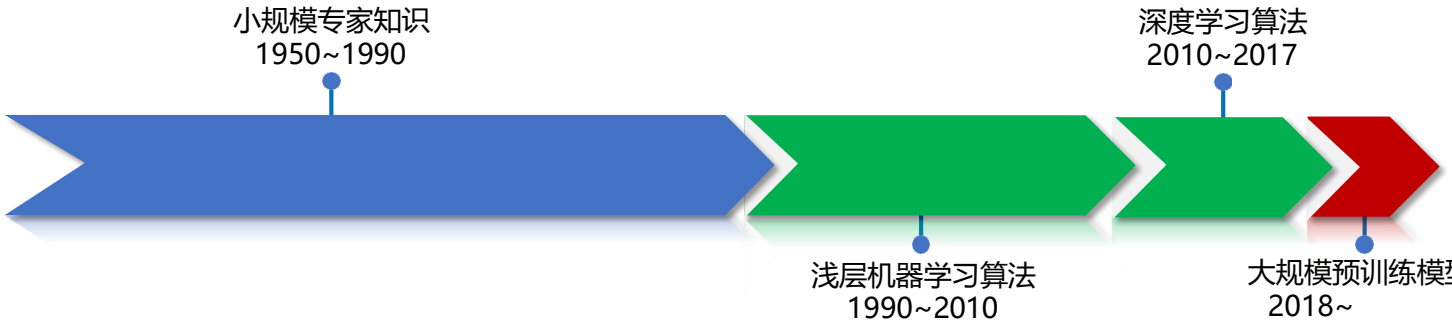
自然语言处理范式变迁



预训练 + 精调 =
自然语言处理新范式

自然语言处理发展历史

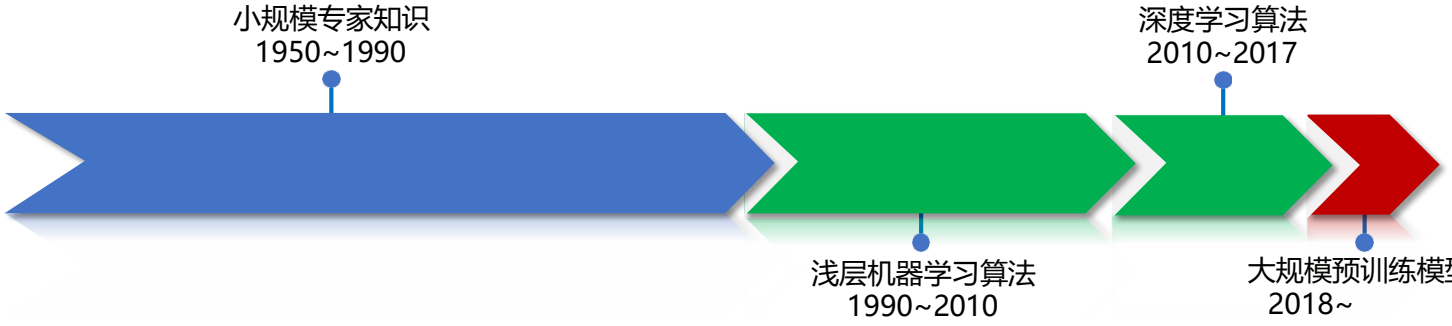
- 自然语言处理范式变迁

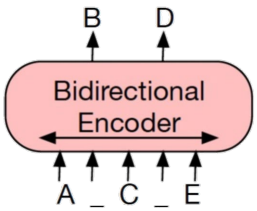


- 利用语言天然的顺序性
  - 我 喜欢 吃 土豆 炖 ***XX***
- 两种任务类型
  - 语言模型
    - 通过历史词序列预测**下一个词**
  - 完形填空
    - 通过周围的词预测**中间的词**

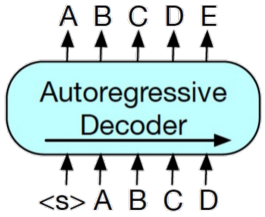
自然语言处理发展历史

自然语言处理范式变迁

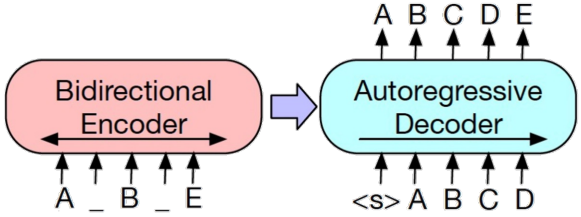




双向掩码模型
(2018)



单向自回归生成模型
(2018)

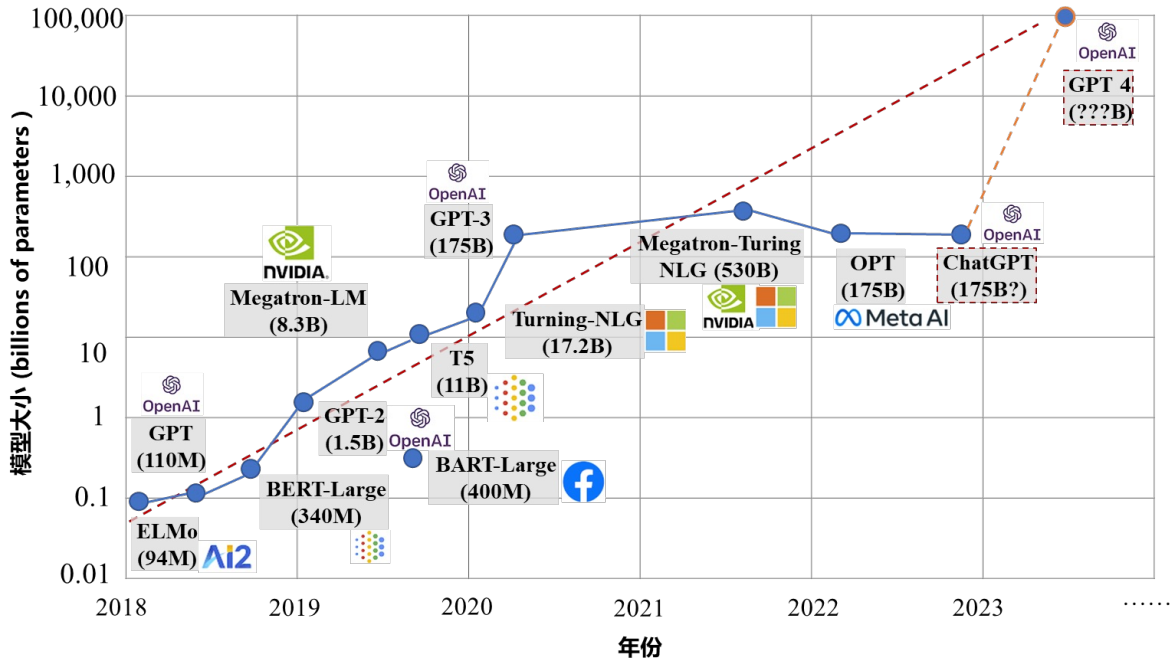


编码器 - 解码器架构
(2019)

预训练模型

预训练语言模型成为自然语言处理领域全新的技术范式

- 模型规模与表现**正相关**，因此不停追求越来越大的规模
- 随着模型规模越来越大，“**涌现**”出了令人惊讶的“智能”



Large Language Models: A New Moore's Law? - Hugging Face

发展历程

GPT-1: Improving Language **Understanding** by Generative Pre-Training

Keyword: unsupervised pre-training, supervised fine-tuning, auxiliary objective

2018

GPT-2: Language Models are Unsupervised **Multitask** Learners

Keyword: multi-task

2019

GPT-3: Language Models are Few-Shot Learners

Keyword: few-shot, one-shot, zero-shot

2020

InstructGPT: Training language models to follow instructions with human feedback

Keyword: instruct learning, **labeler-written prompts**, **reinforcement learning from human feedback**

2022.3

ChatGPT: Optimizing Language Models for Dialogue

2022.11

GPT-4

2023.3

ChatGPT is a **sibling** model to InstructGPT

过去

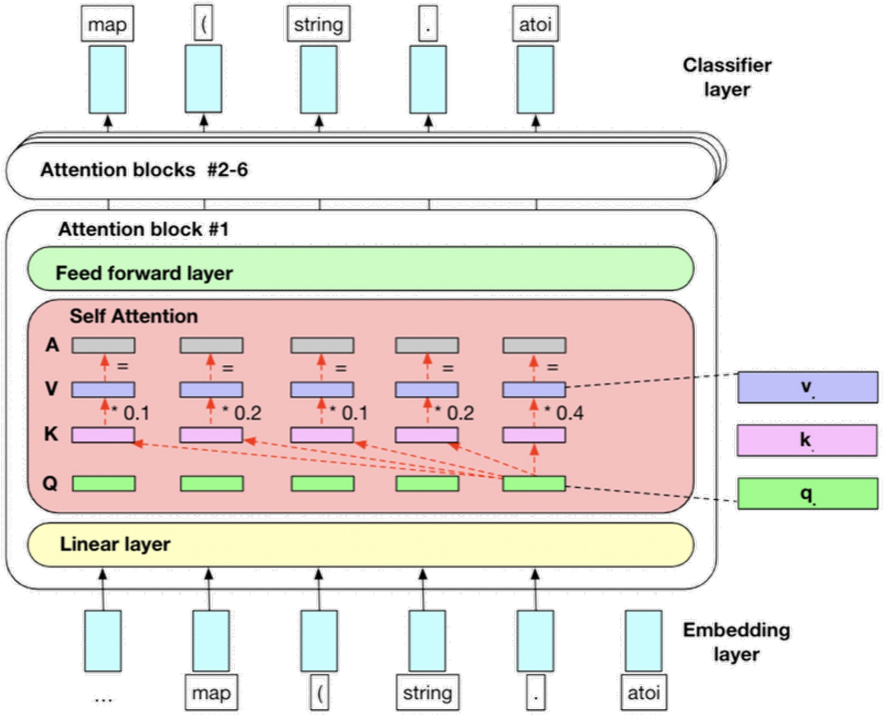
现在

模型结构与规模

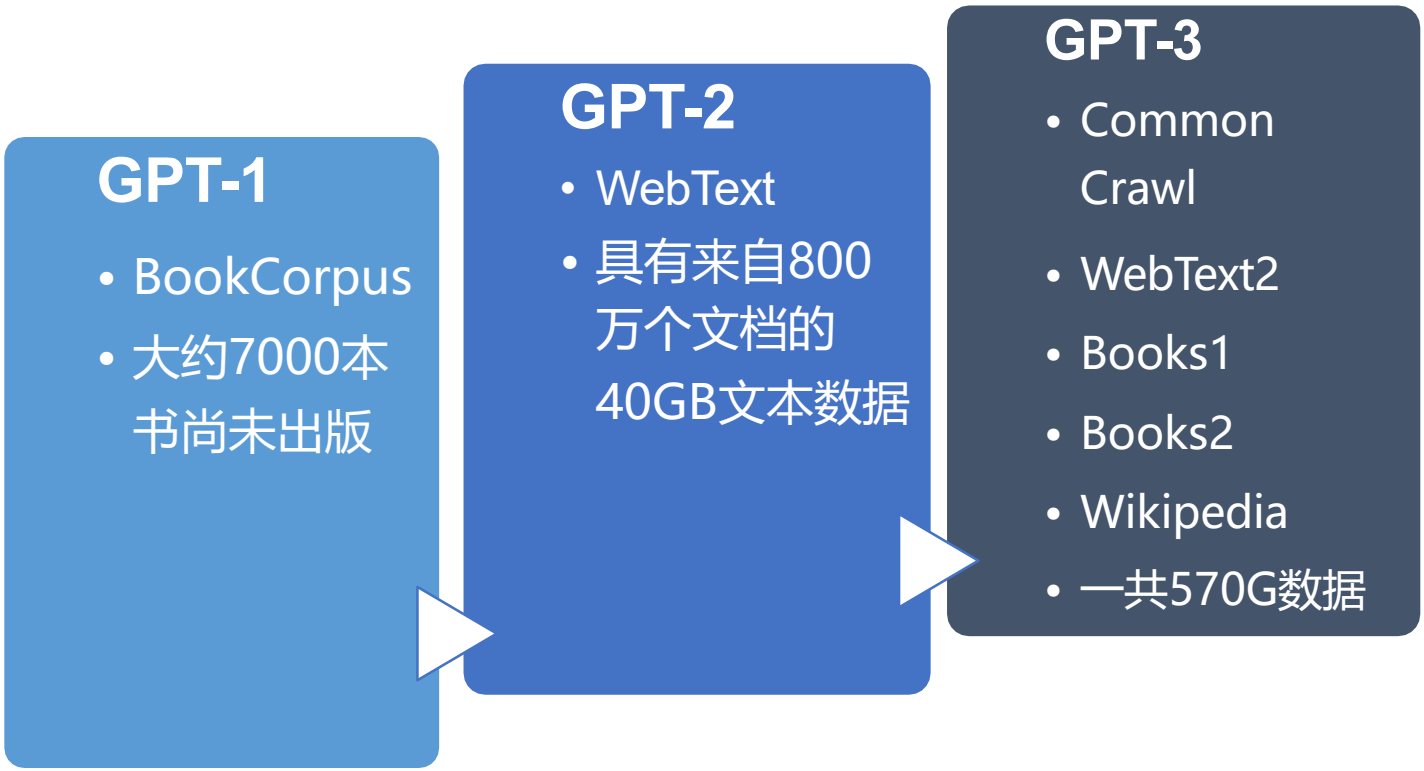
模型	规模
GPT	d_{model}=768, context_size=512, layer_num=12, attention_num=12
GPT-2	d_{model}=1600, context_size=1024, layer_num=48, attention_num=12, param=1.5B, size=774M
GPT-3	d_{model}=12288, context_size=2048, layer_num=96, attention_num=96, param=175B, size=70G

十倍

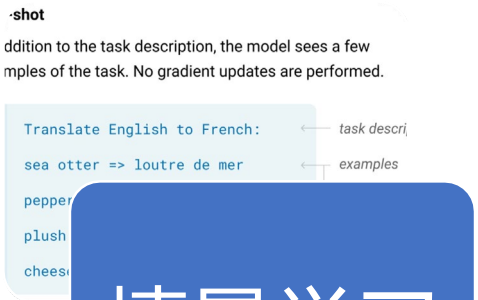
百倍



训练数据

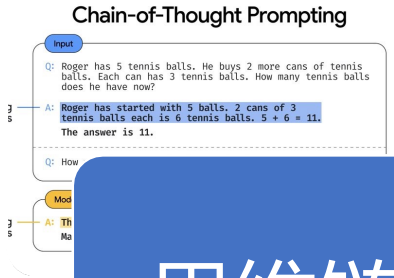


ChatGPT/InstructGPT的成功之处



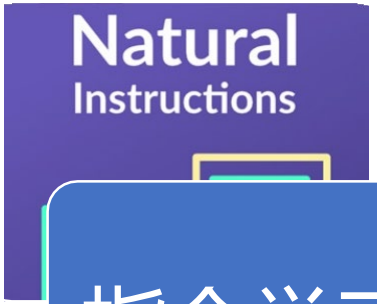
情景学习

大模型的涌现能力
改变传统学习范式



思维链

大模型的涌现能力
打破模型参数约束



指令学习

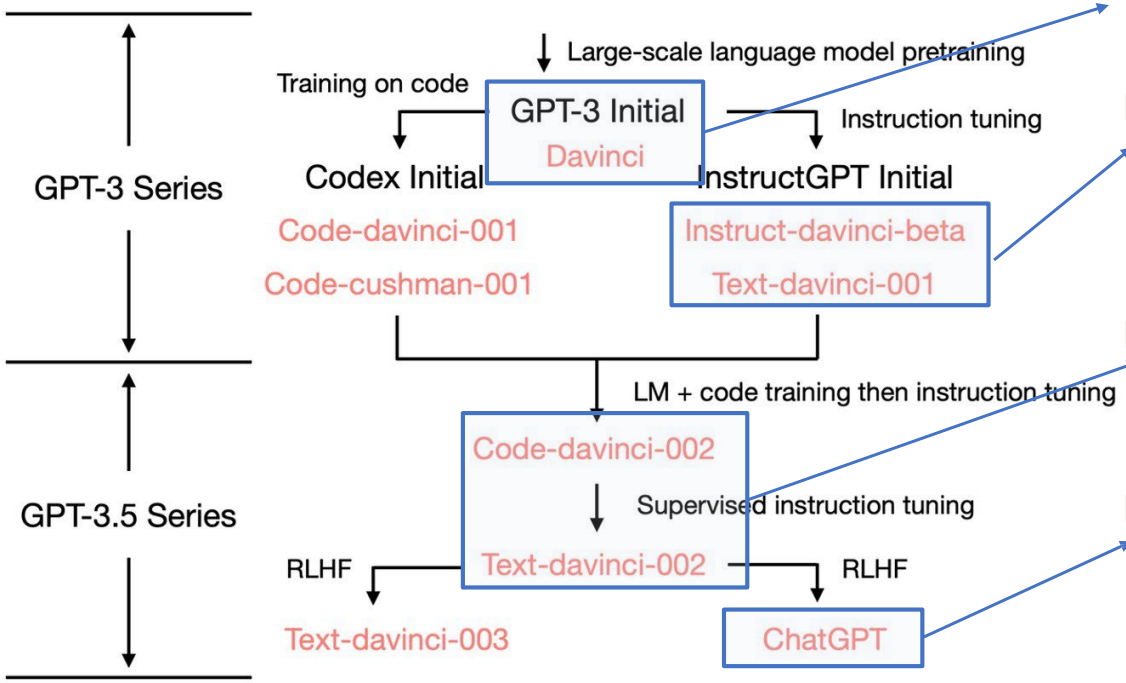
人在环路增强
对齐人类意图

ChatGPT的三个关键能力

InstructGPT演进路径（能力猜测）

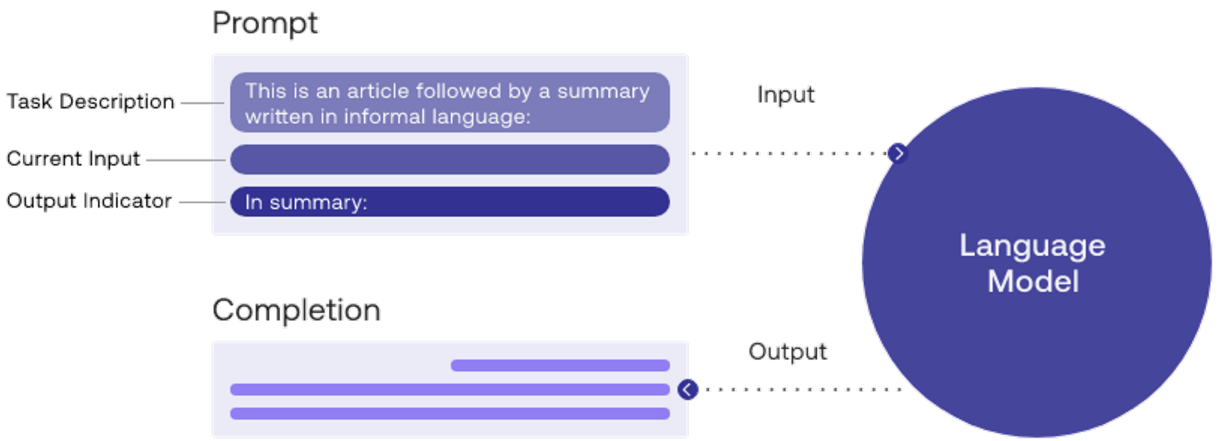
InstructGPT 的进化树

https://beta.openai.com/docs/model-index-for-researchers

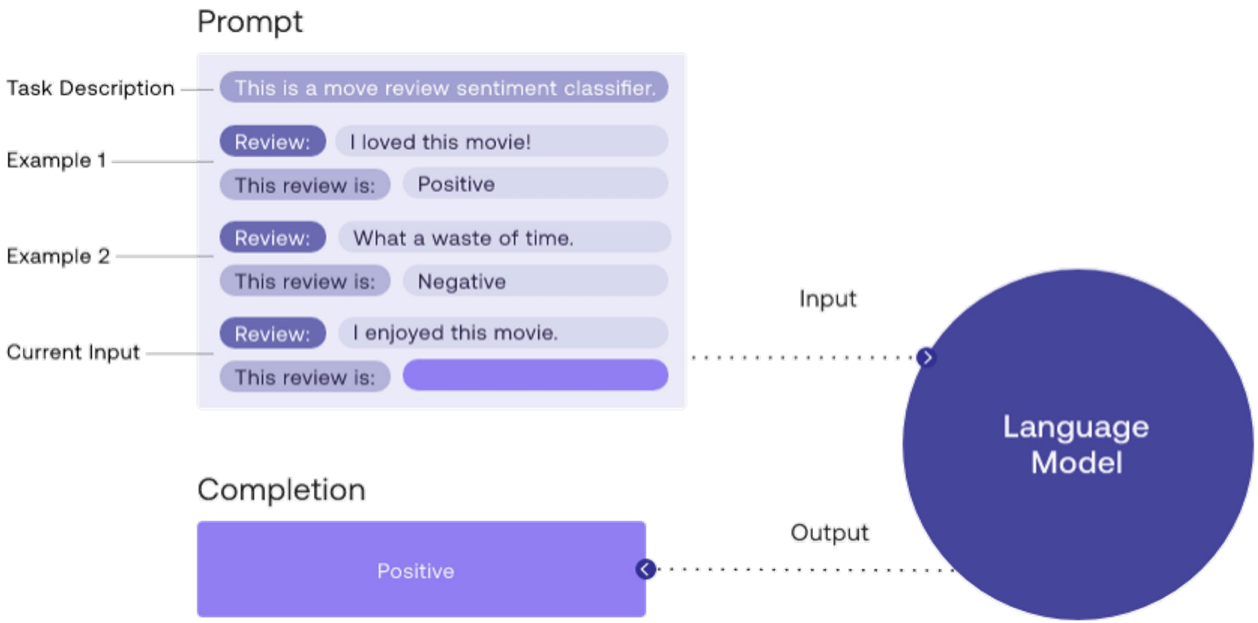


- 底座能力：大规模预训练模型
  - 模型规模足够大才能有“涌现”的潜力
- 情景学习：Instruction Tuning
  - 将任务用Prompt形式统一
  - 精调语言模型（Instruction Tuning）
  - 模型能够处理未见任务
- 思维链能力：在代码上进行继续预训练
  - 代码分步骤、模块解决问题
  - 涌现出逐步推理能力（COT）
- 和人类对齐能力：RLHF
  - 结果更符合人类的预期（多样性、安全性）
  - 利用真实用户的反馈（AI正循环、飞轮）

指令微调（Instruction Tuning）



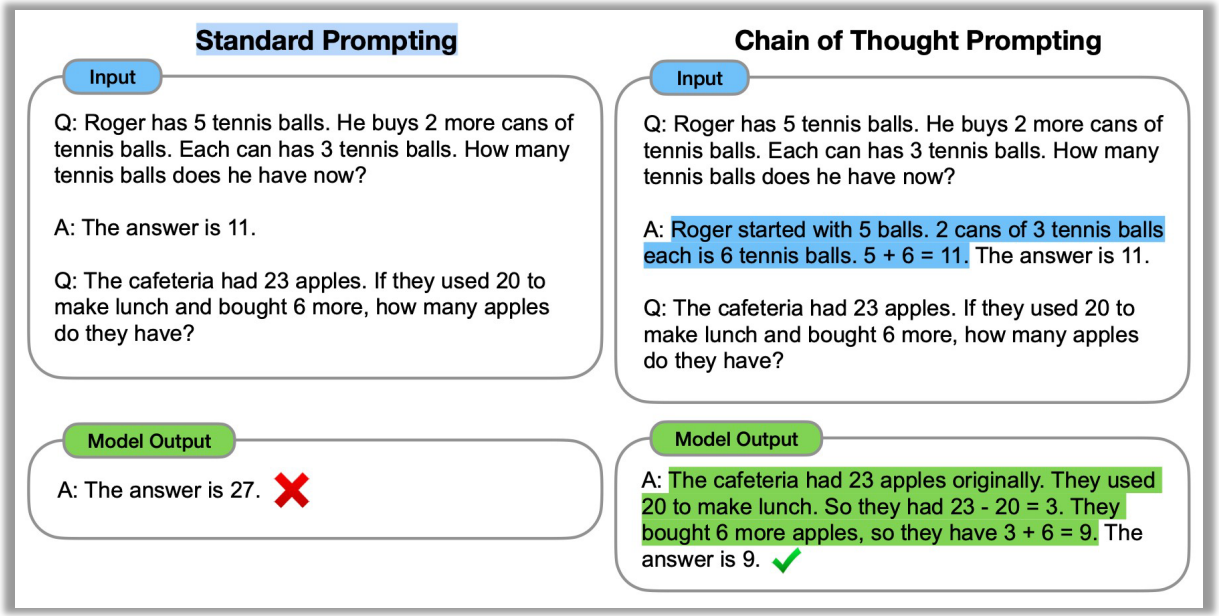
情景学习（In-Context Learning）



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

思维链（Chain-of-Thought, COT）

- 思维链表示一系列中间推理步骤，相当于在求解问题过程中将解题步骤也写出来



Reinforcement Learning from Human Feedback (RLHF)

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.

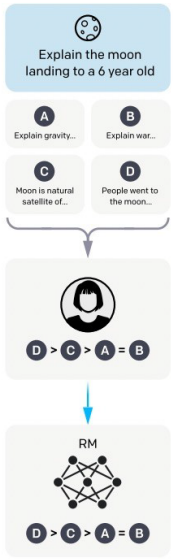
Step 2

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

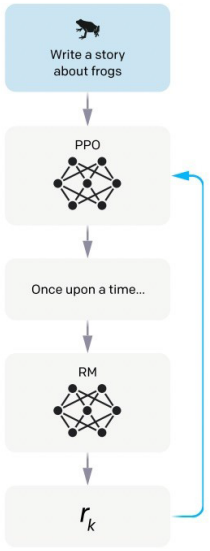
Optimize a policy against the reward model using reinforcement learning.

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.



RLHF

InstructGPT models

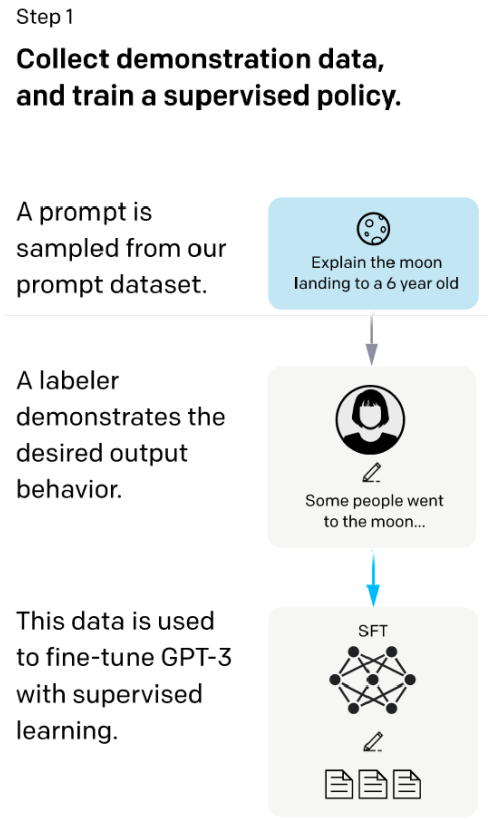
We offer variants of InstructGPT models trained in 3 different ways:

TRAINING METHOD	MODELS
SFT Supervised fine-tuning on human demonstrations	`davinci-instruct-beta`[1]
FeedME Supervised fine-tuning on human-written demonstrations and on model samples rated 7/7 by human labelers on an overall quality score	`text-davinci-001`, `text-davinci-002`, `text-curie-001`, `text-babbage-001`
PPO Reinforcement learning with reward models trained from comparisons by humans	`text-davinci-003`

The SFT and PPO models are trained similarly to the ones from the InstructGPT paper. FeedME (short for "feedback made easy") models are trained by distilling the best completions from all of our models. Our models generally used the best available datasets at the time of training, and so different engines using the same training methodology might be trained on different data.

第一步: 有监督微调（SFT）-1

- 训练 InstructGPT-beta 版本
  - SFT (Supervised fine-tuning)：在人工书写的示例上进行有监督微调，该方式得到的模型有 davinci-instruct-beta
- 标注人员手写 prompts
  - Plain: 标注人员提出任意一个任务，同时保证任务的多样性
  - Few-shot: 要求标注人员提出一个指令，以及在该指令下的多轮“查询-回复”
  - User-based: 根据用户在 OpenAI API 各种应用程序中提交过的用例（涵盖GPT3 API）



数据集的构建：用户 prompts 数据样例

- 数据样例

Use Case	Example
brainstorming	List five ideas for how to regain enthusiasm for my career
brainstorming	What are some key points I should know when studying Ancient Greece?
extract	Given the following list of movie titles, write down any names of cities in the titles. {movie titles}
generation	Write a creative ad for the following product to run on Facebook aimed at parents: Product: {product description}
chat	This is a conversation with an enlightened Buddha. Every response is full of wisdom and love. Me: How can I achieve greater peace and equanimity? Buddha:

第一步: 有监督微调（FeedME）-2

- FeedME (Feedback Made Easy)：在人工书写的示例以及标注者选择的模型最佳输出上进行有监督微调，该方式得到的模型有 text-davinci-001,text-davinci-002
  - 标注人员手写 prompts，为labeler
  - 通过开源text-davinci-001收集了更多的prompts，customer
  - FeedME (Feedback Made Easy)：选择模型最佳输出，无需标注,7/7（具体细节未知）

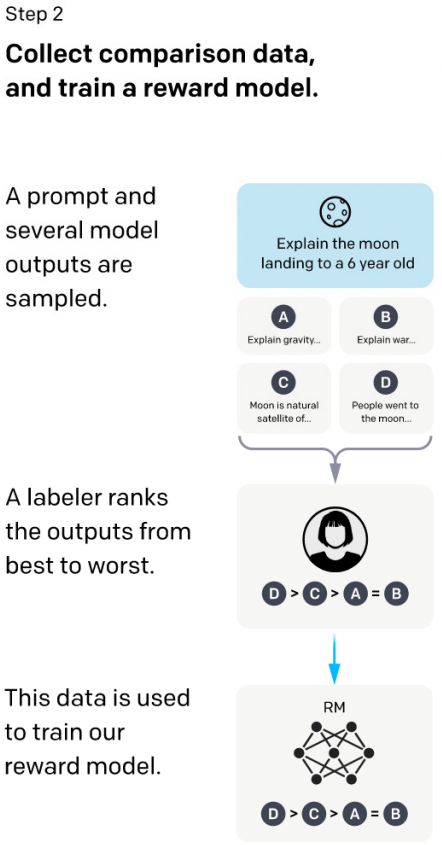
SFT Data		
split	source	size
train	labeler	11,295
train	customer	1,430
valid	labeler	1,550
valid	customer	103

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%

第二步：训练奖励模型

- 收集排序数据，训练奖励模型
  - 采样出一条 prompt 以及第一步模型的多条输出
  - 标注人员对模型的输出进行由好到坏的排序
  - 奖励模型由参数量为**6B的SFT模型初始化**，输入prompt以及第一阶段模型的回复，输出是0-1之间的分数。利用排序好的数据，根据Pairwise Ranking Loss优化奖励模型来模拟标注人员的偏好

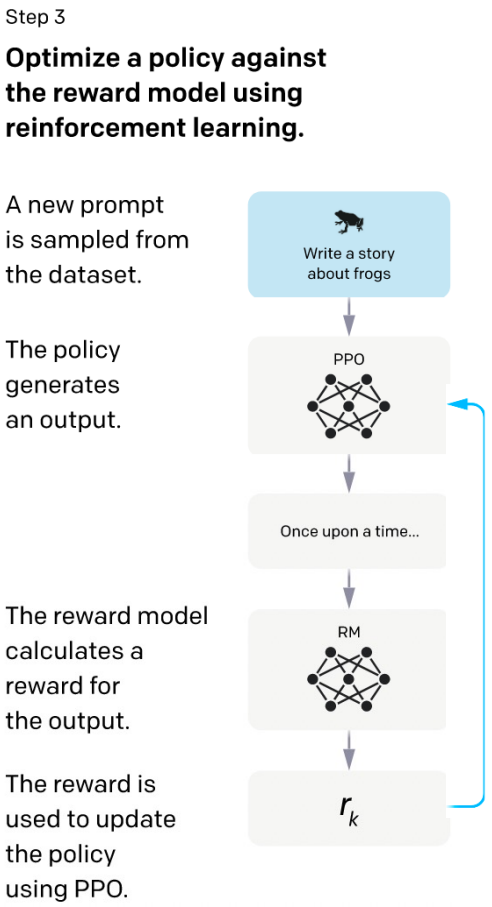
RM Data		
split	source	size
train	labeler	6,623
train	customer	26,584
valid	labeler	3,488
valid	customer	14,399



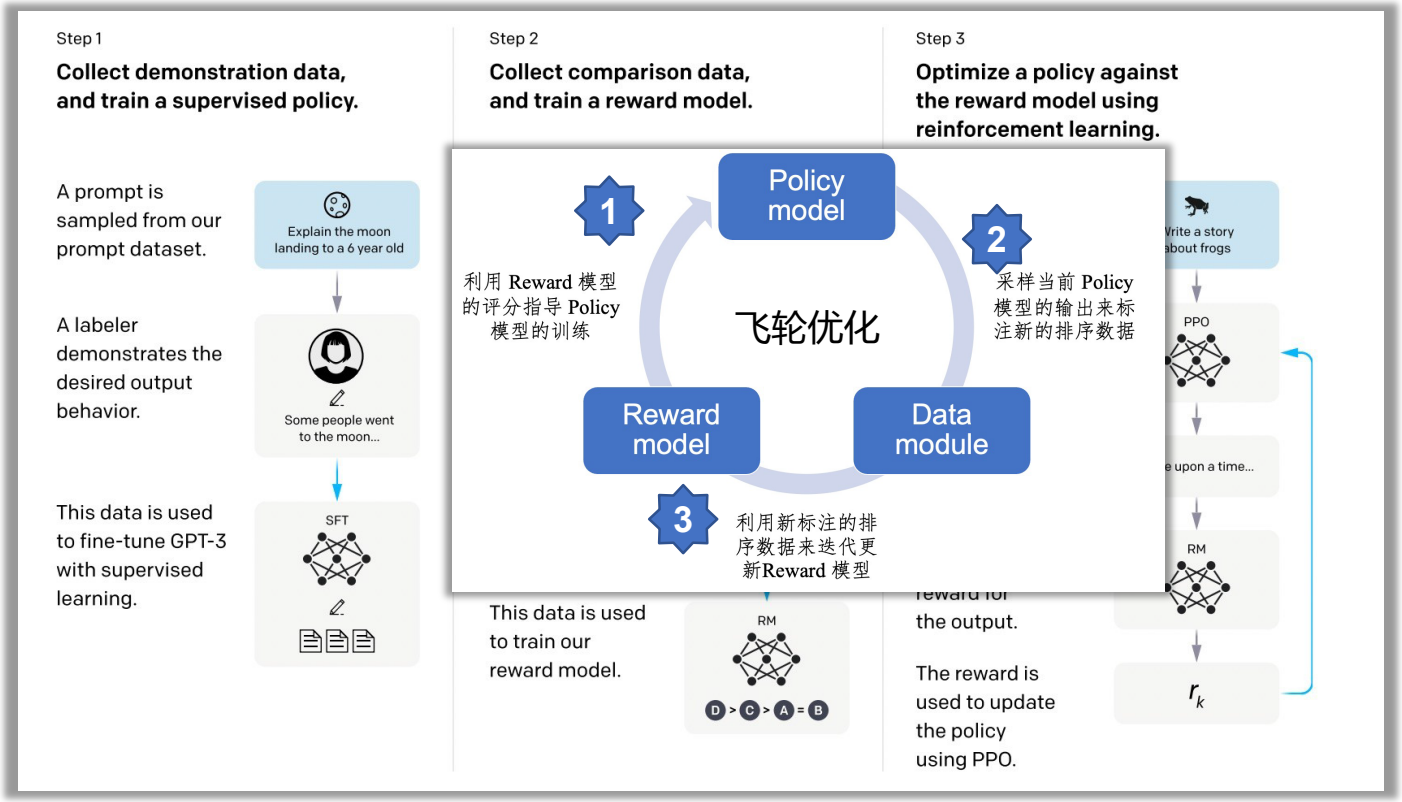
第三步：强化学习

- 使用强化学习PPO算法优化policy
  - 从数据集中采样出一条新的prompt
  - Policy模型首先利用第一阶段微调得到的SFT模型初始化，然后根据prompt生成对应的模型输出
  - 第二步训练得到的奖励模型对该输出计算reward，并利用该reward通过proximal policy optimization (PPO) 算法优化Policy

PPO Data		
split	source	size
train	customer	31,144
valid	customer	16,185

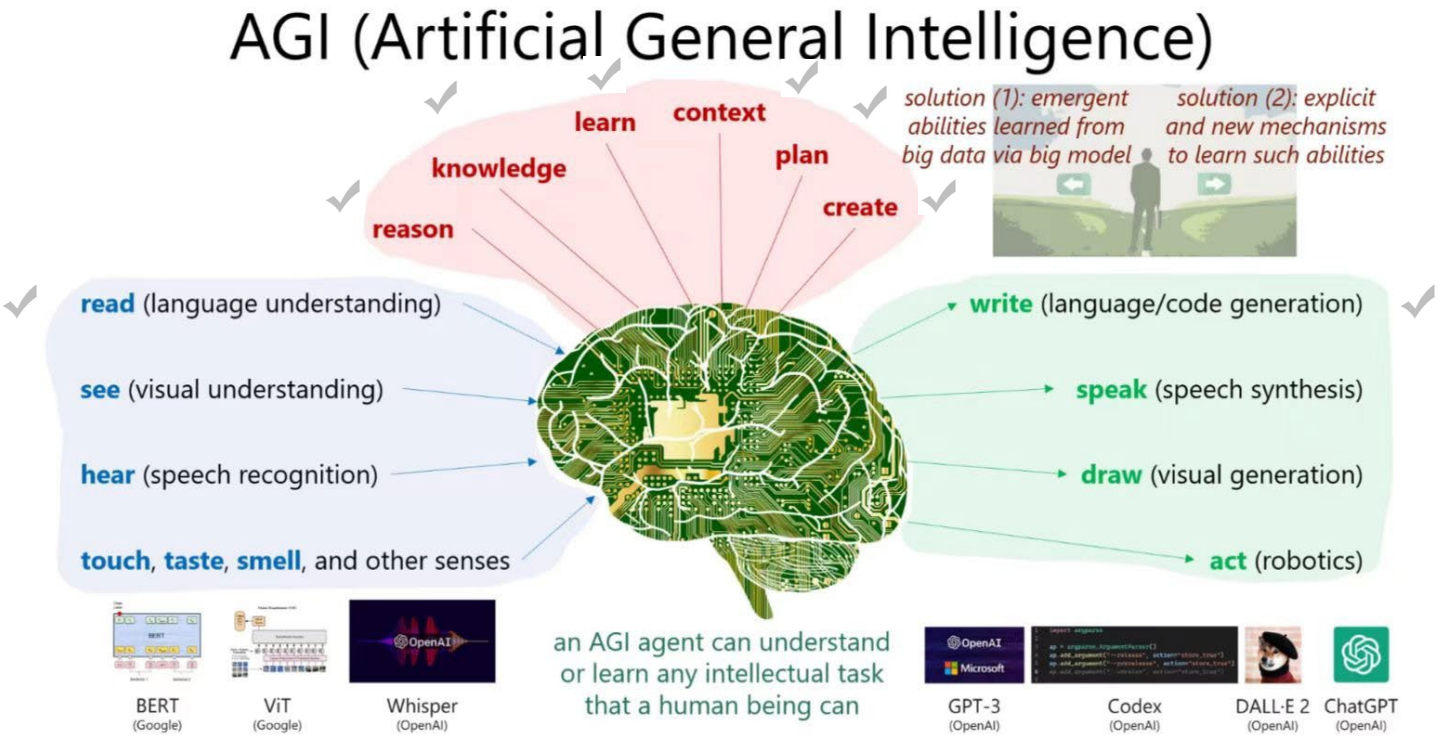


第四步：飞轮优化



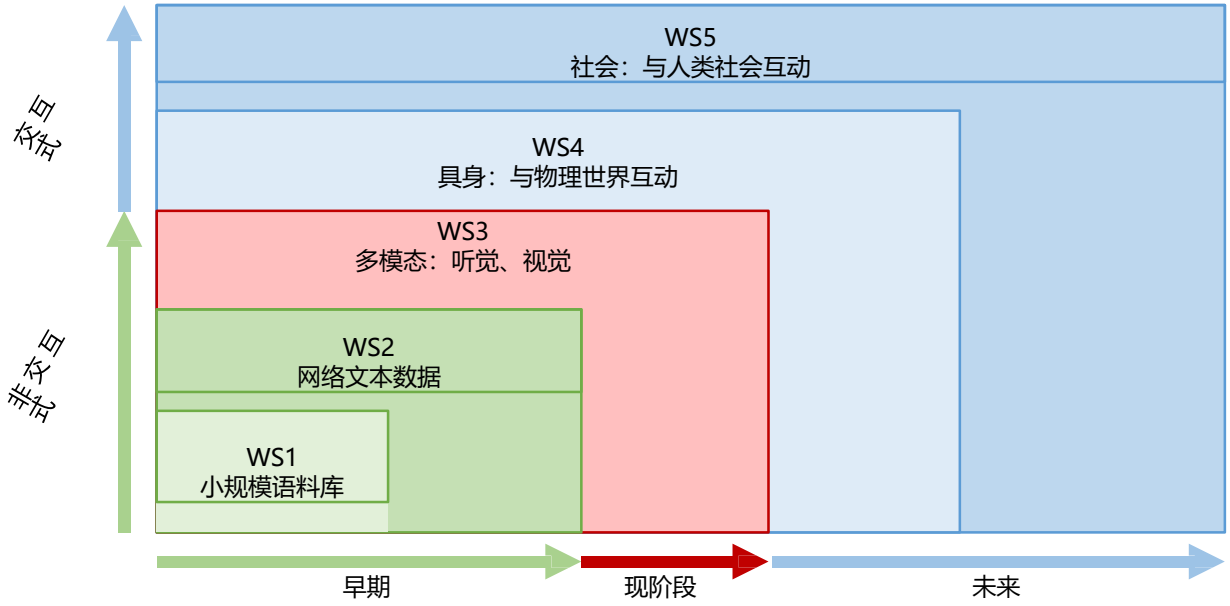
点亮更多的技能

ChatGPT将加速通用人工智能的实现。



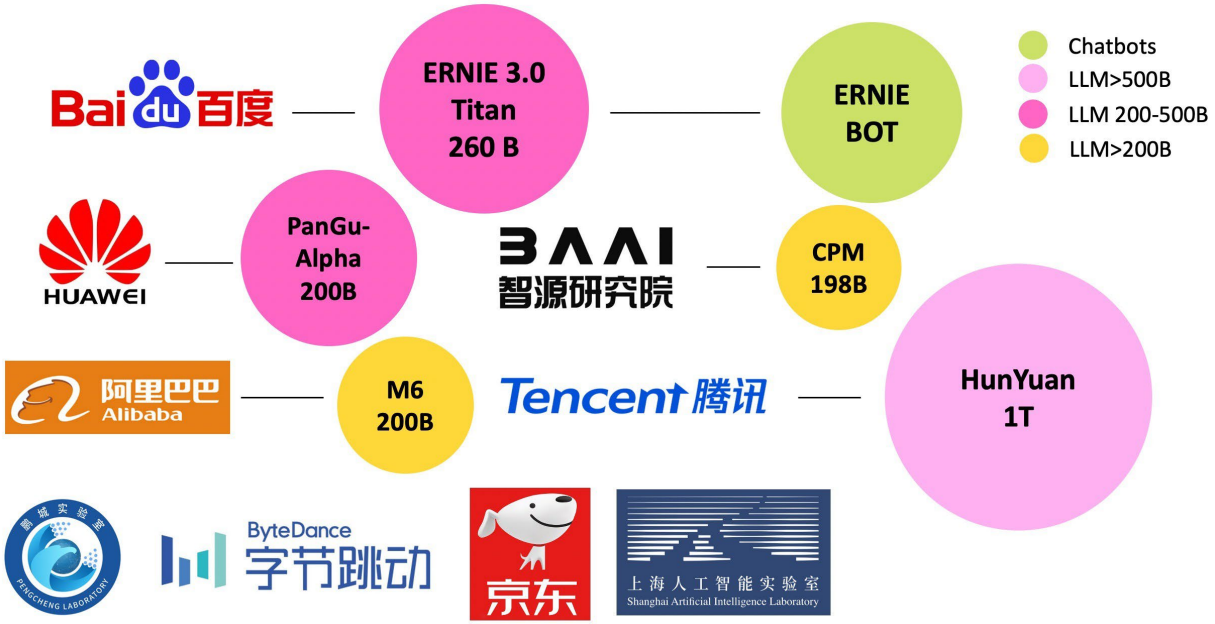
扩大交互范围

- 模型将继续沿着“**同质化**”和“**规模化**”的道路发展
- 拓展除语言之外的认知能力，寻找新的“**知识**”来源
  - 规则 → 算法 → 数据 → **体验**（Experience）
  - Bisk等人（2020）将其称为“**世界范围**”（World Scope，WS）



我们如何应对ChatGPT的挑战?

- 自主可控
  - 联合企业或超算中心，训练自己的大模型
  - 以开源大模型（OPT、BLOOM等）为基础继续预训练



谢谢！