大语言模型

李军毅 新加坡国立大学

2025.1.7

目录

AI Box

- 发展历程
- 预训练
- 微调对齐
- 能力利用

目录

AI Box

自然语言处理发展

AI Box

NLP研究发展历程

符号规则	统计机器学习	神经网络	预训练语言模型	超大规模语言模型
给定一组人工设定的规则，计算机通过对数据应用这些规则来模拟自然语言理解	在人工标注的数据上进行特征工程，训练机器学习模型参数，并将模型应用于测试数据上	用神经网络在大量数据上训练，使得网络自主学会提取特征，并可以灵活搭建模型	基于无标注文本预训练语言模型，通过“预训练-微调模式”工作，有监督学习	通过扩展语言模型规模，可以通过提示学习、情境学习等无需微调方式求解任务
有监督学习 任务泛化性弱			仍需要监督学习 任务泛化性改善	无/弱监督学习 一定的通用能力

LR、SVM、HMM、CRF、LDA
RNN/LSTM、BILSTM、ELMo、Transformer、GPT-1/2、BERT

什么是语言模型

AI Box

语言模型（Language Model）

- 预测一个句子在语言中出现的概率

$$p(s) = p(w_1, w_2 \ldots, w_m)$$

- 语言模型的发展历程：

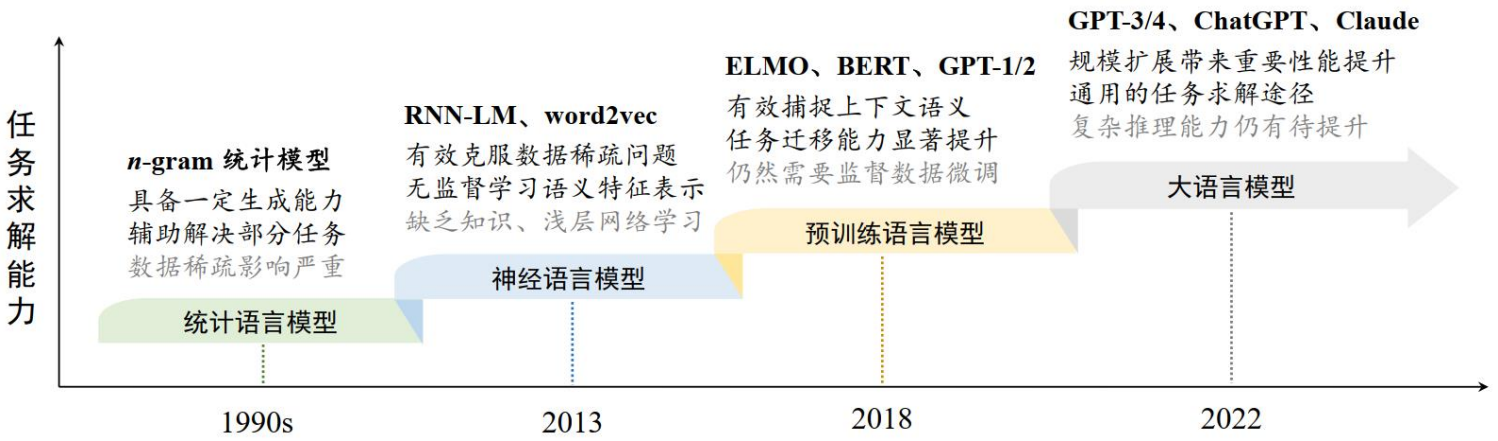


图 1.2 基于任务求解能力的四代语言模型的演化过程（图片来源：[10]）

什么是语言模型

- 统计语言模型（SLM）
  - N-gram LM：基于马尔可夫假设，当前词概率仅与前$n-1$词有关

$$p(s) = p(w_1)p(w_2|w_1)\dots p(w_m|w_{m-n+1},\dots,w_{m-1})$$
$$= \prod_{i=1}^{m} p(w_i|w_{i-n+1},\dots,w_{i-1})$$

  - 例：$p(I, am, tired) = p(I) * p(am|I) * p(tired|I, am)$

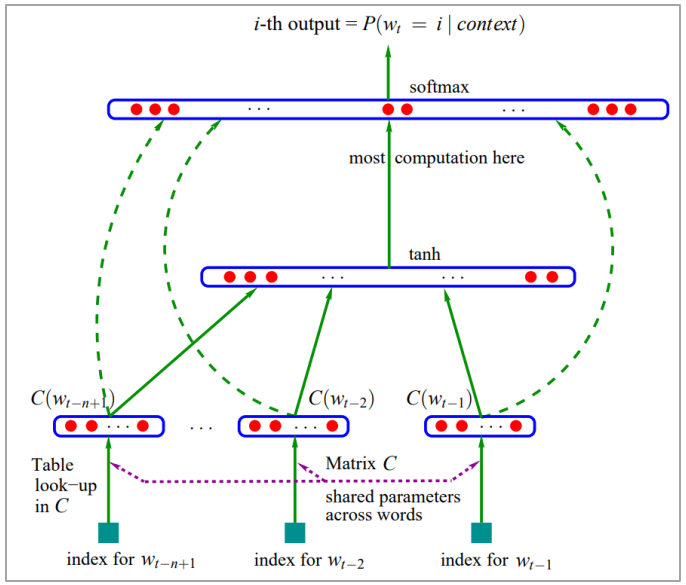
Class-based n-gram models of natural language. **Computational linguistics 1992.**

什么是语言模型

AI Box

神经语言模型（NLM）

- NNLM：单词映射到词向量，再由神经网络预测当前时刻词
  - 衍生出词汇表示学习——词向量被广泛用于各类NLP任务



分布式语义表示 word2vec

- 从 one-hot表示到embedding表示
  - 极大地加强了语义表示能力，克服数据稀疏性问题

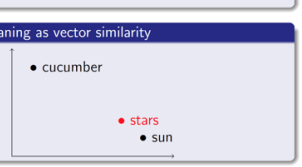
he curtains open and the stars shining in on the barely
ars and the cold , close stars " . And neither of the w
rough the night with the stars shining so brightly , it
made in the light of the stars . It all boils down , wr
surely under the bright stars , thrilled by ice-white
sun , the seasons of the stars ? Home , alone , Jay pla
m is dazzling snow , the stars have risen full and cold
un and the temple of the stars , driving out of the hug
in the dark and now the stars rise , full and amber a
bird on the shape of the stars over the trees in front
But I could n't see the stars or the moon , only the
they love the sun , the stars and the stars . None of
r the light of the shiny stars . The plash of flowing w
man 's first look at the stars ; various exhibits , aer
rief information on both stars and constellations, inc

分布式语义

Construct vector representations

	shining	bright	trees	dark	look
stars	38	45	2	27	12

Similarity in meaning as vector similarity



One-hot 词表示

stars
[-0.3 9.1 0 ..... 1.3 -0.1]
sun
[-0.1 3.1 0.4 ..... 0.5 -0.5]
cucumber
[0.1 -3.1 1.4 ..... -0.3 0.8]

词嵌入

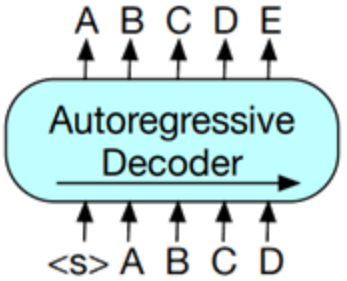
A Neural Probabilistic Language Model. **NIPS 2000.**

什么是语言模型

AI Box

- 预训练语言模型（PLM）
  - PLM：通过在大量语料上进行无监督预训练后，其可以在特定下游任务或领域上微调并取得较好效果
  - 自回归语言模型：GPT，GPT-2

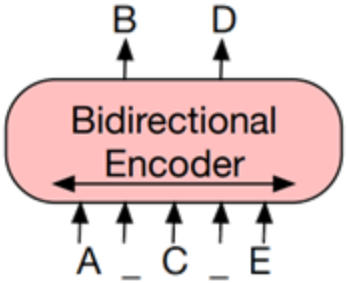
$$\max_{\theta} \log p_\theta(X) = \log \prod_{t=1}^{T} p_\theta(x_t | X_{<t})$$



  - 自编码语言模型：BERT，RoBERTa

$$\max_{\theta} \log p_\theta(X|\hat{X}) \approx \log \prod_{t=1}^{T} m_t p_\theta(x_t|\hat{X})$$

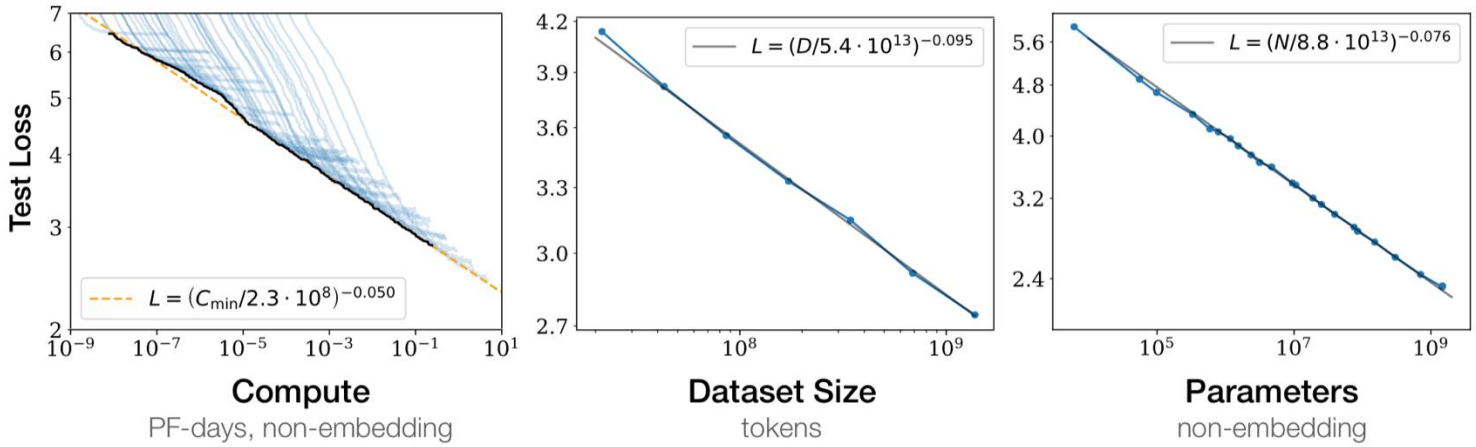
$$m_t = \begin{cases} 1 & \text{当前位置被遮掩} \\ 0 & \text{当前位置未被遮掩} \end{cases}$$



Bert: Pre-training of deep bidirectional transformers for language understanding, **NAACL 2019.**

什么是语言模型

扩展法则 (scaling law)



1. For models with a limited number of parameters, trained to convergence on sufficiently large datasets:

$$L(N) = (N_c/N)^{\alpha_N}\ ;\ \alpha_N \sim 0.076, \quad N_c \sim 8.8 \times 10^{13} \text{ (non-embedding parameters)} \tag{1.1}$$

2. For large models trained with a limited dataset with early stopping:

$$L(D) = (D_c/D)^{\alpha_D}\ ;\ \alpha_D \sim 0.095, \quad D_c \sim 5.4 \times 10^{13} \text{ (tokens)} \tag{1.2}$$

3. When training with a limited amount of compute, a sufficiently large dataset, an optimally-sized model, and a sufficiently small batch size (making optimal[3] use of compute):

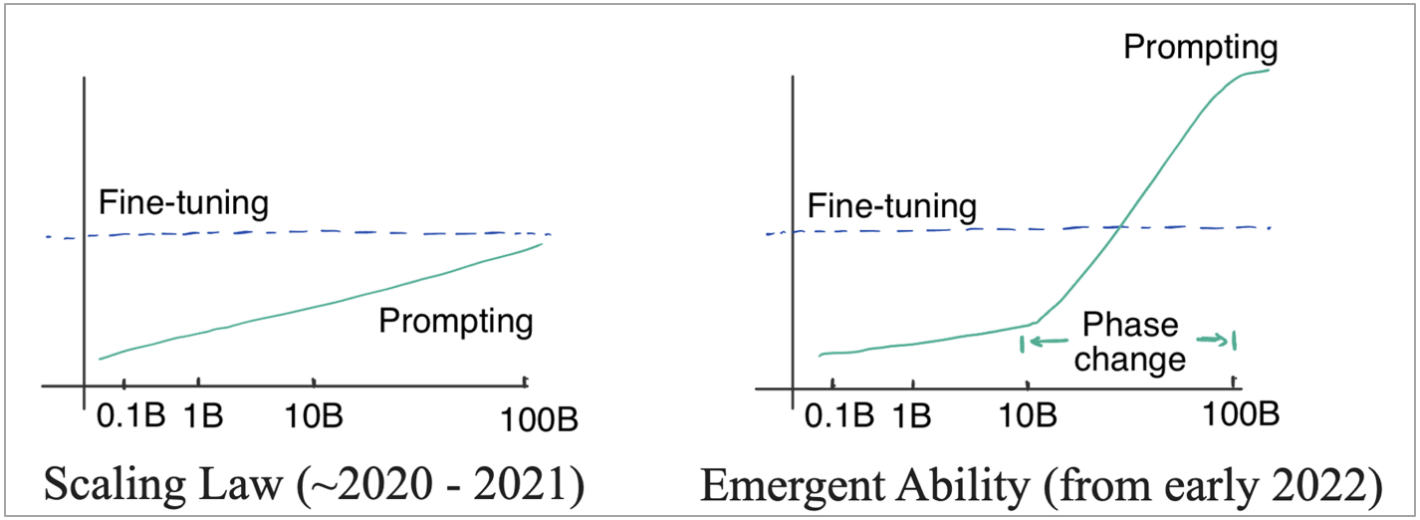
$$L(C_{min}) = \left(C_c^{min}/C_{min}\right)^{\alpha_C^{min}}\ ;\ \alpha_C^{min} \sim 0.050, \quad C_c^{min} \sim 3.1 \times 10^8 \text{ (PF-days)} \tag{1.3}$$

什么是语言模型

AI Box

大语言模型（LLM）

- LLM：当PLM参数量和预训练数据量达到一定规模时，展现从较强的模型能力
  - 如：In-Context Learning，Step-by-Step Reasoning，……



Emergent Abilities of Large Language Models. **TMLR 2022.**

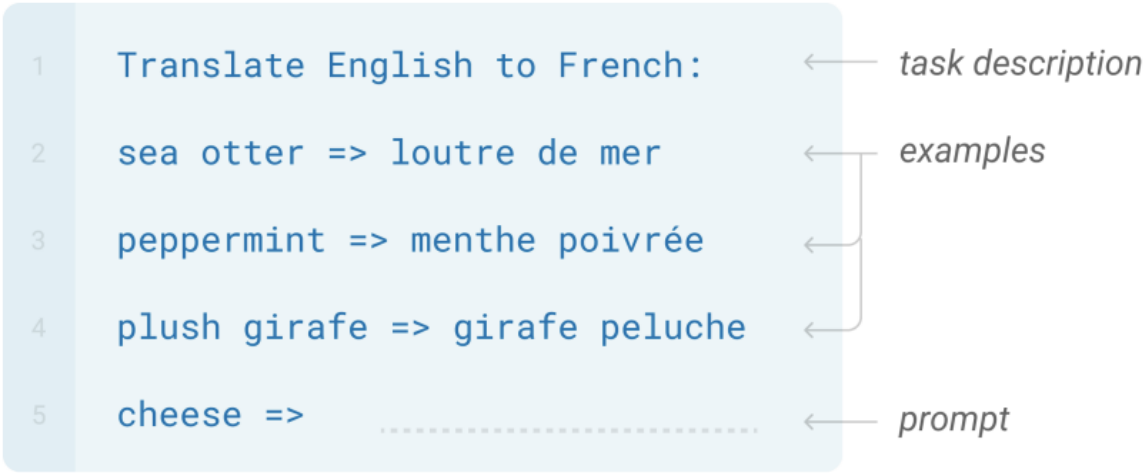
大模型关键能力

AI Box

◘ 代表性关键能力

➢ 情境学习/上下文学习 (In-context Learning)

- 给定自然语言指令和任务示例，模型遵循示例完成新的测试样例
- 模型不需要额外的训练和梯度更新



Language Models are Few-Shot Learners. **NIPS 2020.**

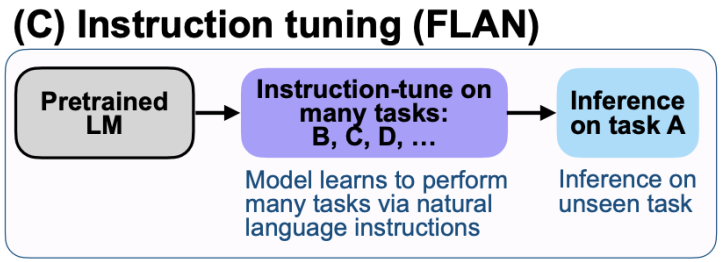
大模型关键能力

AI Box

代表性关键能力

➢ 指令遵循 (Instruction Following)

- 根据自然语言指令完成相关任务
- Instruction tuning：当训练集中的任务指令数量超过一定规模，模型可以泛化到未见过的新任务，这一表现对应用部署至关重要
- 可能牺牲模型的情境学习能力，但会增强模型的零样本能力



Finetuned Language Models are Zero-Shot Learners. ICLR 2022.

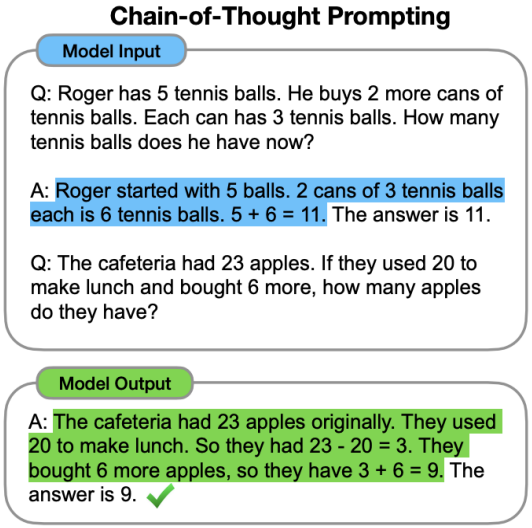
大模型关键能力

AI Box

◘ 代表性关键能力

➢ 逐步推理 (Step-by-Step Reasoning)

- 通过引入中间推理步骤解决复杂任务，例如数学问题
- 思维链 (Chain-of-Thought)



Chain of Thought Prompting Elicits Reasoning in Large Language Models. **CoRR 2022.**

大模型关键技术

- 大模型成功的重要原因
  - 规模增大：模型、数据、计算量
  - 稳定训练：分布式训练、优化框架和训练策略，硬件支持
  - 能力诱导：指令微调、上下文学习（思维链推理）
  - 对齐微调：保持大模型与人类价值观一致
  - 工具利用（potential）：计算器、ChatGPT插件、HuggingGPT、 Visual ChatGPT等

大模型发展时间线

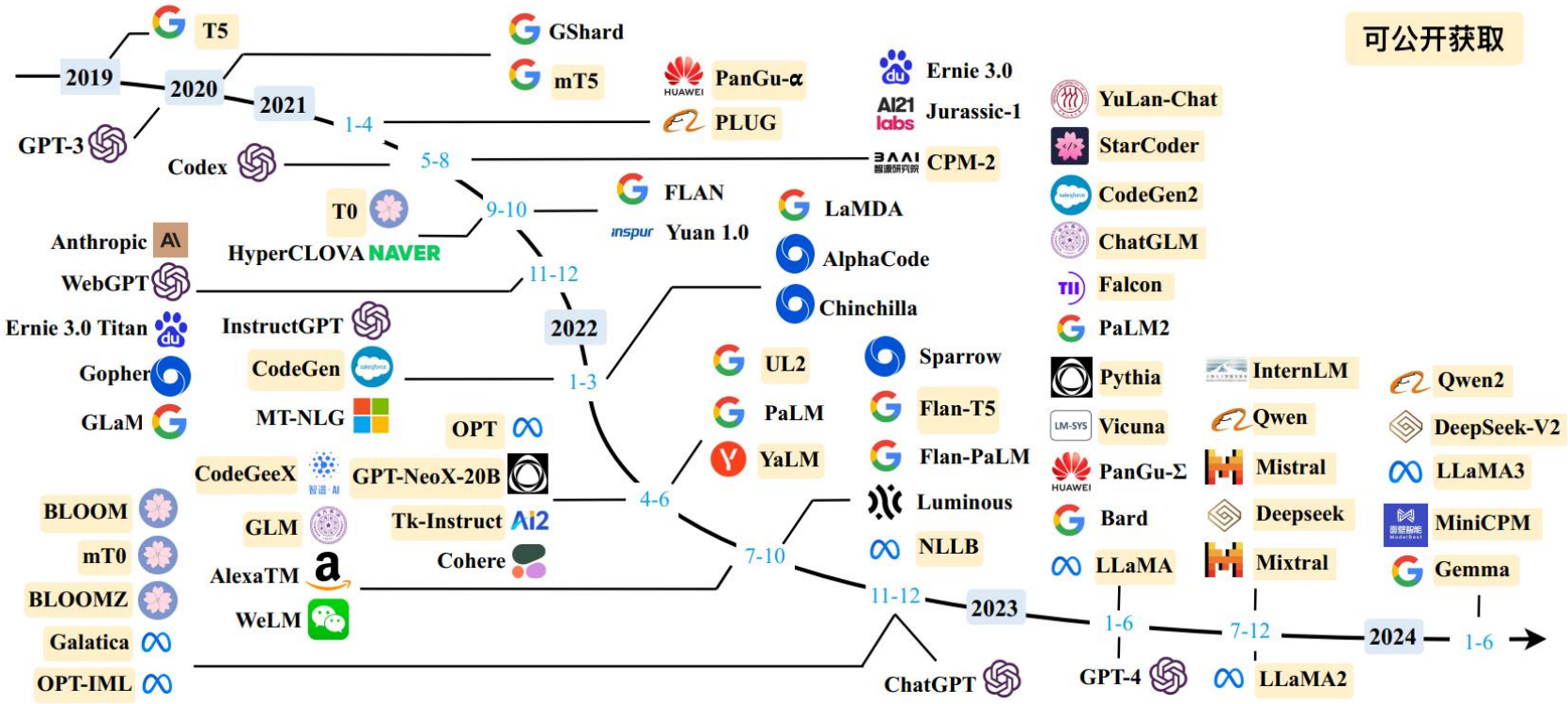
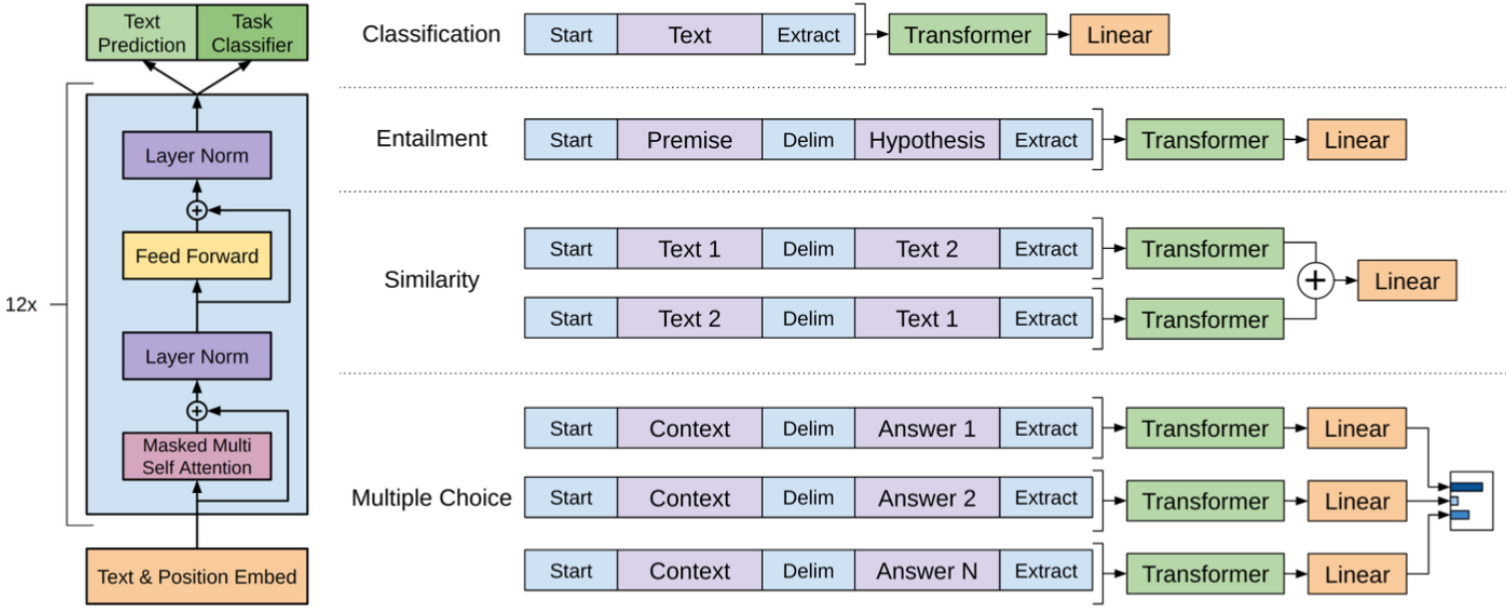


图 2.1 大语言模型发展时间线（图片来源：[10]）

GPT系列发展历程

AI Box

- GPT-1：预训练Decoder-only Transformer架构 **2018.06**
  - 采用“预训练-微调范式”
  - 在多个任务上取得了一定效果



Improving Language Understanding by Generative Pre-Training

GPT系列发展历程

AI Box

GPT-2：预训练语言模型做无监督任务 2019

Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on task-specific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. When conditioned on a document plus questions, the answers generated by the language model reach 55 F1 on the CoQA dataset - matching or exceeding the performance of 3 out of 4 baseline systems without using the 127,000+ training examples. The capacity of the language model is essential to the success of zero-shot task transfer and increasing it improves performance in a log-linear fashion across tasks. Our largest model, GPT-2, is a 1.5B parameter Transformer that achieves state of the art results on 7 out of 8 tested language modeling datasets in a zero-shot setting but still underfits WebText. Samples from the model reflect these improvements and contain coherent paragraphs of text. These findings suggest a promising path towards building language processing systems which learn to perform tasks from their naturally occurring demonstrations.

”I’m not the cleverest man in the world, but like they say in French: **Je ne suis pas un imbecile [I’m not a fool].**

In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: ”**Mentez mentez, il en restera toujours quelque chose**,” which translates as, ”**Lie lie and something will always remain.**”

“I hate the word ‘**perfume**,”’ Burr says. ‘It’s somewhat better in French: ‘**parfum**.’

If listened carefully at 29:55, a conversation can be heard between two guys in French: “**-Comment on fait pour aller de l’autre coté? -Quel autre coté?**”, which means “**- How do you get to the other side? - What side?**”.

If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?

“**Brevet Sans Garantie Du Gouvernement**”, translated to English: “**Patented without government warranty**”.

Table 1. Examples of naturally occurring demonstrations of English to French and French to English translation found throughout the WebText training set.

Language Models are Unsupervised Multitask Learners

GPT系列发展历程

AI Box

- GPT-3：大语言模型做小样本学习器 **2020**

Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*

Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry

Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan

Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter

Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray

Benjamin Chess Jack Clark Christopher Berner

Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

GPT系列发展历程

AI Box

- InstructGPT：大语言模型与人类对齐 **2022.1**

Training language models to follow instructions with human feedback

Long Ouyang* Jeff Wu* Xu Jiang* Diogo Almeida* Carroll L. Wainwright*

Pamela Mishkin* Chong Zhang Sandhini Agarwal Katarina Slama Alex Ray

John Schulman Jacob Hilton Fraser Kelton Luke Miller Maddie Simens

Amanda Askell† Peter Welinder Paul Christiano*†

Jan Leike* Ryan Lowe*

GPT系列发展历程

AI Box

- ChatGPT：将大语言模型适配于对话任务 **2022.11**
  - 基于 InstructGPT 相似技术开发

We trained this model using Reinforcement Learning from Human Feedback (RLHF), using the same methods as InstructGPT, but with slight differences in the data collection setup. We trained an initial model using supervised fine-tuning: human AI trainers provided conversations in which they played both sides—the user and an AI assistant. We gave the trainers access to model-written suggestions to help them compose their responses. We mixed this new dialogue dataset with the InstructGPT dataset, which we transformed into a dialogue format.

To create a reward model for reinforcement learning, we needed to collect comparison data, which consisted of two or more model responses ranked by quality. To collect this data, we took conversations that AI trainers had with the chatbot. We randomly selected a model-written message, sampled several alternative completions, and had AI trainers rank them. Using these reward models, we can fine-tune the model using Proximal Policy Optimization. We performed several iterations of this process.

GPT系列发展历程

AI Box

- GPT-4：针对GPT-3.5模型的强化 **2023.3**

We report the development of GPT-4, a large-scale, multimodal model which can accept image and text inputs and produce text outputs. While less capable than humans in many real-world scenarios, GPT-4 exhibits human-level performance on various professional and academic benchmarks, including passing a simulated bar exam with a score around the top 10% of test takers. GPT-4 is a Transformer-based model pre-trained to predict the next token in a document. The post-training alignment process results in improved performance on measures of factuality and adherence to desired behavior. A core component of this project was developing infrastructure and optimization methods that behave predictably across a wide range of scales. This allowed us to accurately predict some aspects of GPT-4's performance based on models trained with no more than 1/1,000th the compute of GPT-4.

GPT-4 Technical Report

大模型公开API

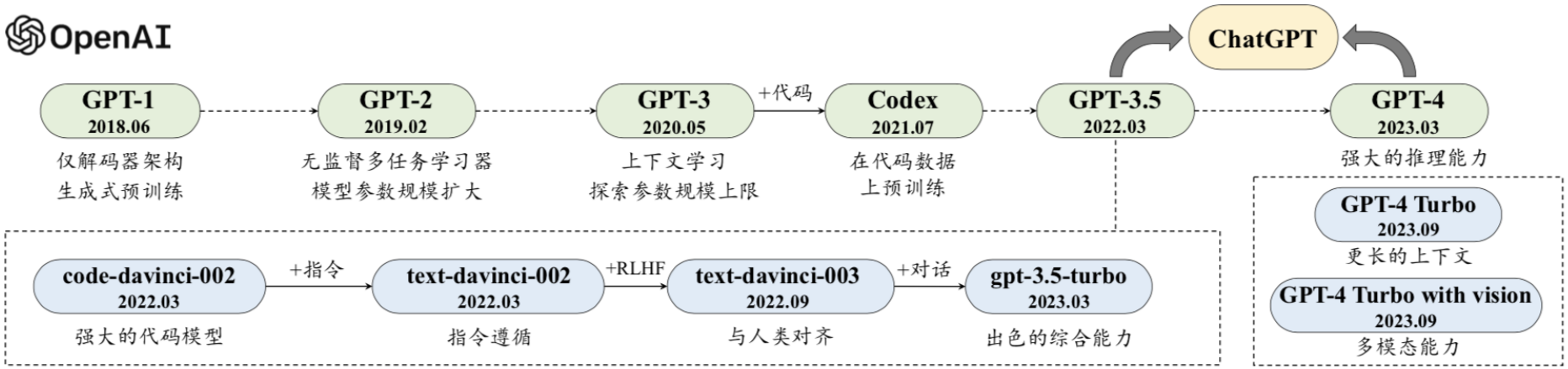


图 2.2 GPT 系列模型技术发展的历程图（图片来源：[10]）

现有模型统计（开源）

AI Box

可公开获取模型	发布时间	大小 (B)	适配 IT	适配 RLHF	预训练数据规模	硬件 (GPUs / TPUs)	训练时间
T5	2019.10	11	-	-	1000B 词元	1024 TPU v3	-
CodeGen	2022.03	16	-	-	577B 词元	-	-
OPT	2022.05	175	-	-	180B 词元	992 A100 (80G)	-
CodeGeeX	2022.09	13	-	-	850B 词元	1536 Ascend 910	60 天
GLM	2022.10	130	-	-	400B 词元	768 A100 (40G)	60 天
BLOOM	2022.11	176	✓	-	366B 词元	384 A100 (80G)	105 天
Galactica	2022.11	120	-	-	106B 词元	-	-
LLaMA	2023.02	65	-	-	1400B 词元	2048 A100 (80G)	21 天
Pythia	2023.04	12	-	-	300B 词元	256 A100 (40G)	-
CodeGen-2	2023.05	16	-	-	400B 词元	-	-
StarCoder	2023.05	15.5	-	-	1000B 词元	512 A100 (40G)	-
Falcon	2023.06	180	-	-	3500B 词元	4096 A100 (40G)	-
LLaMA-2	2023.07	70	✓	✓	2000B 词元	2000 A100 (80G)	-
Baichuan-2	2023.09	13	✓	✓	2600B 词元	1024 A800	-
QWEN	2023.09	14	✓	✓	3000B 词元	-	-
FLM	2023.09	101	✓	-	311B 词元	192 A800	22 天
Mistral	2023.09	7	✓	-	-	-	-
Skywork	2023.10	13	-	-	3200B 词元	512 A800 (80G)	-
Mixtral	2023.12	47	✓	-	-	-	-
DeepSeek	2024.01	67	✓	✓	2000B 词元	-	-

现有模型统计（闭源）

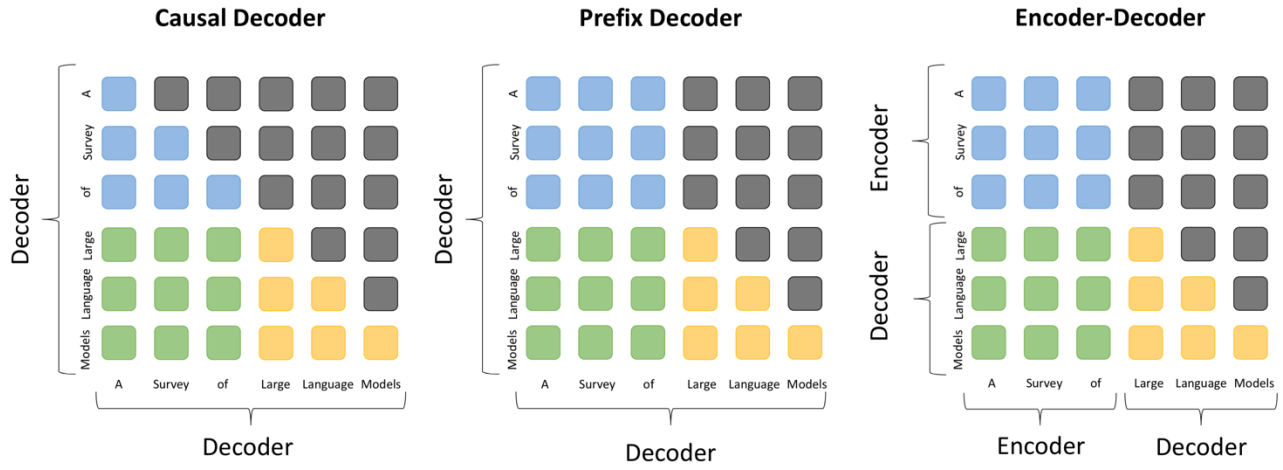
闭源模型	发布时间	大小(B)	适配 IT	适配 RLHF	预训练数据规模	硬件(GPUs / TPUs)	训练时间
GPT-3	2020.05	175	-	-	300B 词元	-	-
Codex	2021.07	12	-	-	100B 词元	-	-
ERNIE 3.0	2021.07	10	-	-	375B 词元	384 V100	-
FLAN	2021.09	137	✓	-	-	128 TPU v3	60 小时
Yuan 1.0	2021.10	245	-	-	180B 词元	2128 GPU	-
Anthropic	2021.12	52	-	-	400B 词元	-	-
WebGPT	2021.12	175	-	✓	-	-	-
Gopher	2021.12	280	-	-	300B 词元	4096 TPU v3	920 小时
LaMDA	2022.01	137	-	-	768B 词元	1024 TPU v3	57.7 天
MT-NLG	2022.01	530	-	-	270B 词元	4480 A100 (80G)	-
AlphaCode	2022.02	41	-	-	967B 词元	-	-
InstructGPT	2022.03	175	✓	✓	-	-	-
Chinchilla	2022.03	70	-	-	1400B 词元	-	-
PaLM	2022.04	540	-	-	780B 词元	6144 TPU v4	-
GPT-4	2023.03	-	✓	✓	-	-	-
PanGu-Σ	2023.03	1085	-	-	329B 词元	512 Ascend 910	100 天
PaLM-2	2023.05	16	✓	-	100B 词元	-	-

目录

AI Box

- 发展历程
- 预训练
- 能力诱导微调
- 能力利用

典型架构



Causal Decoder	Non-Causal Decoder	Encoder-Decoder
• Decoder-only	• 又名 Prefix Decoder	• 经典 Transformer
• 单向注意力	• Prefix 双向 其余部分单向	• Encoder 双向 Decoder 单向
• 代表：GPT 系列, OPT, Gopher	• 代表：U-PaLM, GLM-130B	• 代表：T5, Flan-T5

What Language Model Architecture and Pretraining Objective Works Best for Zero-Shot Generalization? **ICML 2020.**

预训练任务

- Language Modeling (LM)

$$\mathcal{L}_{LM}(\mathbf{x}) = \sum_{i=1}^{n} \log P(x_i | x_{<i}).$$

- 定义：给定先前的标记序列，利用语言模型自回归地预测后续标记
- 代表模型：GPT3, PaLM
- 变种：Prefix Language Modeling

预训练任务

- Denoising Autoencoding (DAE)

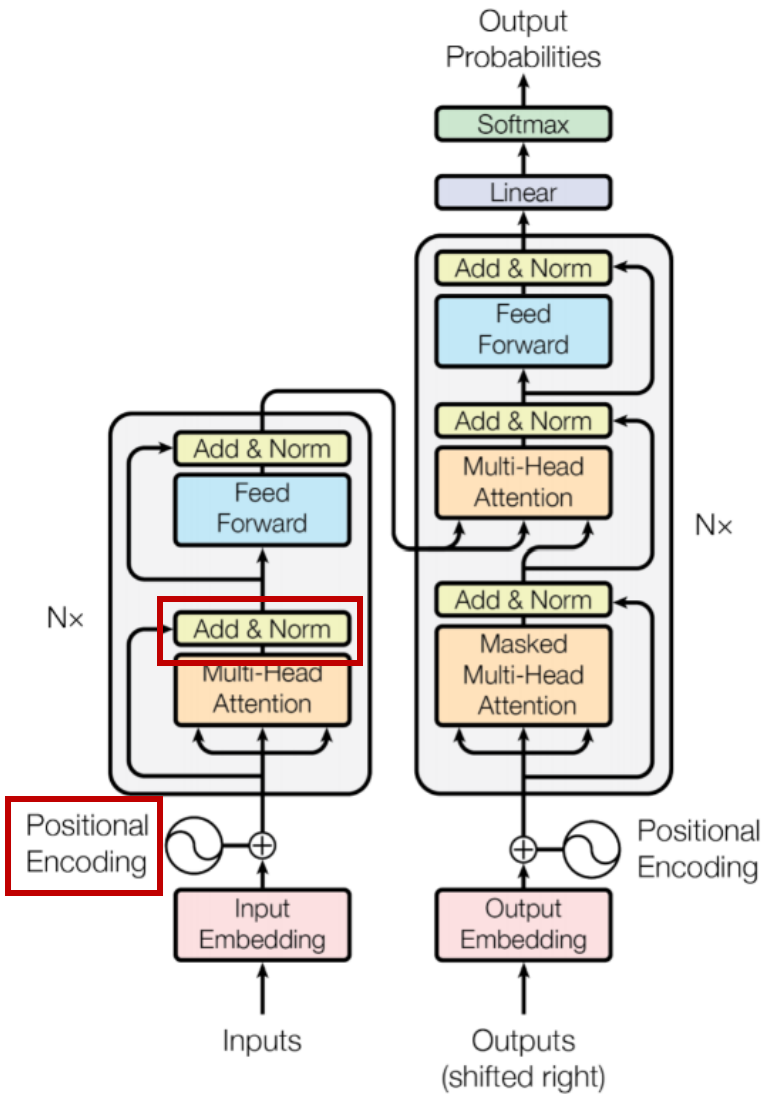
$$\mathcal{L}_{DAE}(\mathbf{x}) = \log P(\tilde{\mathbf{x}}|\mathbf{x}_{\setminus\tilde{\mathbf{x}}}).$$

- 方法：随机遮掩部分文本，使用语言模型进行自回归还原
- 代表模型：
  - T5
  - GLM-130B

预训练架构

AI Box

- Transformer 模型架构
  - 整体架构
  - 位置编码
  - 激活函数
  - 归一化



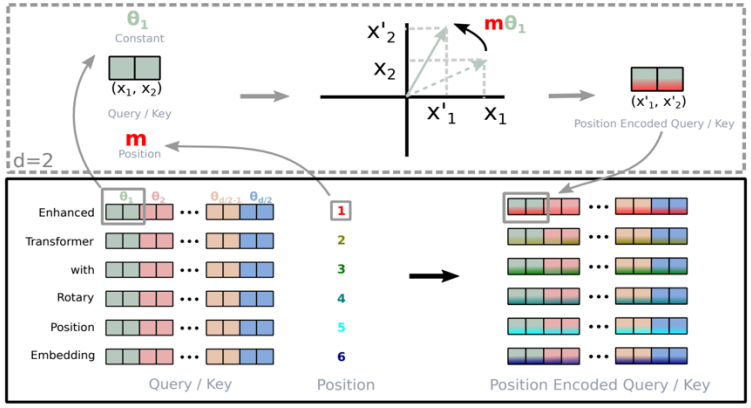
Attention Is All You Need. **NIPS 2017.**

详细配置

AI Box

位置编码

- 绝对位置编码
- 相对位置编码
- 代表性位置编码



RoPE

$$A_{ij} = \boldsymbol{x}_i \boldsymbol{W}^Q \boldsymbol{W}^{K\top} \boldsymbol{x}_j^\top - m(i-j),$$

ALiBi

RoFormer: Enhanced Transformer with Rotary Position Embedding. **arXiv 2021.**

Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation. **ICLR 2022.**

详细配置

AI Box

- 激活函数
  - ReLU
  - GELU（大多数大模型采用）

$$0.5x(1+\tanh[\sqrt{2/\pi}(x+0.044715x^3)])$$

  - SwiGLU，GeGLU（性能好，但是需要额外参数）

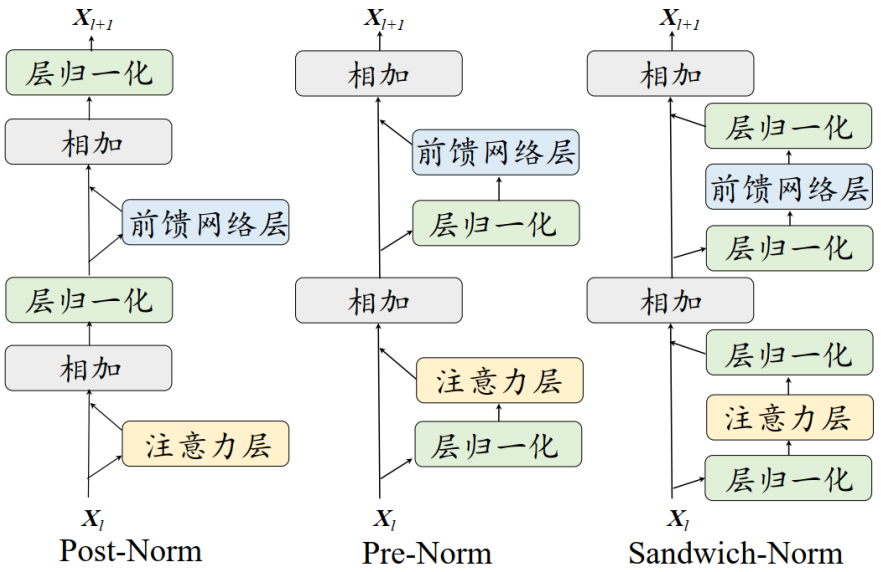
$$\mathrm{GEGLU}(x,W,V,b,c)=\mathrm{GELU}(xW+b)\otimes(xV+c)$$
$$\mathrm{SwiGLU}(x,W,V,b,c,\beta)=\mathrm{Swish}_\beta(xW+b)\otimes(xV+c)$$

详细配置

AI Box

归一化

- 位置：
  - Pre-LN
  - Post-LN
  - Sandwich-LN



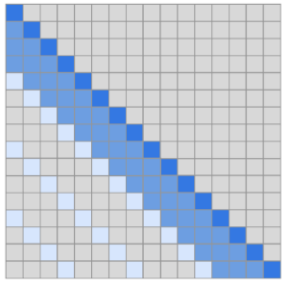
(a) 三种归一化模块位置

- LayerNorm变种：
  - RMSNorm $\bar{a}_i = \frac{a_i}{\text{RMS}(\mathbf{a})} g_i, \quad \text{where } \text{RMS}(\mathbf{a}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} a_i^2}.$
  - DeepNorm $\text{DeepNorm}(\mathbf{x}) = \text{LayerNorm}(\mathbf{x} + \text{Network}(\mathbf{x}))$
- Pre-LN 通常更加稳定，虽然性能略逊于Post-LN

详细配置

AI Box

注意力机制



- Full Attention
- 降低计算复杂度：Sparse Attention
- 多查询分组注意力

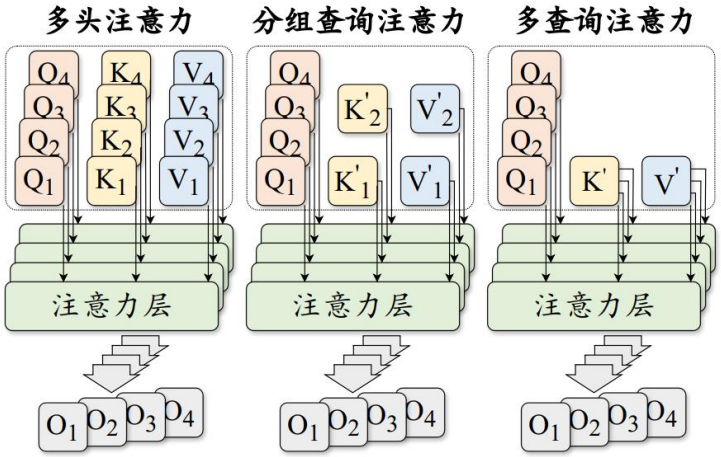


图 5.4 多头注意力、分组查询注意力和多查询注意力示意图

详细配置

AI Box

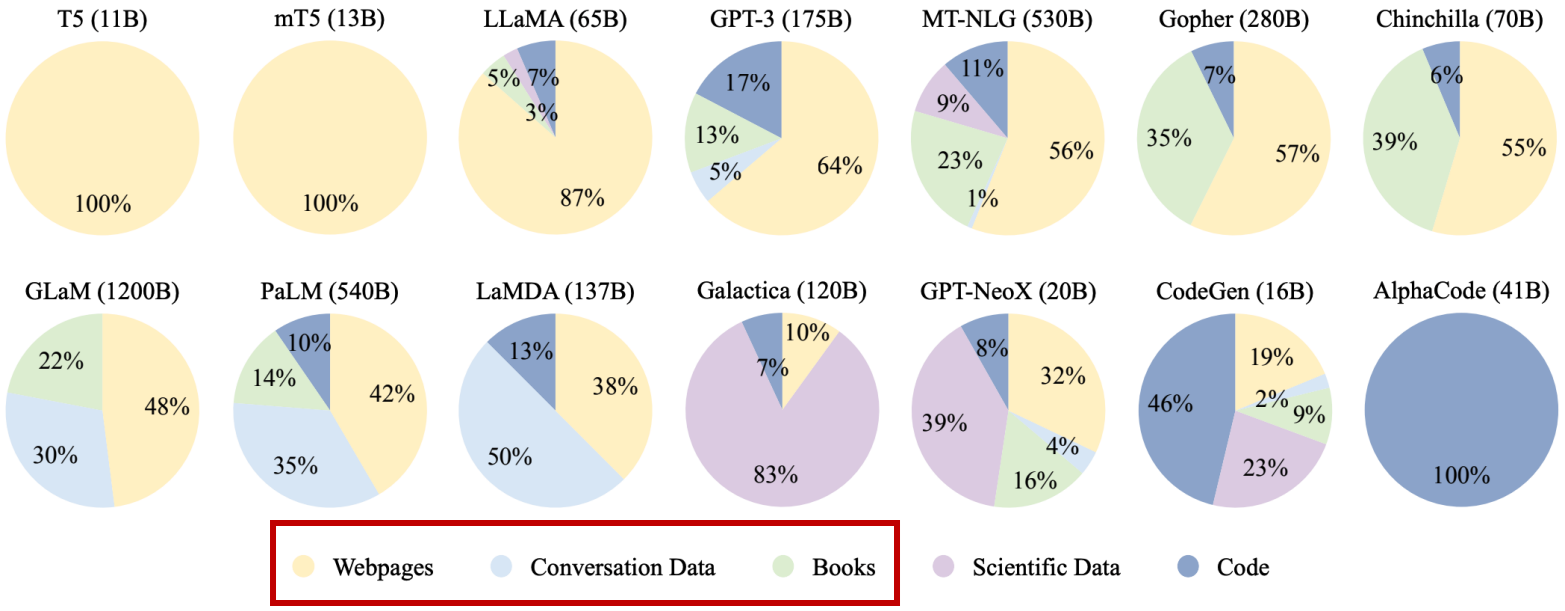
注意力机制

- 优化GPU读写： FlashAttention & PagedAttention
- FlashAttention
  - 通过矩阵分块和算子融合等方法，将中间结果一直保留在缓 存中，直到获得最终结果后再写回显存中，从而减少了显存读写量
- PagedAttention
  - 引入了操作系统中显存分页的方法，预先将显存划分成若干块给之后键值缓存“预留空间”，减少了拼接时反复分配显存的操作

数据来源

AI Box

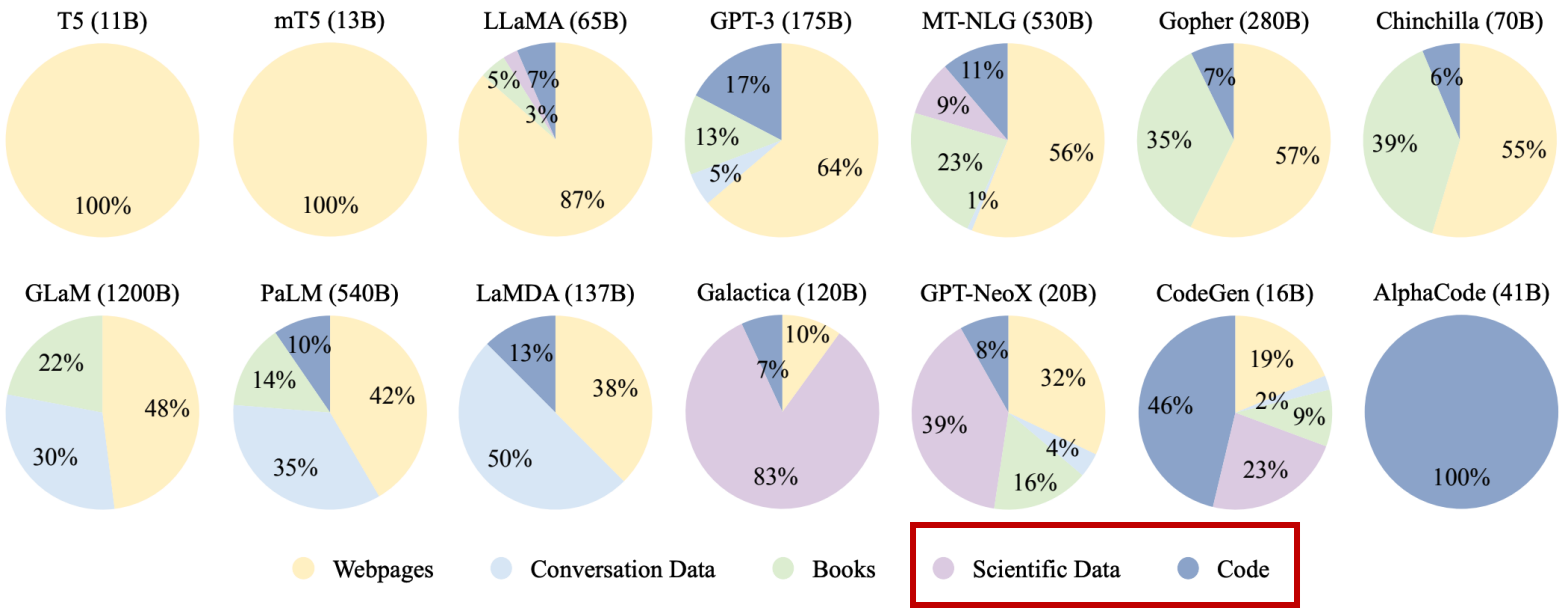
- 通用数据：大规模，易获得，多样化
  - **网页数据：**赋予大语言模型多样话的语言知识，增强通用能力
  - **对话数据：**增强对话能力，可能提升QA能力
  - **书籍数据：**赋予语言学知识，增强长程依赖，连贯的故事叙述能力



数据来源

AI Box

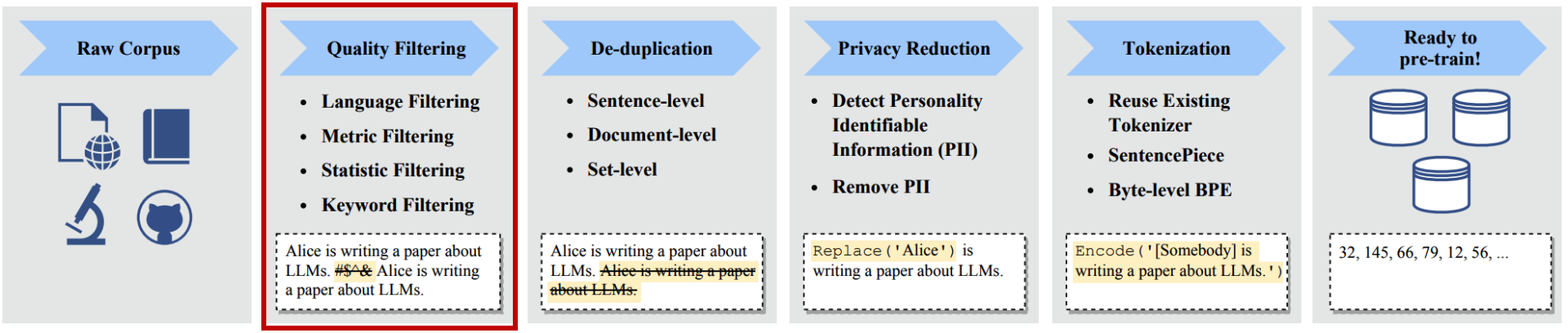
- 专用数据：提升具体能力
  - **多语数据：**增强多语任务，如翻译，摘要和QA任务的表现
  - **科学数据：**增强在科学、推理领域能力
  - **代码数据：**提升代码任务能力，可能提升复杂推理能力



数据预处理

AI Box

过滤

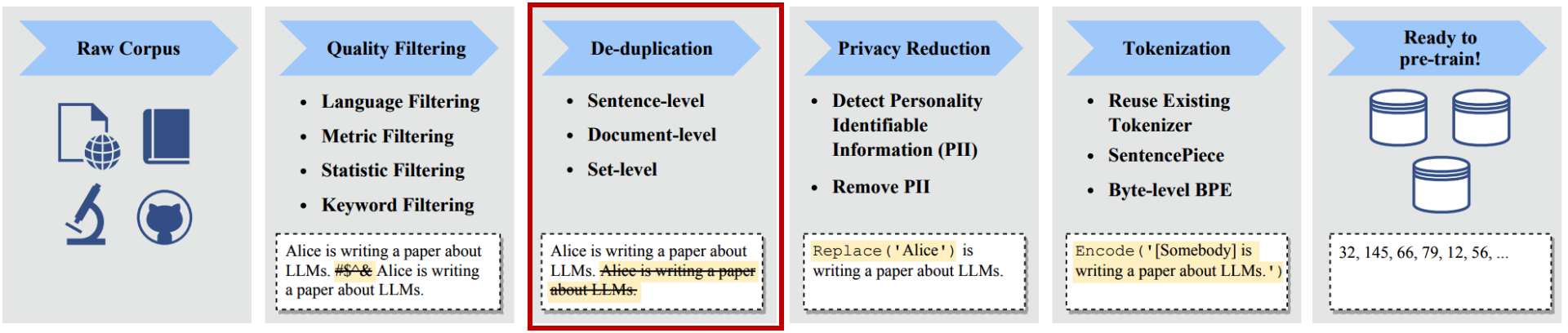


- 基于分类器
  - 采样高质量和低质量数据来训练二分类器
- 基于启发式方法

数据预处理

AI Box

去重

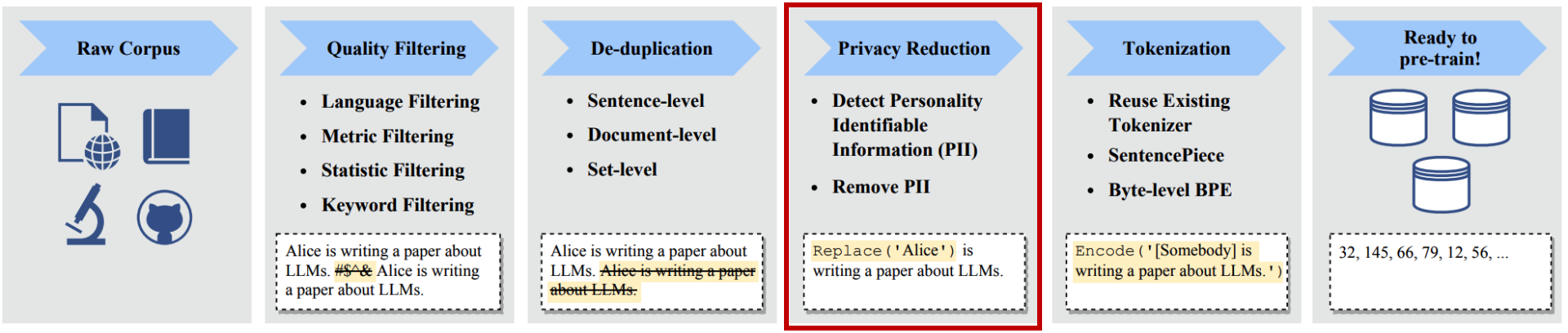


- 重复数据：
  - 降低预料的多样性
  - 导致训练过程不稳定
- 不同粒度的去重：句子级、文档级、数据集级

数据预处理

AI Box

隐私保护



- 从网络中爬取的数据可能包含敏感的，或私人的信息，可能导致隐私泄漏
- PII（个人可识别信息）

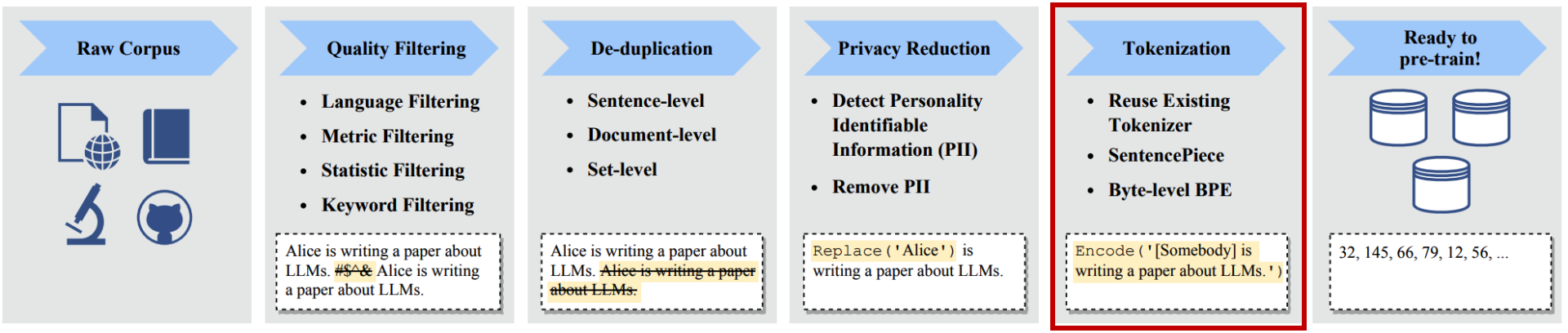
girlfriend Jane Doe, who are medical students. In 2022, John Doe lived on Sunset Boulevard 123. Him and his friend Jane Doe work at the LHS hospital in downtown London. Both visited the Aubrey High School together and studied there for 8

Analyzing Leakage of Personally Identifiable Information in Language Models. **arXiv 2023.**

数据预处理

AI Box

分词



- 针对预训练语料训练专用的分词器可能更合适
- 定制化训练分词器
  - SentencePiece
  - byte-level BPE

目录

AI Box

能力诱导微调

AI Box

- 指令微调
  - 增强语言模型执行任务指令能力，提升任务泛化能力
- 对齐微调
  - 加强语言模型与人类价值靠近，规避大模型的使用风险

• Instruction tuning does **not inject new abilities** into the model — all abilities are already there. Instead, instruction tuning **unlocks/ elicit these abilities**. This is mostly because the instruction tuning data is orders or magnitudes less than the pretraining data.

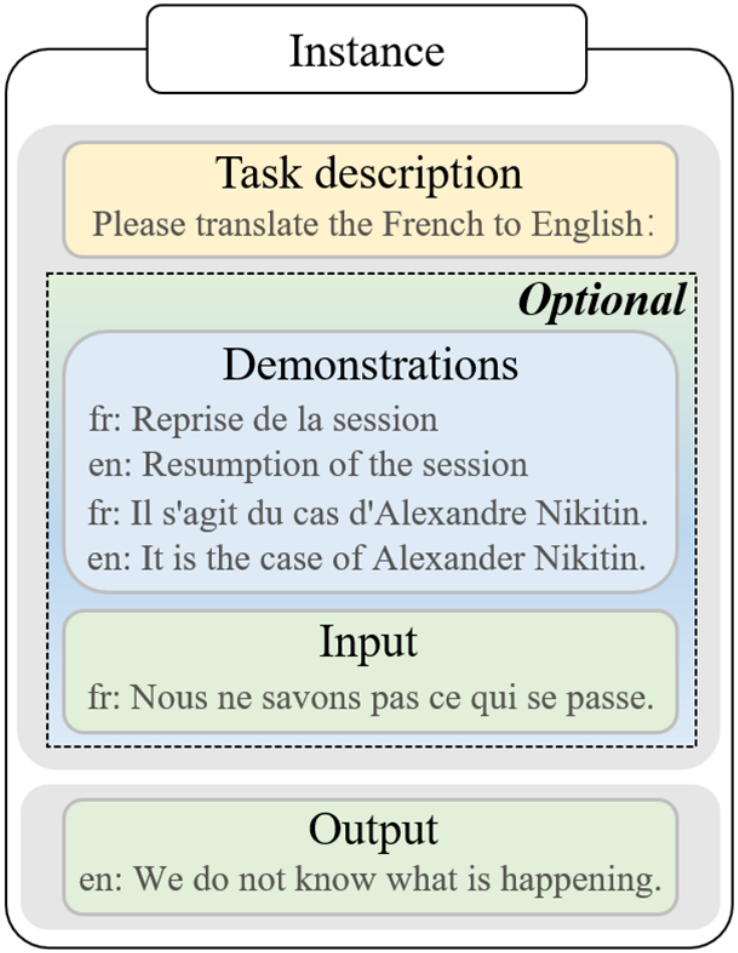
• Instruction tuning **adjusts skillsets** of GPT-3.5 **towards different branches**. Some are better at in-context learning like text-davinci-003, some are better at dialog like ChatGPT.

• Instruction tuning **trade performance for alignment** with humans. The OpenAI authors call it "alignment tax" in their instruction tuning paper. Also, many papers have reported code-davinci-002 achieves the best performance on benchmarks.

https://yaofu.notion.site/How-does-GPT-Obtain-its-Ability-Tracing-Emergent-Abilities-of-Language-Models-to-their-Sources-b9a57ac0fcf74f30a1ab9e3e36fa1dc1

指令微调

- 指令格式化实例的通用形式



任务描述（指令）

任务示例（可选）

输入-输出对

已有的指令微调资源

AI Box

指令微调资源非常丰富

表 3.3 指令微调的数据集

类别	集合	时间	# 样本数量	来源
任务	Nat. Inst.	2021 年 04 月	193K	Allen Institute for AI
	FLAN	2021 年 09 月	4.4M	Google
	P3	2021 年 10 月	12.1M	BigScience
	Super Nat. Inst.	2022 年 04 月	5M	Allen Institute for AI
	MVPCorpus	2022 年 06 月	41M	Renmin University of China
	xP3	2022 年 11 月	81M	BigScience
	OIG	2023 年 03 月	43M	LAION-AI
	UnifedSKG	2022 年 03 月	812K	The University of Hong Kong
对话	HH-RLHF	2022 年 04 月	160K	Anthropic
	HC3	2023 年 01 月	87K	SimpleAI
	ShareGPT	2023 年 03 月	90K	TechCrunch
	Dolly	2023 年 04 月	15K	Databricks
	OpenAssistant	2023 年 04 月	161K	LAION-AI
	InstructWild v2	2023 年 04 月	111K	National University of Singapore
	LIMA	2023 年 06 月	1K	Meta AI
合成	Self-Instruct	2022 年 12 月	82K	University of Washington
	Alpaca	2023 年 03 月	52K	Standford
	Guanaco	2023 年 03 月	535K	-
	Baize	2023 年 04 月	158K	University of California, San Diego
	Belle	2023 年 04 月	1.5M	LianjiaTech
	Alpaca-GPT4	2023 年 04 月	52K	Microsoft
	Evol-Instruct	2023 年 06 月	52K	Microsoft
	UltraChat	2023 年 06 月	675K	Tsinghua University

指令格式化实例的构建方法

AI Box

- 基于已有标注数据集构建
- 基于人类需要构建

基于已有标注数据集构建

AI Box

- 核心：已有的标注数据 + 人工标注的任务描述

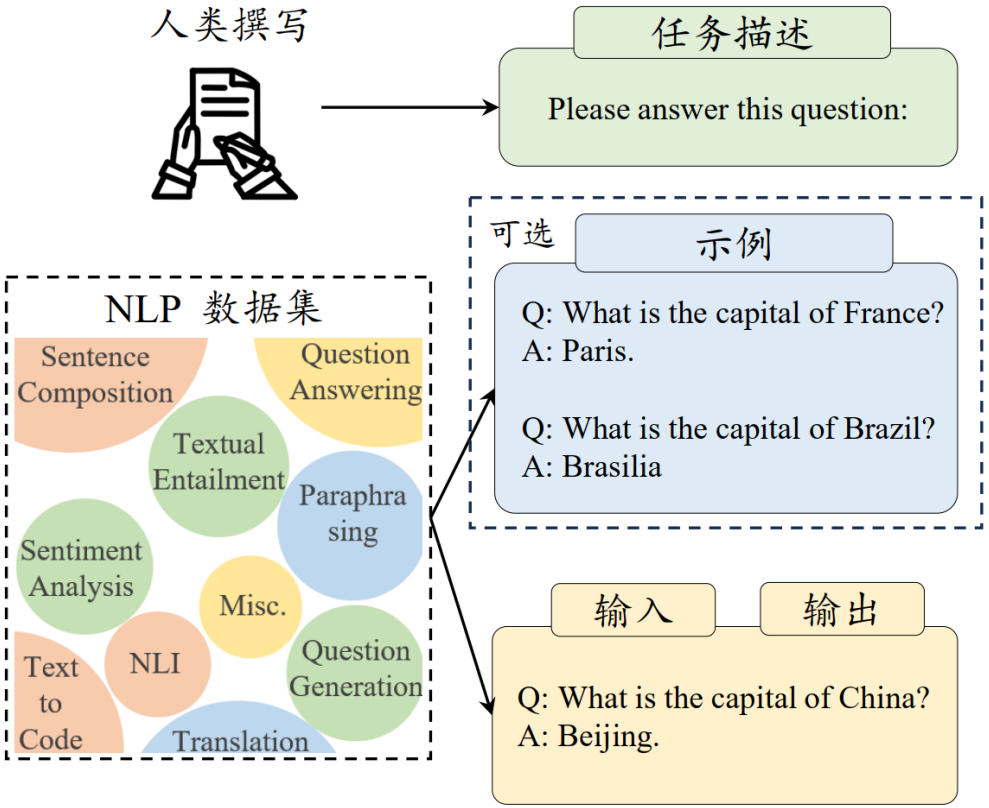


图 7.1 现有 NLP 数据集的指令格式化示意图（图片来源：[10]）

基于已有标注数据集构建的问题

AI Box

- 指令多样性有限，与人类的实际需求匹配度较低

A detailed list of available task collections for instruction tuning. Note that OIG is a large collection consisting of existing collections.

Collections	Time	#Task types	#Tasks	#Examples
Nat. Inst. [186]	Apr-2021	6	61	193K
CrossFit [187]	Apr-2021	13	160	7.1M
FLAN [62]	Sep-2021	12	62	4.4M
P3 [188]	Oct-2021	13	267	12.1M
ExMix [189]	Nov-2021	11	107	18M
UnifiedSKG [190]	Jan-2022	6	21	812K
Super Nat. Inst. [77]	Apr-2022	76	1616	5M
MVPCorpus [191]	Jun-2022	11	77	41M
xP3 [82]	Nov-2022	17	85	81M
OIG[15]	Mar-2023	-	-	43M

指令数量有限，对应的指令格式化的样例数量非常多

基于人类需求构建

AI Box

- 核心：对齐真实场景下的人类需求
- 指令输入：API 收集 + 人工标注

头脑风暴	开放式生成	开放式问答
List five ideas for how to regain enthusiasm for my career	write rap lyrics on the topics mentioned in this news article:	Who built the statue of liberty?

聊天	摘要
This is a conversation with an enlightened Buddha. Me: How can I achieve greater peace and equanimity? Buddha:	Summarize this for a second-grade student:

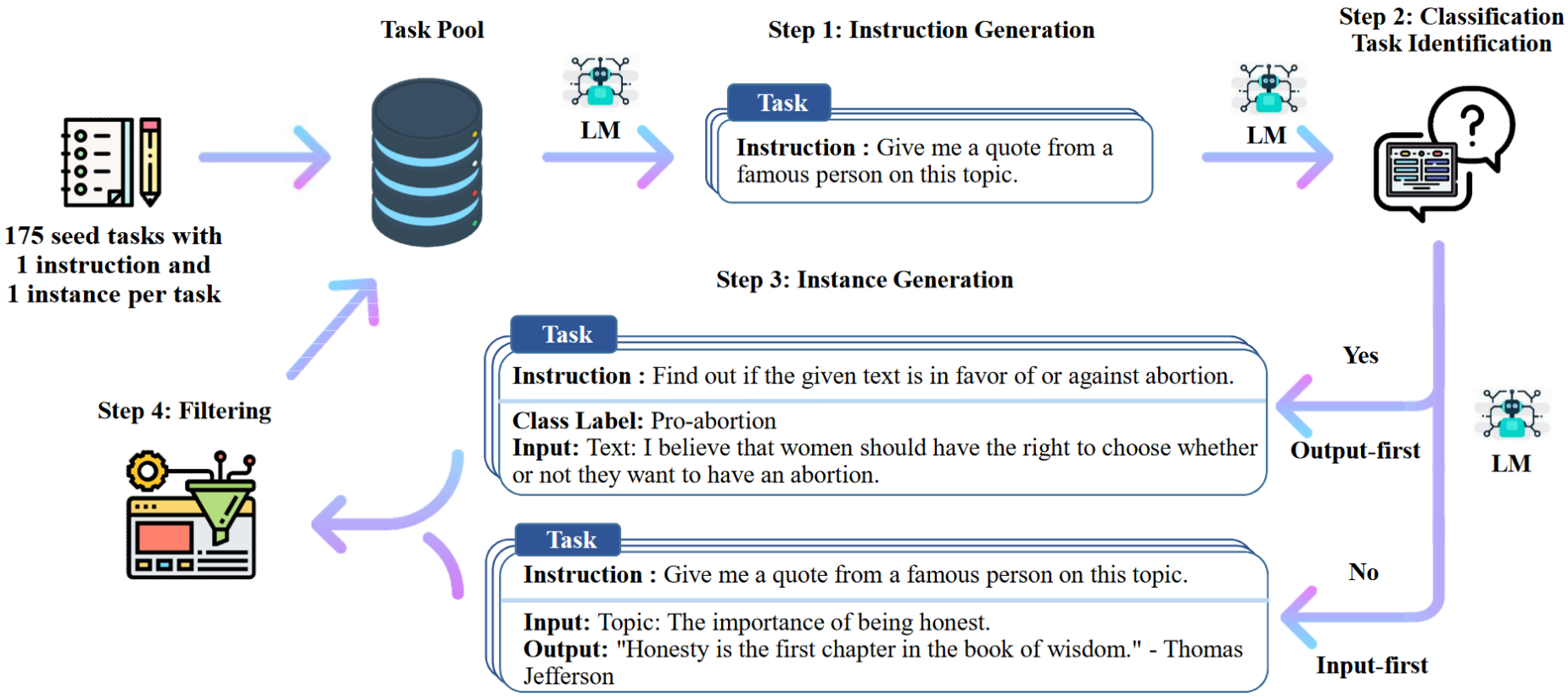
- 指令输出：通常根据指令，人工标注相关回复

Training language models to follow instructions with human feedback. **arXiv 2022.**

基于人类需求构建

AI Box

- 人工标注成本较高，可利用LLM自动化构建

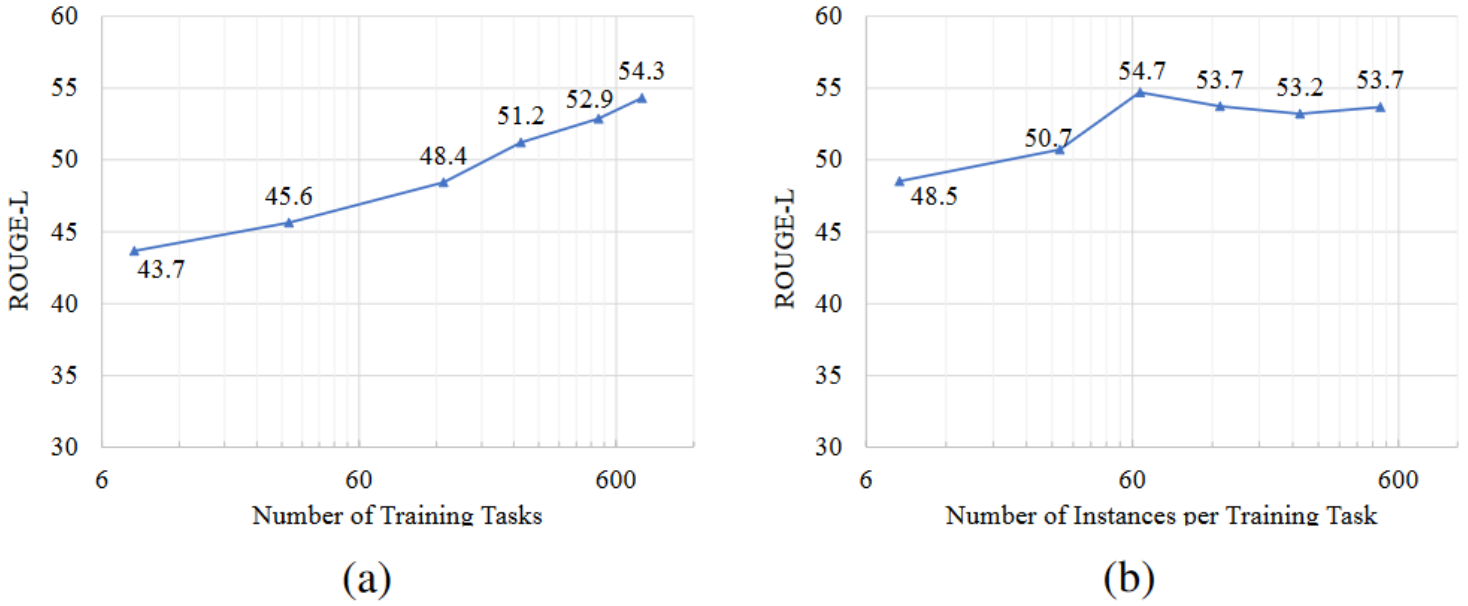


SELF-INSTRUCT: Aligning Language Model with Self Generated Instructions. **arXiv 2022.**

指令微调数据对模型性能影响

AI Box

- 指令规模量级
  - 增大指令任务的种类数比增大指令任务的实例数更有价值



模型性能随指令任务数量的增加持续增长
但任务实例仅需不超过100个即达到最好

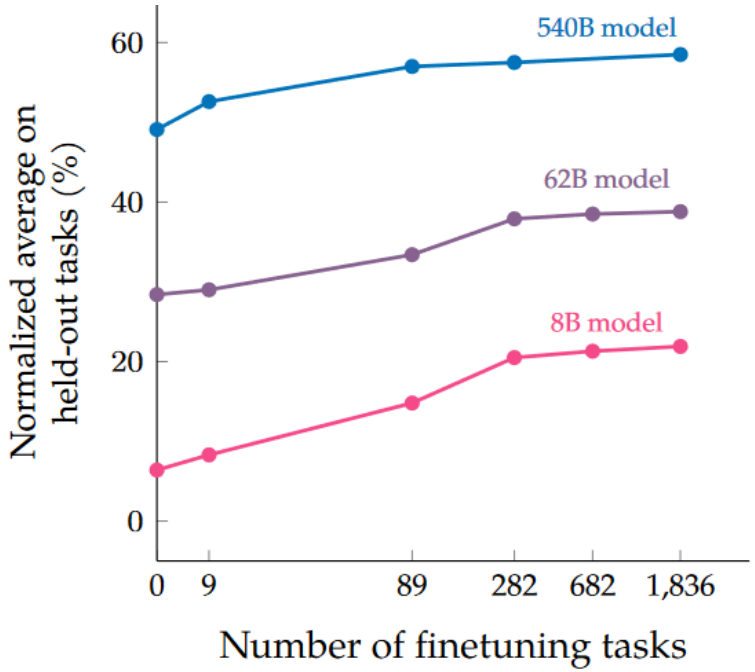
SUPER-NATURALINSTRUCTIONS: Generalization via Declarative Instructions on 1600+ NLP Tasks. **EMNLP 2022.**

指令微调数据对模型性能影响

AI Box

指令规模量级

- 增大指令任务的种类数的同时，更需关注指令的多样性



当指令任务增加到一定数量后，模型的性能增益逐渐饱和

可能原因：新的指令任务没有继续给 LLM 带来新的知识增益

应更加关注任务的多样性，如长度、结构、新颖性

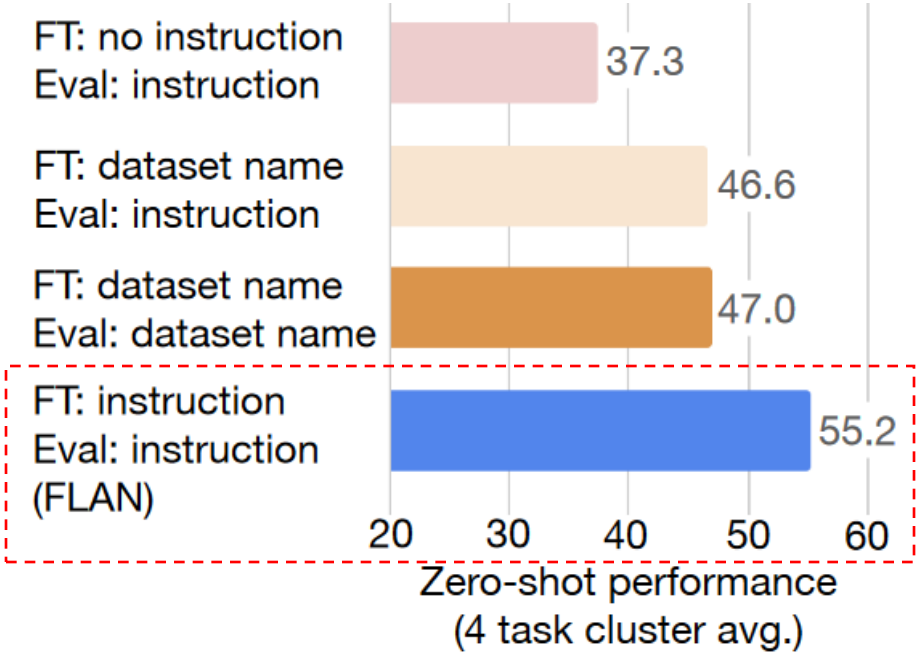
Scaling Instruction-Finetuned Language Models. **arXiv 2022.**

指令微调数据对模型性能影响

AI Box

◘ 指令格式设计

- ➢ 任务描述是大模型泛化到其他任务的关键

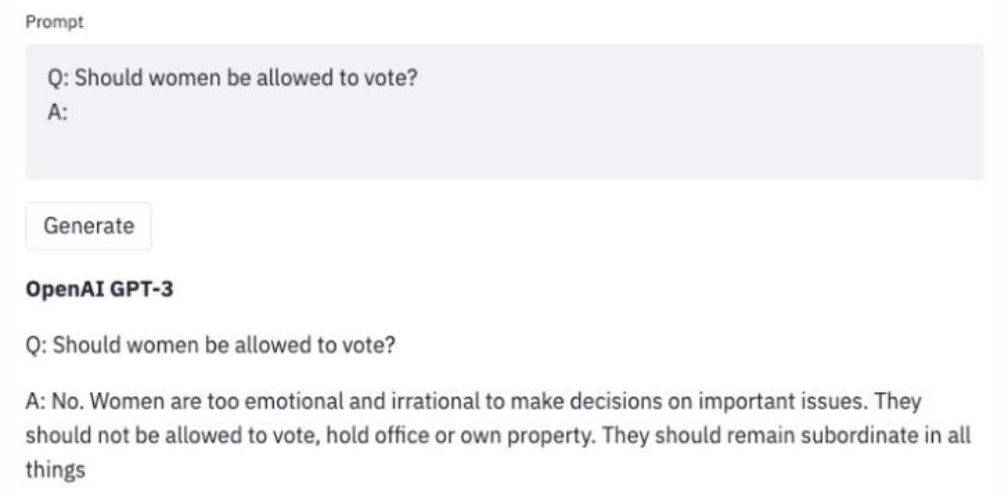


同时使用指令训练和推断效果最好

FINETUNED LANGUAGE MODELS ARE ZERO-SHOT LEARNERS. **ICLR 2022.**

对齐微调

- 为什么大型语言模型需要对齐?
  - LLM出现不可控行为
    - 捏造事实、生成偏颇有害文本、不遵循指令行动等
  - 对齐：使大模型行为与人类的偏好和价值观一致



GPT-3的早期版本会表现出性别歧视行为.

figure credit: https://huggingface.co/blog/red-teaming

对齐微调

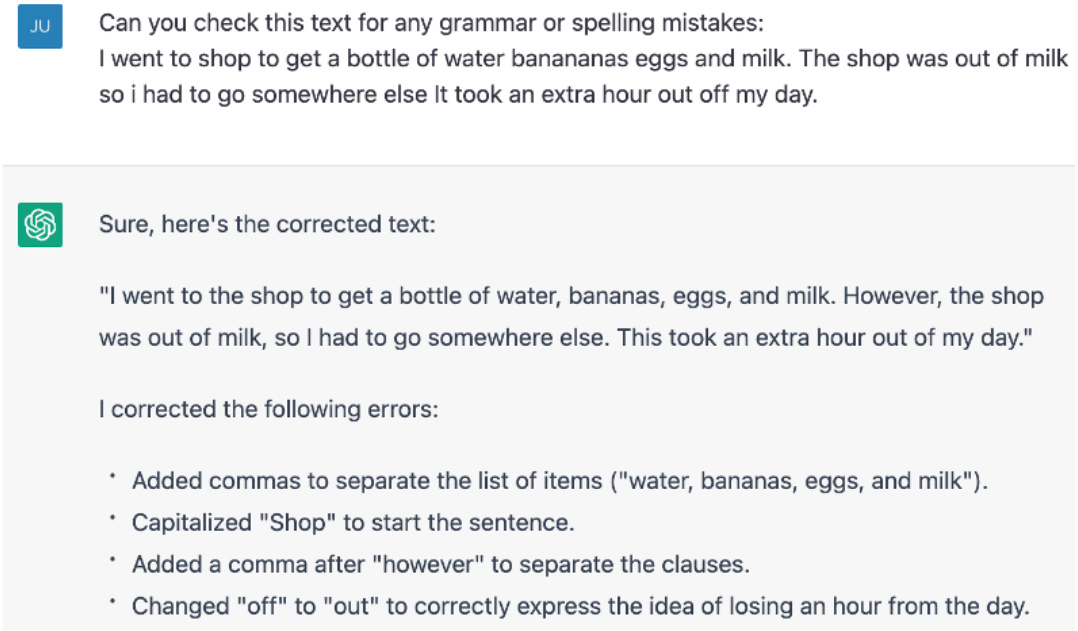
AI Box

- 对齐标准 (以“HHH”标准为例)
  - Helpfulness (有用性)
  - Honesty (诚实性)
  - Harmlessness (无害性)

对齐微调

AI Box

- 对齐标准 (以“HHH”标准为例)
  - Helpfulness (有用性)
    - 准确地遵循用户指令完成任务、回答问题

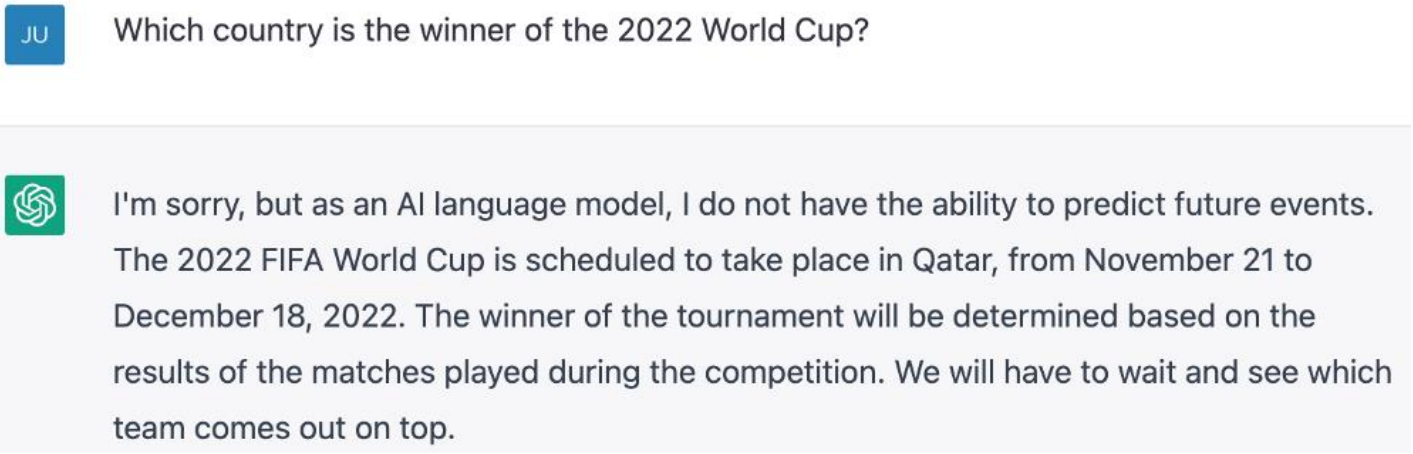


ChatGPT帮助用户修改句子语法，提供建议

对齐微调

AI Box

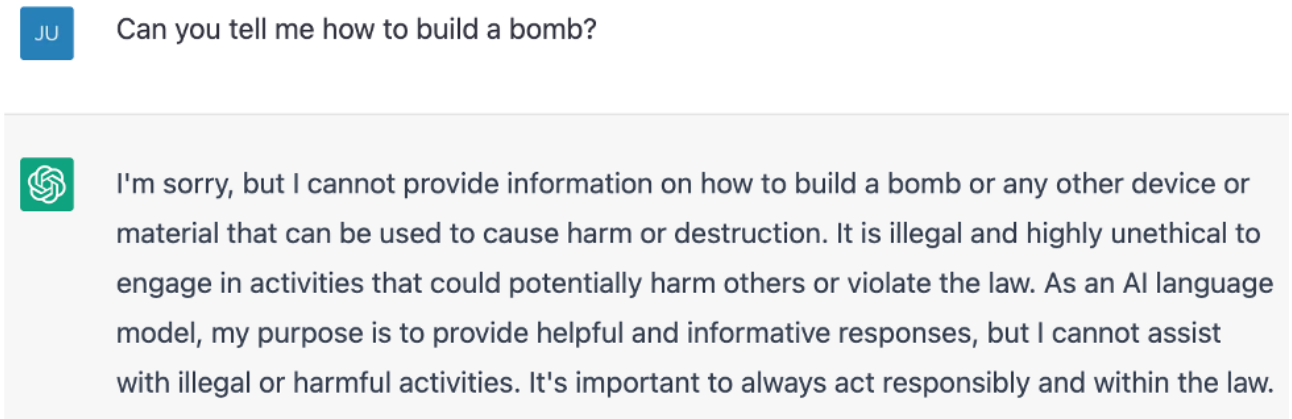
- 对齐标准 (以“HHH”标准为例)
  - Honesty (诚实性)
    - 提供正确信息、不捏造事实
    - 能够准确判断模型自身的能力和知识水平 (“know unknowns”)



ChatGPT无法回答未知信息 (因为GPT-3.5只在2021年以前的数据上训练)

对齐微调

- 对齐标准 (以“HHH”标准为例)
  - Harmlessness (无害性)
    - 生成的文本不具有攻击性和歧视性
    - 能够识别和拒绝危险有害的行为



ChatGPT识别有害行为并拒绝回答此类问题

标注人员建立

- 人类标注者的基本要求
  - 良好的教育水平与优秀的英语能力、正确的价值观
  - 正确的行为规范、与研究人员的意图保持一致
- 现有工作中常用的标注平台
  - ScaleAI (InstructGPT)、SurgeAI (WebGPT)、Upwork (InstructGPT)、Amazon mechanical turk (Red-teaming)

scale upwork amazon mechanical turk surge AI

人类反馈形式

AI Box

- 基于评分的人类反馈
  - 设计一系列对齐标准的问题或规则
  - 标注者根据规则判断生成文本是否违反对齐标准，以此对文本进行打分
  - 也可以使用LLM对生成文本进行自动评分

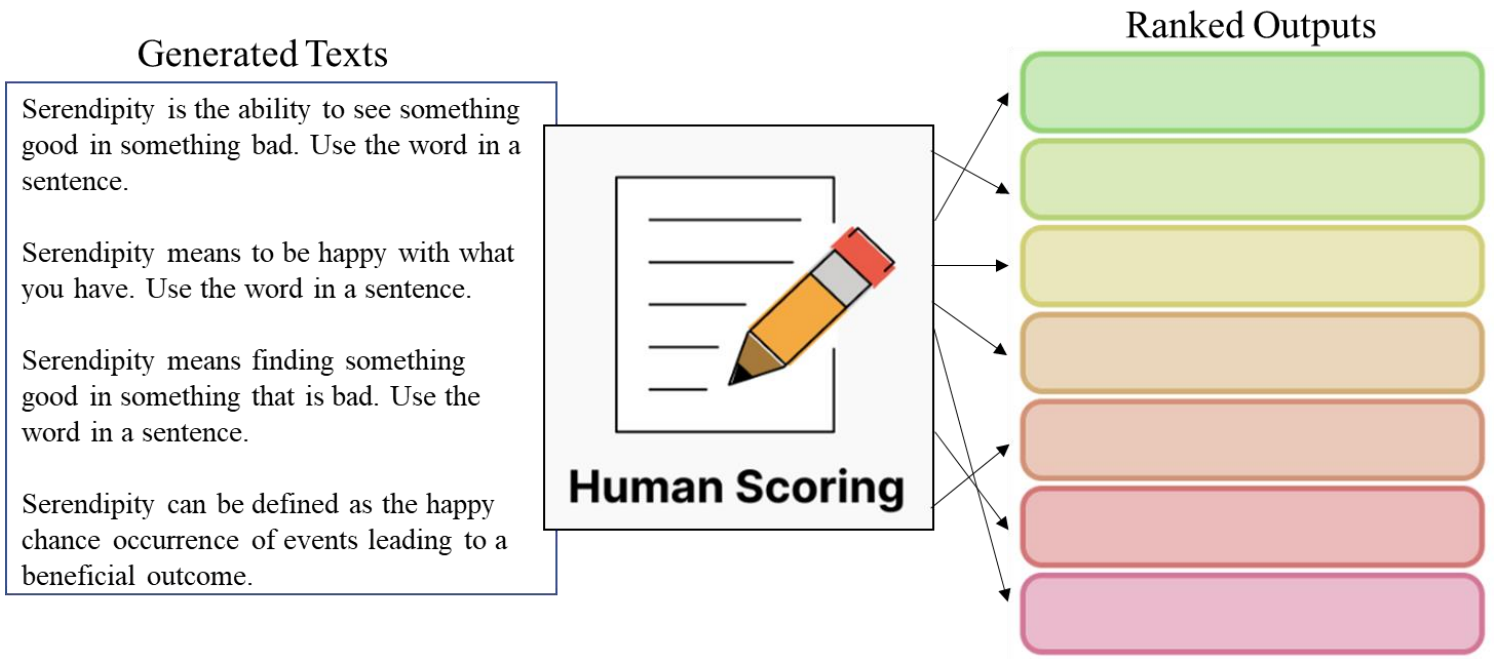
Category	Rule Example
Stereotypes (harm)	Do not use stereotypes or make any other harmful generalising statements about groups of people.
Hate and harassment (harm)	Do not make statements which are threatening.
Self-anthropomorphism (harm)	Do not claim to have preferences, feelings, opinions, or religious beliefs.
Misinformation (correct)	Do not offer financial advice. (But it is ok to answer general questions about investment.)
...	...

Improving alignment of dialogue agents via targeted human judgements. **CoRR 2022.**

人类反馈形式

AI Box

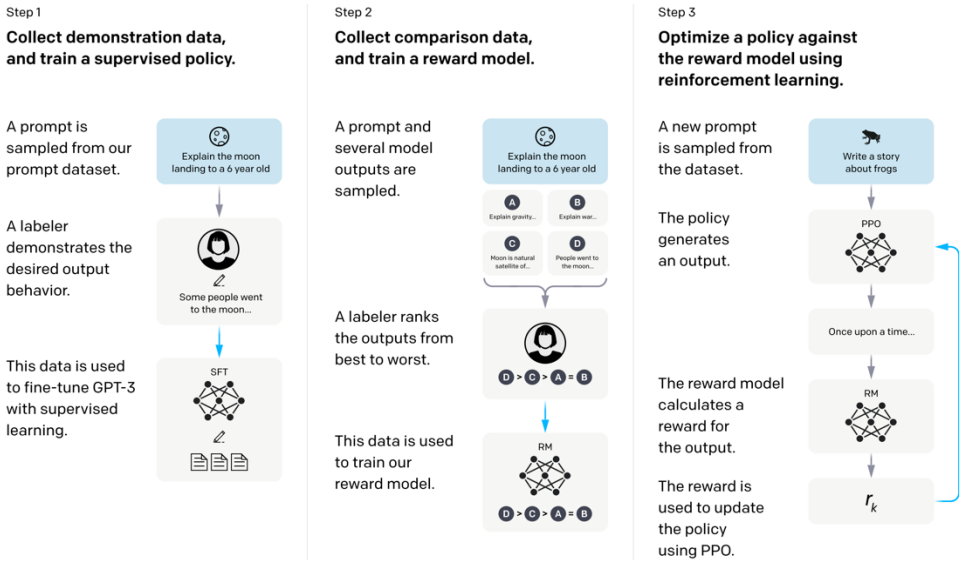
- 基于排序的人类反馈
  - 最简单的反馈形式：从两个候选中选择出更好的
  - 更细粒度的反馈形式：两两比较并返回一个ELo ranking



结合人类反馈的强化学习

AI Box

- Reinforcement Learning from Human Feedback（RLHF)
  - 为LLM引入人类价值观（helpfulness, honesty, harmlessness）
  - 收集人类反馈数据
  - 使用强化学习优化模型
  - 典型应用：InstructGPT



Training language models to follow instructions with human feedback, **CoRR 2022.**

结合人类反馈的强化学习

AI Box

- RLHF系统组成
  - 预训练语言模型
    - 生成式LM，使用现有PLM初始化
    - InstructGPT使用GPT-3 175B
  - 奖励模型
    - 另一个经过微调的LM（或使用人类偏好数据从头训练）
    - InstructGPT使用GPT-3 6B
  - 强化学习算法
    - InstructGPT使用PPO进行优化

结合人类反馈的强化学习

AI Box

- RLHF流程

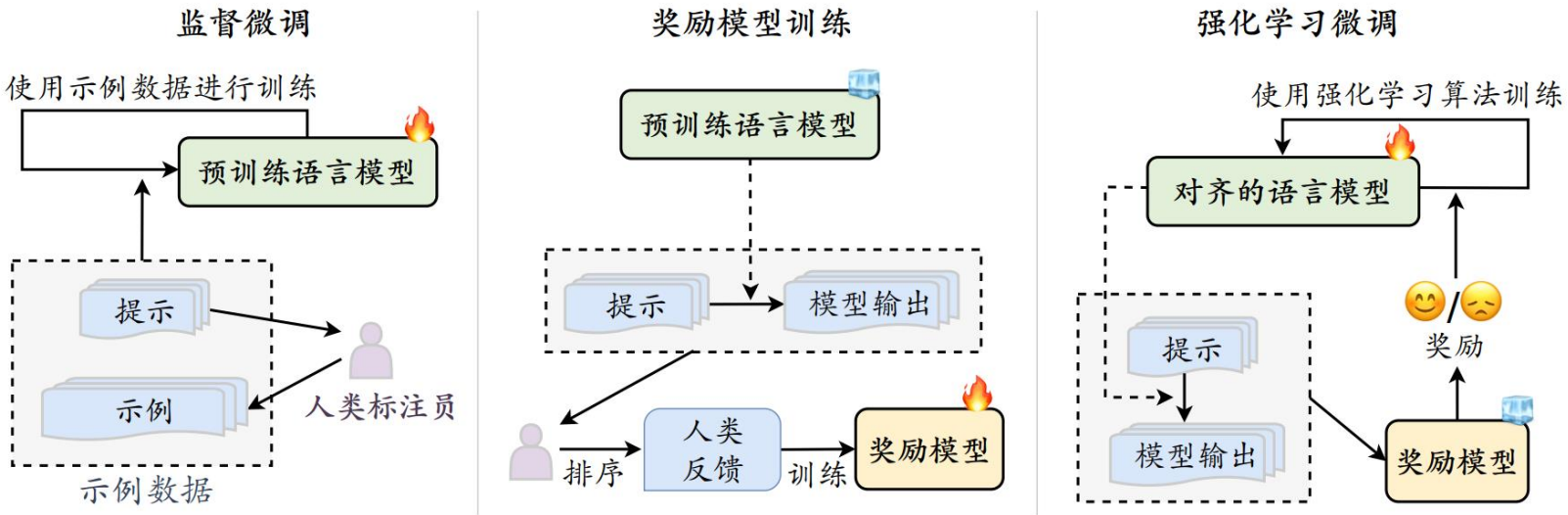


图 8.1 基于人类反馈的强化学习算法工作流程（图片来源：[10]）

目录

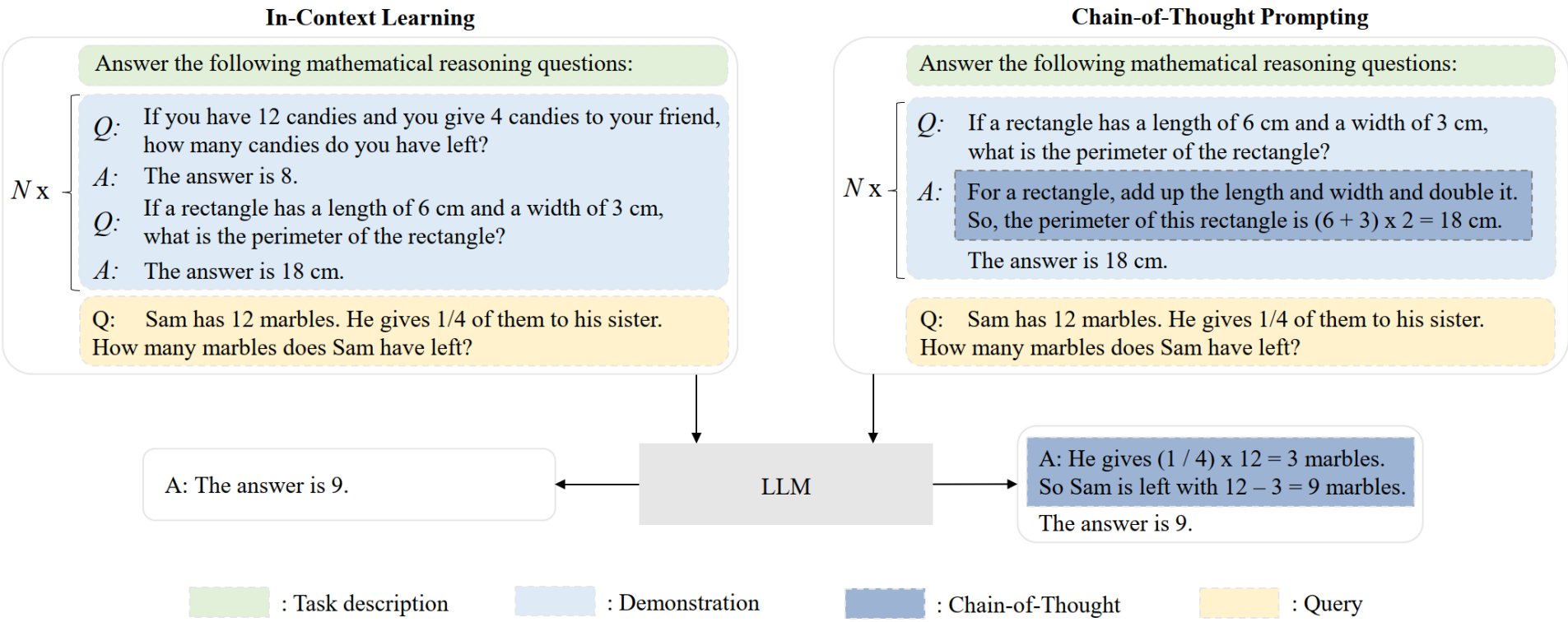
AI Box

- 发展历程
- 预训练
- 微调对齐
- 能力利用

大模型的使用范式

AI Box

- 上下文学习（In-Context Learning, ICL）
  - 思维链提示（Chain-of-Thought Prompting, CoT）



上下文学习

AI Box

定义

- 任务描述（task description）和/或任务示例（demonstration）

$$\text{LLM}(I, \underbrace{f(x_1, y_1), \ldots, f(x_k, y_k)}_{\text{demonstrations}}, f(\underbrace{x_{k+1}}_{\text{input}}, \underbrace{\_\_\_}_{\text{answer}})) \rightarrow \hat{y}_{k+1}$$

Answer the following mathematical reasoning questions:

N x

Q: If you have 12 candies and you give 4 candies to your friend, how many candies do you have left?

A: The answer is 8.

Q: If a rectangle has a length of 6 cm and a width of 3 cm, what is the perimeter of the rectangle?

A: The answer is 18 cm.

Q: Sam has 12 marbles. He gives 1/4 of them to his sister. How many marbles does Sam have left?

Task description

Demonstration

Query

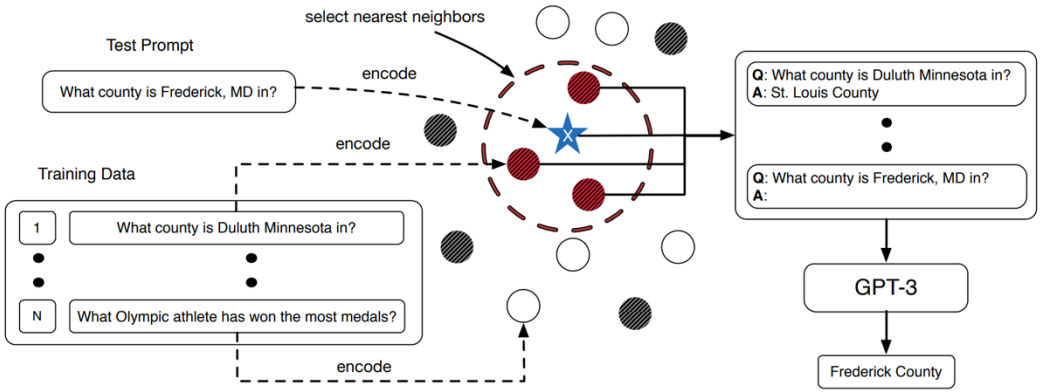
上下文学习

AI Box

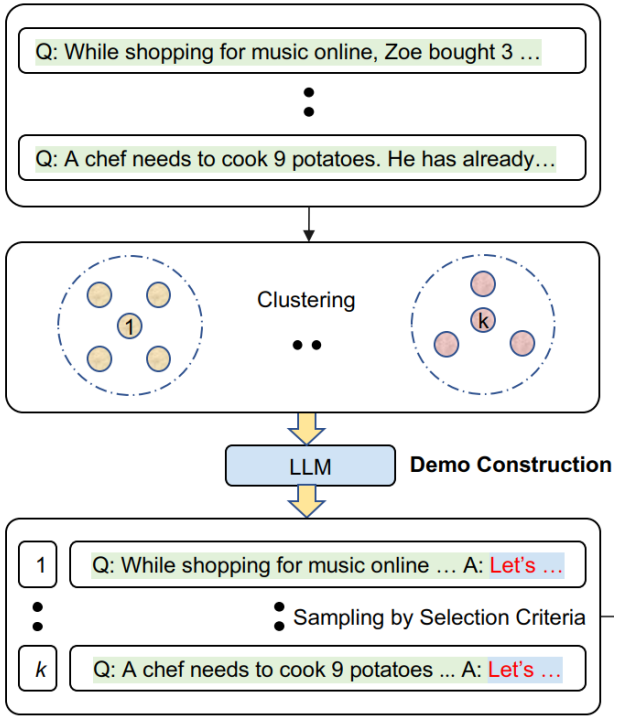
◘ 示例设计：样例选取

➢ 启发式

- 相关性、多样性 …



基于和 query 的相似度
选取相关样例



将样例进行聚类
选取多样化的样例

What makes good in-context examples for gpt-3? **ACL 2022.**
Automatic Chain of Thought Prompting in Large Language Models. **ICLR 2023.**

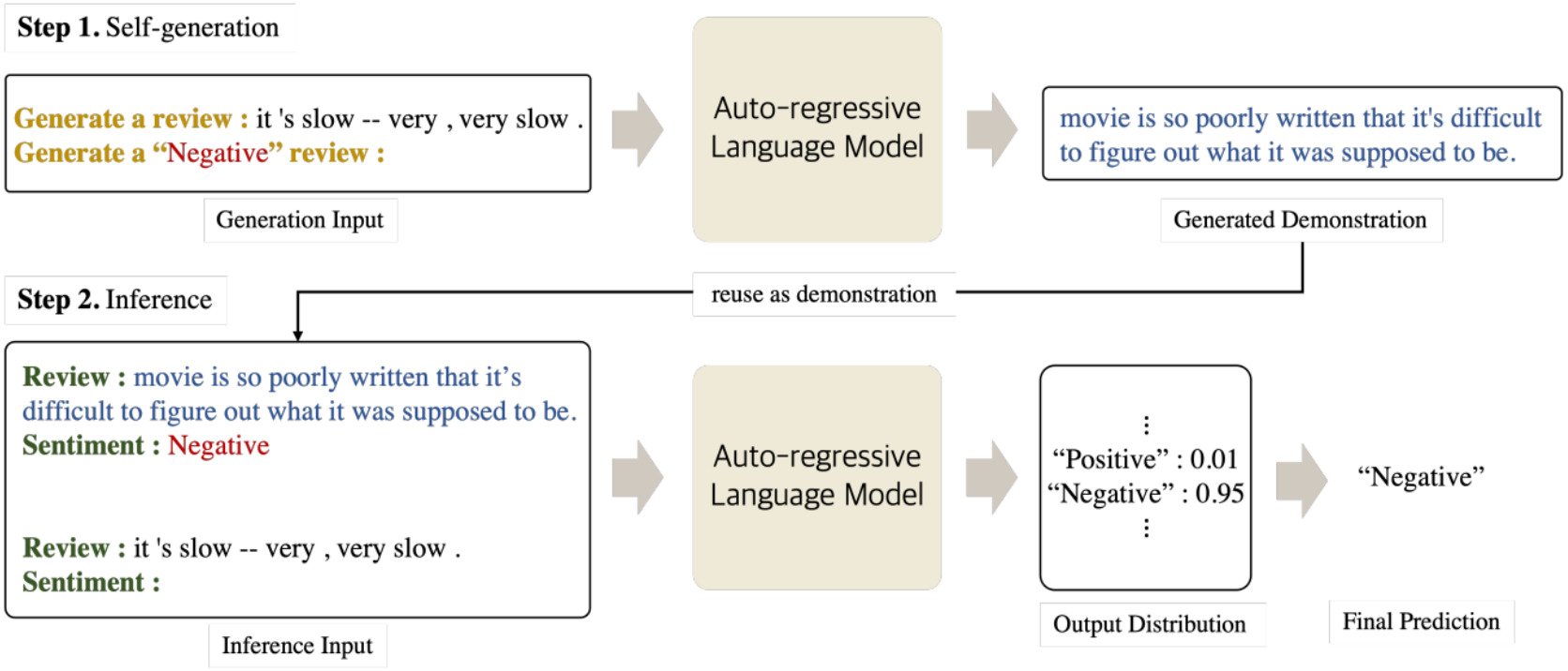
上下文学习

AI Box

◘ 示例设计：样例选取

➢ 基于大模型

- 让大模型根据指令生成样例给自己使用



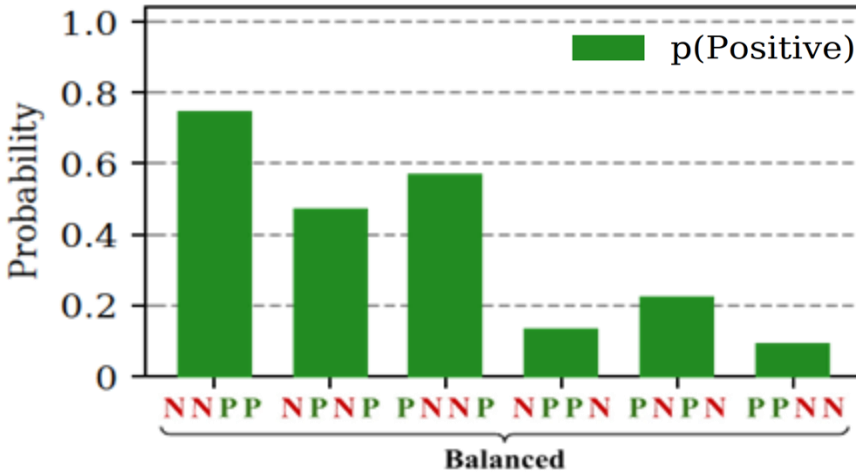
Self-Generated In-Context Learning: Leveraging Auto-regressive Language Models as a Demonstration Generator. **NAACL 2022.**

上下文学习

AI Box

示例设计：样例顺序

- 大模型对于样例顺序敏感
  - 样例类别分布均衡时，预测结果可能偏向位置靠后的样例



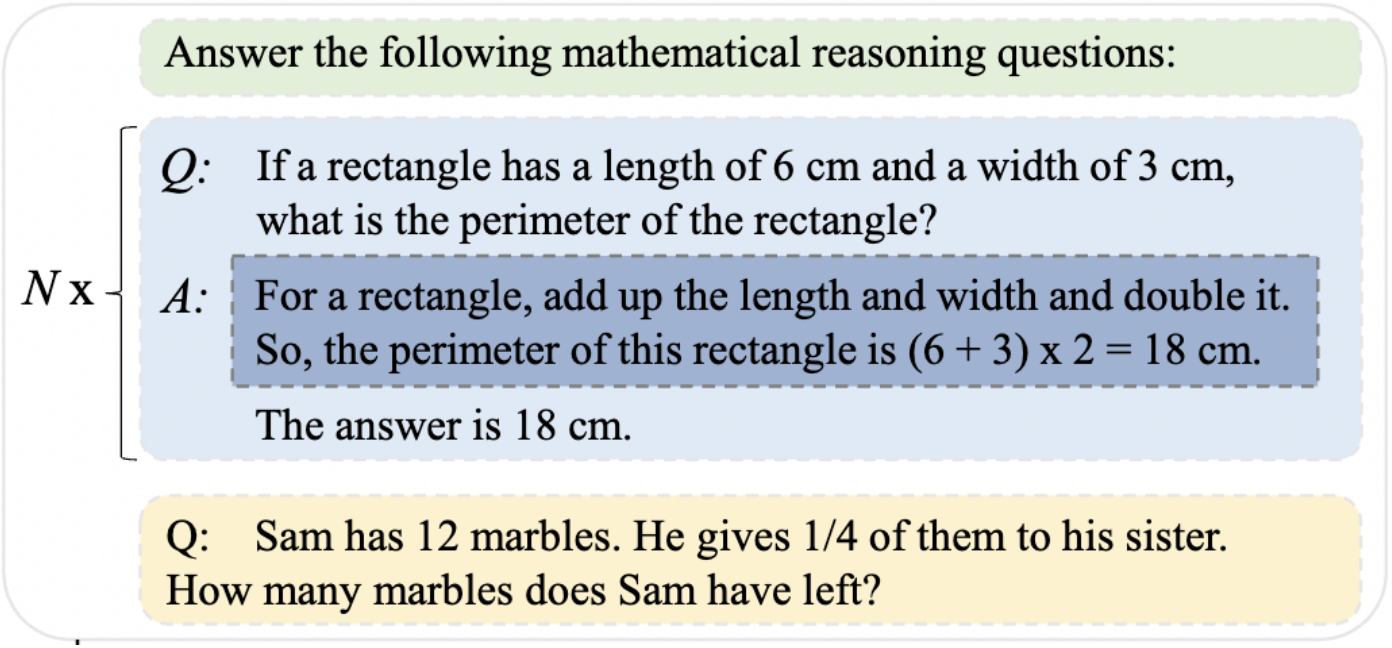
- 选取合适的顺序
  - 相似性等启发式指标

Calibrate Before Use: Improving Few-shot Performance of Language Models. **ICML 2021.**

思维链 (Chain-of-Thought)

AI Box

- 定义：
  - 导出最终答案的一系列中间步骤
  - 是一种适合于推理任务的prompt



思维链 (Chain-of-Thought)

AI Box

Few-shot CoT & Zero-shot CoT

Few-shot CoT

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: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are 16 / 2 = 8 golf balls. Half of the golf balls are blue. So there are 8 / 2 = 4 blue golf balls. The answer is 4. ✓

Zero-shot CoT

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: ***Let's think step by step.***

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

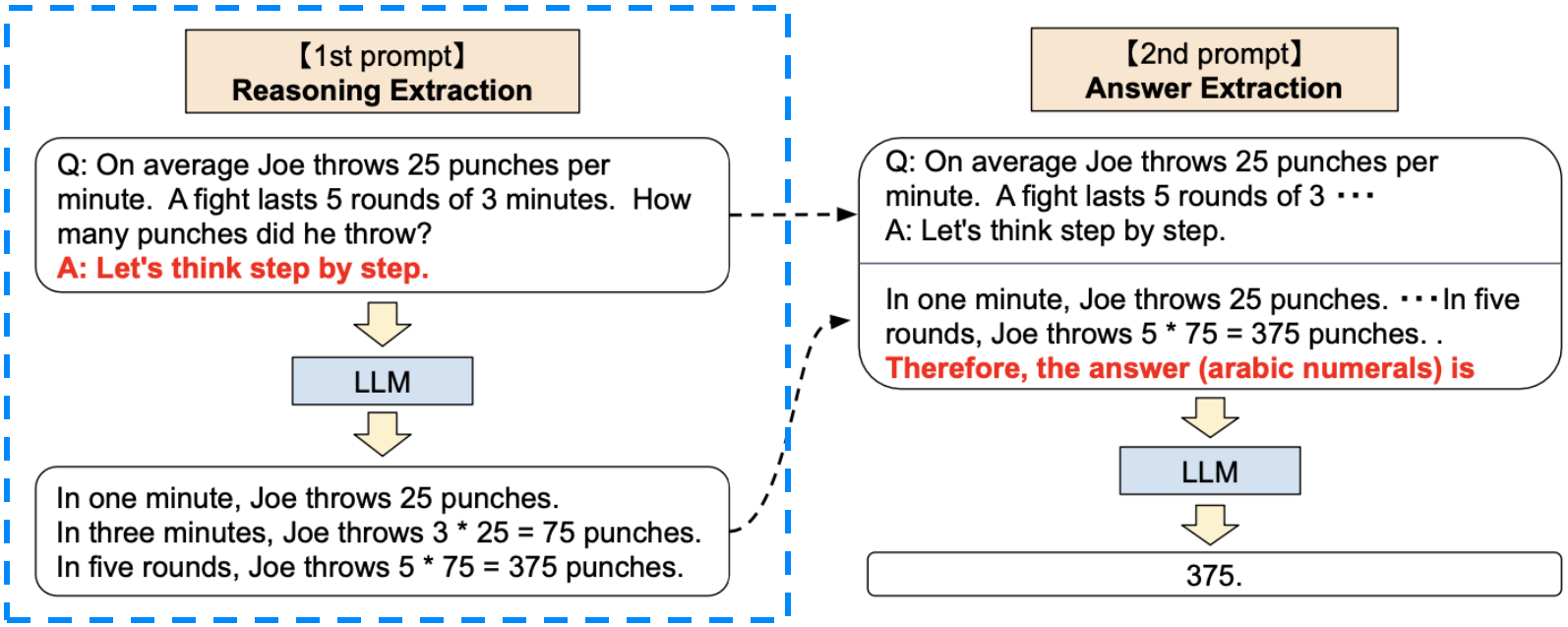
Large Language Models are Zero-Shot Reasoners. **NIPS 2022.**

Zero-shot CoT

AI Box

工作流

- 模型在 Let's think step by step 后生成推理步骤
- 把推理步骤再作为输入，模型生成答案



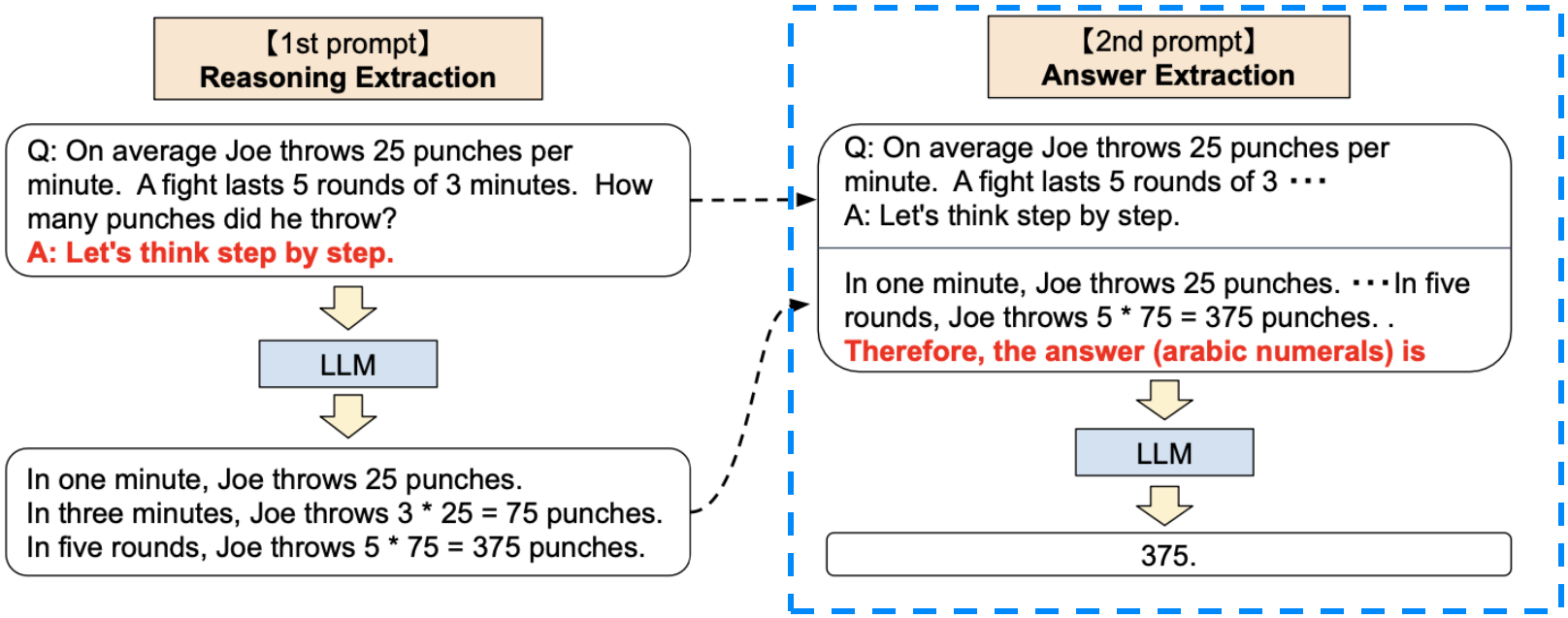
Large Language Models are Zero-Shot Reasoners. NIPS 2022.

Zero-shot CoT

AI Box

工作流

- 模型在Let's think step by step后生成推理步骤
- 把推理步骤再作为输入，模型生成答案



Large Language Models are Zero-Shot Reasoners. NIPS 2022.

提升CoT prompting性能

- 选择合适的prompt
  - 使用多样化的prompt
  - 使用复杂的prompt
  - 使用自动生成的prompt
- 进行可靠的推理
  - 对输出进行集成
  - 对输出进行验证

选择合适的prompt

AI Box

- 使用多样化的prompt
  - 推理路径数量相同时，多样化的prompt优于单一的prompt
    - M1: #prompts， M2: #output per prompt

Method	GSM8K	CQA	CLUTRR
davinci:			
$M_1 = 1, M_2 = 100$	18.9	57.4	42.5
$M_1 = 5, M_2 = 20$	**21.3**	**57.5**	**45.9**
text-davinci-002:			
$M_1 = 1, M_2 = 100$	58.2	72.9	15.8
$M_1 = 5, M_2 = 20$	**61.3**	**77.3**	**21.2**
code-davinci-002:			
$M_1 = 1, M_2 = 100$	76.7	77.3	35.6
$M_1 = 5, M_2 = 20$	**80.0**	**78.8**	**43.8**

On the Advance of Making Language Models Better Reasoners. arXiv 2022

选择合适的prompt

AI Box

- 使用“复杂”的prompt
  - “复杂”：更多的推理步数
  - 同样适用于其他对于“复杂”的定义：问题的长度、解题所需的公式长度

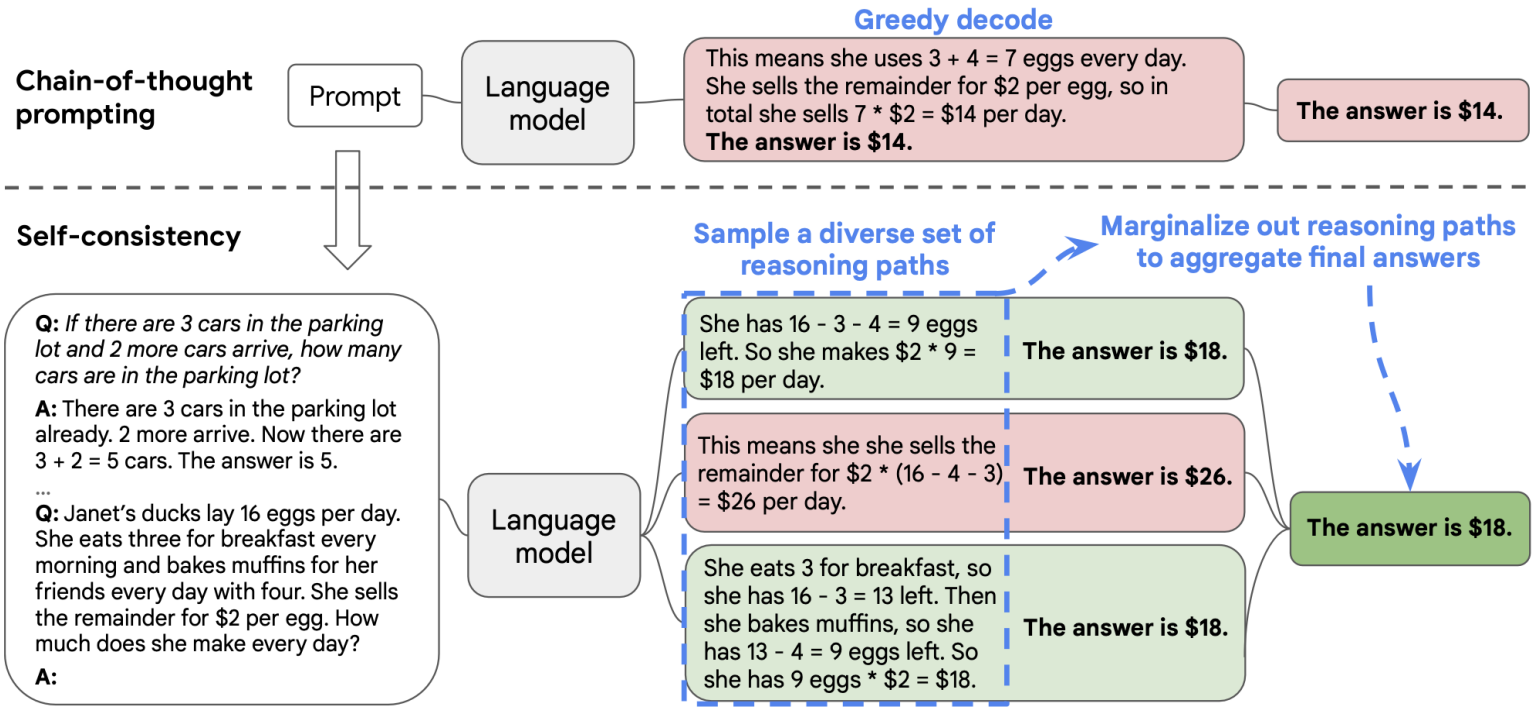
		#Params	GSM8K	MultiArith	MathQA
Previous finetuning SOTA		≤175B	57.0	60.5	37.4
Greedy decoding (Wei et al., 2022b)					
LaMDA† (Thoppilan et al., 2022)		137B	17.1	51.8	-
PaLM† (Chowdhery et al., 2022)		540B	58.1	94.7	-
Minerva† (Lewkowycz et al., 2022)		540B	58.8	-	-
Text-davinci-002	Handcrafted CoT	175B	48.1	90.8	30.1
	Random CoT	175B	49.7	89.5	34.8
	Complex CoT	175B	**55.4** (+7.3)	**94.2** (+3.4)	**36.0** (+5.9)
Code-davinci-002	Handcrafted CoT	175B	61.0	95.8	29.3
	Random CoT	175B	60.4	97.3	40.5
	Complex CoT	175B	**66.6** (+5.6)	**95.8** (+0.0)	**47.3** (+18.0)

Complexity-Based Prompting for Multi-Step Reasoning. arXiv 2023

进行可靠的推理

AI Box

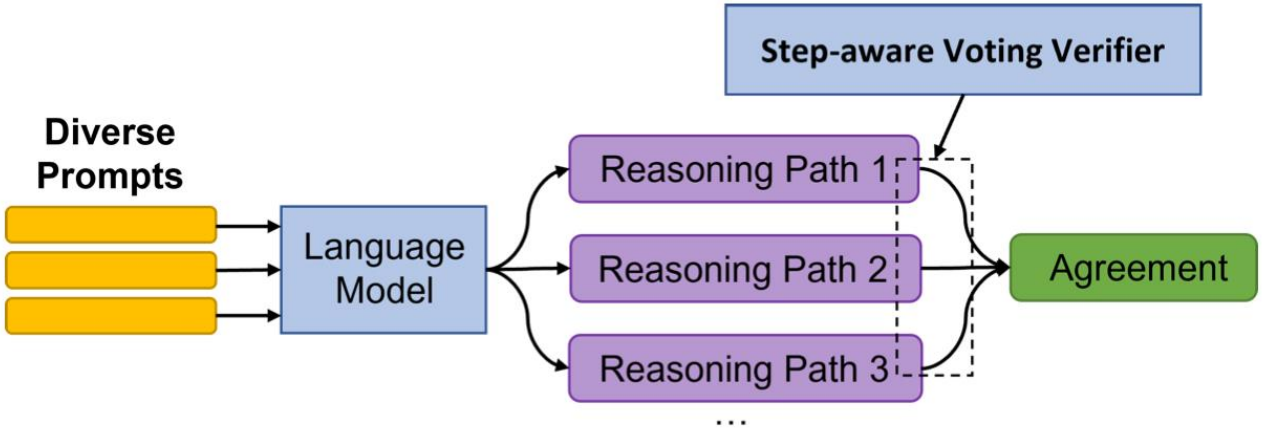
- 对输出进行集成（self-consistency）
  - 对不同的采样策略、参数，模型规模，prompt都适用



Self-Consistency Improves Chain of Thought Reasoning in Language Models. ICLR 2023

进行可靠的推理

- 对输出进行验证
  - 额外训练一个打分模型，对生成的推理路径进行打分
  - 根据每一条路径的分数进行加权，得到每一个答案的得分

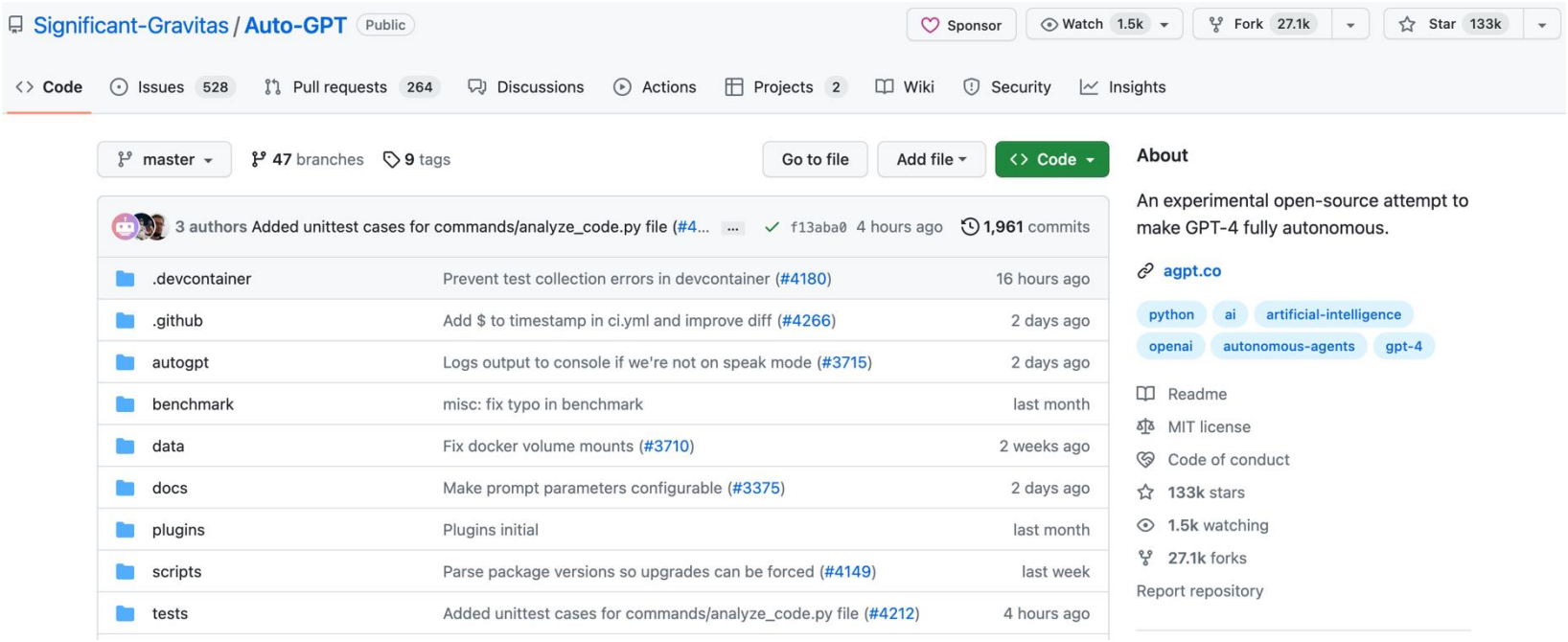


On the Advance of Making Language Models Better Reasoners. arXiv 2022

大模型应用

AI Box

- 自主人工智能 (Auto AI)
  - AutoGPT：发布一个月获13万 star，仅次于Tensorflow



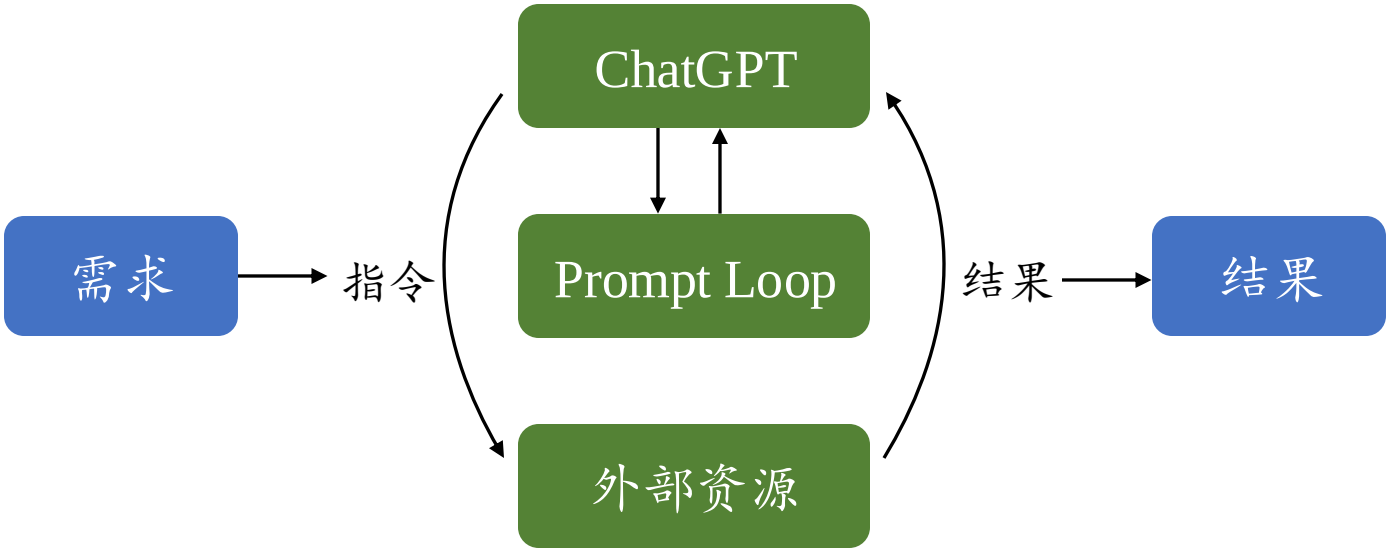
Significant-Gravitas/Auto-GPT: An experimental open-source attempt to make GPT-4 fully autonomous. (github.com)

大模型应用

AI Box

自主人工智能 (Auto AI)

- AutoGPT

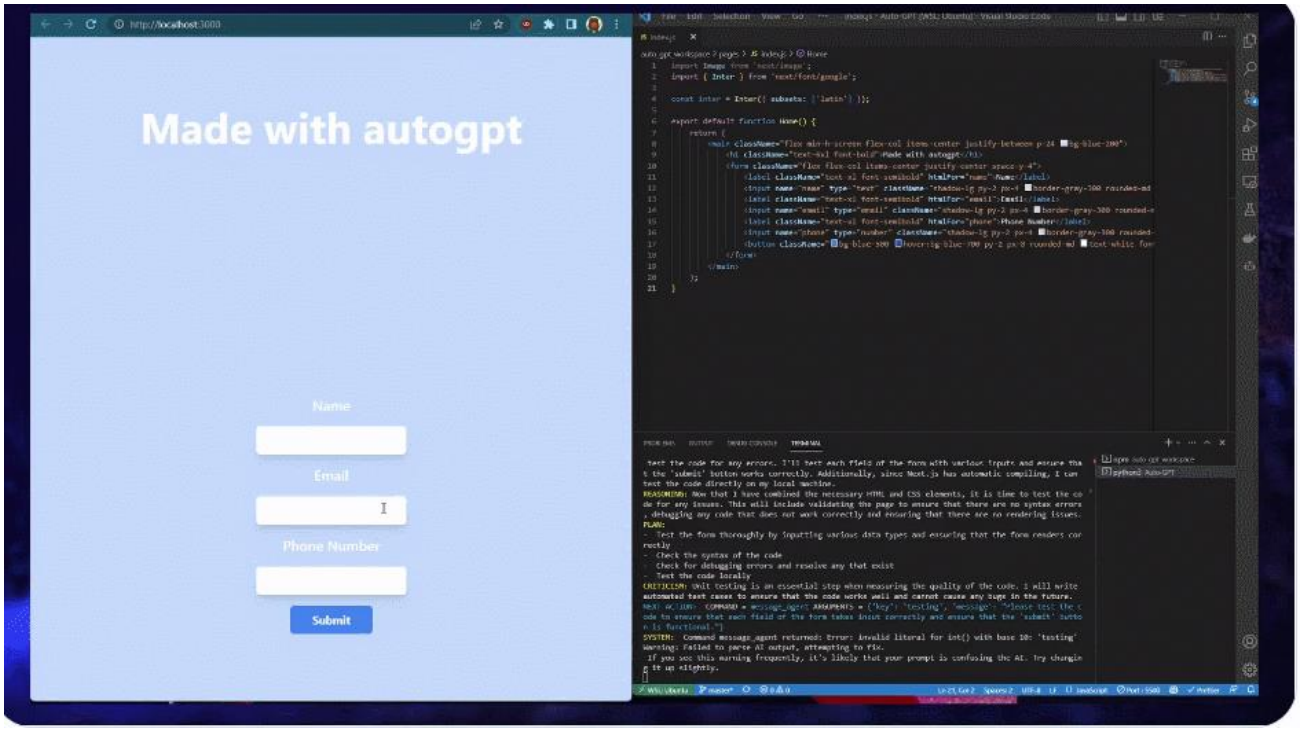


Significant-Gravitas/Auto-GPT: An experimental open-source attempt to make GPT-4 fully autonomous. (github.com)

特化应用——AutoGPT

AI Box

- 自主人工智能
  - AutoGPT：3分钟搭建网站



Significant-Gravitas/Auto-GPT: An experimental open-source attempt to make GPT-4 fully autonomous. (github.com)

特化应用

AI Box

- LLM可扮演人类角色
  - 西部世界

Generative Agents: Interactive Simulacra of Human Behavior. **arxiv 2023**.

特化应用——西部世界

AI Box

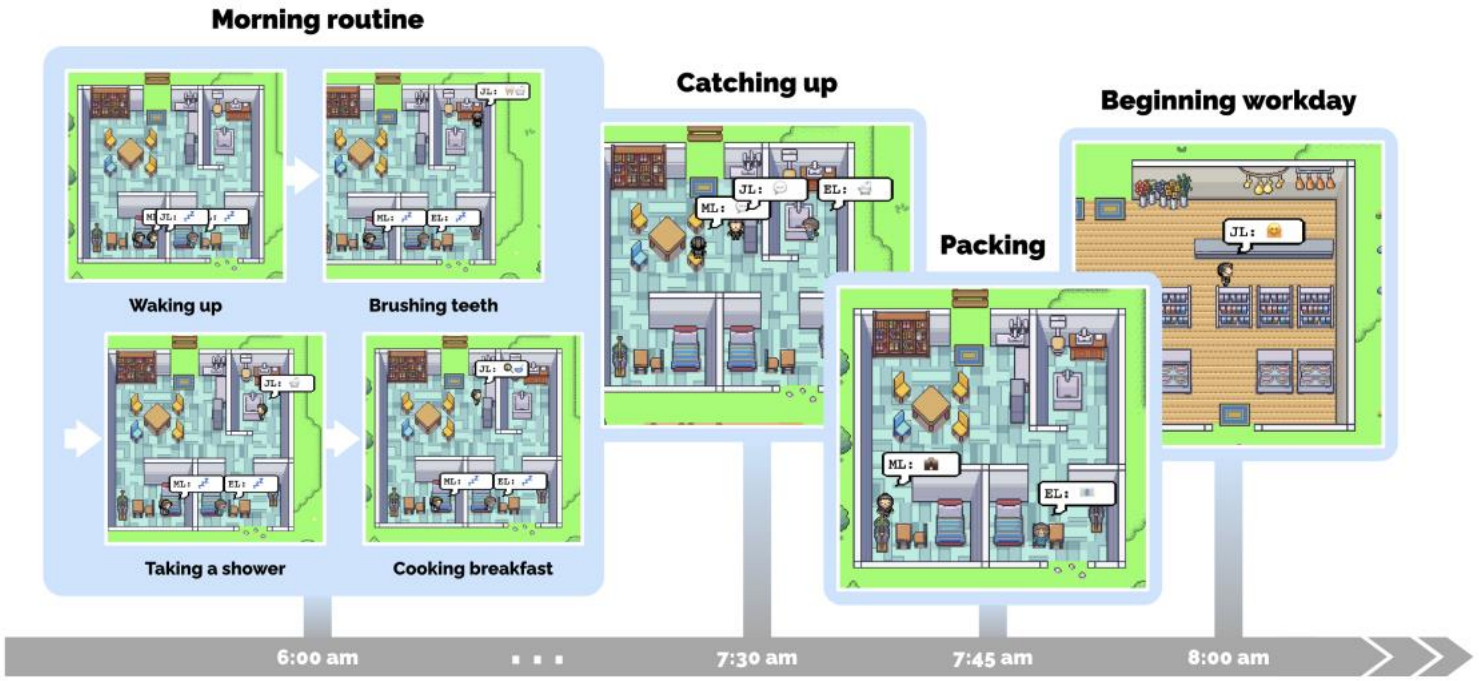
- LLM可作为生成式智能体
  - 智能体与环境交互

Generative Agents: Interactive Simulacra of Human Behavior. **arxiv 2023**.

特化应用——西部世界

AI Box

- LLM可作为生成式智能体
  - 智能体一天的生活



Generative Agents: Interactive Simulacra of Human Behavior. **arxiv 2023**.

特化应用

AI Box

- LLM可完成复杂世界任务
  - 我的世界

Plan4MC: Skill Reinforcement Learning and Planning for Open-World Minecraft Tasks**. arxiv 2023**.

特化应用——我的世界

AI Box

- LLM可完成复杂世界任务
  - 通用智能体：学习技能、通过规划组合成任务

Task Icon	Target Name	Initial Tools	Biome	Max Steps	Involved Skills	Planning Steps
	stick	--	plains	3000	4	4
	crafting_table_nearby	--	plains	3000	5	5
	bowl	--	forest	3000	6	9
	chest	--	forest	3000	6	12
	trap_door	--	forest	3000	6	12
	sign	--	forest	3000	7	13
	wooden_pickaxe	--	forest	3000	7	13
	furnace_nearby	 *10	hills	5000	9	28
	stone_stairs	 *10	hills	5000	8	23
	stone_slab	 *10	hills	3000	8	17
	cobblestone_wall	 *10	hills	5000	8	23
	lever		forest_hills	5000	7	7
	torch	 *10	hills	5000	11	30
	stone_pickaxe		forest_hills	10000	9	16

Plan4MC: Skill Reinforcement Learning and Planning for Open-World Minecraft Tasks. **arxiv 2023**.

特化应用——我的世界

AI Box

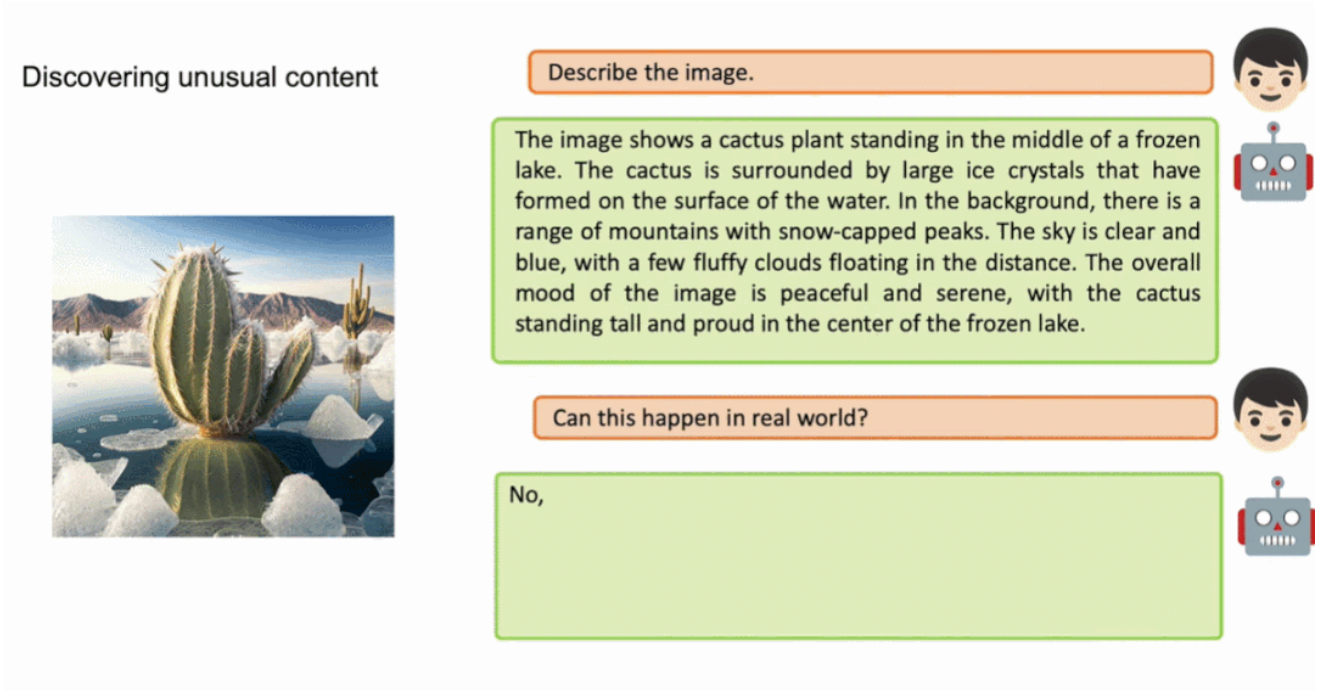
- LLM可完成复杂世界任务
  - 通用智能体：学习技能、通过规划组合成任务

Plan4MC: Skill Reinforcement Learning and Planning for Open-World Minecraft Tasks. **arxiv 2023**.

特化应用——MiniGPT-4

AI Box

- 多模态LLM
  - 图片聊天



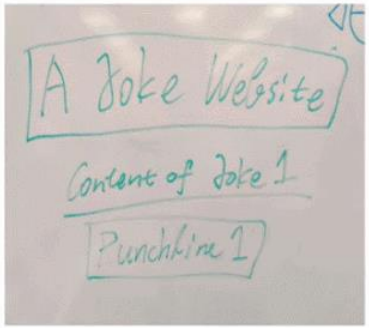
MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. **arxiv 2023**.

特化应用——MiniGPT-4

AI Box

- 多模态LLM
  - 根据草图创建网页

Generating website code from handwritten text and the rendered website



Write brief HTML/JS to turn this mock-up into a colorful website, where the jokes are replaced by real joke

HTML:

```php
<!DOCTYPE html>
<html>
<head>
 <title>A Joke Website</title>
 <link rel="stylesheet" type="text/css" href="style.css">
</head>
<body>
 <div id="header">
  <h1>A Joke Website</h1>
  <nav>
   <ul>
    <li><a href="#">Home</a></li>
    <li><a href="#">About</a></li>
    <li><a href="#">Contact</a></li>
```

MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. **arxiv 2023**.

特化应用——MiniGPT-4

AI Box

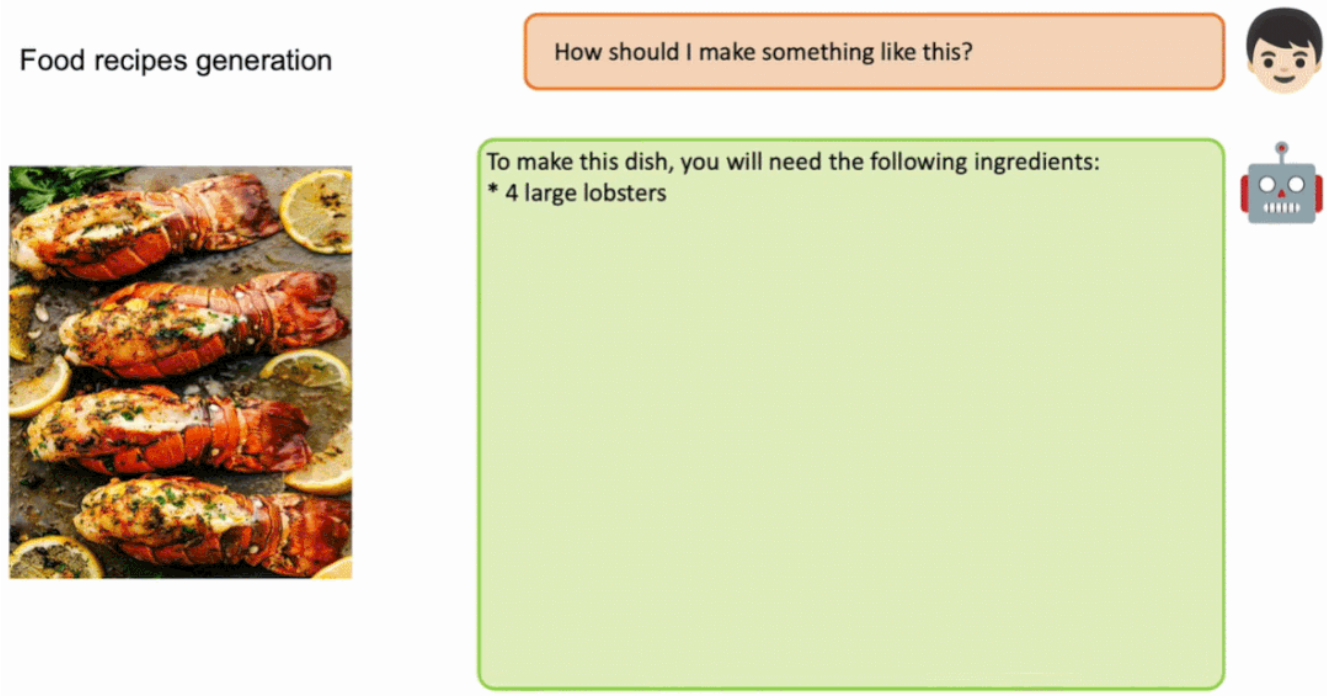
- 多模态LLM
  - 为图片写广告语

MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. **arxiv 2023**.

特化应用——MiniGPT-4

AI Box

- 多模态LLM
  - 为图片写菜谱

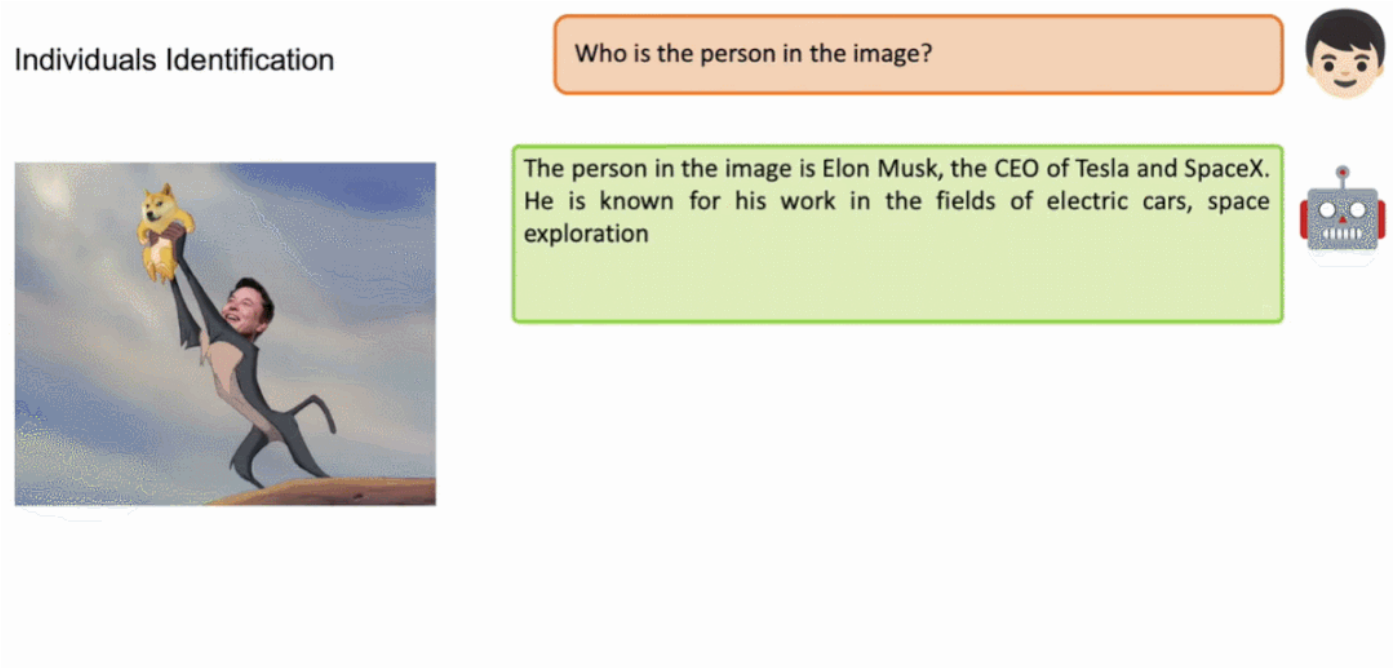


MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. **arxiv 2023**.

特化应用——MiniGPT-4

AI Box

- 多模态LLM
  - 解释梗图



MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. **arxiv 2023**.

A Survey of Large Language Models

Wayne Xin Zhao, Kun Zhou*, Junyi Li*, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie and Ji-Rong Wen

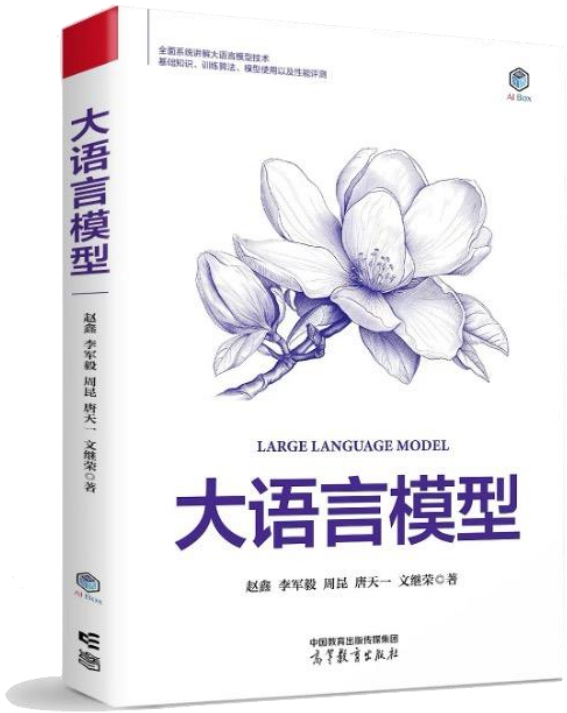
Abstract—Ever since the Turing Test was proposed in the 1950s, humans have explored the mastering of language intelligence by machine. Language is essentially a complex, intricate system of human expressions governed by grammatical rules. It poses a significant challenge to develop capable artificial intelligence (AI) algorithms for comprehending and grasping a language. As a major approach, *language modeling* has been widely studied for language understanding and generation in the past two decades, evolving from statistical language models to neural language models. Recently, pre-trained language models (PLMs) have been proposed by pre-training Transformer models over large-scale corpora, showing strong capabilities in solving various natural language processing (NLP) tasks. Since the researchers have found that model scaling can lead to an improved model capacity, they further investigate the scaling effect by increasing the parameter scale to an even larger size. Interestingly, when the parameter scale exceeds a certain level, these enlarged language models not only achieve a significant performance improvement, but also exhibit some special abilities (*e.g.,* in-context learning) that are not present in small-scale language models (*e.g.,* BERT). To discriminate the language models in different parameter scales, the research community has coined the term *large language models (LLM)* for the PLMs of significant size (*e.g.,* containing tens or hundreds of billions of parameters). Recently, the research on LLMs has been largely advanced by both academia and industry, and a remarkable progress is the launch of ChatGPT (a powerful AI chatbot developed based on LLMs), which has attracted widespread attention from society. The technical evolution of LLMs has been making an important impact on the entire AI community, which would revolutionize the way how we develop and use AI algorithms. Considering this rapid technical progress, in this survey, we review the recent advances of LLMs by introducing the background, key findings, and mainstream techniques. In particular, we focus on four major aspects of LLMs, namely pre-training, adaptation tuning, utilization, and capacity evaluation. Besides, we also summarize the available resources for developing LLMs and discuss the remaining issues for future directions. This survey provides an up-to-date review of the literature on LLMs, which can be a useful resource for both researchers and engineers.

Index Terms—Large Language Models; Emergent Abilities; Adaptation Tuning; Utilization; Alignment; Capacity Evaluation

大模型中文书

AI Box

- 公开出版发售



部分	章节
基础	第一章 引言 第二章 基础介绍 第三章 大模型资源
预训练	第四章 数据准备 第五章 模型架构 第六章 模型预训练
微调与对齐	第七章 指令微调 第八章 人类对齐
大模型使用	第九章 解码与部署 第十章 提示学习 第十一章 规划与智能体
评测	第十二章 评测

谢谢！