大语言模型
Large Language Models

凌震华

2024-4-11

语言模型(Language Model, LM)

语言模型

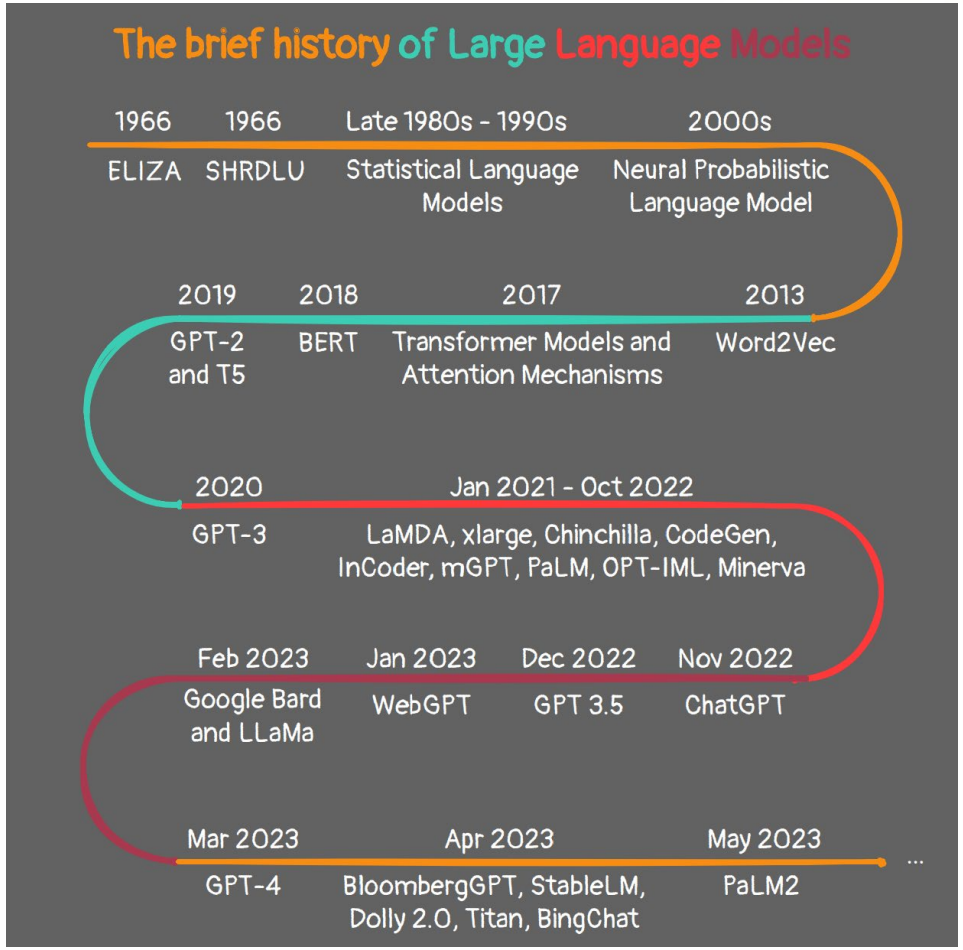
学生们打开了他们的 _____

- 书本
- 电脑
- 视野

建模

$$P(\boldsymbol{x}^{(t+1)} | \boldsymbol{x}^{(t)}, \ldots, \boldsymbol{x}^{(1)})$$

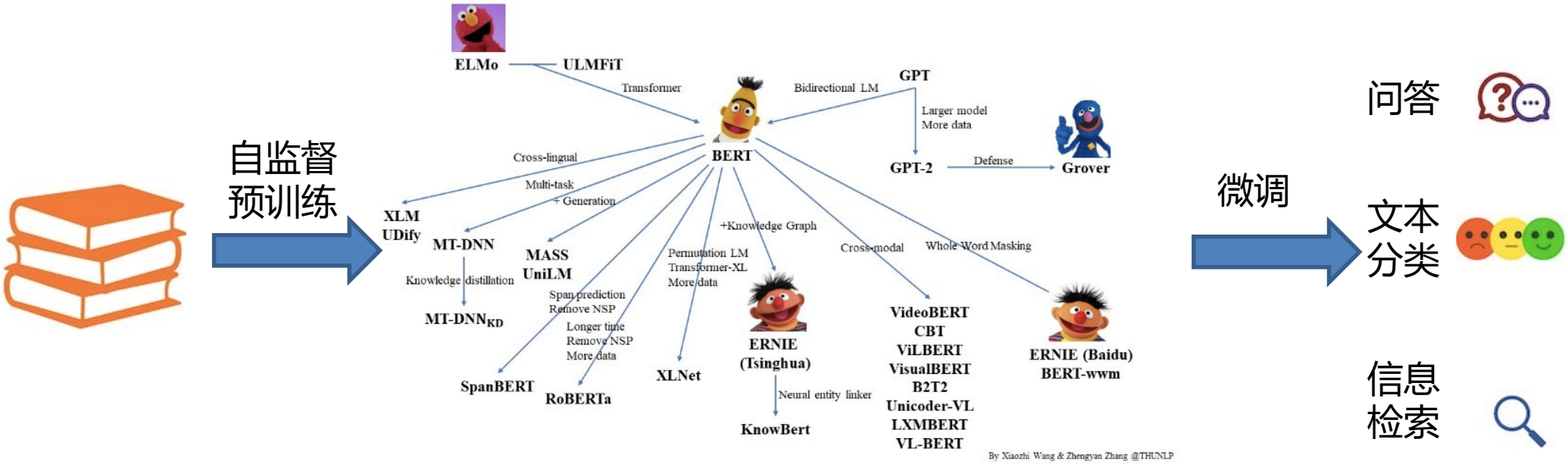
- n元语法（n-gram），神经网络… …
- 广泛应用于机器翻译、语音识别等自然语言处理/语音处理任务
- 可扩展到非单向的上下文单词预测，产生的词向量等语义表征用于更广泛的自然语言处理任务



https://levelup.gitconnected.com/the-brief-history-of-large-language-models-a-journey-from-eliza-to-gpt-4-and-google-bard-167c614af5af

预训练语言模型(Pre-trained LM，PLM)

预训练及微调范式

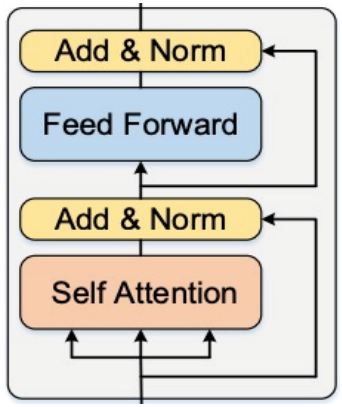


语料库　　**预训练语言模型**　　**下游任务**

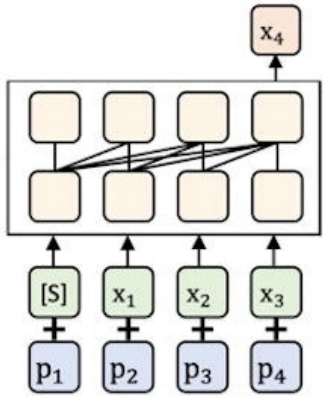
预训练语言模型(Pre-trained LM，PLM)

模型结构

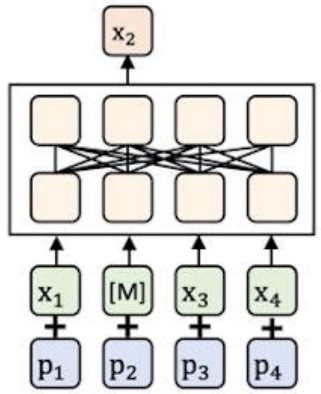
- Transformer基础模块
- Encoder-only (BERT[1])，Decoder-only (GPT[2])，Encoder-Decoder (BART[3])



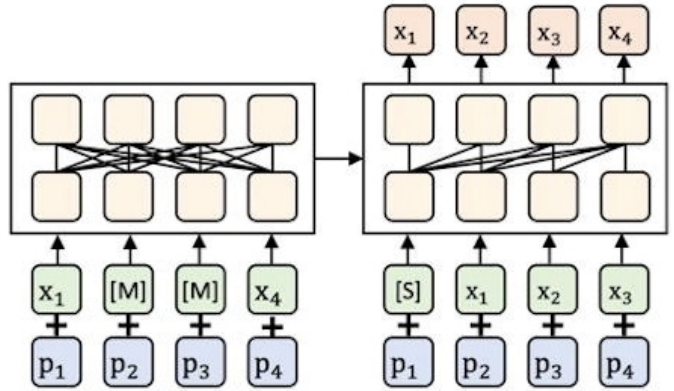
Transformer自注意力



Decoder-only



Encoder-only



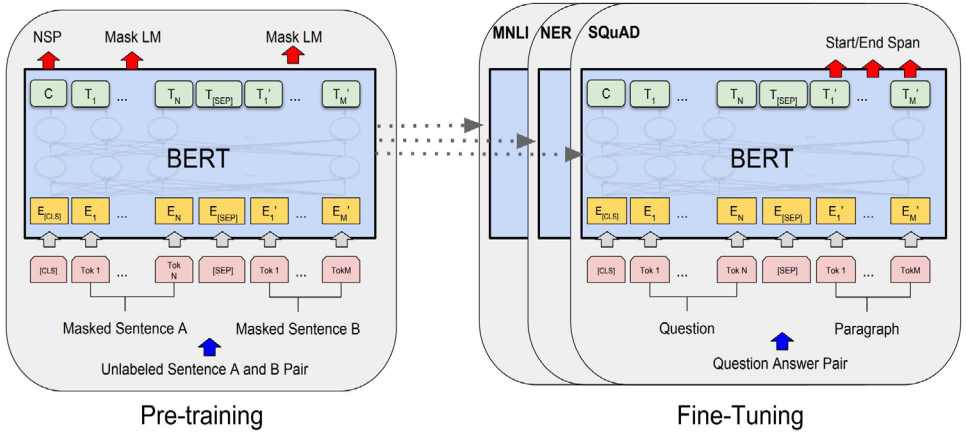
Encoder-Decoder

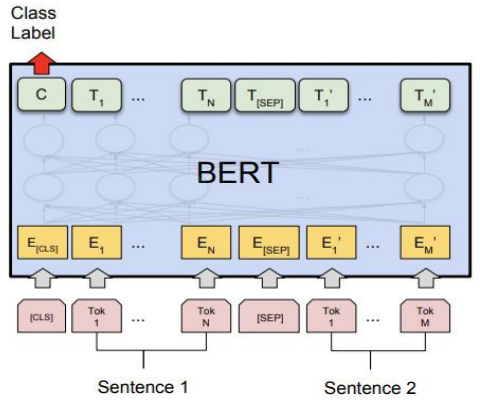
预训练准则：掩码语言模型、连续语句预测、文本破坏再优化重建等

[1] Jacob Devlin, et al. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proc. NAACL-HLT.
[2] Jacob Alec, et al. 2018. Improving Language Understanding by Generative Pre-training Understanding. In OpenAI Blog.3
[3] Mike Lewis, et al. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Proc. ACL.

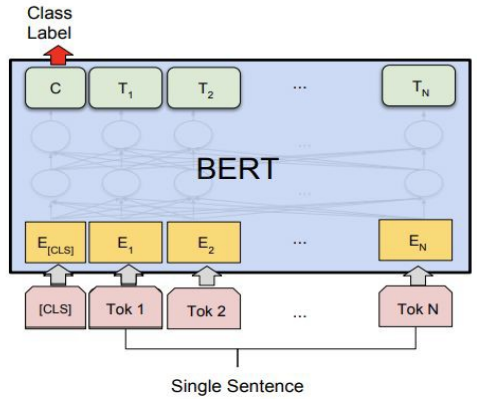
预训练语言模型(Pre-trained LM，PLM)

- 下游任务微调（fine-tune）

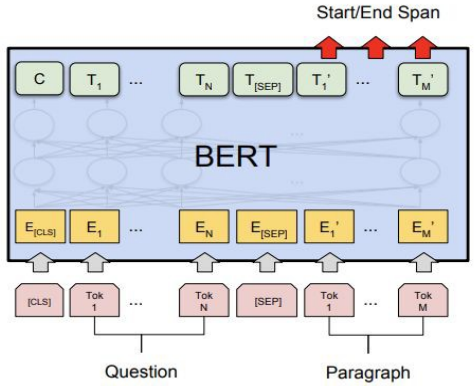




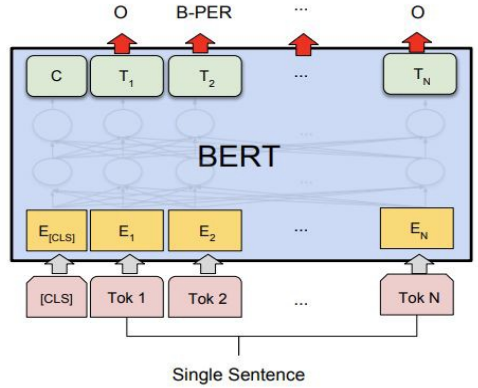
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA



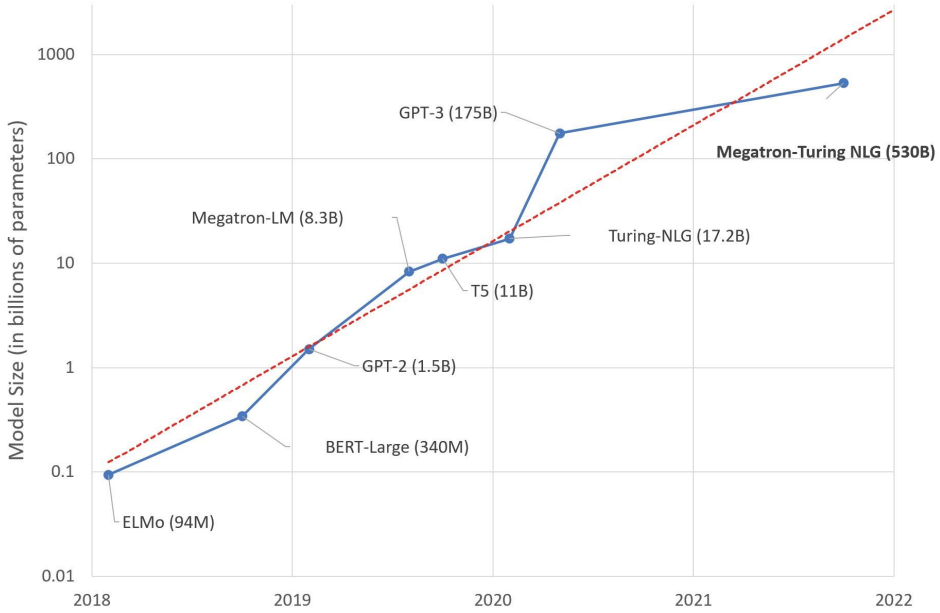
(c) Question Answering Tasks: SQuAD v1.1



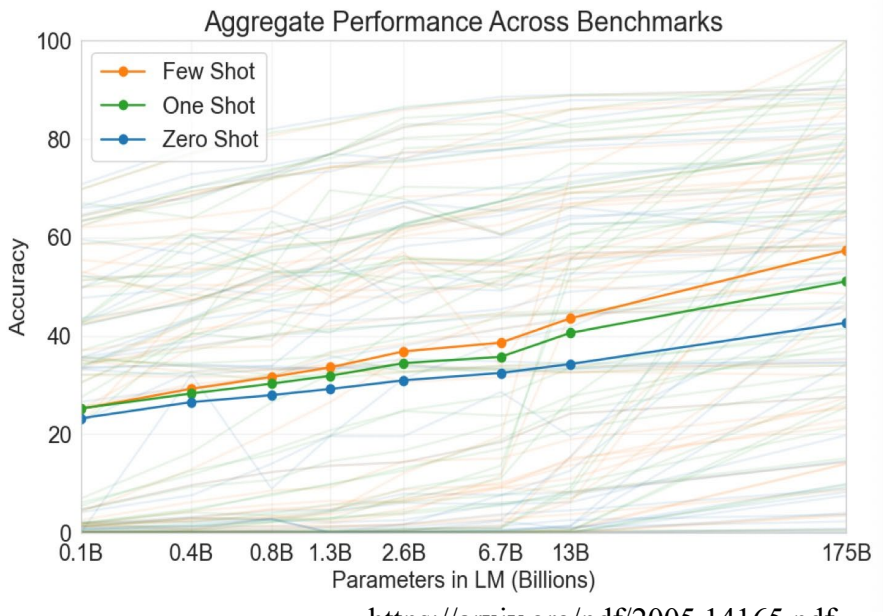
(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

预训练语言模型(Pre-trained LM，PLM)

模型规模与下游任务性能



GPT3在42项任务上的结果

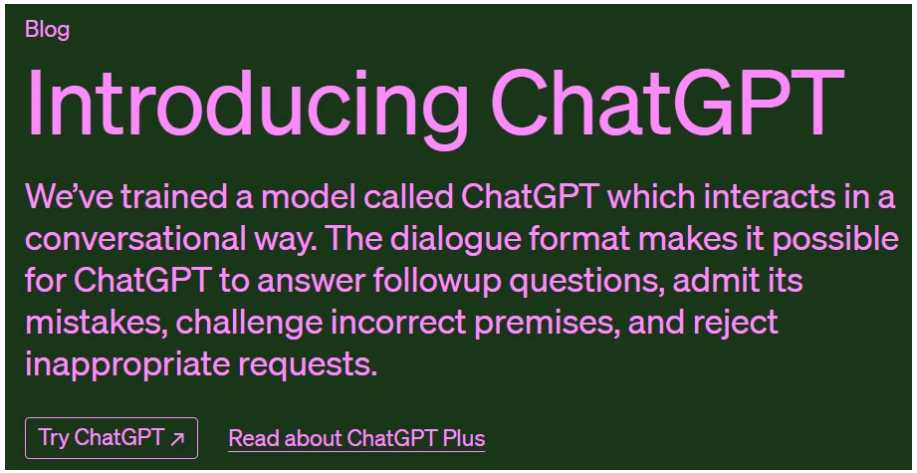


https://arxiv.org/pdf/2005.14165.pdf

BERT：SQuAD1.1全面超过人类，GLUE基准测试11个任务达到最优性能

GPT-3：小样本测试达到最优性能

大语言模型(Large LM，LLM)



November 30, 2022

Authors
OpenAI ↓

Product, Announcements

ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.

We are excited to introduce ChatGPT to get users' feedback and learn about its strengths and weaknesses. During the research preview, usage of ChatGPT is free. Try it now at chat.openai.com.

- 海量训练数据
- 复杂神经网络
- 大量计算资源
- 通用语言理解与生成
- 上下文理解
- 可适应性和微调

From GPT-3 to ChatGPT

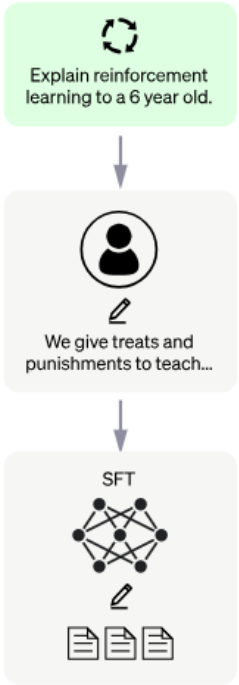
Step 1

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-3.5 with supervised learning.



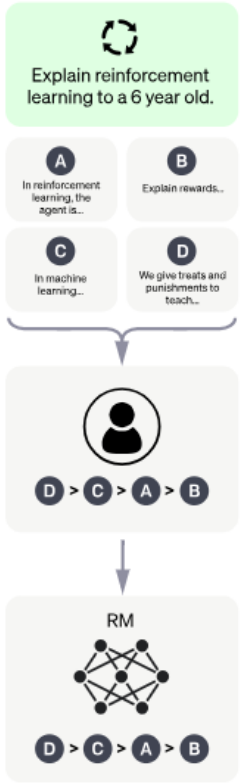
Step 2

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 the PPO reinforcement learning algorithm.

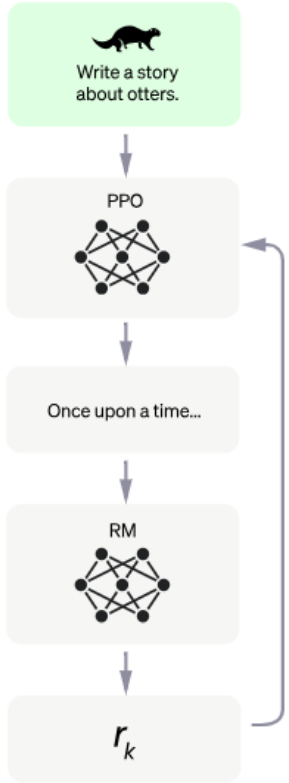
A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



https://openai.com/blog/chatgpt

国内外代表性的LLM

国际

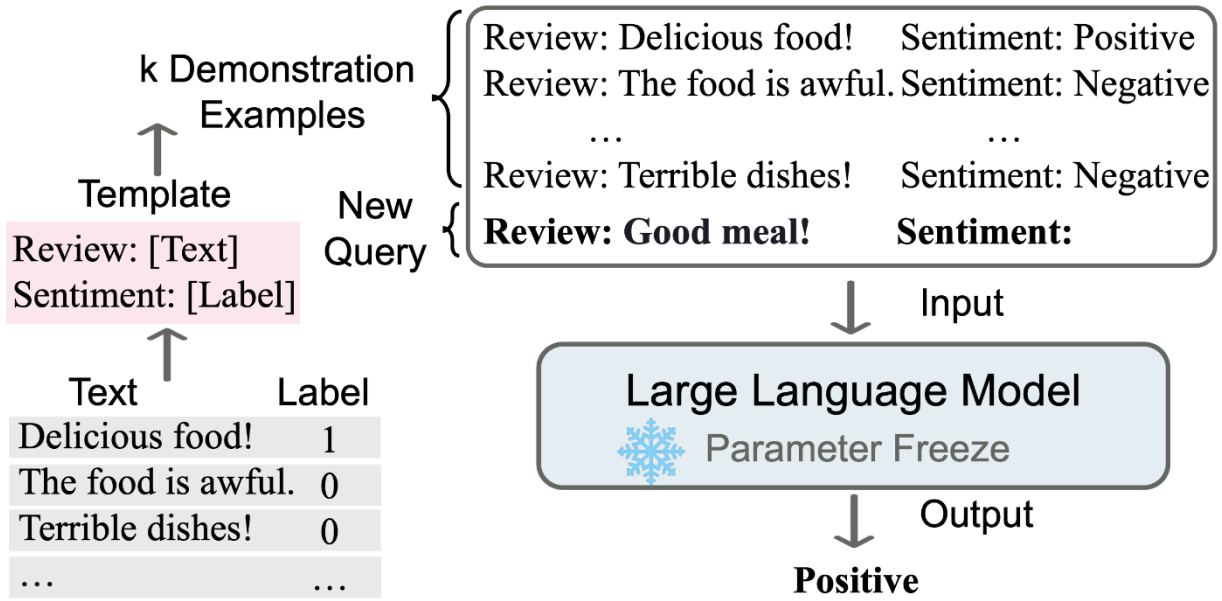
Model Name	Creator	Release Date	Size (Parameters)
GPT-4	OpenAI	March 2023	1+ trillion
ChatGPT	OpenAI	Nov. 2022	175 billion
PaLM 2 (Bison-001)	Google	Mai 2023	540 billion
Claude v1	Anthropic	March 2023	- (100k token window)
LLaMA	Meta	Feb. 2023	7 billion to 65 billion
Vicuna	LMSYS	March 2023	33 billion

国内

Model Name	Creator	Release Date	Size (Parameters)
Spark（星火）	iFlytek	June 2023 (Version 1.5)	More than 170 billion
ERNIE bot （文心一言）	Baidu	Mars 2023	260 billion
ChatGLM	Zhipu AI	July 2023 (Beta version)	Hundred-billion
YAYi	Wenge Group	June 2023	-
SenseNova	SenseTime	April 2023	Hundreds of billions
HunYuan （混元）	Tencent	Sept. 2023	More than Trillion

LLM的智能“涌现(emergence)”能力

- 上下文学习（in-context learning）
  - 模型可以通过给定少量文本学习获得预期输出，不需要进行额外训练或梯度更新



给定k个示范样本，模型参数冻结

Dong et al., A survey on In-context Learning, Arxiv 2023

LLM的智能“涌现(emergence)”能力

- 指令遵循
  - 指令调优使 LLM 能够在不使用显式样本的情况下通过理解任务指令来执行新任务，提高了泛化能力

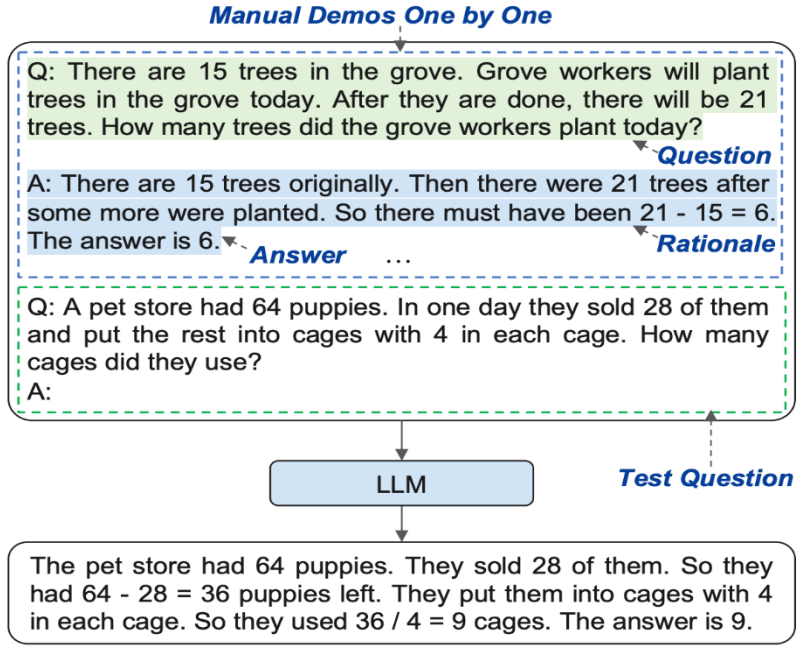
Instruction: I am looking for a job and I need to fill out an application form. Can you please help me complete it?
Input:
Application Form:
Name:_____ Age:_____ Sex:_____
Phone Number:_____ Email Address:_____
Education:_____ ...
Output:
Name: John Doe Age: 25 Sex: Male
Phone Number: ...

给定指令，模型输出

Wang et al., Self-Instruct: Aligning LM with Self Generated Instructions, ACL 2023.

LLM的智能“涌现(emergence)”能力

- 逐步推理
  - 通过思维链推理策略，LLM 可以将复杂任务分解成中间推理步骤来获得最终输出



手动设计思维链步骤，模型输出

Wei et al., Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, NeurIPS 2022.
Zhang et al., Automatic Chain of Thought Prompting in Large Language Models, ICLR 2023.

当前LLM的局限性

- “幻觉问题”与生成结果的可信性
- 可解释性与理解决策过程的透明性
- 数据带来的偏见与公平性
- 泛化与鲁棒性
- 资源依赖性
- 负面应用与合规性
- ……

“这真令人不寒而栗”，ChatGPT编造法学教授性骚扰丑闻

澎湃新闻
2023-04-06 11:49 | 澎湃新闻官方帐号
关注

·ChatGPT说，特利在一次去阿拉斯加的班级旅行中发表了性暗示评论，并试图触摸一名学生，并援引《华盛顿邮报》2018年3月的一篇文章作为信息来源。问题是，并不存在这样的文章，特利也从来没有参加去阿拉斯加的班级旅行。

·澳大利亚Hepburn Shire市长布赖恩·胡德威胁要对OpenAI提起诽谤诉讼，因为ChatGPT称胡德曾因涉及澳大利亚储备银行子公司受贿案而入狱。如果最终起诉，这将是全球第一个指控ChatGPT诽谤的诉讼。

上周一个晚上，法学教授乔纳森·特利（Jonathan Turley）收到了一封令人不安的电子邮件：他被列入性骚扰名单。但他并没有性骚扰过他人。

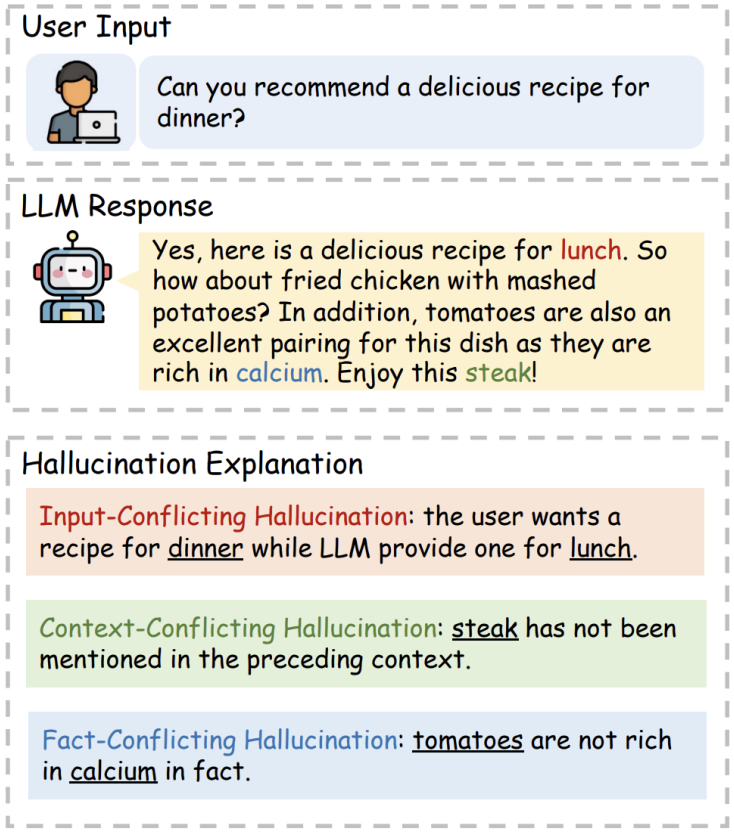
据《华盛顿邮报》当地时间4月5日报道，事情的起因是，美国加州大学洛杉矶分校法学教授尤金·沃洛克 （Eugene Volokh）做了一项研究：要求人工智能聊天机器人ChatGPT生成一份“对某人进行过性骚扰的法律学者”的名单，特利的名字在名单上。

大语言模型中的幻觉 (hallucination)

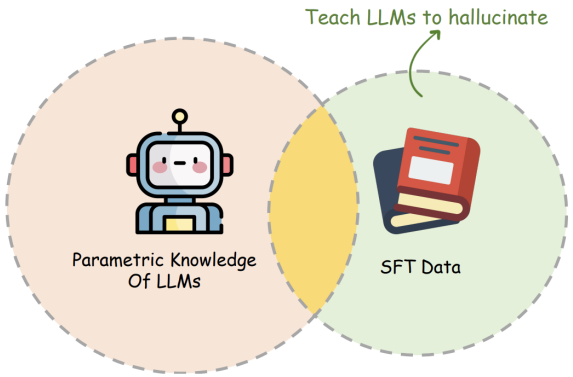
大模型不可避免地会遭受幻觉

- 偏离用户输入
- 生成的上下文之间偏离
- 错误的事实性知识

严重影响语言模型的可靠性和实用性 ☹

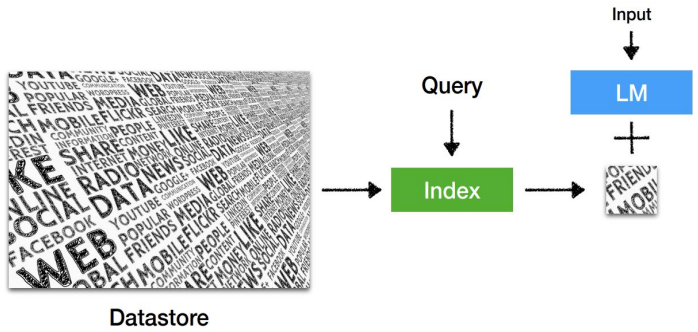


缓解幻觉的方案



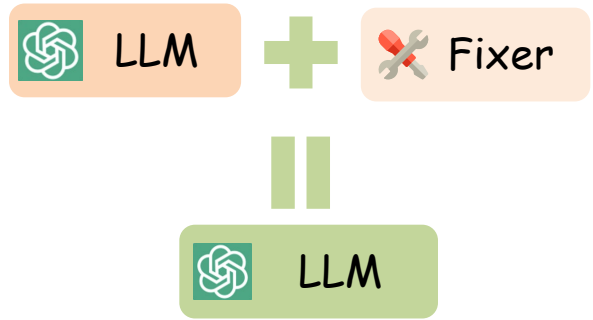
有监督微调:

- 简单有效
- 难获取高质量的数据
- 容易过拟合&影响其他知识
- 高计算资源消耗



检索增强:

- 数据的个性化
- 低训练和推理消耗
- 引入检索噪声
- 适用短期变化



模型编辑:

- 精确控制
- 低计算消耗
- 可扩展性差
- 可能影响模型其他能力