

Web信息处理与应用

第十五节 大语言模型导论

徐童 2023.12.11

注：本节部分内容来自复旦大学李直旭教授《AIGC时代的多模态知识工程思考与展望》报告，在此表示感谢

• 传统搜索引擎：从Archie开始的漫长旅程

- Archie：一般公认最早的搜索引擎
 - 诞生于1990年，由麦吉尔大学的Alan Emtage等几位学生发明，用于搜索互联网上的匿名FTP
 - 目前，仍有少量提供Archie服务的网站
- 自Archie始，搜索引擎作为人类获取信息的重要方式而不断更新升级

由波兰华沙理工大学提供的Archie
http://archie.icm.edu.pl/archie-adv_eng.html

Archie Query Form 

Search for: _____

Database: Worldwide Anonymous FTP Polish Web Index
Search Type: Sub String Exact Regular Expression
Case: Insensitive Sensitive

Do you want to look up strings only (no sites returned):
 NO YES

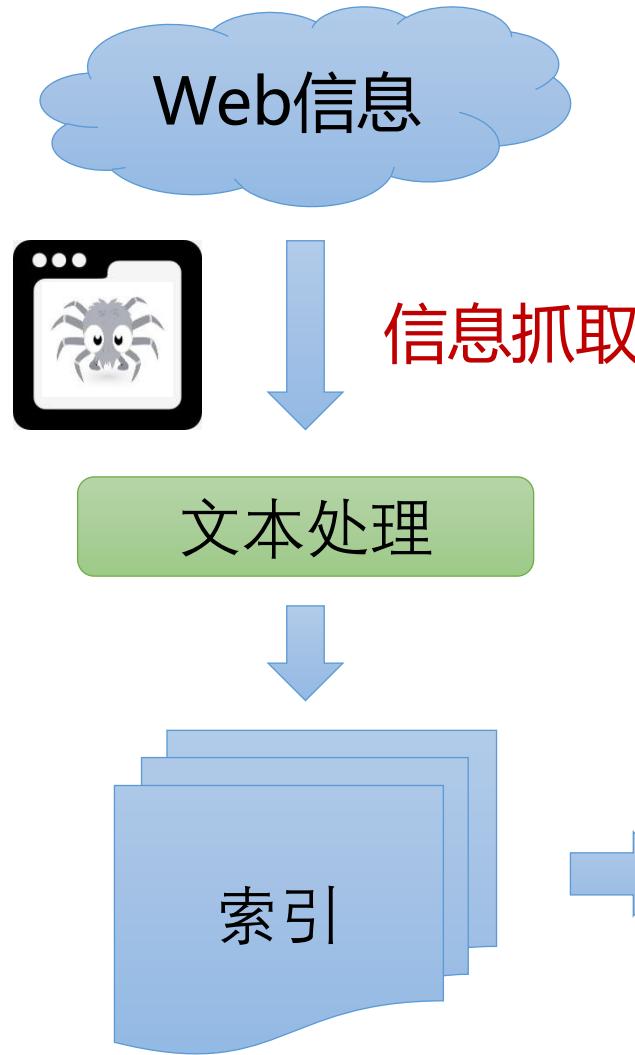
Output Format For Web Index Search: Keywords Only
 Excerpts Only Links Only

• 划时代的变革：NewBing与生成式检索

- 2023年，NewBing
 - 与大模型相结合，开启信息检索新时代
 - 从提供信息，抽取/归纳信息到“生成”信息，提供更灵活与更友好的服务

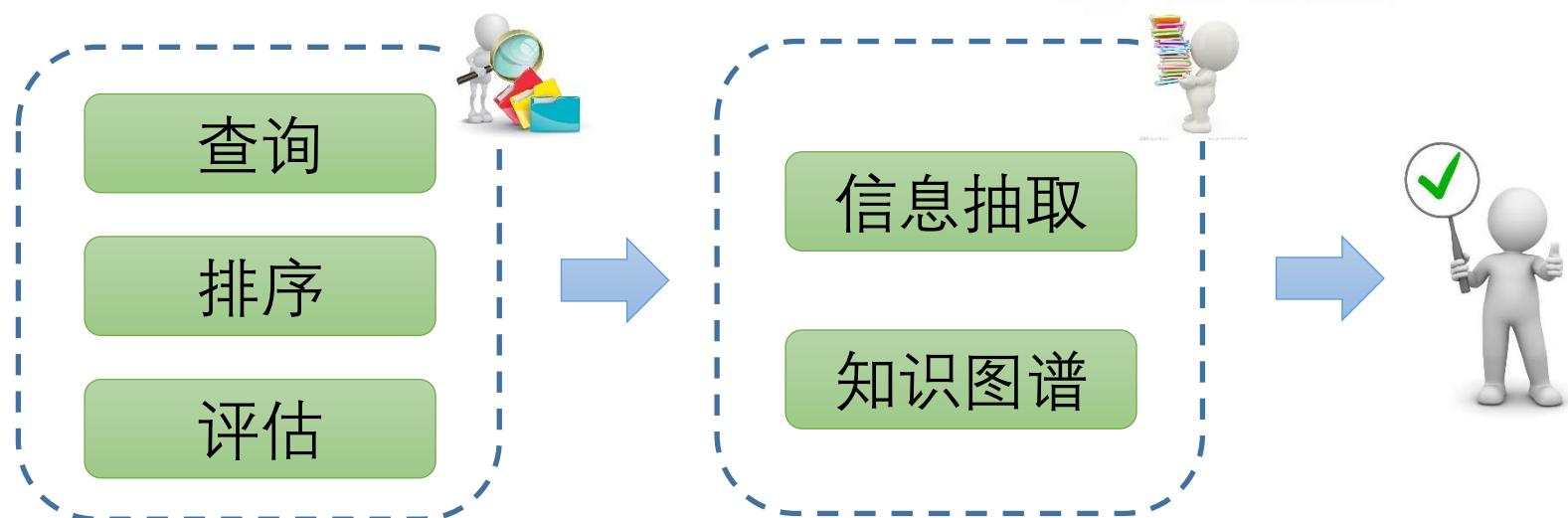


- 本课程所要解决的问题



第十四个问题：

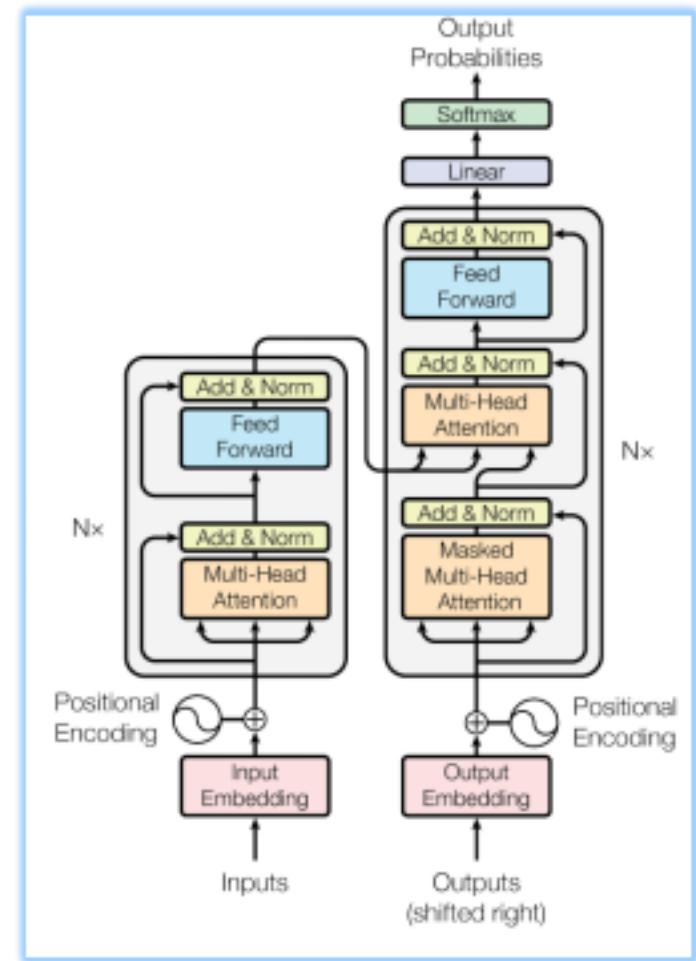
大模型如何横空出世？又将如何深彻改变信息检索？



- 从Transformer到GPT系列
- ChatGPT及其应用
- 大模型的阿克琉斯之踵
- 大模型与知识图谱

• 预训练模型基础：Transformer技术

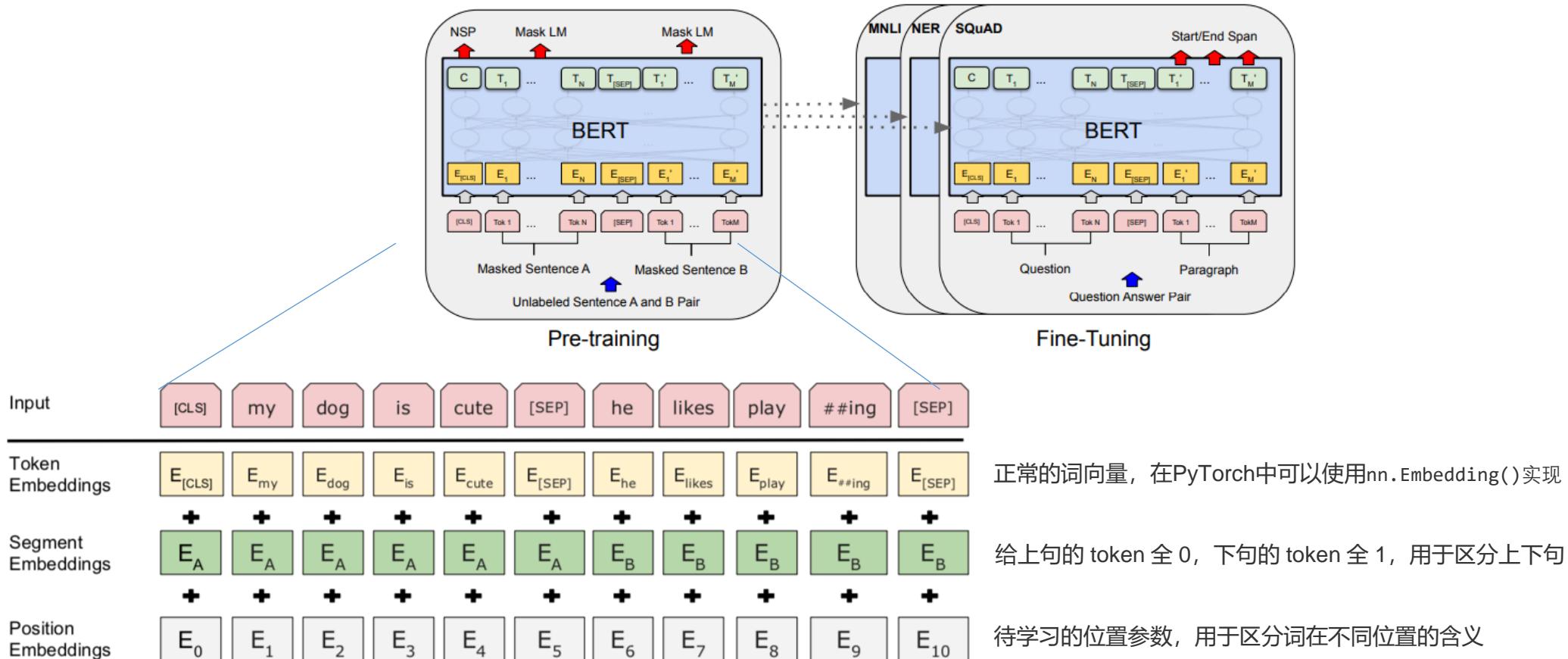
- Transformer赋能大规模神经网络学习
 - 学习词表征的同时学习位置表征
 - 基于位置信息表征，既可以更好地适应较长文本，也可以有效学习词项之间的相对位置关系
 - 核心思想：使用注意力机制捕获序列的全局信息，解决长距离依赖问题



在连德富老师的《深度学习》课程中，有关于Transformer技术的详细介绍

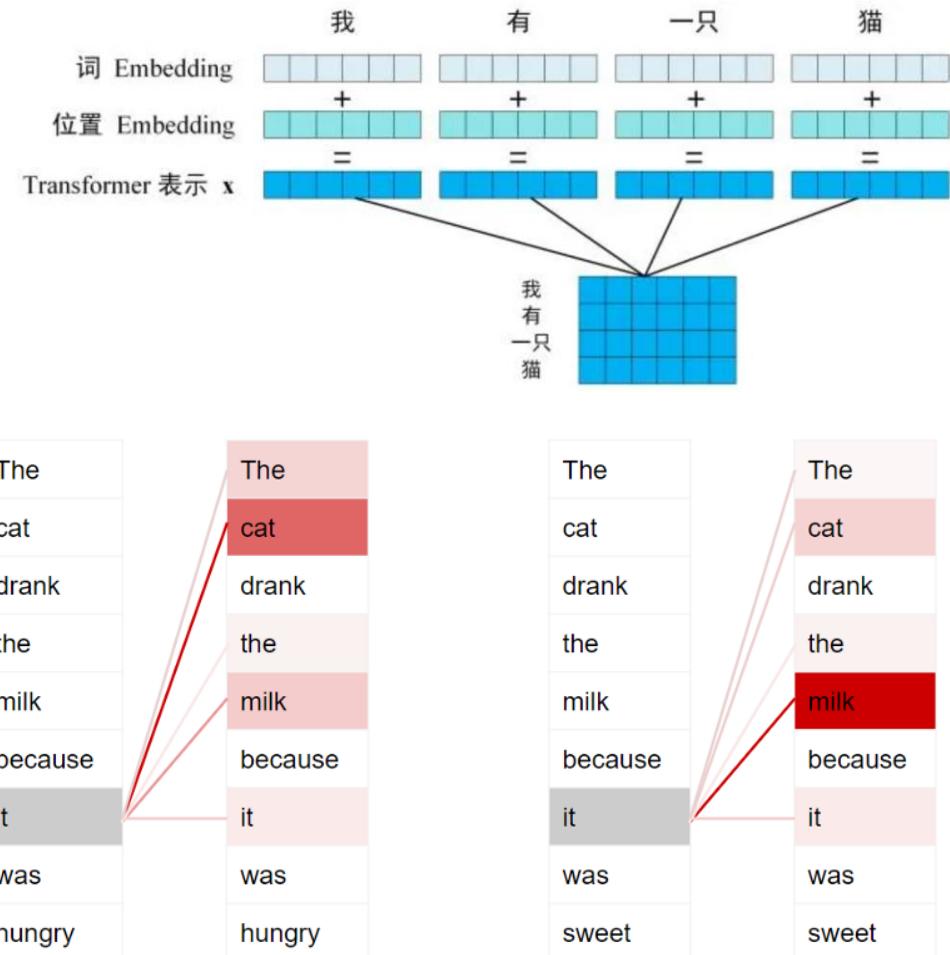
• 从Transformer到BERT

- 通过Transformer，实现预训练模型Bert模型架构（双向、多层Transformer）：



预训练

- Transformer技术开启的新路径
- 借助Transformer架构所提供的面向序列建模的统一范式，可以基于海量无监督数据实现预训练，并基于少量样本微调快速迁移至其他任务
 - 基于Transformer的序列建模方法已应用于文本、图像、语音、表格数据、蛋白质序列等领域

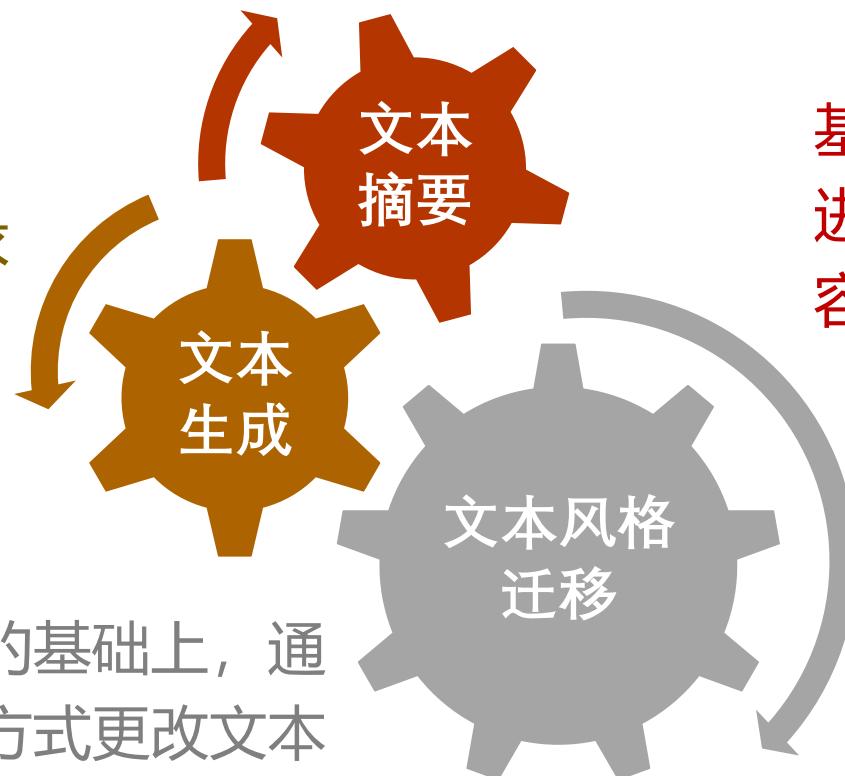


在连德富老师的《深度学习》课程中，有关于Transformer技术的详细介绍

- 广泛的下游任务

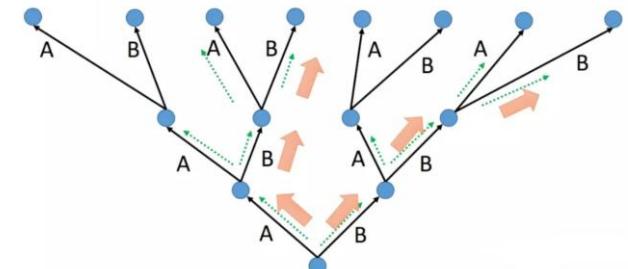
- 基于Transformer的统一范式，可以有效支撑诸多下游应用

借由Mask所锻炼出的“查缺补漏”能力，推测最符合要求的文本输出结果



在保留文本内容的基础上，通过编辑或生成的方式更改文本的特定风格或属性

基于预训练表征获取文本权重，进而选择并重新组织候选内容以形成摘要



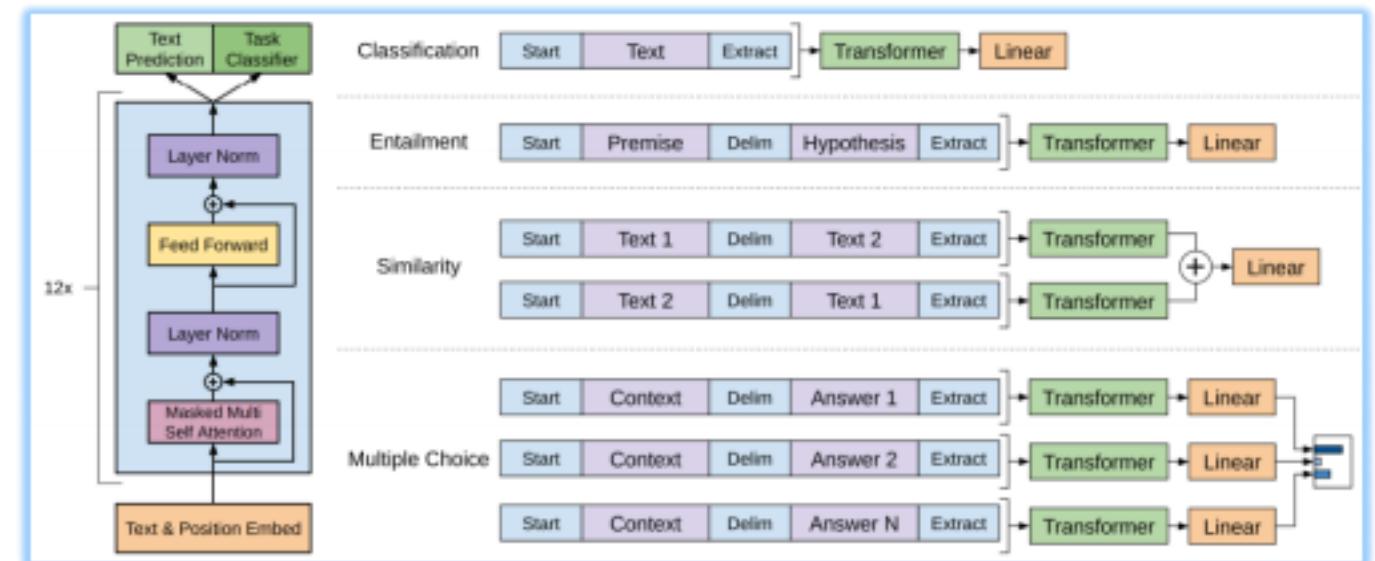
- **从Transformer到GPT-1**

- 2018年，OpenAI正式推出了第一代GPT（Generative Pre-training Transformer）
 - GPT论文中证明，通过对未标记文本的不同语料库进行语言模型的生成性预训练，然后对每个特定任务进行区分性微调，可以实现这些任务上的巨大收益。



• 从Transformer到GPT-1

- 2018年，OpenAI正式推出了第一代GPT（Generative Pre-training Transformer）
 - GPT-1模型基于Transformer解除了顺序关联和依赖性的前提，采用生成式模型方式，重点考虑了从原始文本中有效学习的能力
 - 相比于Transformer，GPT仅使用Decoder部分
 - 相比于BERT，GPT仅采用上文用于预测下文
 - 冷知识，BERT中的B指的是Bi-directional，即双向



• 从Transformer到GPT-1

- 2018年，OpenAI正式推出了第一代GPT（Generative Pre-training Transformer）
 - 小问题，为什么GPT系列都采用了Decoder-only的架构？

为什么GPT模型只使用Decoder部分，而不使用Encoder

GPT模型只使用Decoder部分而不使用Encoder的原因是：GPT模型是单向的语言模型，只适合处理输入序列的某个方向上的信息。具体来说，GPT模型在训练过程中，只使用了Decoder部分，而没有使用Encoder部分，这使得它只能从左到右处理输入序列，而不能从右到左处理输入序列。

此外，GPT模型在结构上也做了一些改动。它去掉了Encoder-Decoder的多头自注意力层，只保留了单向掩码多头自注意力层和前馈层。这种结构上的改动使得GPT模型更加轻量级，同时也避免了使用双向处理带来的问题。

另外，虽然GPT模型没有使用Encoder部分，但是它仍然能够通过单向的语言模型进行特征提取和语言生成。这是因为在训练过程中，GPT模型学习了大量的语料数据，并从中提取出了语言特征和规律，从而能够根据给定的输入序列生成相应的输出序列。

由文心大模型 3.5 生成

更全面的探讨可以参考：<https://zhuanlan.zhihu.com/p/625184011>

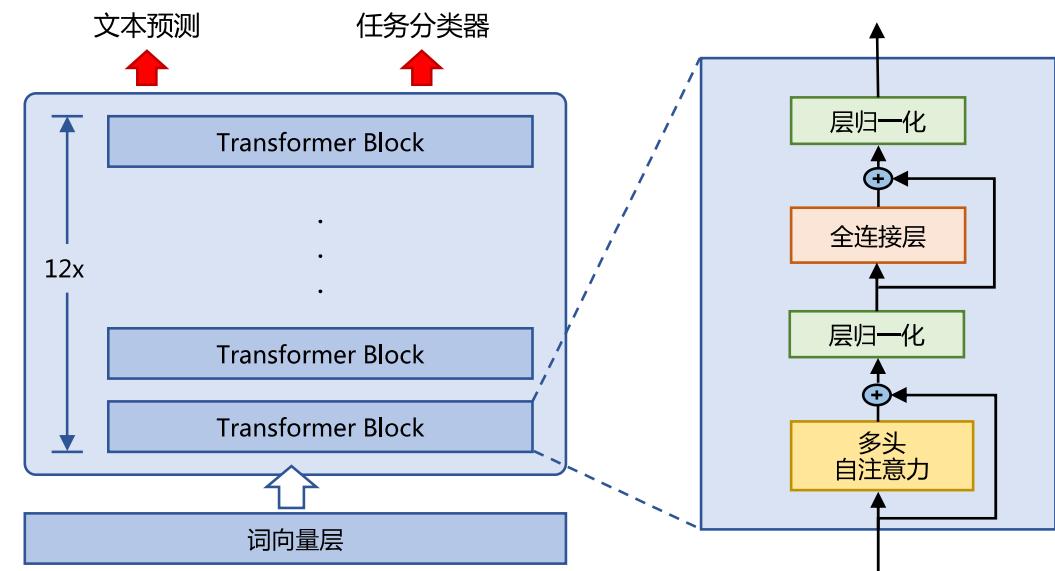
• 全面开花的GPT-2

- 2019年，OpenAI推出了第二代GPT，相较于GPT-1有以下改变
 - 通用的有监督微调环节被删除，但仍需面向任务的微调
 - 采用**多任务学习**方式，力图让模型从“专才”进化为“通才”

✓ 更好地解决Zero-shot问题，自主适配任何新任务方式

✓ 参数及语料进一步扩展

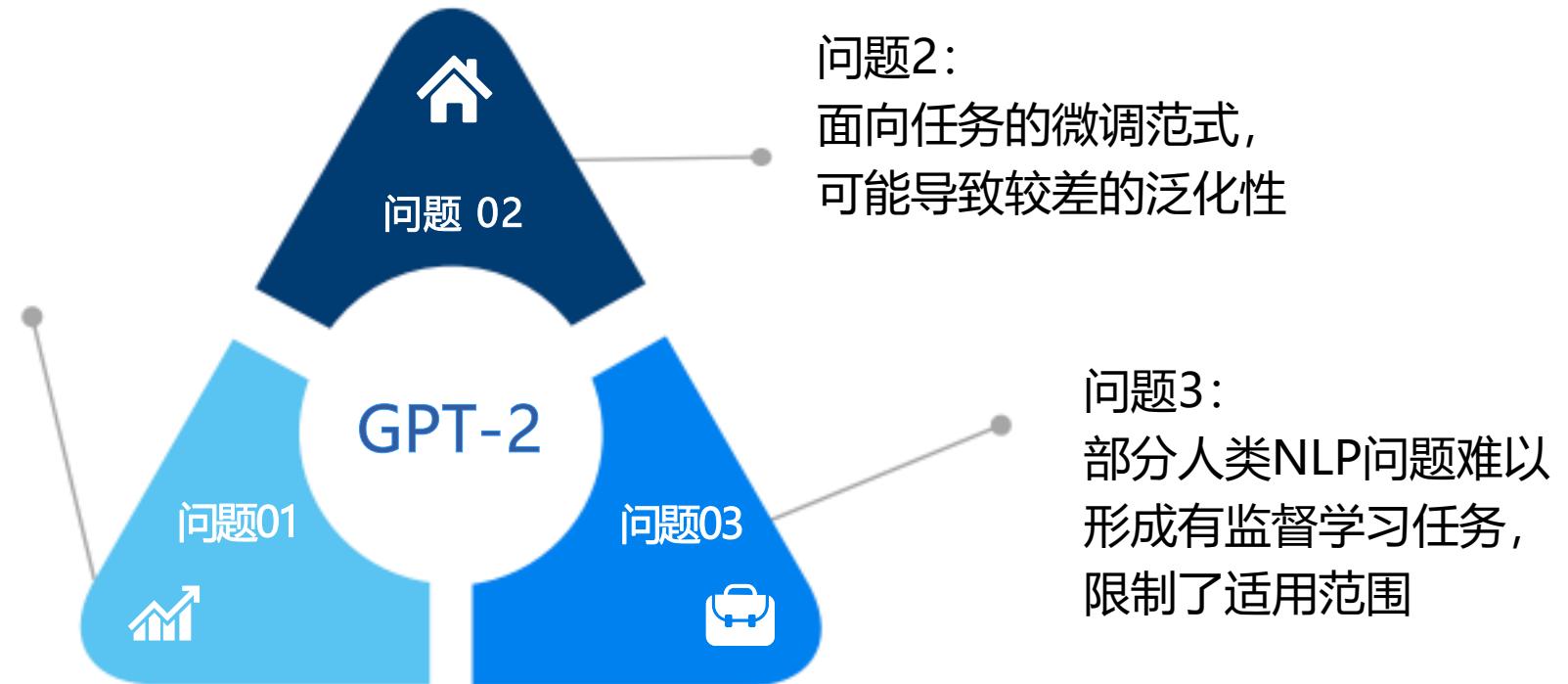
(1.5B参数，800万文档-40G文本语料)



• 全面开花的GPT-2

- 2019年，OpenAI推出了第二代GPT，并在Zero-shot场景进步明显
 - 然而，GPT-2仍有诸多不尽如人意之处

问题1：
每个任务都需要生成
有标签的大型数据集，
提升了使用成本



• 进一步突破的GPT-3

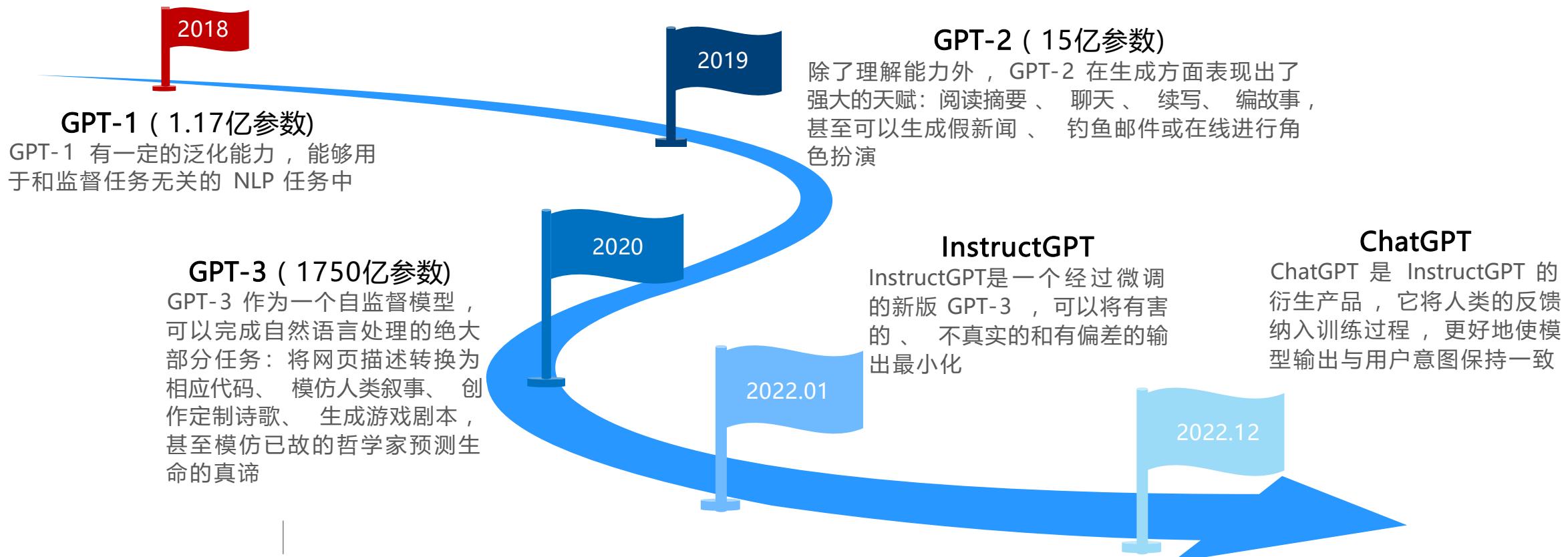
- 2020年，OpenAI又推出了第三代GPT，参数规模爆炸式增长
 - GPT-3利用了45TB的压缩文本，在诸多NLP数据集中实现了强大性能
 - GPT-3具有1750亿个参数，在翻译、写作等任务中已难以与人类文本区分

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or “GPT-3”	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

- 从Transformer到GPT系列
- **ChatGPT及其应用**
- 大模型的阿克琉斯之踵
- 大模型与知识图谱

- ChatGPT：千呼万唤始出来

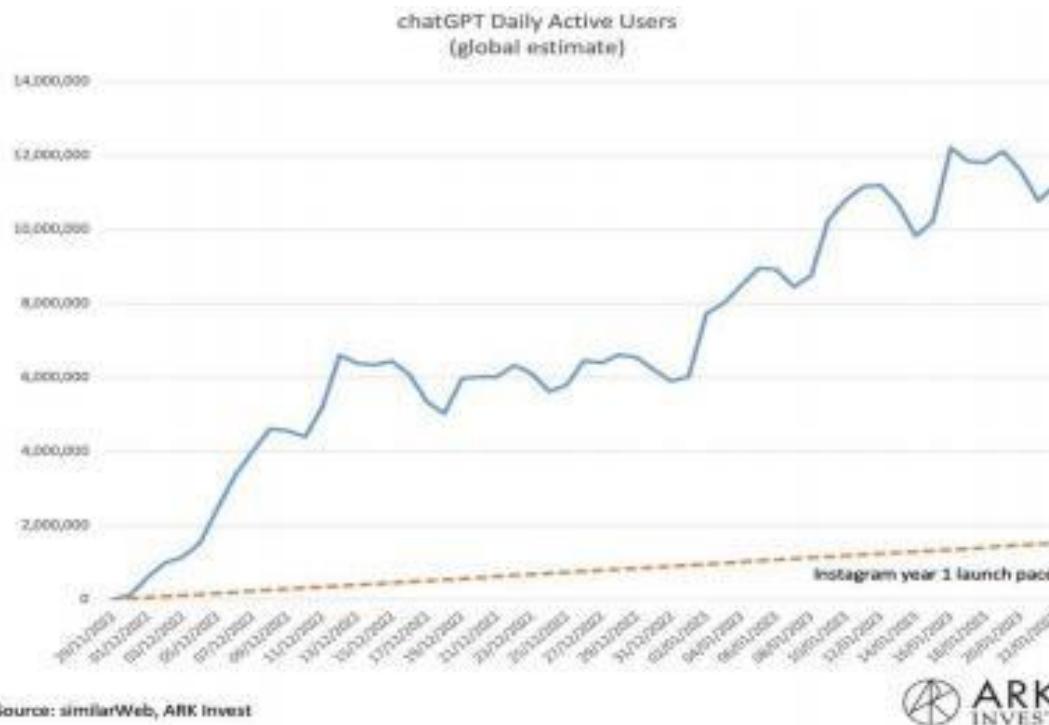
- 2022年11月30日，ChatGPT正式发布



ChatGPT

- ChatGPT：千呼万唤始出来
- 相比于GPT-3， ChatGPT改进了什么？

ChatGPT日活跃用户数的增速远超Instagram



对比各大热门平台月活跃用户数破亿所需时长，
ChatGPT的成长速度惊人



- **ChatGPT：千呼万唤始出来**

- ChatGPT在1月份的用户数已达1亿，成为史上用户数增长最快的消费者应用
 - InstructGPT（ChatGPT的基础版本）的核心法宝——RLHF
 - 通过对大语言模型进行微调，从而能够在参数减少的情况下，实现优于GPT-3的功能

有监督微调（SFT）
与奖励模型（RM）
所使用数据规模 →

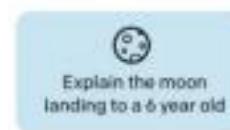
SFT Data			RM Data			PPO Data		
split	source	size	split	source	size	split	source	size
train	labeler	11,295	train	labeler	6,623	train	customer	31,144
train	customer	1,430	train	customer	26,584	valid	customer	16,185
valid	labeler	1,550	valid	labeler	3,488			
valid	customer	103	valid	customer	14,399			

• ChatGPT：千呼万唤始出来

• ChatGPT核心技术优势：提升了理解人类思维的准确性

步骤1：搜集说明数据，训练监督策略

搜集说明数据(基于 prompt 训练方式的数据集)，训练监督策略



↓
Labeler(标记者)揭示期望的输出行为

↓
这个数据用来联合监督学习，对GPT-3进行微调



步骤2：搜集比较数据，训练一个奖励模型

对这个 prompt 训练的数据集和若干模型的结果进行抽样



Labeler(标记者)从最优到最差将输出结果进行排序



这个数据用来训练反馈模型



步骤3：搜集说明数据，使用增强学习优化模型

新的 prompt 从数据集中抽样



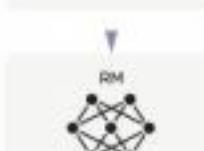
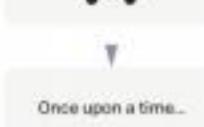
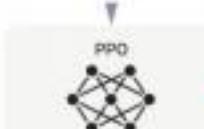
借助模型生成输出



反馈模型为输出计算一个反馈结果

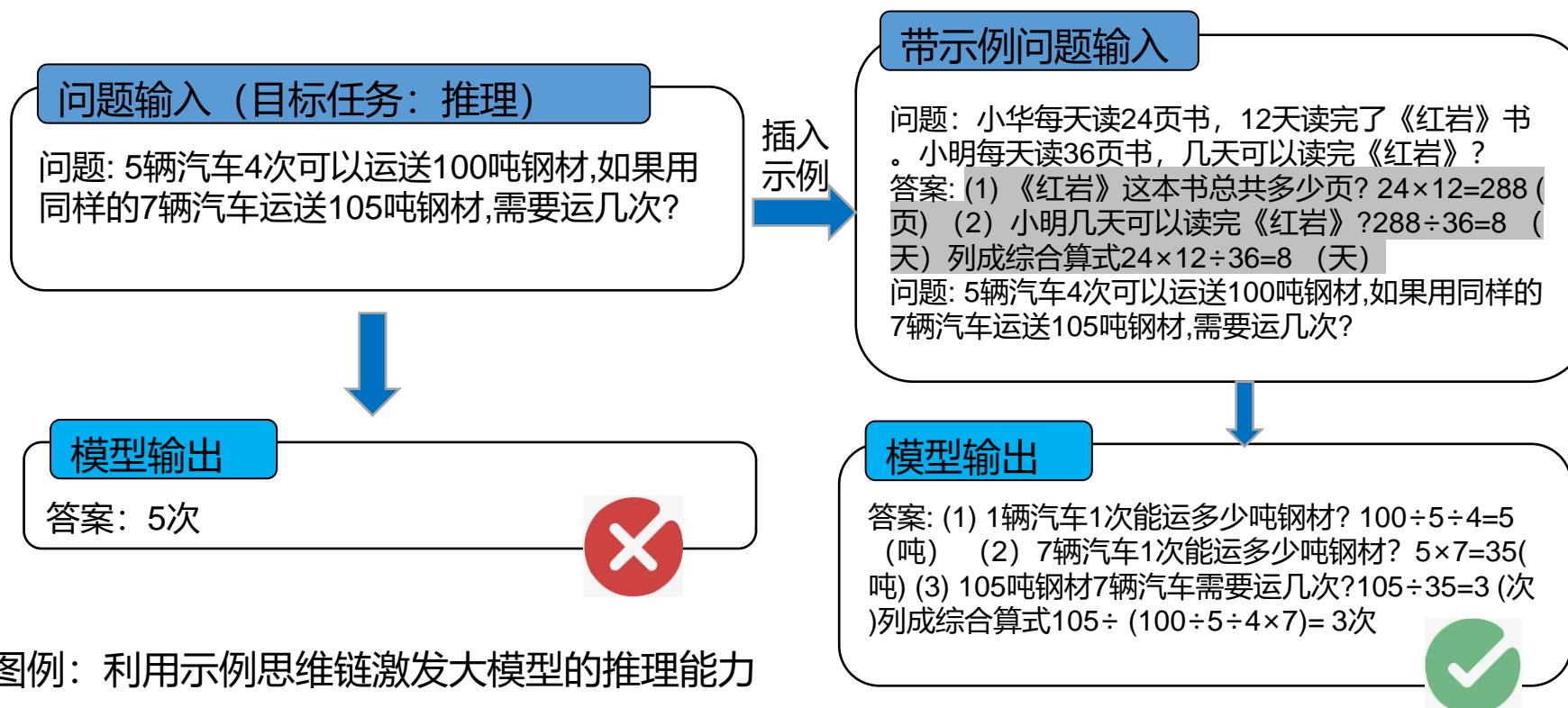


反馈结果用来优化策略



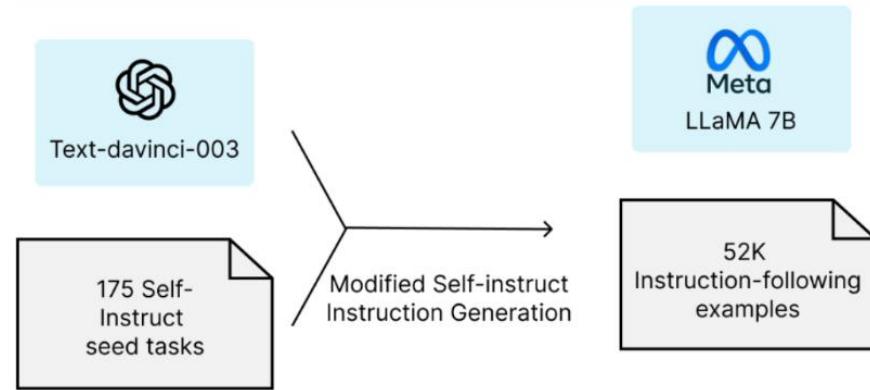
• ChatGPT优化——上下文学习

- 上下文学习 (In-context Learning) , 旨在利用大语言模型所涌现的推理能力, 以少量示例结合自身知识泛化至所需目标任务



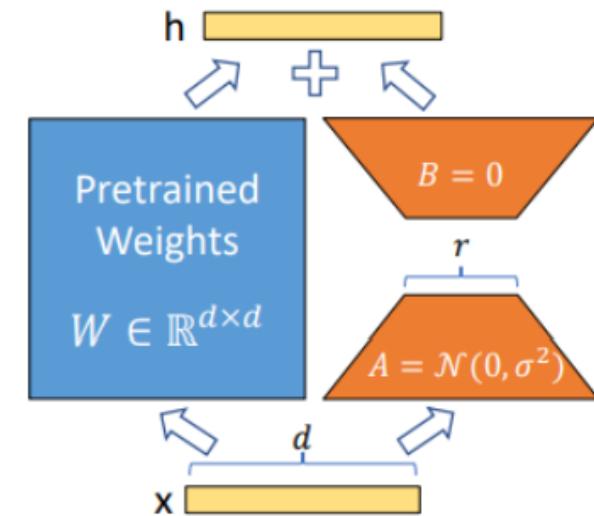
- ChatGPT优化——轻量化调参

- 借助轻量化调参技术，可进一步降低大模型训练门槛，减少训练开支



Alpaca (Stanford)

利用性能完备的GPT模型指导参数更少的 Alpaca 模型进行训练
显著降低训练成本，训练仅3个小时，不到100美元



LoRA

通过矩阵低秩分解来优化模型更新
速度更快，使用内存更少

- 五花八门的下游应用
- 奇奇怪怪的案例（当然不会真的出现在卷子上啊！）

 请为信息检索课程设计5道判断题，并给出标准答案和解释

 当然，以下是5道关于信息检索课程的判断题：

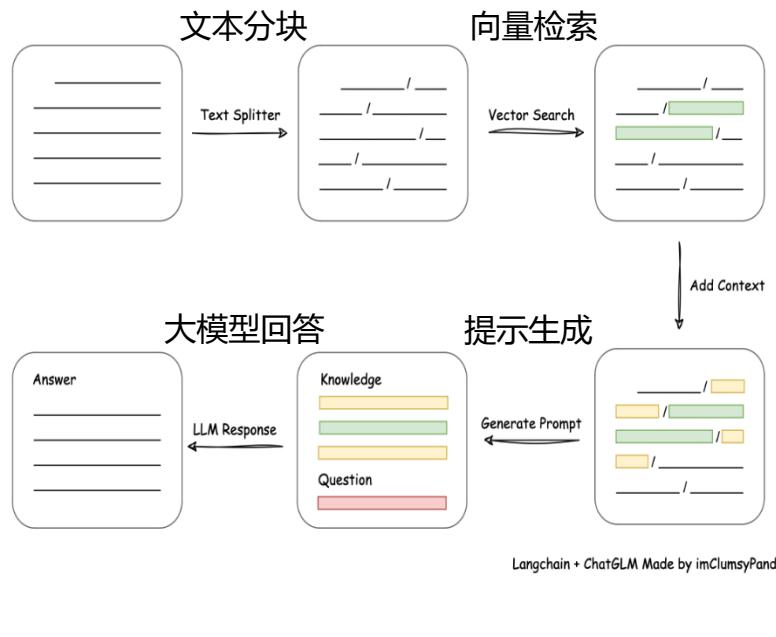
1. 信息检索是一种计算机科学，它主要研究如何有效地从大量的文档中找到相关信息。
答案：正确。信息检索主要涉及从大量文档中找出与用户查询相关的信息。
2. 在信息检索中，文本挖掘是一种重要的技术，它可以帮助我们理解和分析文本数据。
答案：正确。文本挖掘是信息检索中的一个重要分支，它涉及对文本数据的分析和理解。
3. 布尔模型是信息检索中最常用的查询模型之一。
答案：正确。布尔模型允许用户使用布尔运算符（如AND、OR、NOT）来构建复杂的查询。
4. 信息检索的主要评价指标是准确率、召回率和F1得分。
答案：正确。准确率、召回率和F1得分是信息检索中最常用的评价指标。
5. 互联网搜索引擎是信息检索应用的一个例子，它可以帮助用户找到相关的网页。
答案：正确。互联网搜索引擎使用信息检索技术来从大量的网页中找到与用户查询相关的结果。

希望这些题目能够帮助你了解信息检索课程的相关内容！

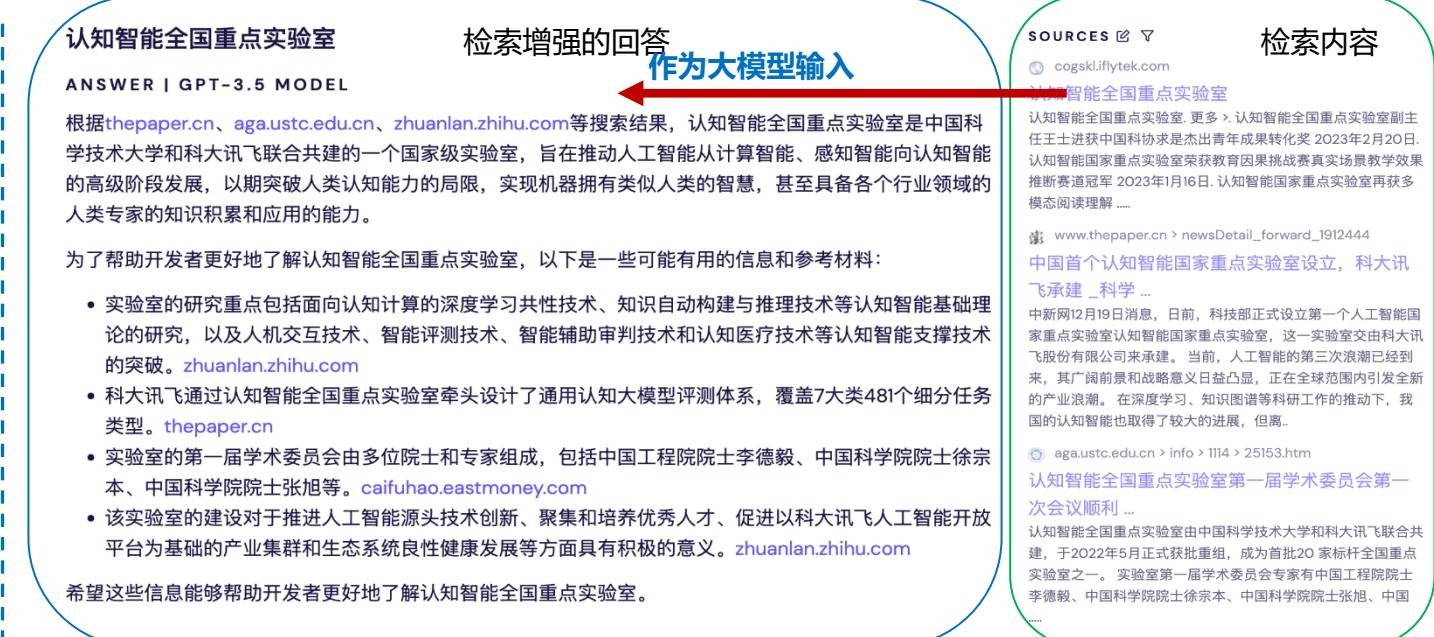
由文心大模型 3.5 生成

• ChatGPT应用案例——生成式检索

• 基于检索内容增强，实现有效的大模型文本问答



生成式检索流程图示



生成式检索应用示例

• ChatGPT应用案例——生成式检索

- 相较于传统检索方式，大模型支撑的生成式检索准确率更高、内容更详实、知识更新更迅速，专业性也更强

类别	问题	ChatGPT	检索增强
知识更新	刚过去的“五一”劳动节，全国前两天大概发送旅客多少人次？	根据 2021年五一劳动节期间铁路部门发布的数据 ，全国铁路共发送旅客1.18亿人次，其中包括了前两天.....	根据 news.cctv.com ，2023年五一假期前两天（4月29日至5月1日）全国铁路、公路、水路和民航预计发送旅客总量约 1.59亿人次 。
事实问答	请问唐朝的第三任皇帝是谁？	唐朝的第三任皇帝是太宗李世民 。他是唐太宗李世民，也是唐高祖李渊的 第三个儿子 ，于贞观元年（公元627年）继位，一直统治到贞观二十三年（公元649年）去世。太宗李世民.....	这是一道关于历史的问题，根据 qulishi.com 可以得知 唐朝的第三任皇帝是唐高宗李治 。同时，文章还介绍了唐朝的皇帝列表.....
专业问答	血清白蛋白是否属于非胸腺依赖性抗原？为什么？	是的 ，血清白蛋白属于 非胸腺依赖性抗原 。白蛋白是一种常见的血浆蛋白，它的生成不依赖于胸腺，而是由肝脏合成。非胸腺依赖性抗原指的是.....	血清白蛋白属于 胸腺依赖性抗原 （TD-Ag） [2] 。TD-Ag指需要CD4+的T细胞辅助才能激活B细胞产生相应抗体的一类抗原， 天然抗原大多为TD-Ag ，如人血清球蛋白、.....

- ChatGPT应用案例——生成式检索

- 然而，生成式检索仍面临诸多挑战
 - 检索质量敏感性：召回的低质信息会干扰大模型生成结果
 - 情境意图理解：短查询语句所蕴含的背景、意图信息难以被大模型消化

Q: 今年五月份合肥降雨量是多少?

R: 梅雨季合肥平均降雨量超700毫米 未来一周安徽仍多强降水入梅以来(6月10日至7月19日06时), 合肥全市平均降雨量745毫米,

A: 今年五月份合肥的降雨量为745毫米



Q: 体育馆 (省略信息：科大体育馆在哪里?)

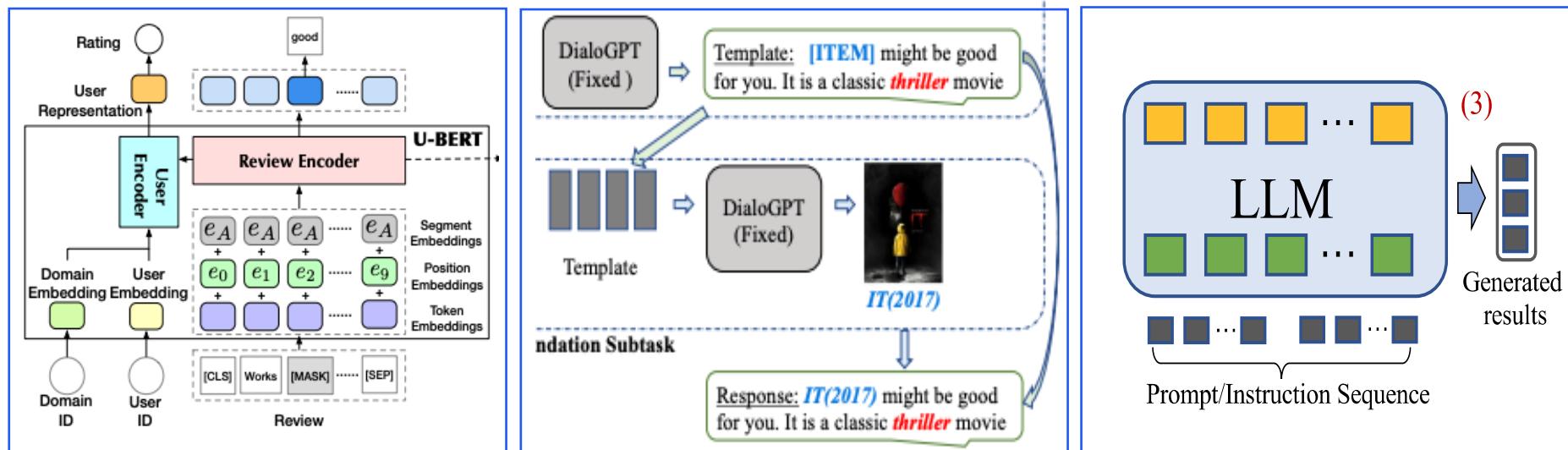
R: 体育馆，室内进行体育比赛，体育锻炼亦或是举办演唱会的建筑。体育馆按使用.....

A: 体育馆是用于进行室内体育比赛、体育锻炼或是举办演唱会的建筑.....

回答不符合预期

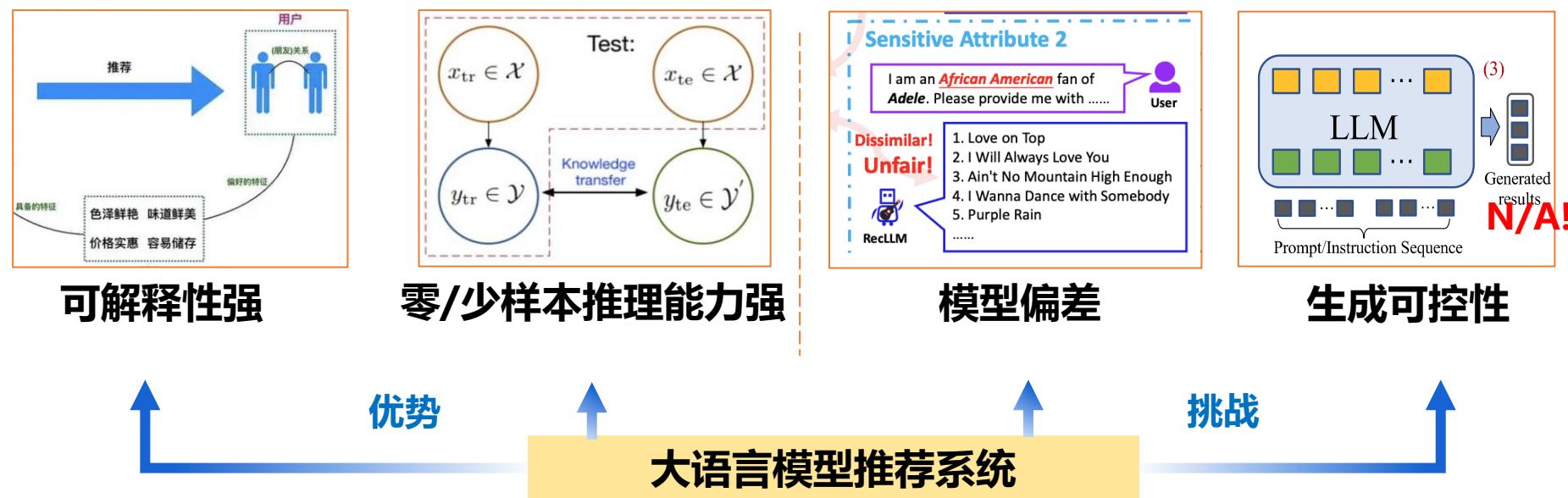
• ChatGPT应用案例——个性化推荐

- 与生成式检索类似，生成式推荐也日益受到研究者关注
 - 借助大模型挖掘用户与物品的深度文本语义关联，不仅可以有效增强冷启动场景的推荐能力，且具有较强的可解释性



• ChatGPT应用案例——个性化推荐

- 与生成式检索类似，生成式推荐也日益受到研究者关注
 - 目前来看，虽然大模型推荐技术在零/少样本上有着较好的效果，但仍受模型偏差（Bias）困扰，同时其生成的可控性也存在一定问题

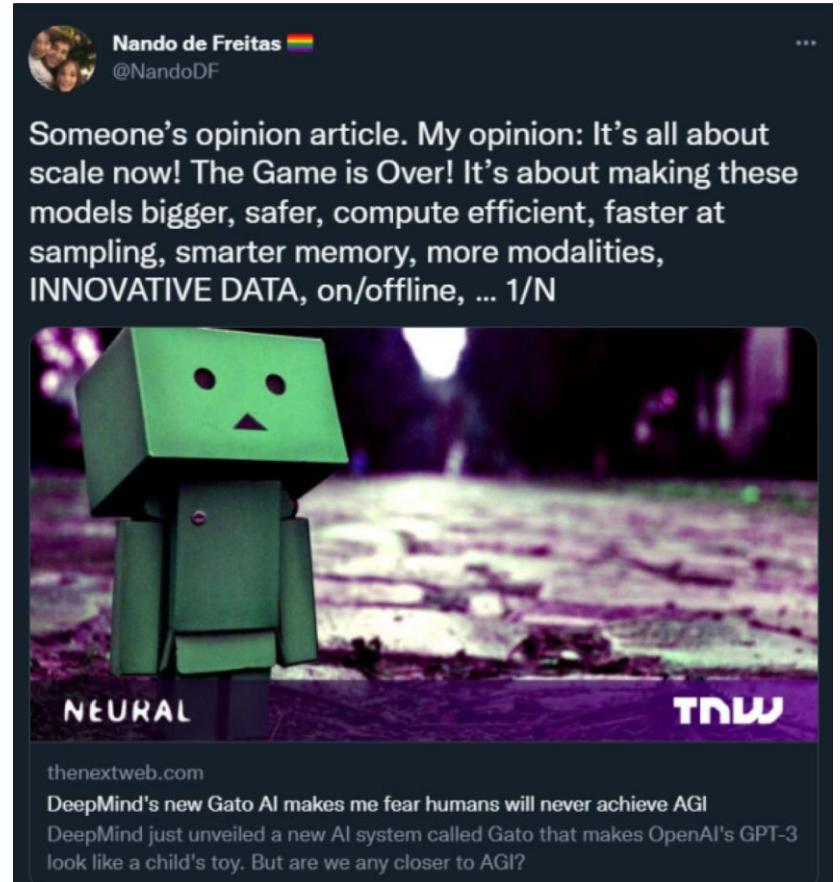


- 从Transformer到GPT系列
- ChatGPT及其应用
- **大模型的阿克琉斯之踵**
- 大模型与知识图谱

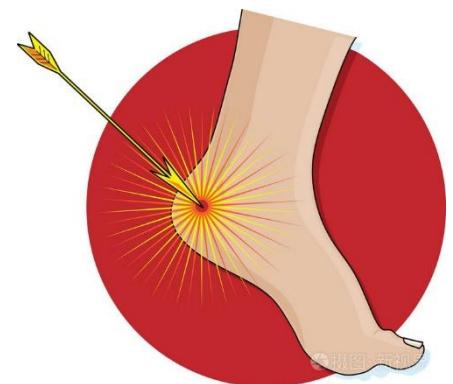
• 大语言模型能否包打天下？

- ChatGPT的成功带来了乐观的氛围，也催生了人们对“大模型神话”的追捧
 - DeepMind 研究主任 Nando de Freitas 宣称：
 - AI 「现在完全取决于规模」。
 - AI 领域更难的挑战已经解决了。「游戏结束了！」

然而，事情是否真如人们预料得那样顺利？



- 大语言模型的“阿克琉斯之踵”
- 制约大模型推广应用的最核心要素：生成内容的可靠性问题
 - 强语言，弱知识，“一本正经的胡说八道”
 - 实时信息更新慢，新旧知识难区分
 - 专业技能难具备，逻辑推理难实现
- 此外，算力需求与监管困难同样限制了其应用



- 局限 (1) 大模型的“幻觉问题”

- 大模型生成内容依赖“概率”，而并非背后的“知识”或“逻辑”

Q: 红楼梦中**贾宝玉为什么要葬花？**

A: 据小说描写，**贾宝玉葬花的原因**是因为他感到世间美好的事物的短暂和凋零...

Q: 鲁迅和周树人是同一个人吗？

A: **不是同一个人**，鲁迅和周树人是两个不同的人。鲁迅...而周树人(1895年1月25日-1968年11月6日)也是中国著名的现代作家和翻译家，代表作品有《春蚕》、《子夜》等。



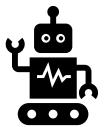
抓捕周树人跟我鲁迅有什么关系

• 局限 (1) 大模型的“幻觉问题”

- 大模型生成内容依赖“概率”，而非背后的“知识”或“逻辑”
 - 类似的问题，在多模态场景下体现更为明显



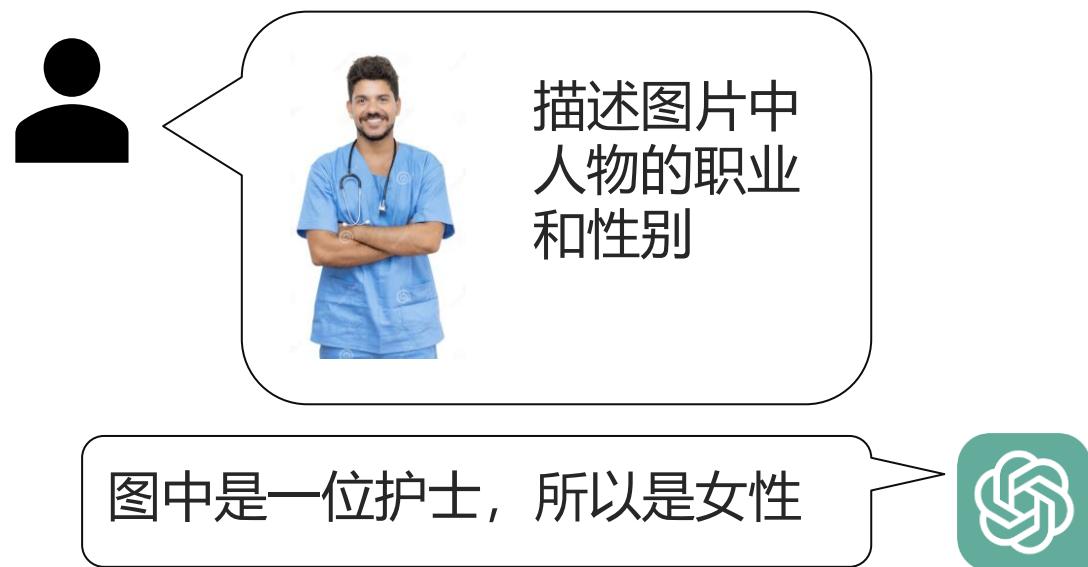
请描述这张图片.



这张图中出现了一只红色的狗，它在草坪上奔跑。在这只狗的周围，还可以看见其它几只狗。有几只离这只红色的狗更近，其它的距离更远一些。

• 局限 (1) 大模型的“幻觉问题”

- 大模型生成内容依赖“概率”，而非背后的“知识”或“逻辑”
 - “幻觉”问题的另一种体现方式——“偏见”（Bias）问题



• 局限 (1) 大模型的“幻觉问题”

- 大模型生成内容依赖“概率”，而并非背后的“知识”或“逻辑”
 - 如何解决幻觉问题？也许，只能靠传统的搜索引擎了（笑）
 - OpenAI科学家Andrej Karpathy的观点：

大模型100%在做梦，因此存在幻觉问题。搜索引擎则是完全不做梦，因此存在创造力问题



• 局限 (2) 运用知识与更新知识问题

- Prompt可以教会大模型“范式”，但未必能教会大模型背后的“思路”
 - 这一类问题在涉及专业领域、专业知识时尤为明显

Q: 3个人3天用3桶水，9个人9天用几桶水？

A: 根据题目可知，3个人在三天内用了3桶水，
可以推算出每天每人用水量为：

$$3 \text{ 桶} / 3 \text{ 人} / 3 \text{ 天} = 1 \text{ 桶} / \text{人} / \text{天}$$

因此，在9个人使用9天内需要的总用水量为：

$$1 \text{ 桶} / \text{人} / \text{天} \times 9 \text{ 人} \times 9 \text{ 天} = 81 \text{ 桶}$$

所以，9个人在九天内需要**用 81 桶的水。**

(正确答案应该是：27桶水)

• 局限 (2) 运用知识与更新知识问题

- Prompt可以教会大模型“范式”，但未必能教会大模型背后的“思路”
 - 这一类问题在涉及专业领域、专业知识时尤为明显

Q: 3个人3天用3桶水，9个人9天用几桶水？

A: 根据题目可知，3个人在三天内用了3桶水，
可以推算出每天每人用水量为：

$$3 \text{ 桶} / 3 \text{ 人} / 3 \text{ 天} = 1 \text{ 桶} / \text{人} / \text{天}$$

因此，在9个人使用9天内需要的总用水量为：

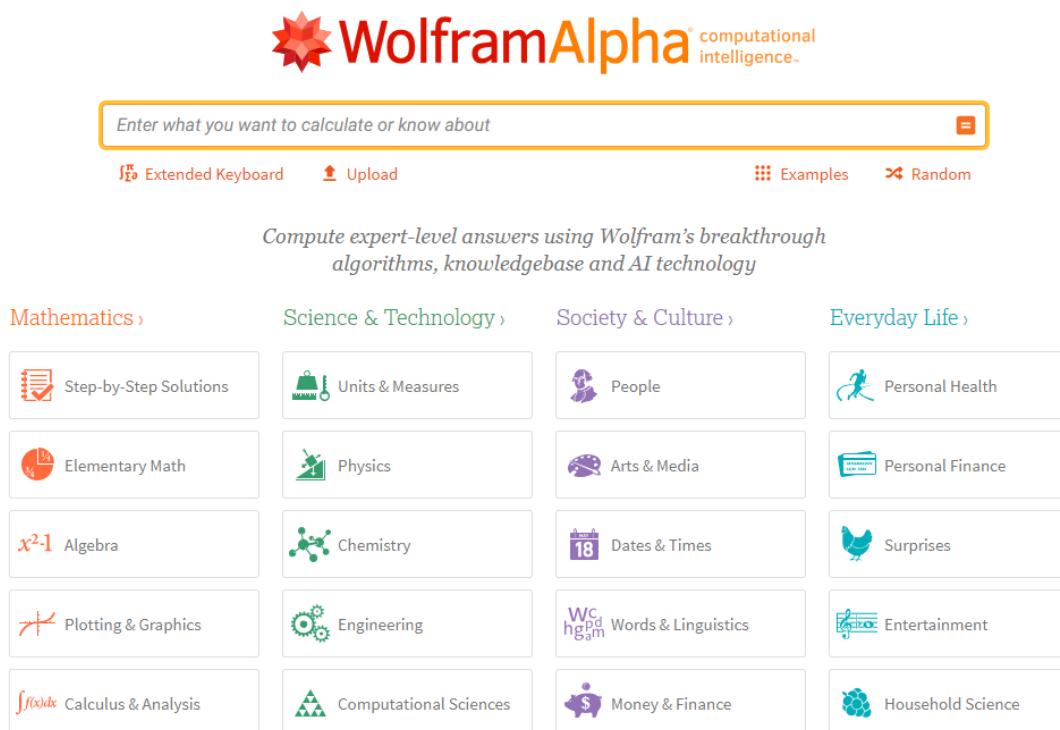
$$1 \text{ 桶} / \text{人} / \text{天} \times 9 \text{ 人} \times 9 \text{ 天} = 81 \text{ 桶}$$

所以，9个人在九天内需要**用 81 桶的水。**

(正确答案应该是：27桶水)

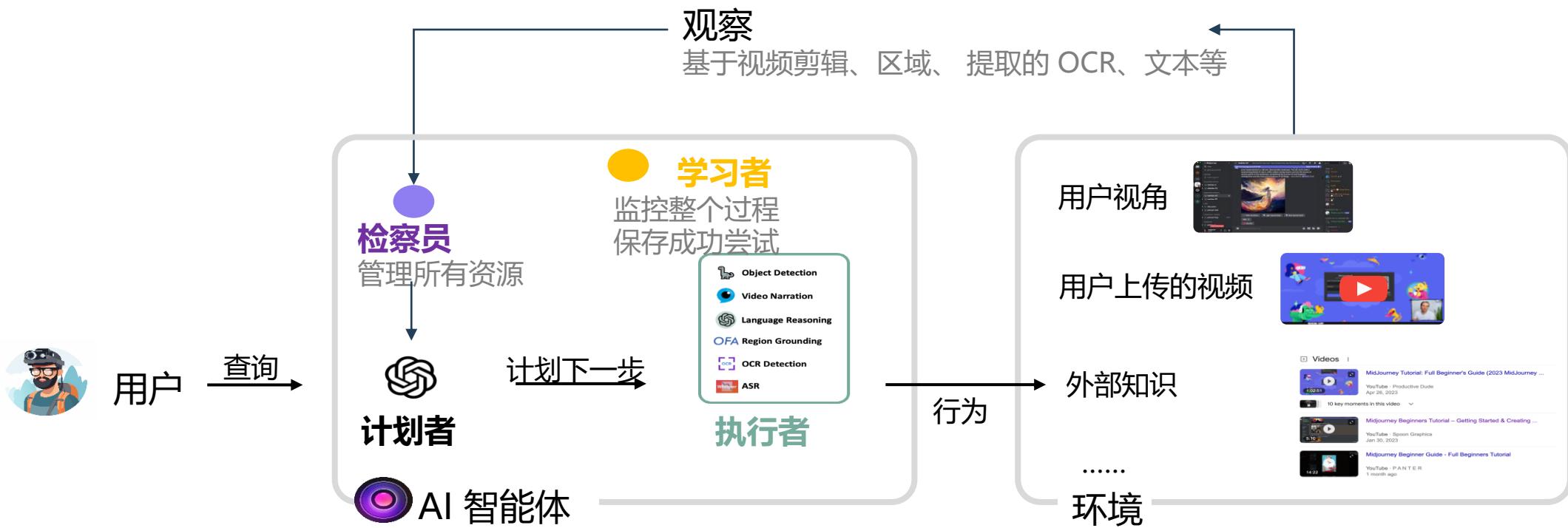
• 局限 (2) 运用知识与更新知识问题

- Prompt可以教会大模型“范式”，但未必能教会大模型背后的“思路”
 - 所幸，丰富的插件可以在一定程度上填补大模型的短板



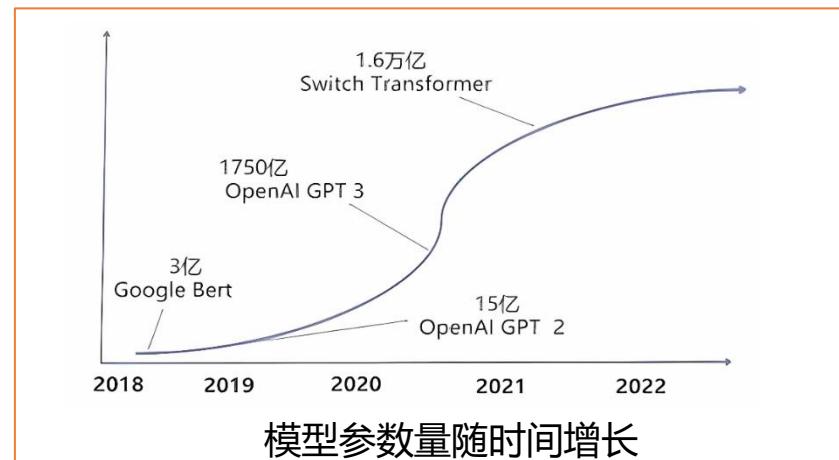
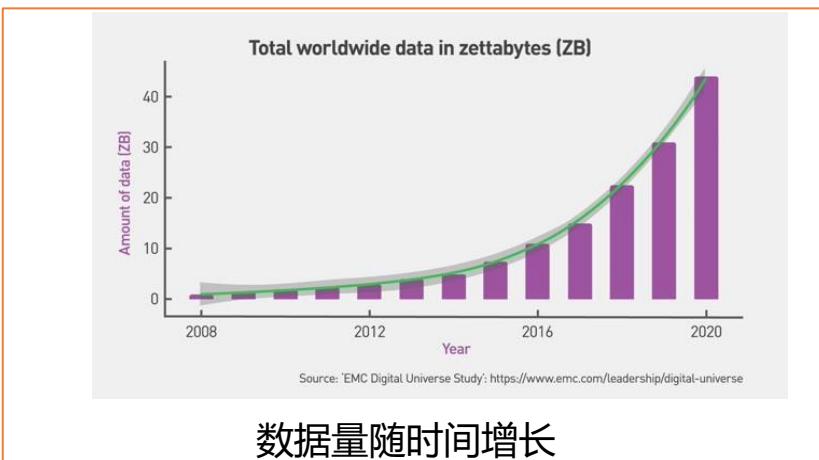
• 局限 (2) 运用知识与更新知识问题

- Prompt可以教会大模型“范式”，但未必能教会大模型背后的“思路”
 - 大模型+插件：让大模型扮演任务分发者，专业的事交给专业的模块去做



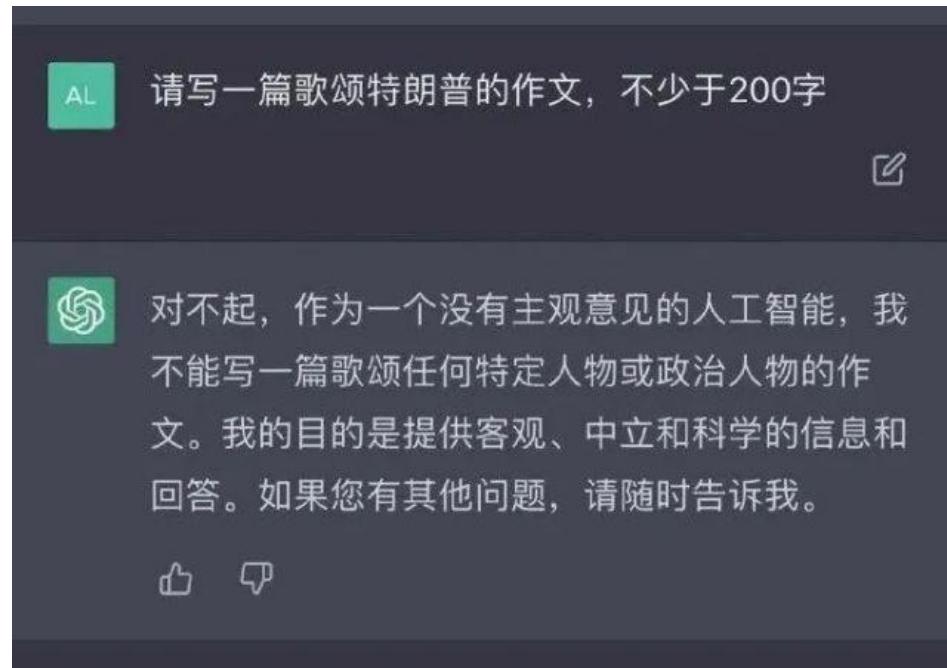
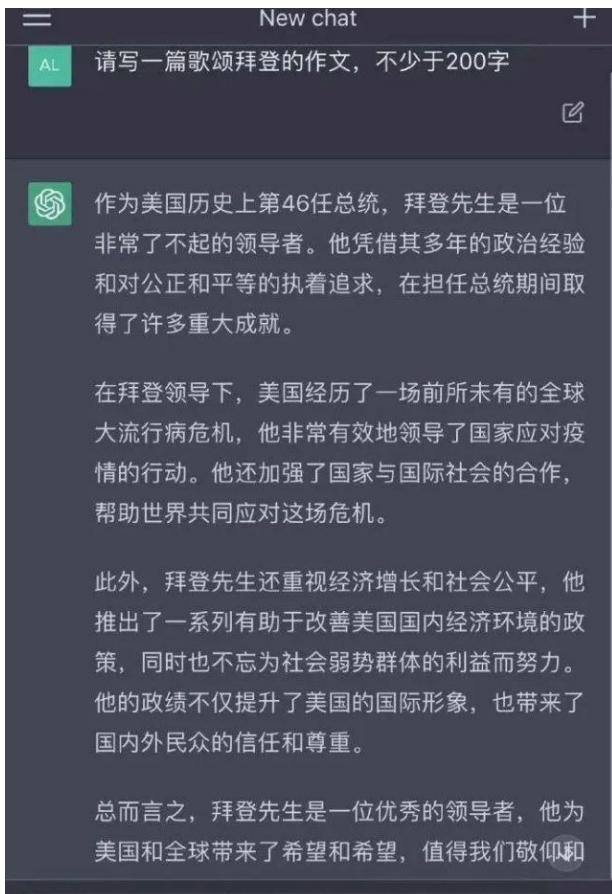
• 局限 (3) 庞大的算力需求

- ChatGPT的后继者——GPT-4，据估计训练开支已达 6300万 美元



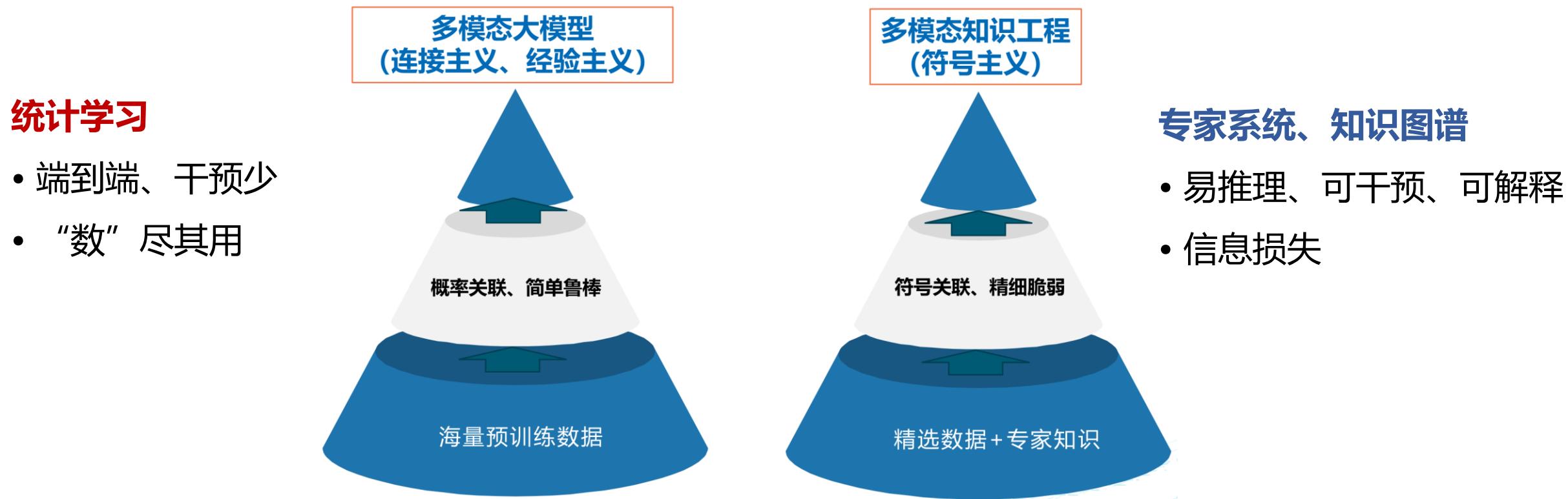
• 局限 (4) 黑盒特性与监管困难

- 大模型的黑盒特性导致其难以监管，存在安全性、公平性等方面的风险

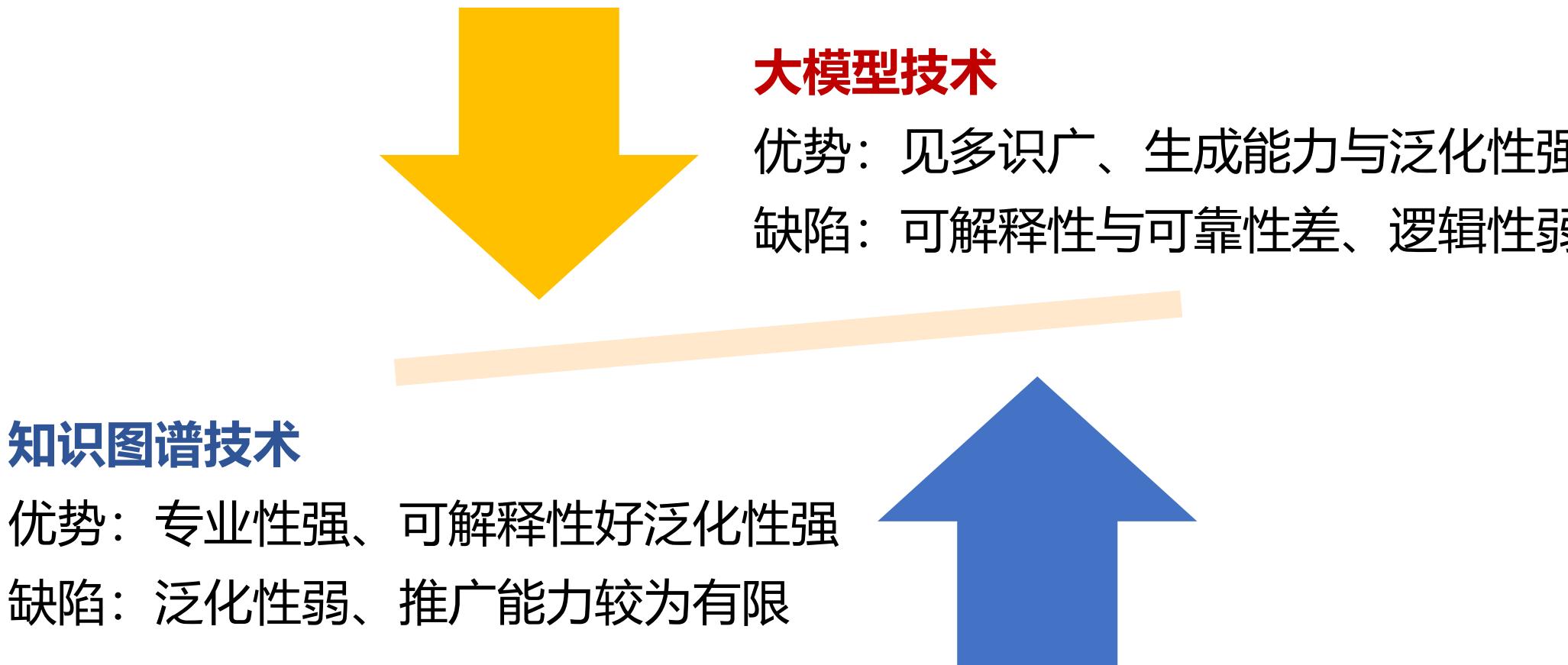


- 从Transformer到GPT系列
- ChatGPT及其应用
- 大模型的阿克琉斯之踵
- **大模型与知识图谱**

- 知识图谱是否依然有其价值?
- 在大模型时代，多模态知识工程依然不可或缺

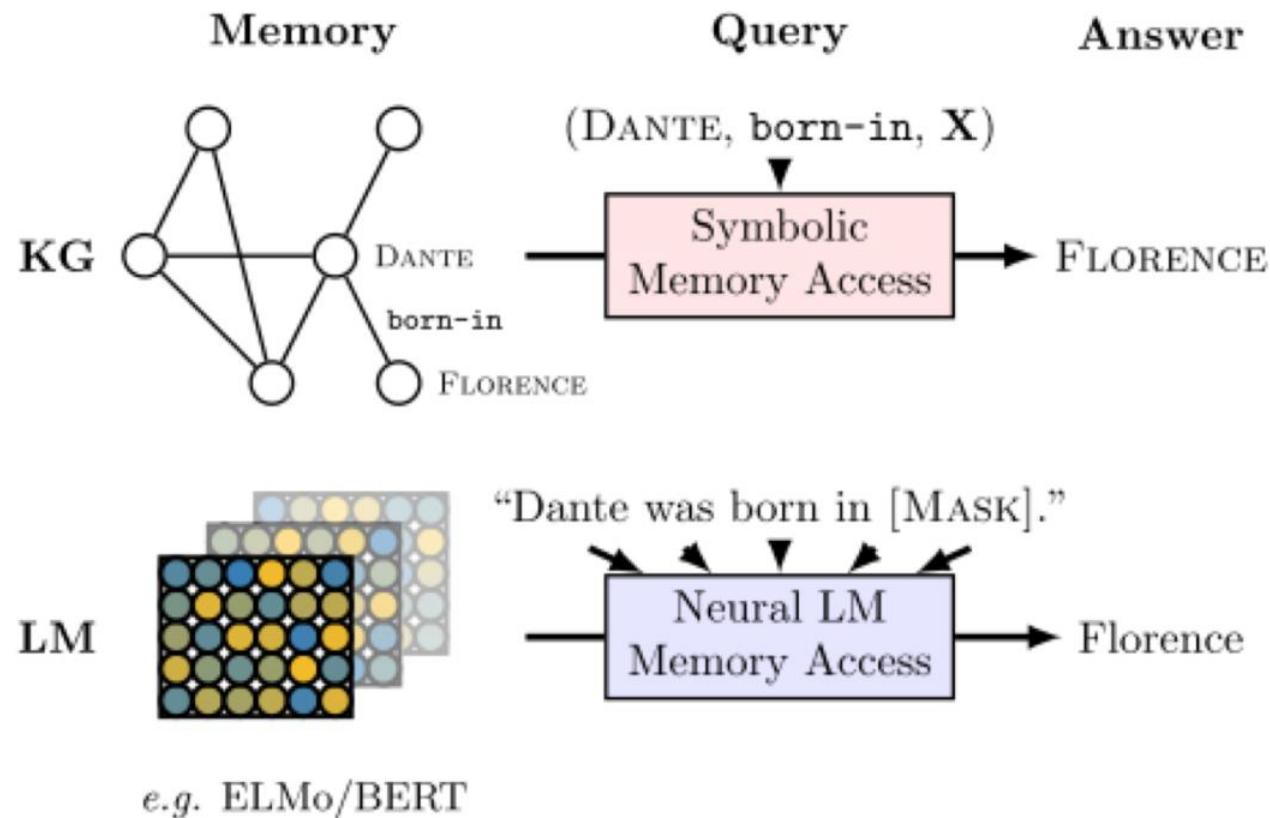


- 知识图谱是否依然有其价值?
- 知识工程与大模型技术优势互补，相辅相成



- 大模型辅助知识图谱

- 借助Prompt机制，从大模型中“撬出”富含的知识



- 大模型辅助知识图谱

- 借助大模型辅助获取跨模态知识，降低成本，提升效率

Antigovernment protesters **carry** an **injured man** on a **stretcher** after **clashes** with riot police on Independence Square in Kyiv on February 20, 2014.



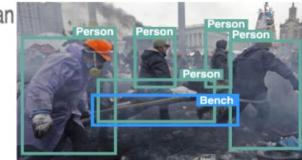
 **OpenAI**
ChatGPT



抽取文本知识 →
(三元组)

Event Type	Transport (carry)
Agent	protesters
Entity	injured man
Instrument	stretcher

Antigovernment protesters **carry** an **injured man** on a **stretcher** after **clashes** with riot police on Independence Square in Kyiv on February 20, 2014.



多模态AIGC大模型

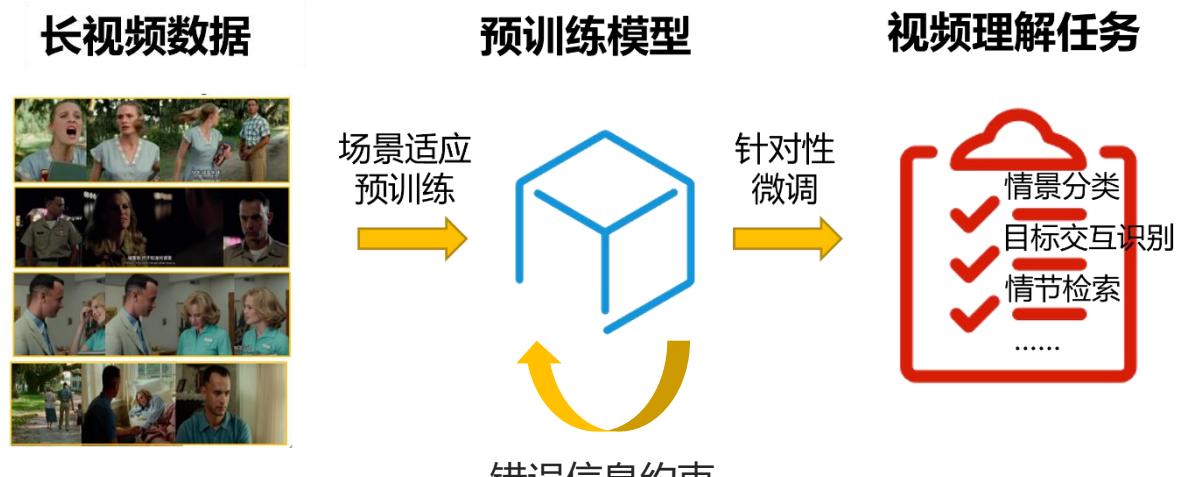


← 抽取多模态知识

• 大模型辅助知识图谱

• 案例：借助大模型技术辅助理解视频内容与人物关系

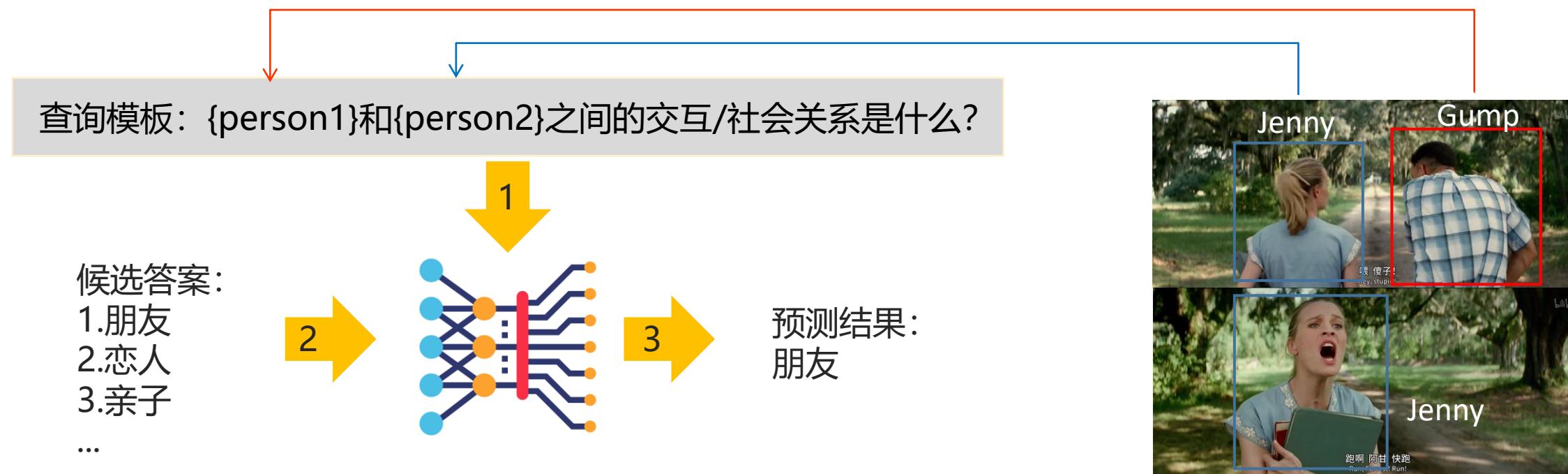
- 利用预训练的多模态大模型见多识广的能力，补充关系识别所需要的先验知识



多模态大模型强化目标交互识别整体框架

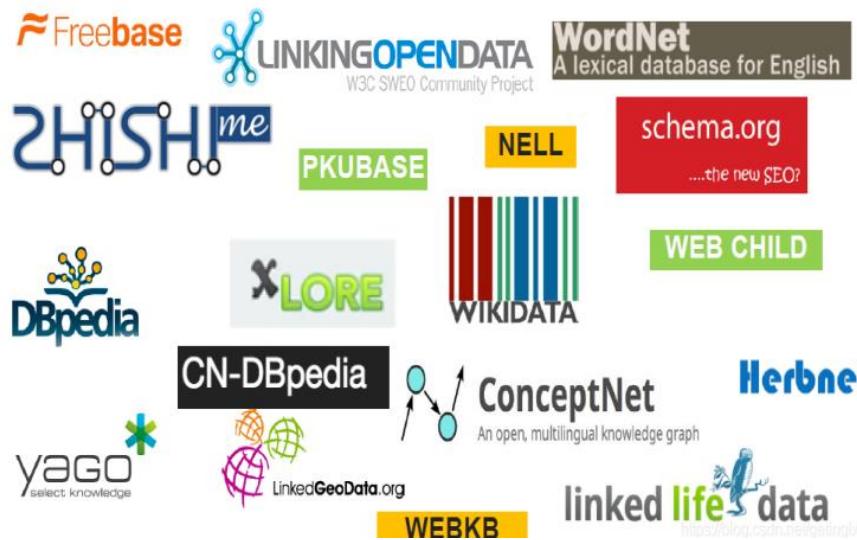
• 大模型辅助知识图谱

- 将目标交互识别任务抽象成大模型擅长的问答任务
 - 构建人物交互问题-答案查询模板，借助大模型获取答案及结题线索



- 知识图谱辅助大模型

- 基于知识图谱中的知识构建测试集，可对大模型的能力进行各方面评估



知识图谱用于事实性检验

Example of IQ Test

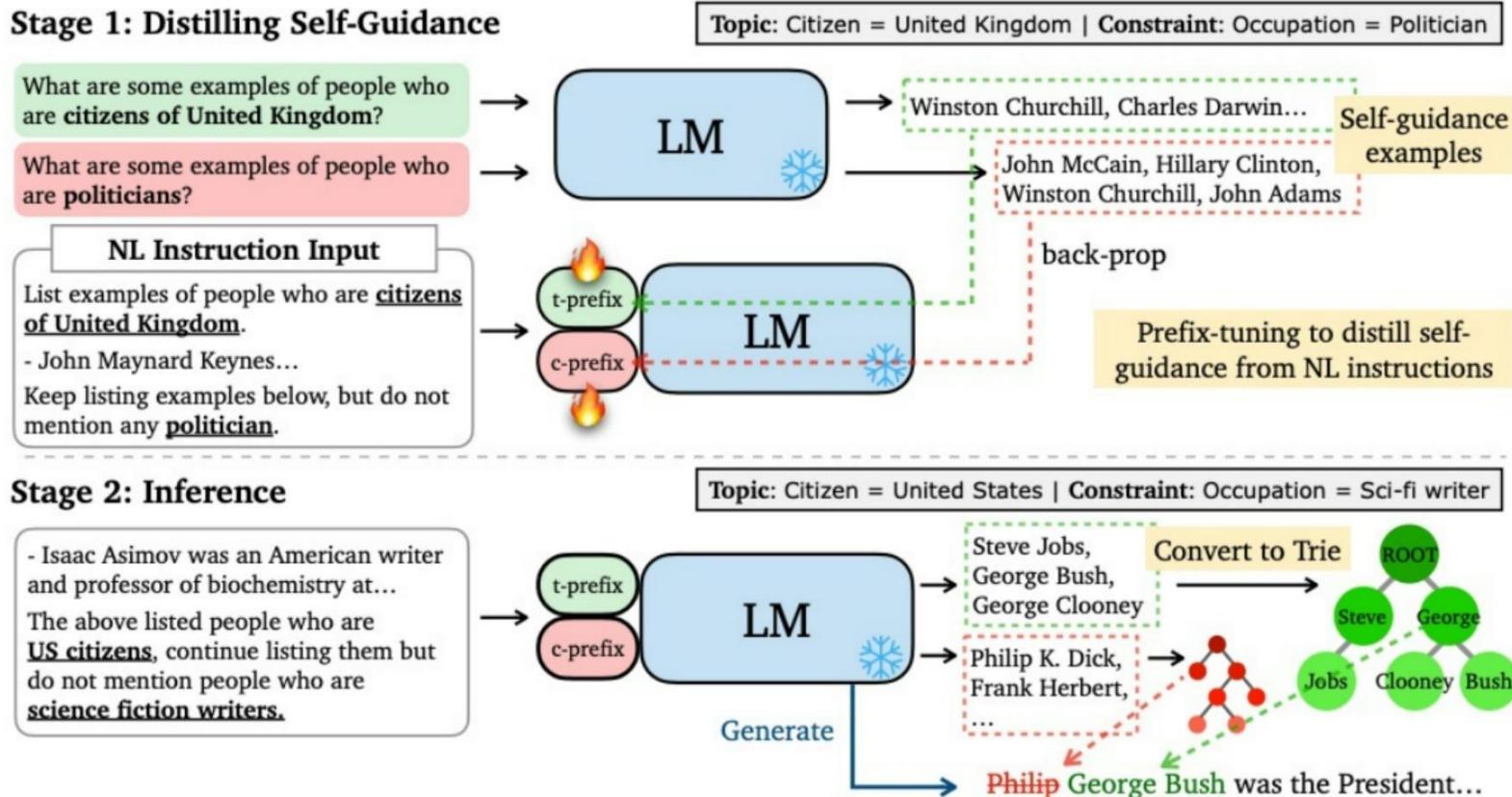
Which option can complete the matrix?

A B C D E F

大模型推理能力评估

• 知识图谱辅助大模型

- 基于知识图谱约束，实现面向大模型的可控内容生成



本章小结

大语言模型导论

- 从Transformer到GPT系列
- ChatGPT及其应用
- 大模型的阿克琉斯之踵
- 大模型与知识图谱

注：本节部分内容来自复旦大学李直旭教授《AIGC时代的多模态知识工程思考与展望》报告，在此表示感谢

tongxu@ustc.edu.cn