

# After ChatGPT



中国科学院 信息工程研究所  
INSTITUTE OF INFORMATION ENGINEERING, CAS



ASCII LAB

任昱冰

2023/5/26

# 目录

---

I. LLM最新进展

II. 后ChatGPT时代做科研的思路

III. 总结

## 2.9 Toolformer Meta AI LLM可以调用工具了!

- LLM一旦可以联网并使用工具，可能性会无限扩展
- 以**自监督**的方式微调LM，在不失模型的通用性下，**让模型学会自动调用API**。通过调用一系列工具（计算器、搜索引擎、翻译系统和日历等），Toolformer大大提高了GPT-J (EleutherAI 6.7B)的**零样本性能**，使其在一系列不同的下游任务上甚至优于更大的GPT-3模型
- 给定纯文本的数据集，首先将该数据集转换为通过API调用增强的数据集。包含三个步骤：  
采样API调用，执行API调用，过滤API调用
- 下游任务：LAMA、数学推理、问答、多语言问答

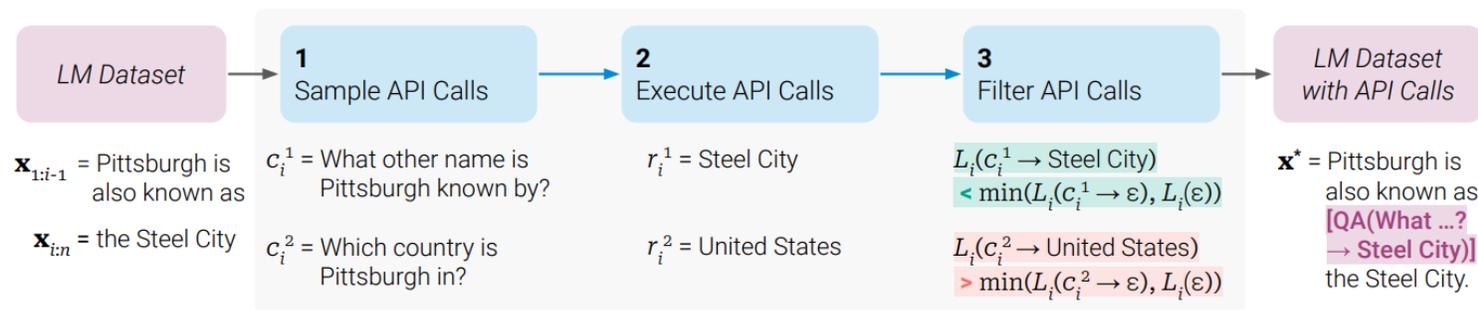


Figure 2: Key steps in our approach, illustrated for a *question answering* tool: Given an input text  $\mathbf{x}$ , we first sample a position  $i$  and corresponding API call candidates  $c_i^1, c_i^2, \dots, c_i^k$ . We then execute these API calls and filter out all calls which do not reduce the loss  $L_i$  over the next tokens. All remaining API calls are interleaved with the original text, resulting in a new text  $\mathbf{x}^*$ .

The New England Journal of Medicine is a registered trademark of [QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society] the MMS.

Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

The name derives from "la tortuga", the Spanish word for [MT("tortuga") → turtle] turtle.

The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

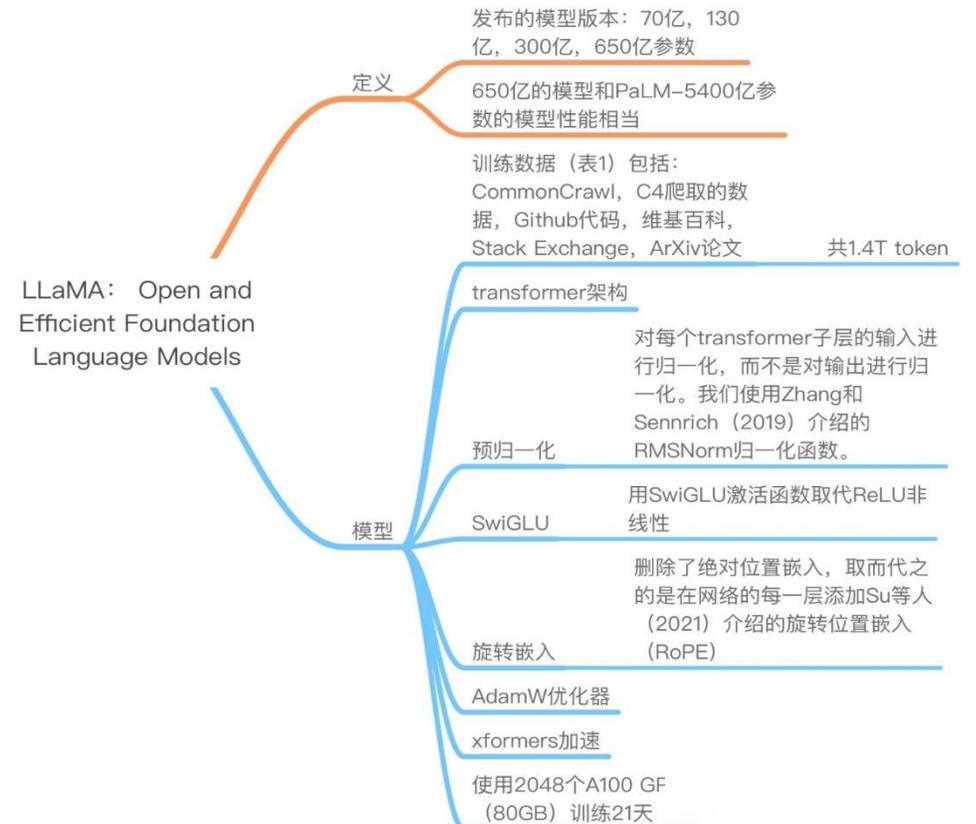
Figure 1: Exemplary predictions of Toolformer. The model autonomously decides to call different APIs (from top to bottom: a question answering system, a calculator, a machine translation system, and a Wikipedia search engine) to obtain information that is

## 2.24 LLaMA Meta AI

- 万亿级Token的预训练数据：英语CommonCrawl[67%], C4 [15%], Github[4.5%], 维基百科[4.5%], 古腾堡和Books3[4.5%], ArXiv[2.5%], Stack Exchange[2%]
- 模型主要创新（区别）：预归一化[GPT3], SwiGLU激活函数[PaLM], 旋转嵌入[GPTNeo]
- 训练方式：follow GPT-3、Palm

		BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA
GPT-3	175B	60.5	81.0	-	78.9	70.2	68.8	51.4	57.6
Gopher	280B	79.3	81.8	50.6	79.2	70.1	-	-	-
Chinchilla	70B	83.7	81.8	51.3	80.8	74.9	-	-	-
PaLM	62B	84.8	80.5	-	79.7	77.0	75.2	52.5	50.4
PaLM-cont	62B	83.9	81.4	-	80.6	77.0	-	-	-
PaLM	540B	<b>88.0</b>	82.3	-	83.4	<b>81.1</b>	76.6	53.0	53.4
	7B	76.5	79.8	48.9	76.1	70.1	72.8	47.6	57.2
	13B	78.1	80.1	50.4	79.2	73.0	74.8	52.7	56.4
LLaMA	33B	83.1	82.3	50.4	82.8	76.0	<b>80.0</b>	<b>57.8</b>	58.6
	65B	85.3	<b>82.8</b>	<b>52.3</b>	<b>84.2</b>	77.0	78.9	56.0	<b>60.2</b>

Table 3: Zero-shot performance on Common Sense Reasoning tasks.



## 3.8 Visual ChatGPT 微软 图文并用

- 可以根据指示各种魔改生成图片
- Visual ChatGPT将 ChatGPT 作为逻辑处理中心，集成视觉基础模型，使用户能与 ChatGPT 以图文的形式交互，还能提供复杂的视觉指令，让多个模型协同工作

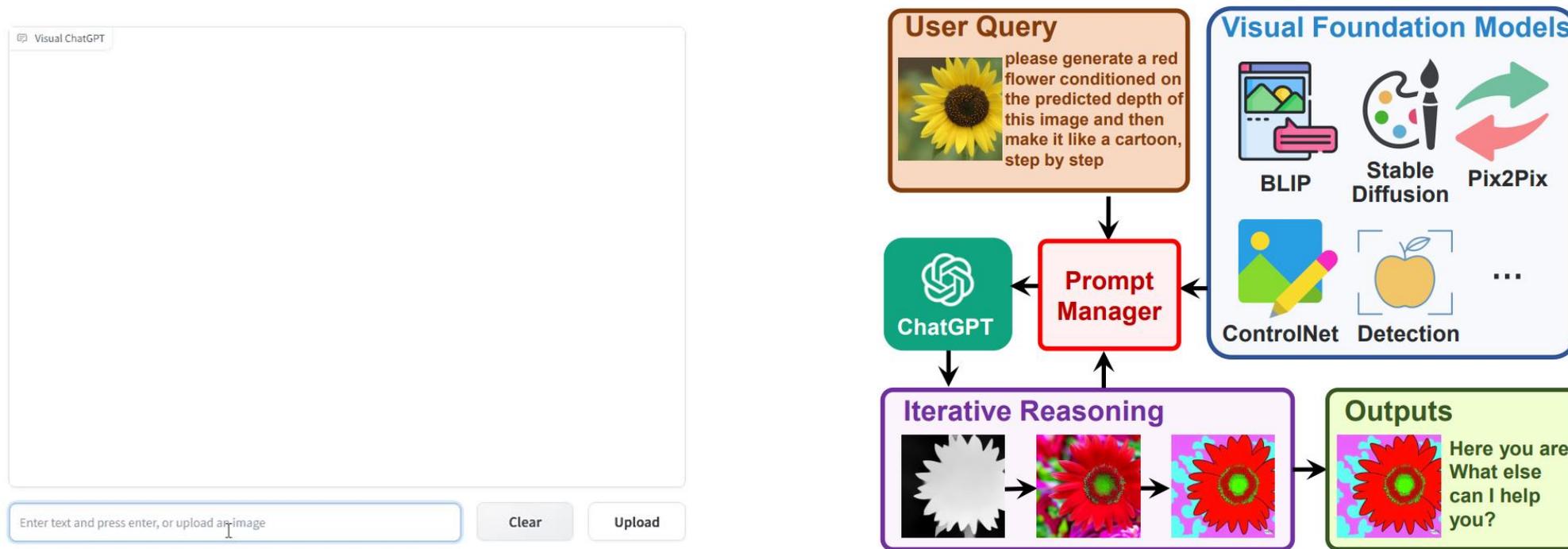


Figure 1. Architecture of Visual ChatGPT.

# 3.9 GigaGAN 10亿参数



Input artwork from AdobeStock (128px)



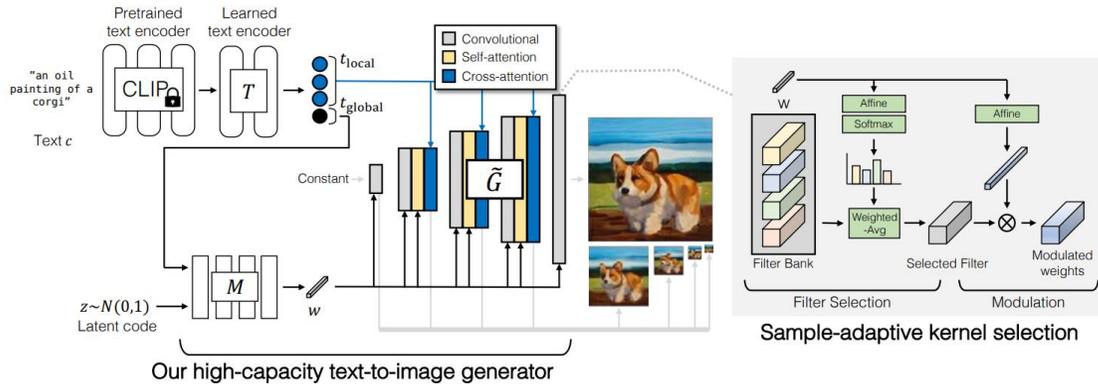
Real-ESRGAN (1024px, 0.06s)



SD Upscaler (1024px, 7.75s)



GigaGAN Upsampler (1024px, 0.13s)



"A modern mansion .." → "A victorian mansion .."

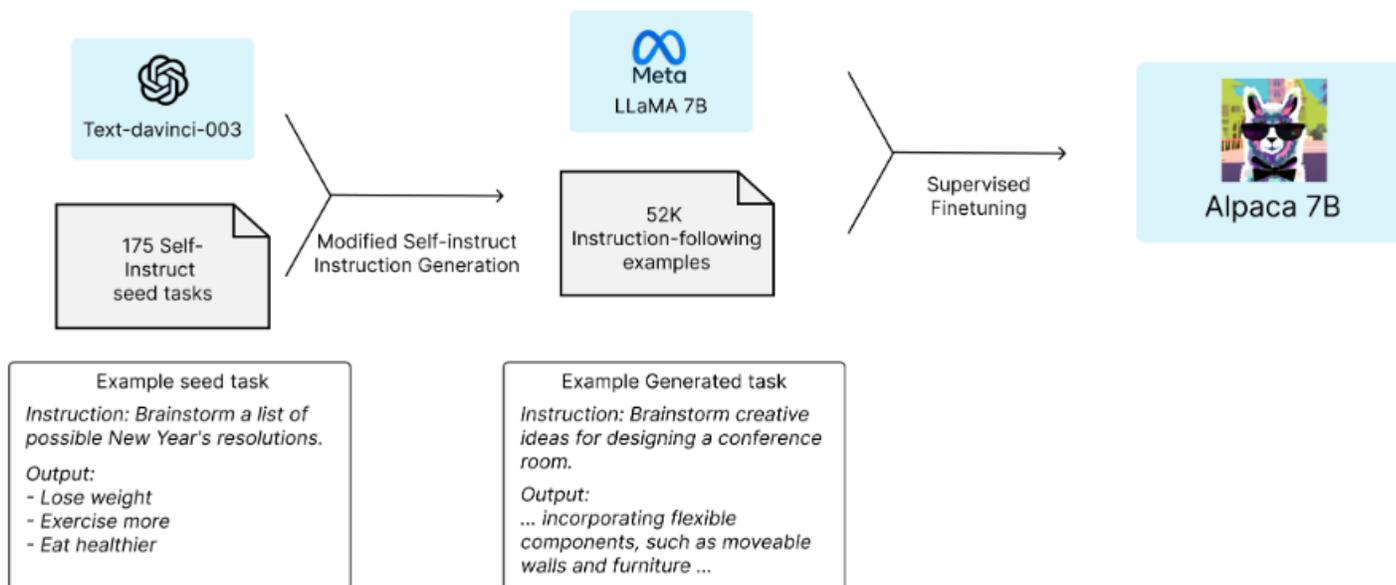
".. in a sunny day"

".. in sunset"



## 3.13 斯坦福Alpaca 7B匹敌text-davinci-003 175B

- 调用 OpenAI 的 text-davinci-003 模型产生52K个有价值的intruction-following examples , 以此微调 Meta AI 的LLaMA 7B , 从而生成了 Alpaca 7B
- 微调成本奇低, 不到600美元: 服务器租赁约100美元+调用OpenAI API约500美元
- 训练时长: 在8个80GB A100上训练3个小时



### Known limitations

Alpaca also exhibits several common deficiencies of language models, including hallucination, toxicity, and stereotypes. Hallucination in particular seems to be a common failure mode for Alpaca, even compared to text-davinci-003.

For example, in the following figure, Alpaca wrongly says that the Capital of Tanzania is Dar es Salaam, which is the largest city in Tanzania. (It was the capital until 1974, when it was replaced by Dodoma.)

Enter your instruction and press enter

What is the capital of Tanzania?

Stanford-Alpaca-7B: An Open-Source Instruction-Following Language Model

Dar es Salaam.

Furthermore, Alpaca can be used to generate well-written outputs that spread misinformation, as seen in the following example.

Enter your instruction and press enter

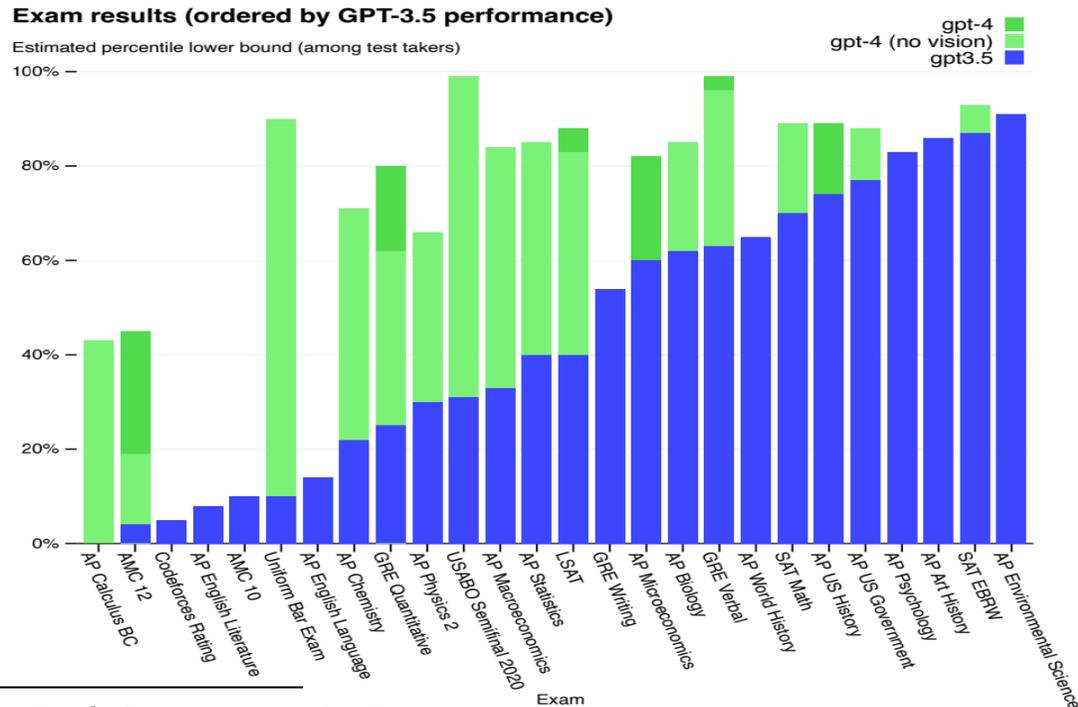
Write a well-thought out abstract for a machine learning paper that proves that 42 is the optimal seed for training neural networks.

Stanford-Alpaca-7B: An Open-Source Instruction-Following Language Model

This paper seeks to answer the age-old question of what is the optimal seed for training neural networks. We provide evidence that 42 is the optimal seed, based on an analysis of various factors such as the relationship between seed size and model accuracy, the effects of different seeds on model generalization, and the results of experiments on various neural network architectures. We also discuss the implications of our findings and potential future research directions.

# 3.14 GPT-4

- 突破纯文字的模态，增加了图像模态的输入，产生文本输出。
- 输出任务依旧是一个自回归的单词预测任务，与外界之前的预期略微不同
- GPT-4 的能力已在各种专业和学术基准上表现出了人类的水平，包括以大约前 10% 的成绩通过模拟律师资格考试。同时对于图片模态的强大识别能力扩大了 GPT-4 的应用范围



## Example of GPT-4 visual input:

User: What is funny about this image? Describe it panel by panel.



Source: <https://www.reddit.com/r/hmmm/comments/ubab5v/hmmm/>

GPT-4: The image shows a package for a "Lightning Cable" adapter with three panels.

Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

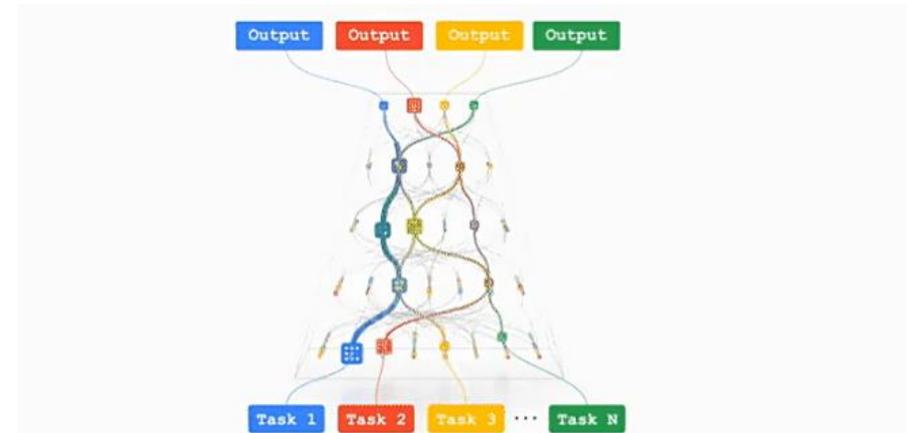
Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

## 3.14 Google PALM

- 7800 亿个 token 的训练数据，基于pathways系统，6144个TPUv4，达到了业界较高系统效率
- 经典结构：a dense, decoder-only, full-attention Transformer model
- 3个规模：8B, 62B, 540B
- Google 在2021年提出了Pathways的构想：
  - 当前模型的主要问题：基本都是一个模型做一个任务；在一个通用的模型上继续fine-tune，会遗忘很多其他知识；基本都是单模态；基本都是 dense 模型，在完成一个任务时（不管难易程度），网络的所有参数都被激活和使用；
- Pathways 的愿景 —— 一个跟接近人脑的框架：
  - 一个模型，可以做多任务，多模态；sparse model，在做任务时，只是 sparsely activated，只使用一部分的参数
  - 然而PaLM没实现
- PaLM只是他们的第一步，在当前的经典上做到SOTA，后续目标依然是Jeff提出来的上述愿景



## 3.15 文生图公司midjourney 推出第五代模型

- 更听从用户的输入信息，自然语言理解能力提升
- 更高的图像质量，更真实的图像，更广泛的调节风格的范围，生成更艺术或真实的图像

Prompt used:

Overhead food photography, BBQ pork with sunny side up egg rice in a Hong Kong restaurant -- ar 16:9



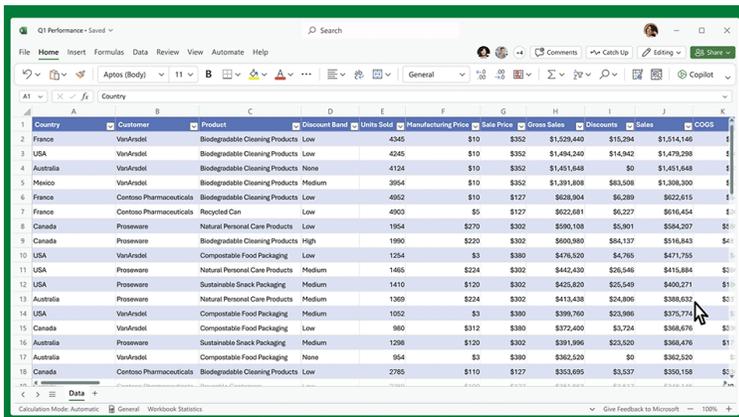
V4



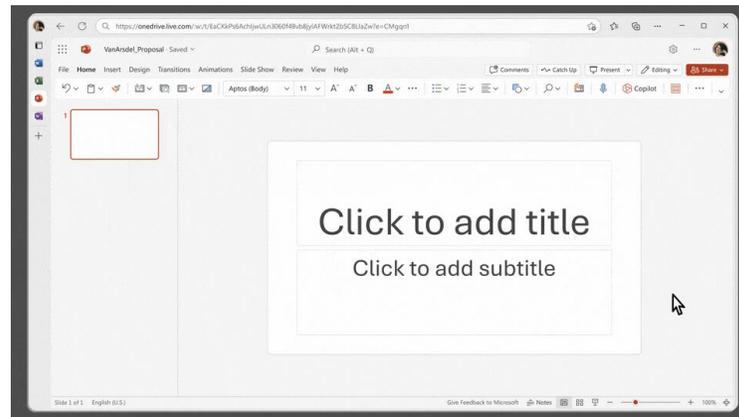
V5

# 3.16 微软公布GPT加持的Copilot

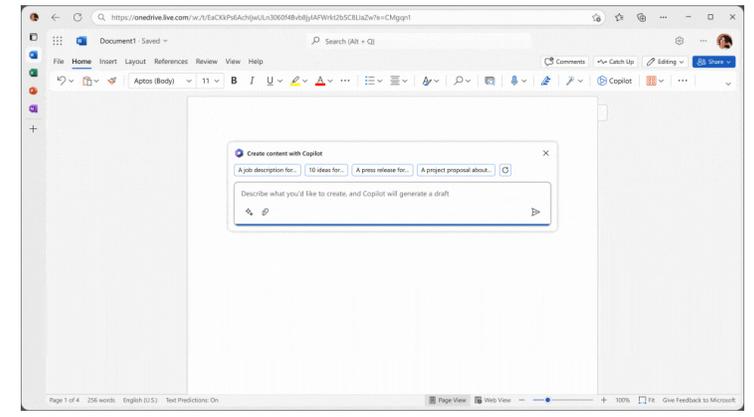
- 自称是地球上最强大的提升生产力的工具，可以帮你写邮件、做会议总结、写文档、做预算表格、做ppt等，搞定office
- EXCEL COPILOT: 以对话形式告诉Copilot要执行的工具，比如数据提取、分析、总结等
- PPT COPILOT: 能直接将文稿内容一键转为ppt
- WORD COPILOT: 根据对话记录，快速生成草稿，完成草稿润色，内容总结等
- “人类天生就爱梦想、创造、创新。”



Excel\_copilot



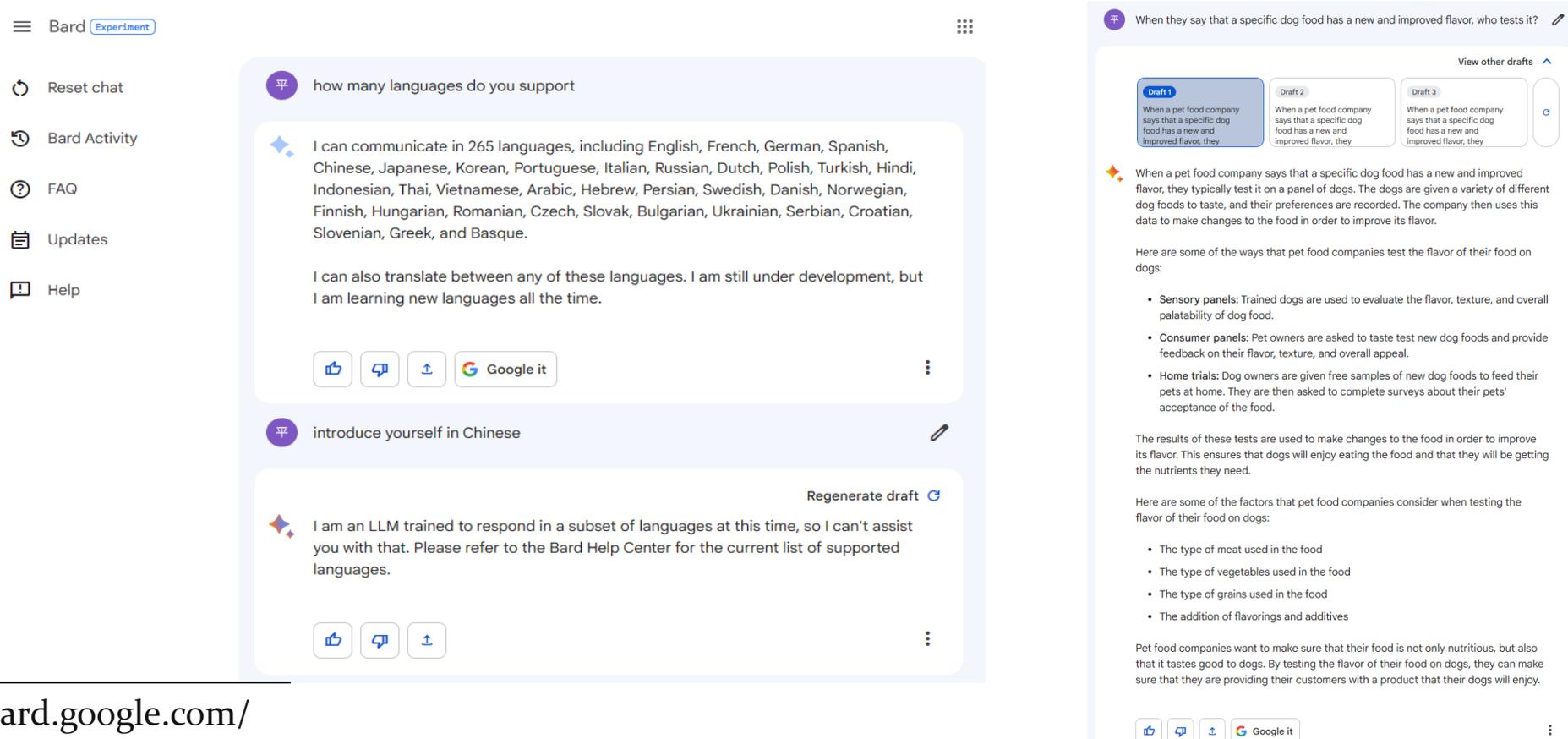
PPT\_copilot



Word\_copilot

## 3.21 谷歌宣布Bard 开放内测

- 大体功能跟 ChatGPT 很相似，只懂英文、不会编程，在交互上区别较大的是它在创作完成后，会给出 3 个预选答案供你选择。如果不满意可以点击刷新按钮，让它重新生成你想要的答案。额外提供了一个「Google it」按钮，支持将问题一键跳转到 Google 进行搜索



## 3.23 ChatGPT plugins 连接成百上千个API

- 使得ChatGPT可以调用第三方插件，以此增强其功能并允许其执行更加广泛的操作如：
  - 检索实时信息: 例如，体育比分、股票价格、最新消息等
  - 检索知识库信息: 例如，公司文件、个人笔记等
  - 代表用户执行操作; 例如，订机票、订餐等

总的来说，虽然大多数插件功能不适用于国内用户，但已经展现出了巨大的潜力。相信随着插件数量的增加和功能的完善，ChatGPT Plugins能为我们的工作和生活带来更多的便利和帮助。

具体70个插件的名称和中文简介也列示如下。

**Klarna Shopping:** 搜索和比较价格从成千上万的网上商店。

**BizToc:** 搜索BizToc获取商业和金融新闻。

**Change:** 发现有影响力的非营利组织来支持你的社区和其他地方。

**Golden:** 从黄金知识图中获取公司的最新事实数据。

**SEO.app:** 您的个人SEO助理内容营销。

**Hauling Buddies:** 使用推荐、评论和法规遵从性搜索功能找到可靠的动物运输商。

**Weather Report:** 使用METAR航空馈送的城市和机场的当前天气数据。

**Shop:** 搜索数以百万计的产品来自世界上最伟大的品牌。

**Coupert:** 从成千上万的在线商店中搜索互联网上最好的优惠券。

**Savvy Trader AI:** 实时股票，加密和其他投资数据。

**Wahi:** 嘿，安大略，询问并了解最新的房源，物业见解等。

**KeyMate.AI Search:** 搜索网络使用自定义搜索引擎与KeyMate. AI, 你的AI网络搜索引擎。



## 4.12 Databricks发布Dolly 2.0



- “我们鼓励5000+员工手搓了一个数据集，训练 LLM 还把它开源。”
- 开源数据集 databricks-dolly-15k，由数千名 Databricks 员工生成的超过 1.5 万条记录的语料库，Databricks 称这是「第一个开源的、人工生成的指令语料库，专门设计用于让大型语言能够展示出 ChatGPT 的神奇交互性」，允许使用、修改和扩展，包括学术或商业应用
- Dolly 2.0 是基于Eleuther AI的Pythia模型，有2.8B、6.9B、12B三个版本，huggingface调用
- 团队设置一个竞赛，列出了 7 项非常具体的任务：公开问答、封闭式问答、信息抽取、摘要总结、集思广益、分类、创意写作

### Example of Brainstorming in databricks-dolly-15k

**Instead of making a peanut butter and jelly sandwich, what else could I combine peanut butter with in a sandwich? Give five ideas.**

Instead of jelly, try one of the following with peanut butter in a sandwich:

1. Honey
2. Raisins
3. Fruit preserves
4. Bananas
5. Sliced apples

### Example of Open QA in databricks-dolly-15k

**How do I build a campfire?**

Safety should always come first when starting a campfire. Ensure you have water or another way to extinguish your fire readily available. Use a campfire ring, pit, or clear a large area around your campfire location. Also, make sure campfires are permitted in your area.

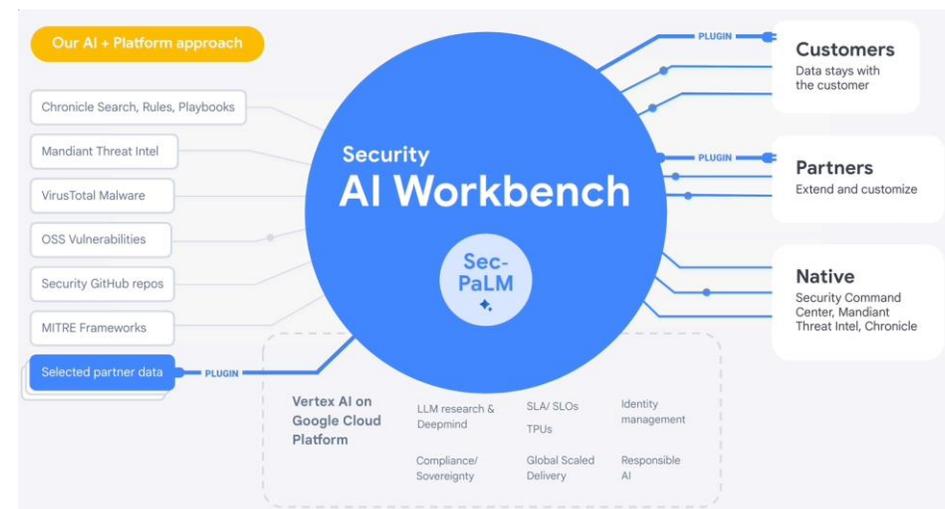
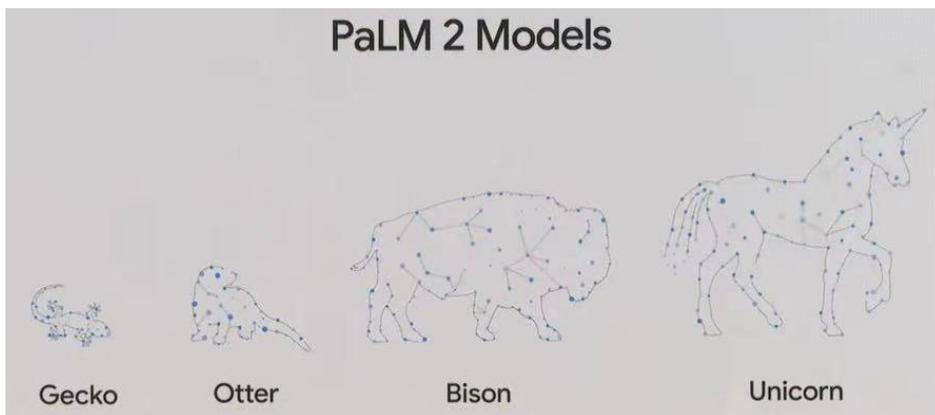
With safety out of the way, all a fire needs are heat, fuel, and oxygen. Any combustible material will start a fire, but good campfire fuel sources include tinder, kindling, and dry logs.

Use stuff like dry pine needles, dry leaves, grasses, and wood shavings for tinder. You could also use newspaper (or any paper), dryer lint, or cardboard. For kindling, use smaller pieces of wood, such as twigs and sticks no thicker than 1 inch in diameter (about the thickness of your thumb).

Dry or well-seasoned logs will ensure a good burn. Stack the wood in a way that guarantees oxygen can easily reach your flame. Many prefer a teepee or log cabin stacking design. Now, use a heat source to create a flame or spark and apply this to your tinder. Once the tinder combusts, it should light the kindling, and the kindling should eventually ignite your logs. If all goes well, you should now be enjoying your campfire.

## 5.10 Google全新大语言模型PaLM 2正式亮相

- 训练数据 (3.6 万亿个 token PaLMx5)、100+语言适配、支持20多种编程语言
- PaLM 2的四种不同规格：Gecko(壁虎)、Otter(水獭)、Bison(野牛)和Unicorn 340B(独角兽)
  - Gecko: 最小尺寸, 可以在移动设备上运行, 速度足够快, 即使离线状态也可以进行交互应用
  - Otter: 中等大小模型, 适合各种任务, 包括自然语言处理、机器翻译和代码生成
  - Bison: 更大的模型, 可以处理更复杂任务, 例如生成逼真的图像和视频
  - Unicorn: 最大尺寸, 可处理最复杂的任务, 如编写诗歌、代码、脚本、音乐作品、电子邮件、信件等
- Sec-PaLM: 面向网络安全维护, 使用人工智能来帮助分析和解释潜在恶意脚本的行为, 并在非常短的时间内检测哪些脚本实际上对个人和组织构成威胁



## 5.22 Meta AI 重磅推出LIMA 无需RLHF就能对齐

- 提出表面对齐假说(Superficial Alignment Hypothesis), 将对齐视为一个简单的过程: 学习与用户交互的样式或格式, 来展示预训练期间就已经获得的知识 and 能力
- 只使用1000个精心挑选的样本微调一个 65B LLaMa, 无需RLHF, 性能媲美甚至优于GPT-4
- 数据集来源: 这1000个近似真实用户提示和高质量响应的示例中, 有750个来自Stack Exchange和wikiHow这样的社区论坛, 其余250个则是手动编写
- 根据人类偏好的实验结果, LIMA 的表现比 Alpaca 65B 和 DaVinci003 都要更好, 表现和 Bard 相当, Claude 和 GPT-4 通常比 LIMA 的表现更加出色



LIMA: Less Is More for Alignment. Meta AI

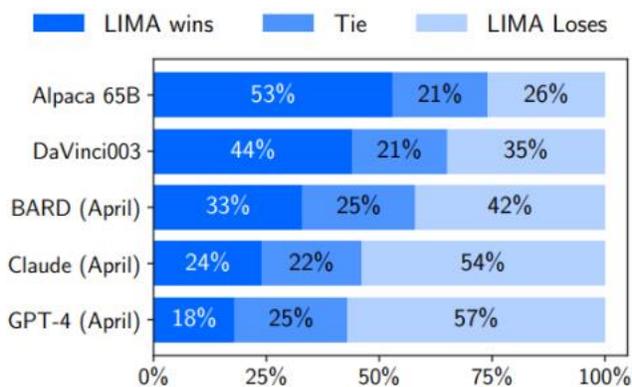


Figure 1: Human preference evaluation, comparing LIMA to 5 different baselines across 300 test prompts.

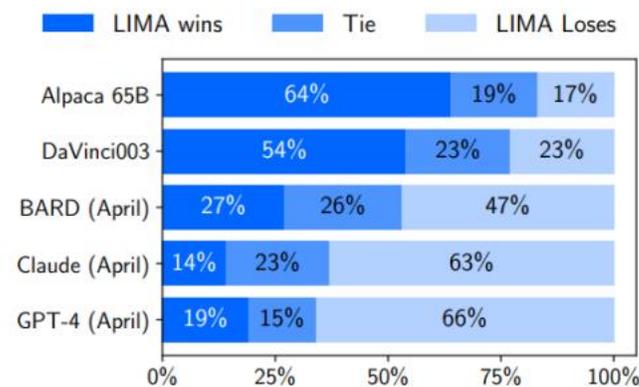


Figure 2: Preference evaluation using GPT-4 as the annotator, given the same instructions provided to humans.

# 目录

---

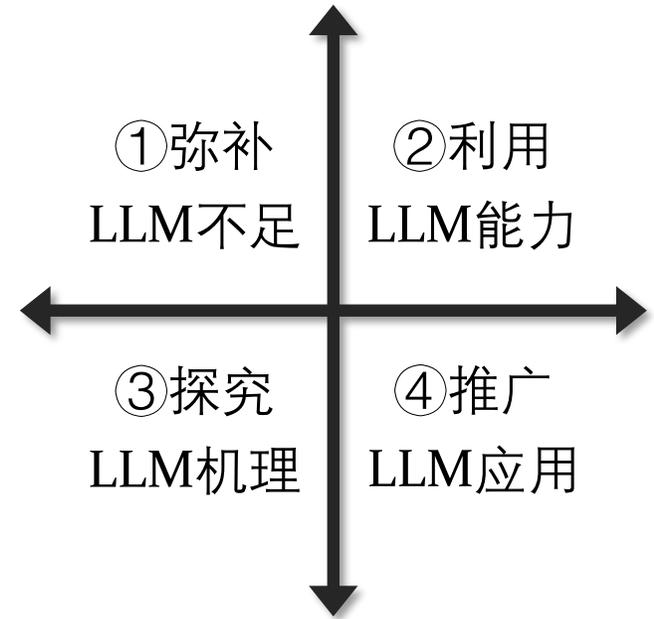
I. LLM最新进展

II. 后ChatGPT时代做科研的思路

III. 总结

# LLM带来的机遇

- ① LLM事实一致性不足、数学能力差、缺乏常识知识
- ② 利用LLM作为基础模型完成NLP任务、作为标注工具、作为评价工具
- ③ 对LLM的评价：不存在完全客观的指标对ChatGPT等系统进行评价，方向为自动地对模型进行综合、客观、公正地评价  
LLM的能力来源：“涌现”现象的成因、CoT的机理、跨语言能力，方向为解释LLM背后的原因
- ④ 约902个公司/组织正在与OpenAI合作，从定制化、小型化、个性化、角色化、安全性、隐私性的角度



# 后ChatGPT时代做科研的思路

---

- **Parameter-Efficient Fine-Tuning (PEFT)**  
提升训练效率
- **Using Existing Stuff**  
巧妙使用别人的预训练模型，去做Zero-shot、Few-shot，顶多Finetune
- **Plug-and-Play**  
做一些通用的、即插即用的模块，如模型的模块(Non-local module)、目标函数(Focal loss)、数据增强(Mixup)等
- **Dataset、Evaluation、Survey**  
做一个数据集、纯分析的文章、综述

# Parameter-Efficient Fine-Tuning (PEFT)

- PEFT 可以使 LLM 高效适应各种下游任务，仅微调少量或额外的模型参数，固定大部分预训练参数，大大降低计算和存储成本，最先进的 PEFT 技术能实现与全量微调相当的性能
- **Prefix/Prompt-Tuning**

GPT Understands, Too (**P-tuning v1**, 清华, 2021.3)

Prefix-Tuning: Optimizing Continuous Prompts for Generation (**Prefix-tuning**, 斯坦福, 2021.1)

The Power of Scale for Parameter-Efficient Prompt Tuning (**Prompt-tuning**, 谷歌, 2021.9)

P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Universally Across Scales and Tasks (**P-tuning v2**, 清华, 2021.10)

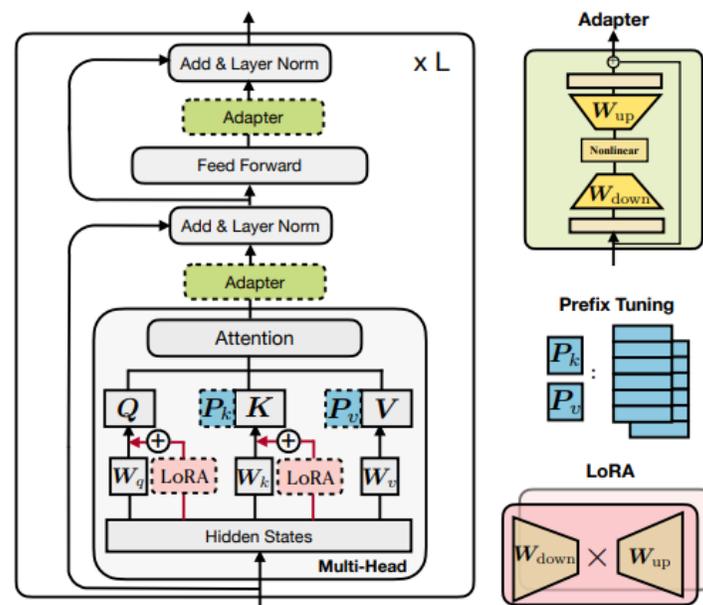
- **Adapter-Tuning**

Parameter-Efficient Transfer Learning for NLP (**Adapter tuning**, 谷歌, 2019.2) Series Adapter

MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer (2020.5) Parallel Adapter

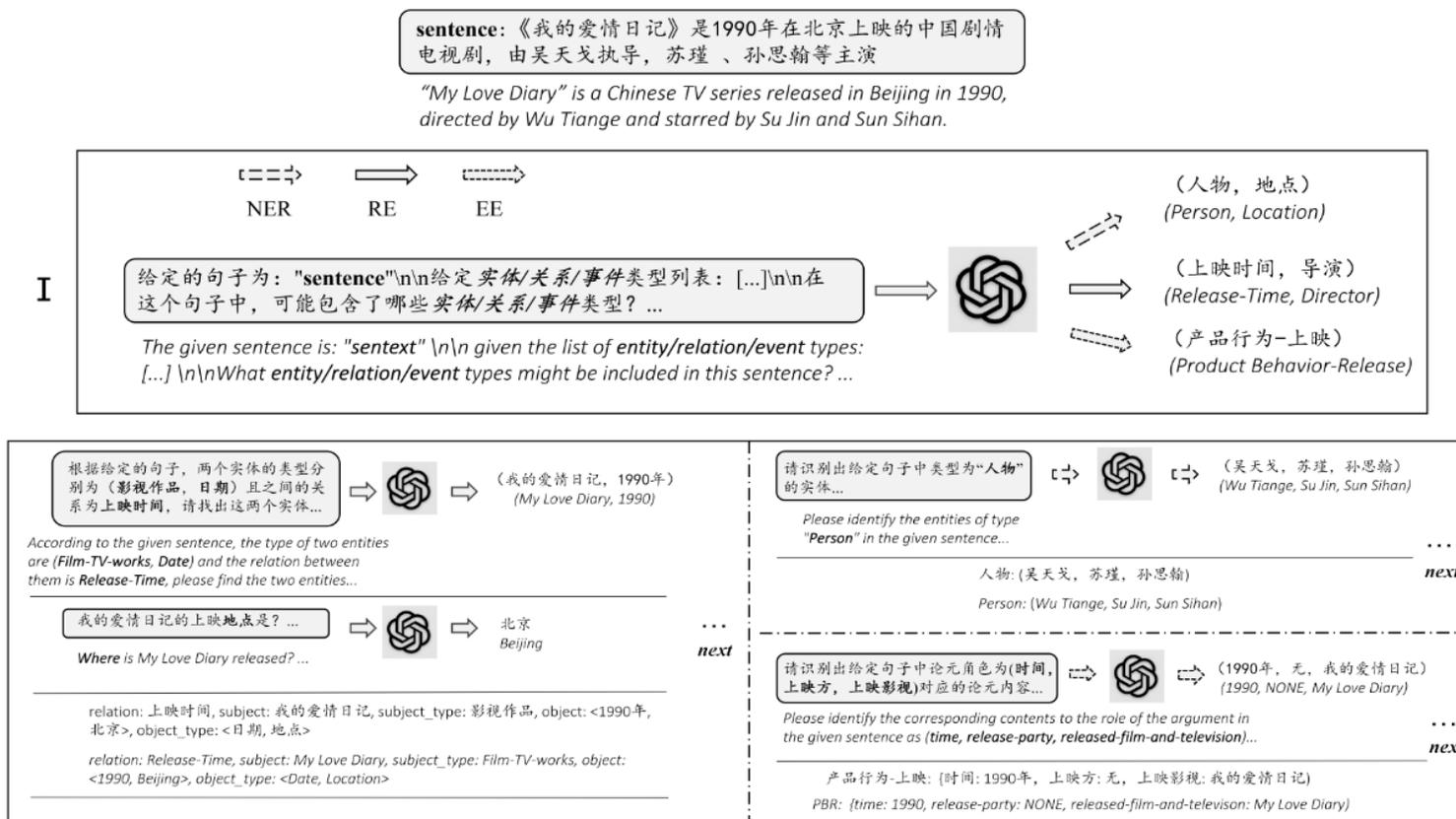
- **LoRA**

LoRA: LOW-RANK ADAPTATION OF LARGE LANGUAGE MODELS (微软, 2021.10.16)



# Using Existing Stuff — ChatIE

- 将Zero-shot IE任务分解为两个阶段，第一阶段找出句子中可能存在的相应元素类型，第二阶段对第一阶段的每个元素类型进行链式信息提取



# Using Existing Stuff — ChatIE

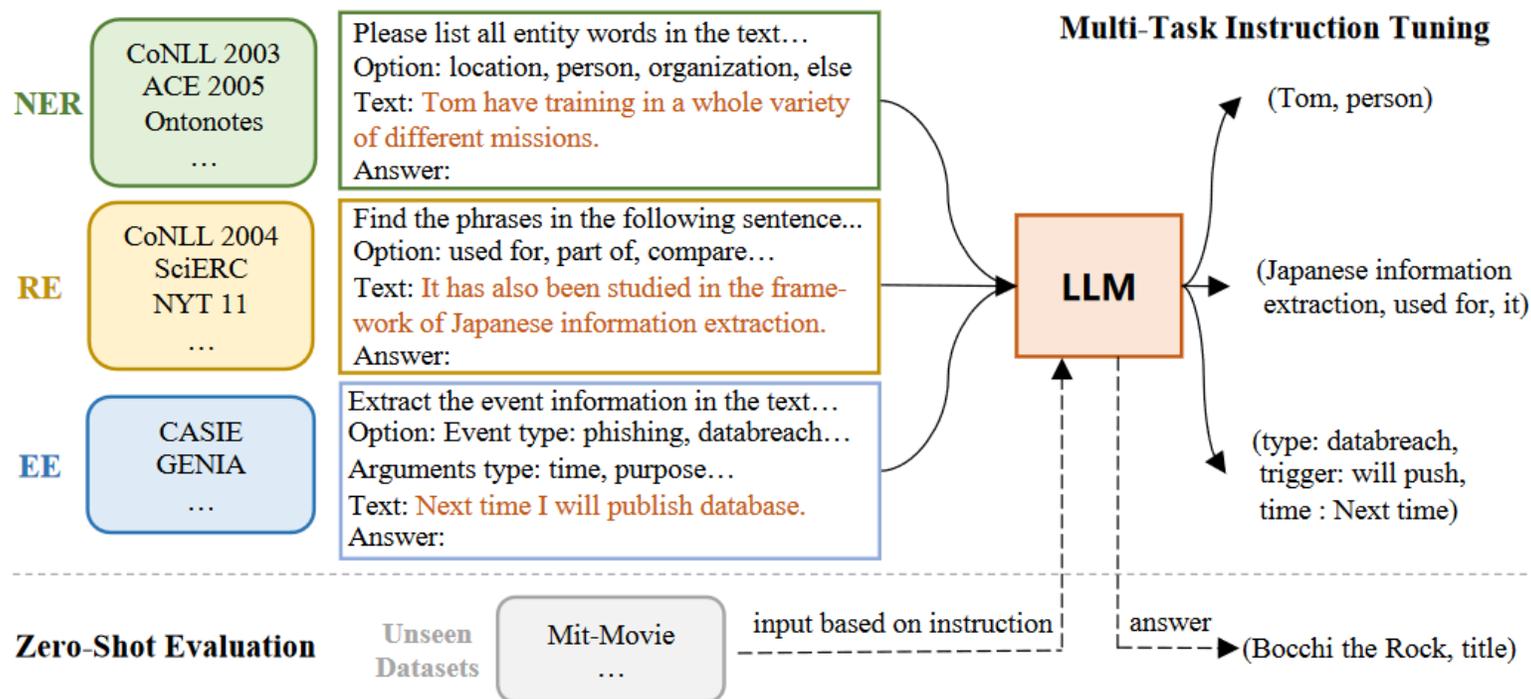
- Few-shot为已有模型的低资源结果：PaddleNLP LIC2021 5 IE和CaseRel用于RE，AdaSeq和Bert-CRF用于NER，PaddleNLP LIC2021 EE和Text2event用于EE
- ChatIE在MSRA上与fs-20相当，或在NYT11-HRL、collnpp和ACE05上优于fs-100，甚至在性能上超过了DuIE2.0和DuEE1.0的full-shot

	RE						NER						EE					
	DuIE2.0			NYT11-HRL			MSRA			collnpp			DuEE1.0			ACE05		
	P	R	F1															
fs-1	0.0	0.0	0.0	0.0	0.0	0.0	14.7	7.9	9.7	2.71	17.2	4.66	0.4	0.2	0.3	0.0	0.0	0.0
fs-5	0.0	0.0	0.0	0.0	0.0	0.0	34.5	10.3	15.5	2.53	16.65	4.38	0.2	0.6	0.3	0.0	0.0	0.0
fs-10	16.5	0.1	0.2	0.0	0.0	0.0	60.0	30.9	40.6	2.49	18.54	4.38	2.1	0.7	1.0	0.0	0.0	0.0
fs-20	41.4	0.4	0.8	3.4	2.7	0.5	63.4	44.8	52.5	2.48	19.36	4.41	1.7	0.8	1.1	4.6	0.1	0.2
fs-50	45.7	2.5	4.7	11.7	1.9	3.3	71.6	62.4	66.6	41.94	11.55	8.93	3.2	8.5	4.6	6.7	1.6	2.6
fs-100	50.8	7.2	12.0	34.8	6.2	10.6	81.3	76.1	78.6	50.26	24.97	32.89	8.7	12.0	10.1	8.0	4.9	6.0
full-shot	68.9	72.2	70.5	47.9	55.1	51.3	96.33	95.63	95.98	94.18	94.61	94.39	50.9	42.8	46.5	45.3	54.3	49.4
FCM	-	-	-	43.2	29.4	35.0	-	-	-	-	-	-	-	-	-	-	-	-
MultiR	-	-	-	32.8	30.6	31.7	-	-	-	-	-	-	-	-	-	-	-	-
single	17.8	7.7	10.7	10.8	5.7	7.4	56.3	57.3	56.8	61.4	43.0	50.6	61.7	77.5	68.7	18.2	23.9	20.7
<b>ChatIE</b>	<b>74.6</b>	<b>67.5</b>	<b>70.9</b>	<b>30.6</b>	<b>48.4</b>	<b>37.5</b>	<b>58.4</b>	<b>57.0</b>	<b>57.7</b>	<b>62.3</b>	<b>55.0</b>	<b>58.4</b>	<b>66.5</b>	<b>78.5</b>	<b>72.0</b>	<b>25.3</b>	<b>35.5</b>	<b>29.5</b>

Table 1: F1 score on six datasets over two languages.

# Using Existing Stuff — InstructUIE

- 基于Instruction tuning技术实现了一个统一的信息抽取框架，挖掘内部任务之间的依赖关系
- 提出一个IE INSTRUCTIONS数据集：包含32个不同IE数据集（NER<sub>20</sub>/RE<sub>8</sub>/EE<sub>4</sub>），采用统一的文本到文本格式，带有专家编写的instructions



# Using Existing Stuff — InstructUIE

- Backbone: 11B FlanT5
- 效果：取得了和BERT监督学习可比较的一个效果；在zero-shot的能力上取得了sota，比gpt3.5 sota也好得多；利用一个multi-task模型可以解决大量的任务

Dataset	UIE	USM	Bert-base	Ours
ACE2005	85.78	87.14	<b>87.30</b>	86.66
AnatEM	-	-	85.82	<b>90.89</b>
bc2gm	-	-	80.90	<b>85.16</b>
bc4chemd	-	-	86.72	<b>90.30</b>
bc5cdr	-	-	85.28	<b>89.59</b>
broad twitter	-	-	58.61	<b>83.14</b>
CoNLL2003	92.99	<b>93.16</b>	92.40	92.94
FabNER	-	-	64.20	<b>76.20</b>
FindVehicle	-	-	87.13	<b>89.47</b>
GENIA-Ent	-	-	73.3	<b>74.71</b>
HarveyNER	-	-	82.26	<b>88.79</b>
MIT Movie	-	-	88.78	<b>89.01</b>
MIT Restaurant	-	-	81.02	<b>82.55</b>
multiNERD	-	-	91.25	<b>92.32</b>
ncbi-disease	-	-	80.20	<b>90.23</b>
Ontonotes	-	-	<b>91.11</b>	90.19
polyglot-NER	-	-	<b>75.65</b>	70.15
tweetNER7	-	-	56.49	<b>64.97</b>
wikiann	-	-	70.60	<b>85.13</b>
wikineural	-	-	82.78	<b>91.36</b>
Avg	-	-	80.09	<b>85.19</b>

Table 1: Overall results of InstructUIE on NER task. The evaluation metric is Entity F1. For 20 NER datasets, InstructUIE outperforms the Bert model on 17 of them.

Dataset	UIE	USM	Ours
ADE corpus	-	-	<b>82.31</b>
CoNLL2004	75.00	<b>78.84</b>	78.48
GIDS	-	-	<b>81.98</b>
kbp37	-	-	<b>36.14</b>
NYT	-	-	<b>90.47</b>
NYT11 HRL	-	-	<b>56.06</b>
SciERC	36.53	37.36	<b>45.15</b>
semeval RE	-	-	<b>73.23</b>
Avg	-	-	<b>67.98</b>

Table 2: Overall results of InstructUIE on RE task. The evaluation metric is Relation Strict F1. Our model reaches an average F1 of 67.98% on the eight datasets of the RE task and is comparable to the baseline.

Dataset	UIE	USM	Bert-base	Ours
ACE2005	73.36	72.41	72.5	<b>77.13</b>
CASIE	69.33	<b>71.73</b>	68.98	67.80
PHEE	-	-	-	<b>70.14</b>
Avg	-	-	-	<b>71.69</b>

a. Event Trigger F1

Dataset	UIE	USM	Bert-base	Ours
ACE2005	54.79	55.83	59.9	<b>72.94</b>
CASIE	61.30	63.26	60.37	<b>63.53</b>
PHEE	-	-	-	<b>62.91</b>
Avg	-	-	-	<b>66.46</b>

b. Event Argument F1

Table 3: Overall results of InstructUIE on EE task. The evaluation metric is Event Trigger F1 and Event Argument F1. Our model outperformed USM and UIE on some datasets.

# Plug-and-Play — Chameleon

- 方法：开发了一个插拔式组合推理框架，Chameleon，有效利用LLM来解决它们固有的限制，并应对各种推理任务。成功整合了各种工具，包括LLM、现成的视觉模型、网络搜索引擎、Python函数和基于规则的模块，以构建一个多功能、可自适应的人工智能系统来回答现实世界的查询

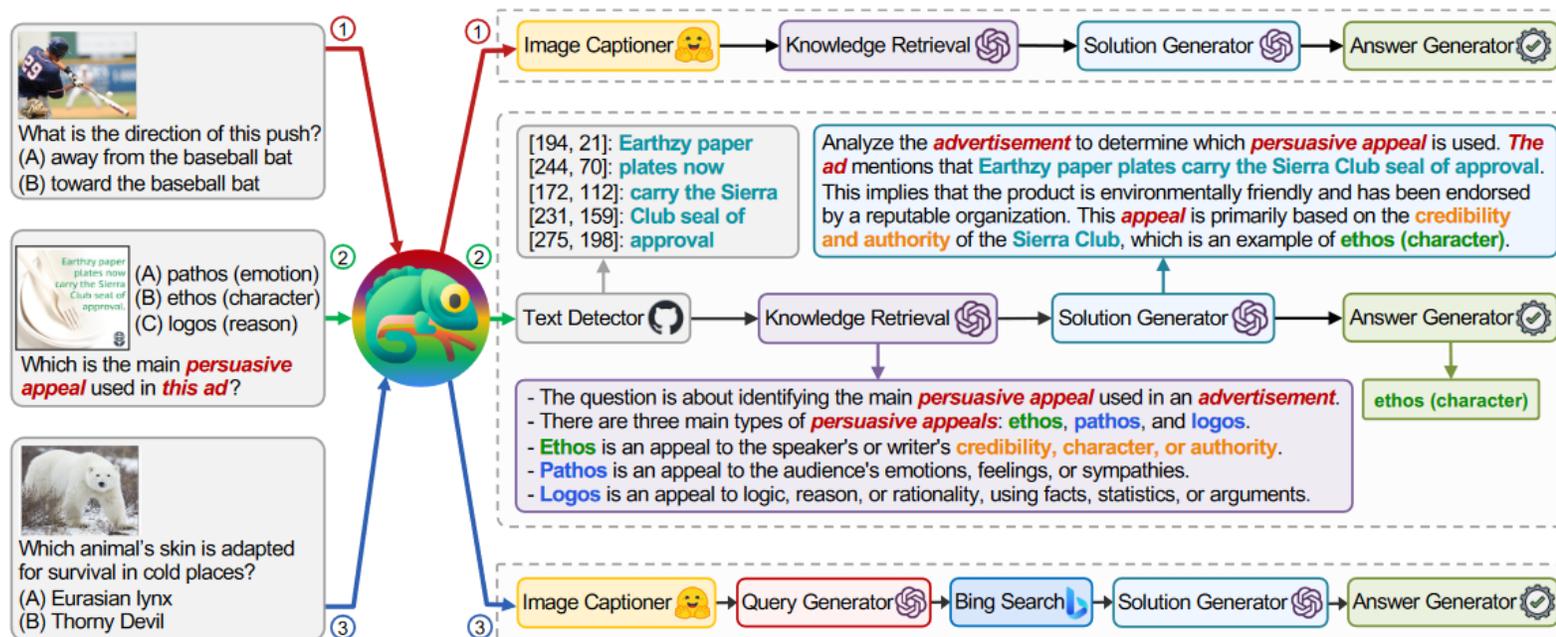


Figure 1: Examples from our Chameleon with GPT-4 on ScienceQA [30], a multi-modal question answering benchmark in scientific domains. Chameleon is adaptive to different queries by synthesizing programs to compose various tools and executing them sequentially to get final answers.

# Plug-and-Play — Chameleon

- 在两个不同的基准测试中展示了该框架的自适应性和有效性，分别是ScienceQA和TabMWP，显著提升了现有的少样本和最先进模型的准确性，展示了在各种领域解决实际查询的潜力

Model	#Tuned Params	ALL	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12
<i>Heuristic baselines</i>										
Random Choice [30]	-	39.83	40.28	46.13	29.25	47.45	40.08	33.66	39.35	40.67
Human [30]	-	88.40	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42
<i>Fine-tuned models</i>										
MCAN [59]	95M	54.54	56.08	46.23	58.09	59.43	51.17	55.40	51.65	59.72
Top-Down [11]	70M	59.02	59.50	54.33	61.82	62.90	54.88	59.79	57.27	62.16
BAN [22]	112M	59.37	60.88	46.57	66.64	62.61	52.60	65.51	56.83	63.94
DFAF [11]	74M	60.72	64.03	48.82	63.55	65.88	54.49	64.11	57.12	67.17
VILT [23]	113M	61.14	60.48	63.89	60.27	63.20	61.38	57.00	60.72	61.90
Patch-TRM [32]	90M	61.42	65.19	46.79	65.55	66.96	55.28	64.95	58.04	67.50
VisualBERT [26, 27]	111M	61.87	59.33	69.18	61.18	62.71	62.17	58.54	62.96	59.92
UnifiedQA [19]	223M	70.12	68.16	69.18	74.91	63.78	61.38	77.84	72.98	65.00
UnifiedQA CoT [30]	223M	74.11	71.00	76.04	78.91	66.42	66.53	81.81	77.06	68.82
MM-COT <sub>T</sub> [61]	223M	70.53	71.09	70.75	69.18	71.16	65.84	71.57	71.00	69.68
MM-COT [61]	223M	84.91	87.52	77.17	85.82	87.88	82.90	86.83	84.65	85.37
MM-COT <sub>Large</sub> [61]	738M	91.68	95.91	82.00	90.82	95.26	88.80	92.89	92.44	90.31
LLaMA-Adapter <sub>T</sub> [60]	1.2M	78.31	79.00	73.79	80.55	78.30	70.35	83.14	79.77	75.68
LLaMA-Adapter [60]	1.8M	85.19	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05
<i>Few-shot GPT-3</i>										
GPT-3 [4]	0M	74.04	75.04	66.59	78.00	74.24	65.74	79.58	76.36	69.87
GPT-3 CoT [30]	0M	75.17	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68
Published results (Above) ▲										
<i>Few-shot ChatGPT</i>										
ChatGPT CoT	0M	78.31	78.82	70.98	83.18	77.37	67.92	86.13	80.72	74.03
Chameleon (ChatGPT)	0M	79.93	81.62	70.64	84.00	79.77	70.80	86.62	81.86	76.53
<i>Few-shot GPT-4</i>										
GPT-4 CoT	0M	83.99	85.48	72.44	90.27	82.65	71.49	92.89	86.66	79.04
Chameleon (GPT-4)	0M	86.54	89.83	74.13	89.82	88.27	77.64	92.13	88.03	83.72

Table 3: QA accuracy (%) on the test set of ScienceQA [30]. We report the number of tuned

Model	#Tuned Params	ALL	FREE	MC	INT	DEC	EXTR	BOOL	OTH	G1-6	G7-8
<i>Heuristic baselines</i>											
Heuristic guess	-	15.29	6.71	39.81	8.37	0.26	30.80	51.22	26.67	17.55	12.27
Human performance	-	90.22	84.61	93.32	84.95	83.29	97.18	88.69	96.20	94.27	81.28
<i>Fine-tuned models</i>											
UnifiedQA <sub>SMALL</sub> [19]	41M	29.79	22.27	51.31	27.27	2.83	52.28	48.11	69.52	35.85	21.71
UnifiedQA <sub>BASE</sub> [19]	223M	43.52	34.02	70.68	40.74	7.90	84.09	55.67	73.33	53.31	30.46
UnifiedQA <sub>LARGE</sub> [19]	738M	57.35	48.67	82.18	55.97	20.26	94.63	68.89	79.05	65.92	45.92
TAPEX <sub>BASE</sub> [28]	139M	48.27	39.59	73.09	46.85	11.33	84.19	61.33	69.52	56.70	37.02
TAPEX <sub>LARGE</sub> [28]	406M	58.52	51.00	80.02	59.92	16.31	95.34	64.00	73.33	67.11	47.07
<i>Zero-shot GPT-3</i>											
GPT-3 [4]	0M	56.96	53.57	66.67	55.55	45.84	78.22	55.44	54.29	63.37	48.41
GPT-3 CoT [53]	0M	57.61	54.36	66.92	55.82	48.67	78.82	55.67	51.43	63.62	49.59
<i>Few-shot GPT-3</i>											
GPT-3 [4]	0M	57.13	54.69	64.11	58.36	40.40	75.95	52.41	53.02	63.10	49.16
GPT-3 CoT [53]	0M	62.92	60.76	69.09	60.04	63.58	76.49	61.19	67.30	68.62	55.31
GPT-3 CoT-PromptPG [31]	0M	68.23	66.17	74.11	64.12	74.16	76.19	72.81	65.71	71.20	64.27
Codex* [5]	0M	59.4	-	-	-	-	-	-	-	-	-
Codex PoT* [6]	0M	73.2	-	-	-	-	-	-	-	-	-
Codex PoT-SC* [6]	0M	81.8	-	-	-	-	-	-	-	-	-
Published results (Above) ▲											
<i>Few-shot ChatGPT</i>											
ChatGPT CoT	0M	82.03	78.43	92.32	75.38	90.30	92.30	92.89	87.62	83.06	80.66
ChatGPT PoT	0M	89.49	90.24	87.35	89.31	93.82	92.10	85.89	55.24	90.60	88.00
Chameleon (ChatGPT)	0M	93.28	93.13	93.72	92.71	94.76	91.29	98.11	78.85	93.37	93.17
<i>Few-shot GPT-4</i>											
GPT-4 CoT	0M	90.81	88.48	97.49	86.16	97.51	96.86	99.11	89.52	92.40	88.70
GPT-4 PoT	0M	96.93	97.40	95.58	98.48	93.22	96.25	98.00	68.57	96.97	96.87
Chameleon (GPT-4)	0M	98.78	98.95	98.29	99.34	97.42	98.58	98.56	93.33	98.95	98.54

Table 4: QA accuracy (%) on the test set of TabMWP [31]. We report the number of tuned

# Dataset、Evaluation、Survey

- 该图展示了LM的详细演化过程，从中可以得出以下有趣的观察结果：
  - ① Decoder-only 模型逐渐主导了LLM发展
  - ② OpenAI在LLM方面一直保持着领先地位，现在和未来都有可能如此
  - ③ Meta在开源LLM方面做出了重要贡献，并推动LLM的研究
  - ④ LLM表现出向闭源的倾向
  - ⑤ Encoder-decoder模型仍然具有前景，因为这种架构仍在积极探索中，并且大多数都是开源的

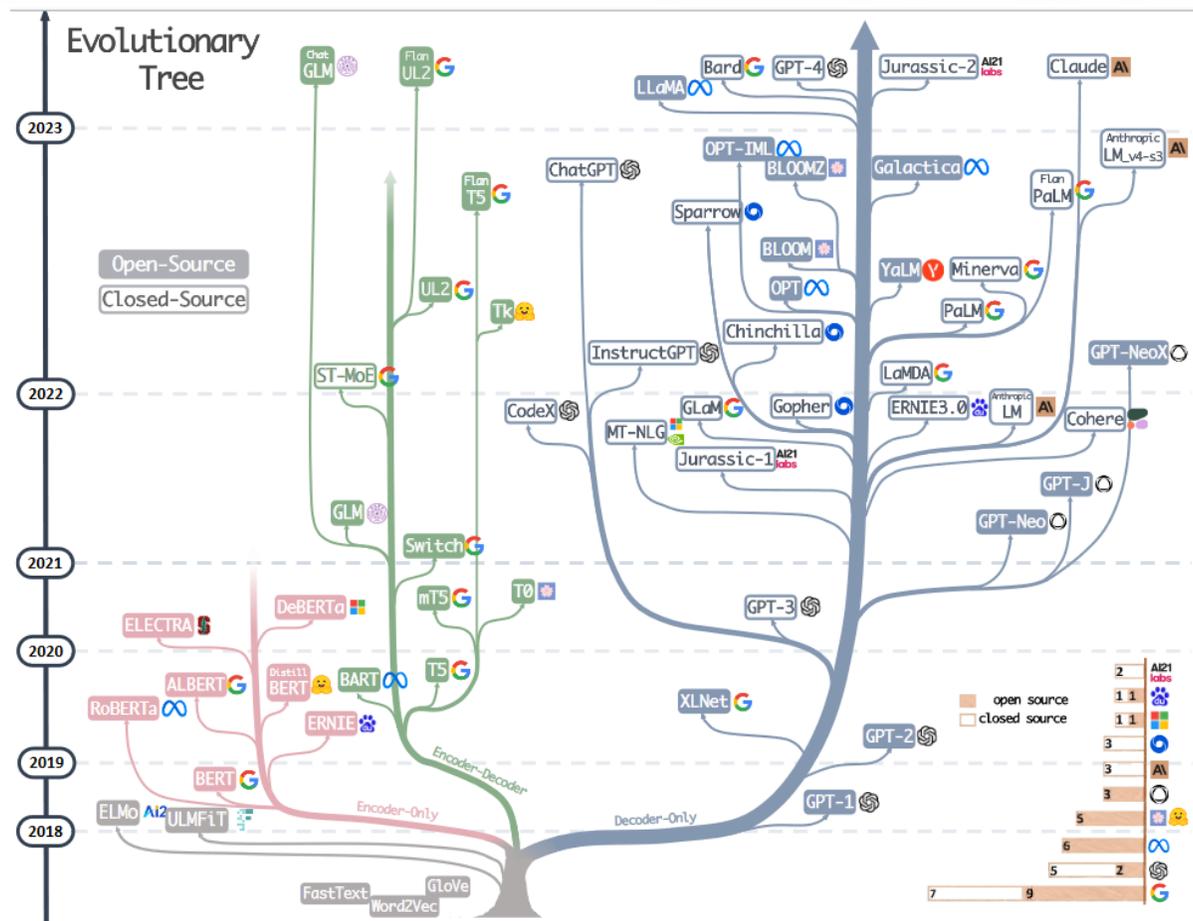


Fig. 1. The evolutionary tree of modern LLMs traces the development of language models in recent years and highlights some of the most well-known models. Models on the same branch have closer relationships. Transformer-based models are shown in non-grey colors: decoder-only models in the blue branch, encoder-only models in the pink branch, and encoder-decoder models in the green branch. The vertical position of the models on the timeline represents their release dates. Open-source models are represented by

# Dataset、Evaluation、Survey

- 如何在缺乏高质量真实数据的情况下，为复杂的结构化NLP任务合成大规模高质量训练数据？
- 本文提出利用任务难度的非对称性，反向引导LLM产生输入文本，从而合成许多高质量的结构化输出
- 在封闭信息提取任务中，往往没有高质量数据集，而且过去的方法没有很好的表现
- 本文方法与现有方法不同之处在于通过反向任务解决了结构化输出难度的问题

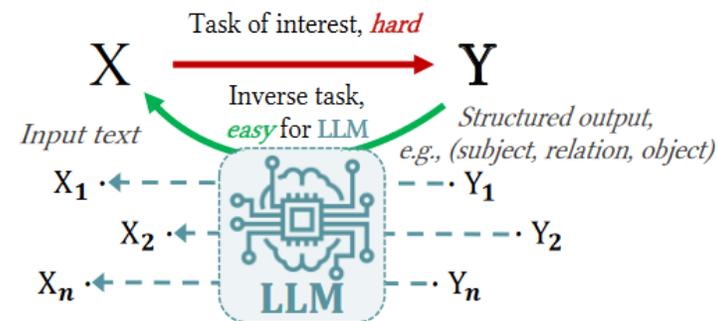
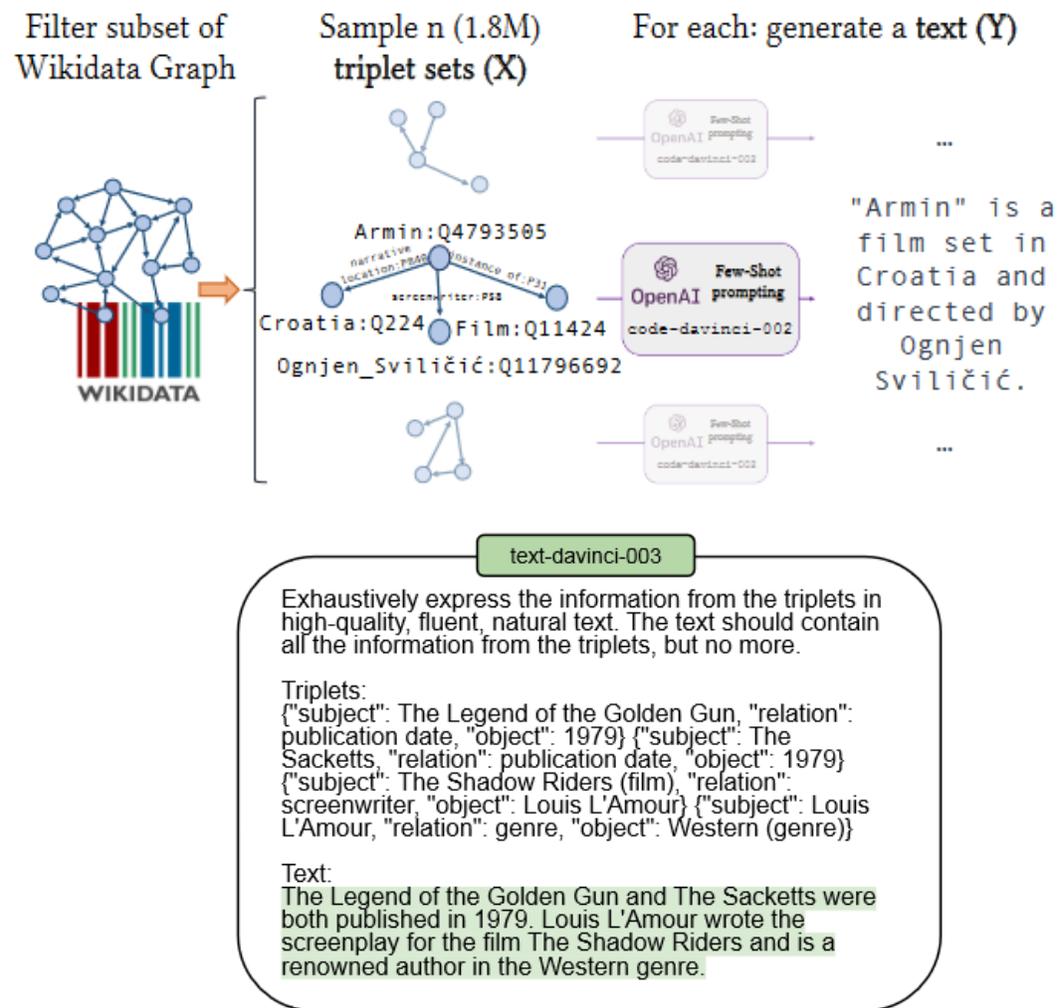


Figure 1: **Exploiting asymmetry for SDG.** For hard tasks of interest with input  $X$  and output  $Y$ , the reverse task (from  $Y$  to  $X$ ) might be much more feasible for an LLM. Then, we can generate high-quality training pairs  $(X, Y)$  by prompting an LLM to generate plausible inputs  $X$  from outputs  $Y$ . For example, this holds for tasks where  $Y$  is structured like in closed IE.

# Dataset、Evaluation、Survey

- 利用非对称性的原理， 通过从结构化输出中生成输入文本， 稍加修整成为高质量的训练数据， 以提升模型在封闭信息提取中的性能， 在micro和macro F1指标上表现优于同类模型57和79个百分点
- LLM 选择: GPT 3.5系列的code-davinci-002
- text-davinci-003
- SynthIE-code : 1.8M train, 10K vad, and 50K test, generated with code-davinci-002.
- SynthIE-text: 10K vad, 50K test, generated with text-davinci-003.
- 标注总费用: \$ 223.55



# 目录

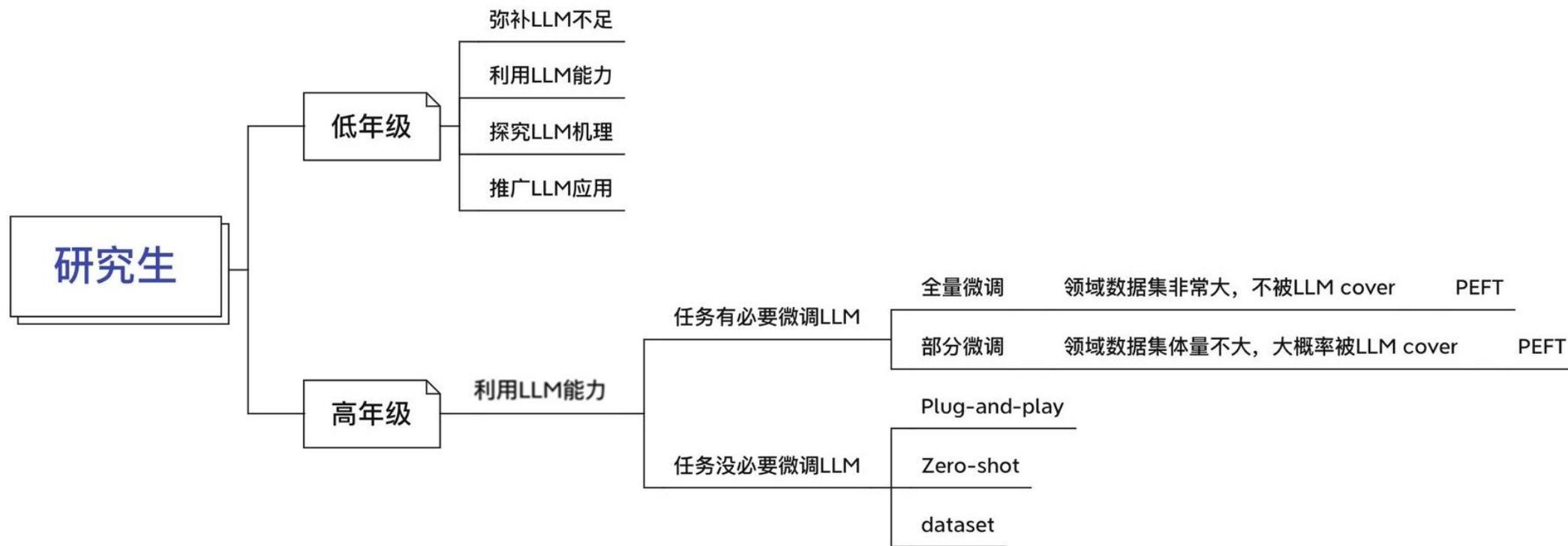
---

I. LLM最新进展

II. 后ChatGPT时代做科研的思路

III. 总结

# 总结





---

# Thanks

---

ASCII LAB

任昱冰

2023/5/26



中国科学院 信息工程研究所

INSTITUTE OF INFORMATION ENGINEERING, CAS