



中国科学院新疆理化技术研究所
Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences

大模型时代Tool Learning的 必要性探讨

米尔阿迪力江 麦麦提
(昌吉, 2024.10.23)

Outline

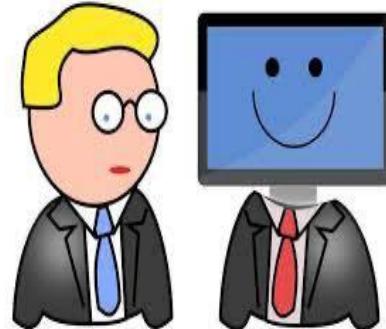
- Introduction of LLM
- LLM Performance in NLP
- Fine-Tuning Strategies & Challenges for LLM
- Categorization of Tools
- Benchmarking Tool Use
- Open-source Tool-Kit

Introduction of LLM

The Three Paths to Achieving AI

输入经验

将知识通过规则等方式教给计算机，进行**符号处理**



实现人脑机制

解明人脑机制，基于相同原理实现人类智能



从数据中学习

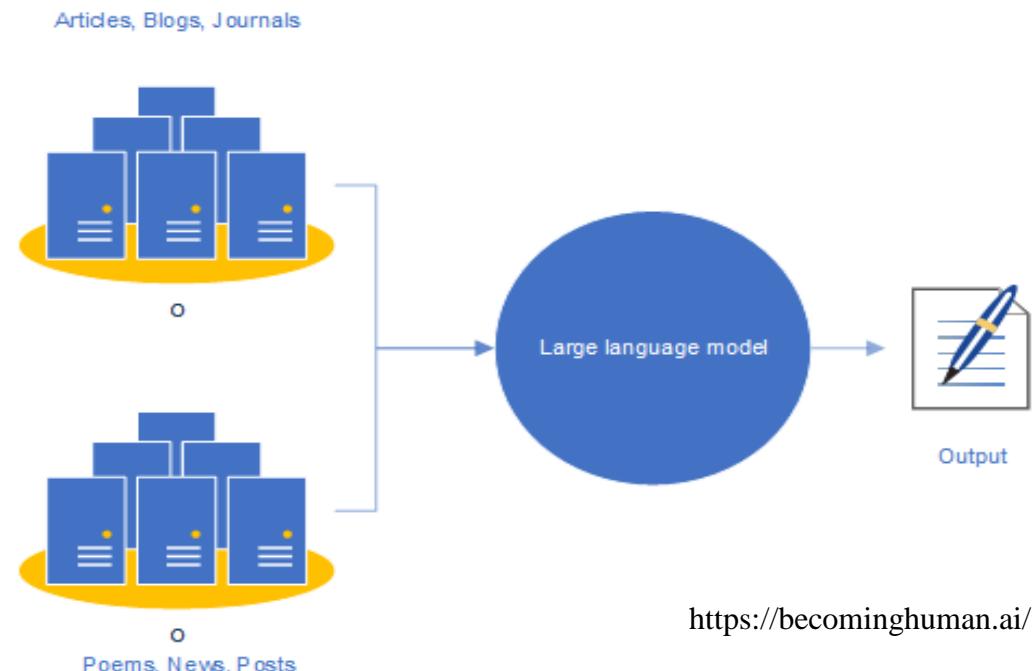
通过数据驱动，ML方法模仿人类智能



LLM = 通过数据学习 + 受人脑启发 + 输入经验知识

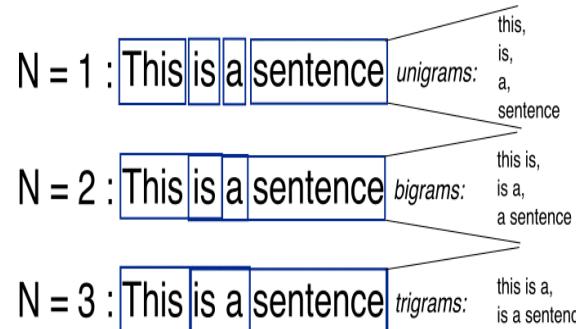
What Are Large Language Models?

大规模语言模型（LLM）是基于Transformer架构的模型，它们通过自监督或半监督学习在大量未标记的文本上进行广泛训练。这些模型旨在生成听起来自然且与上下文相关的文本，适用于各种风格和格式。

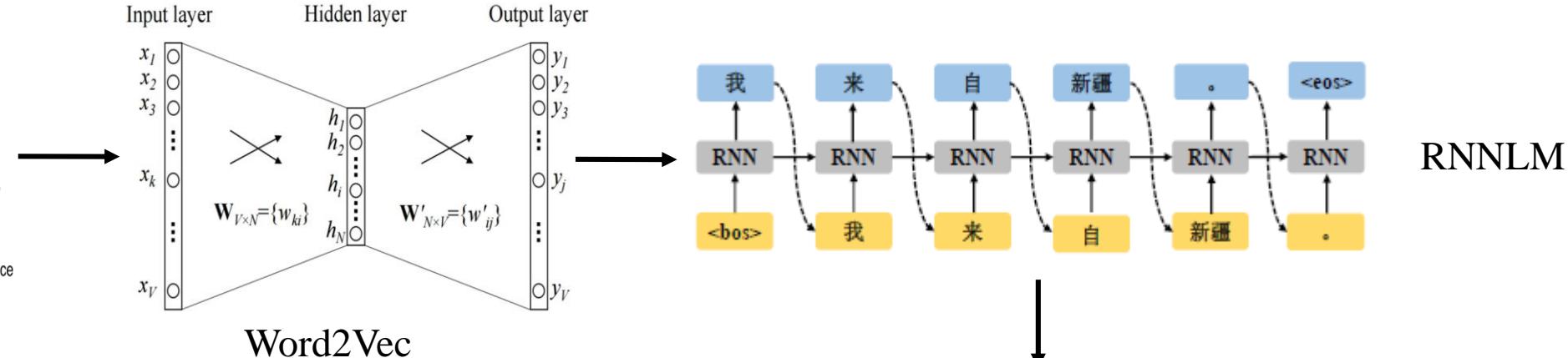


<https://becominghuman.ai/>

Language Models



N-gram LM

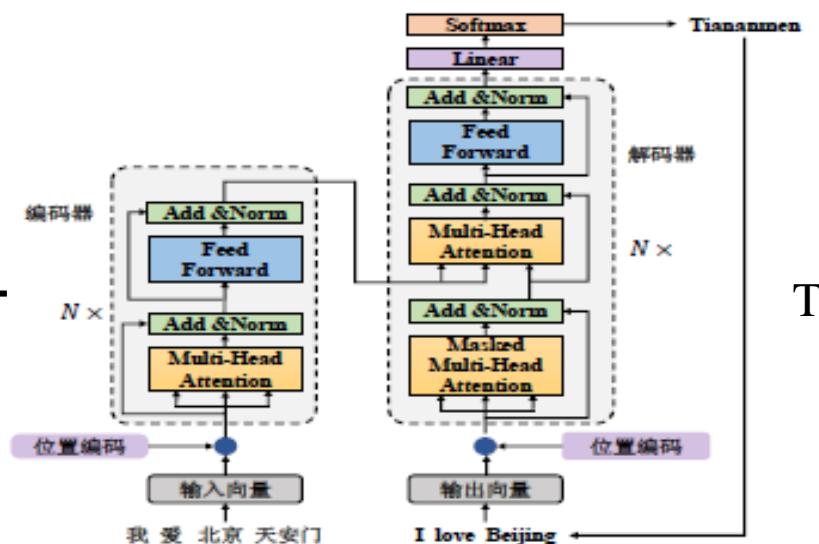


RNNLM

LLM



PLM



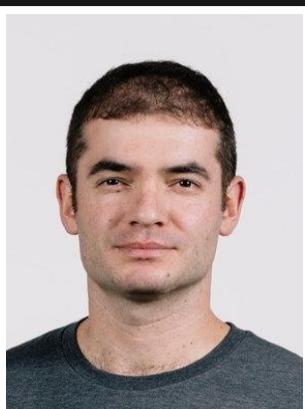
Transformer

Large-scale PLMs become the foundational infrastructure of Modern NLP



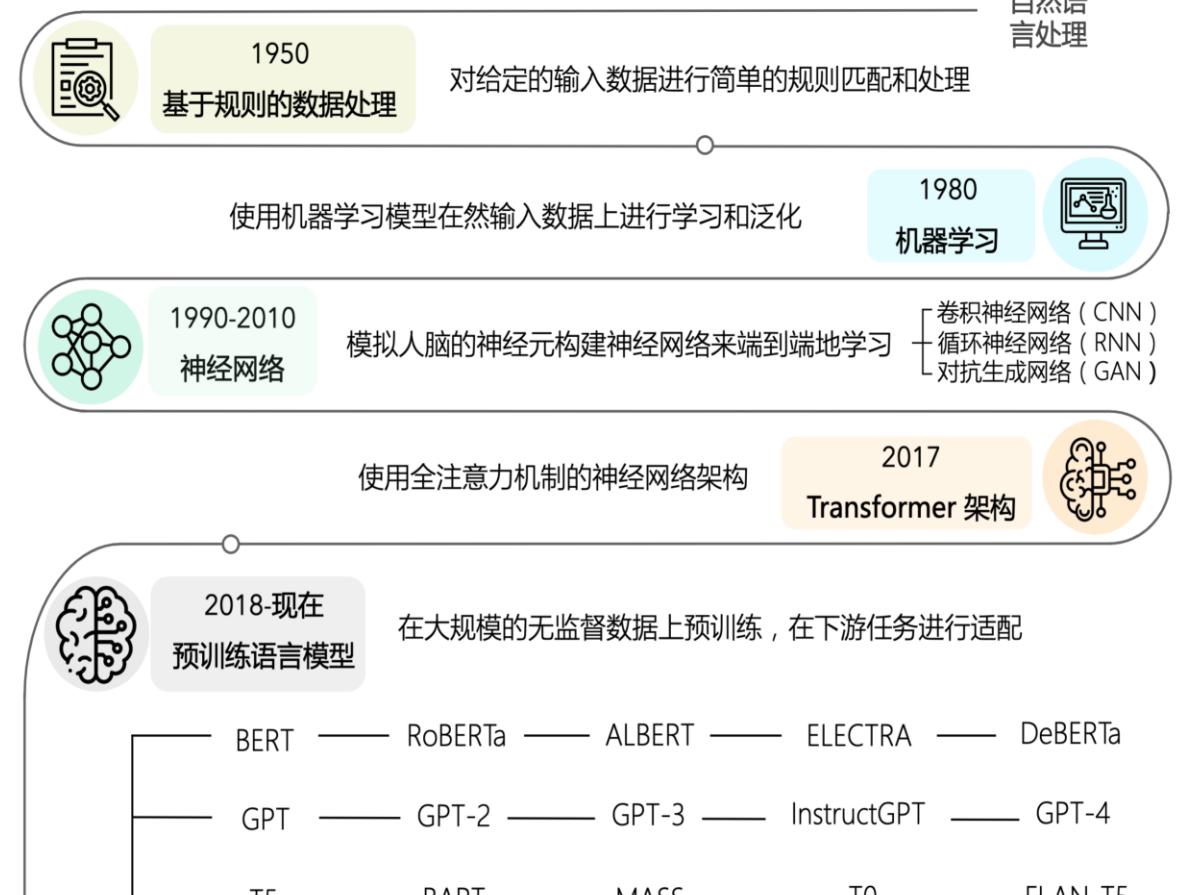
Geoffrey Hinton
(图灵奖获得者)

将GPT-3的壮观表现推断到未来，
表明生命、宇宙和万物的答案只是
4.398万亿个参数。



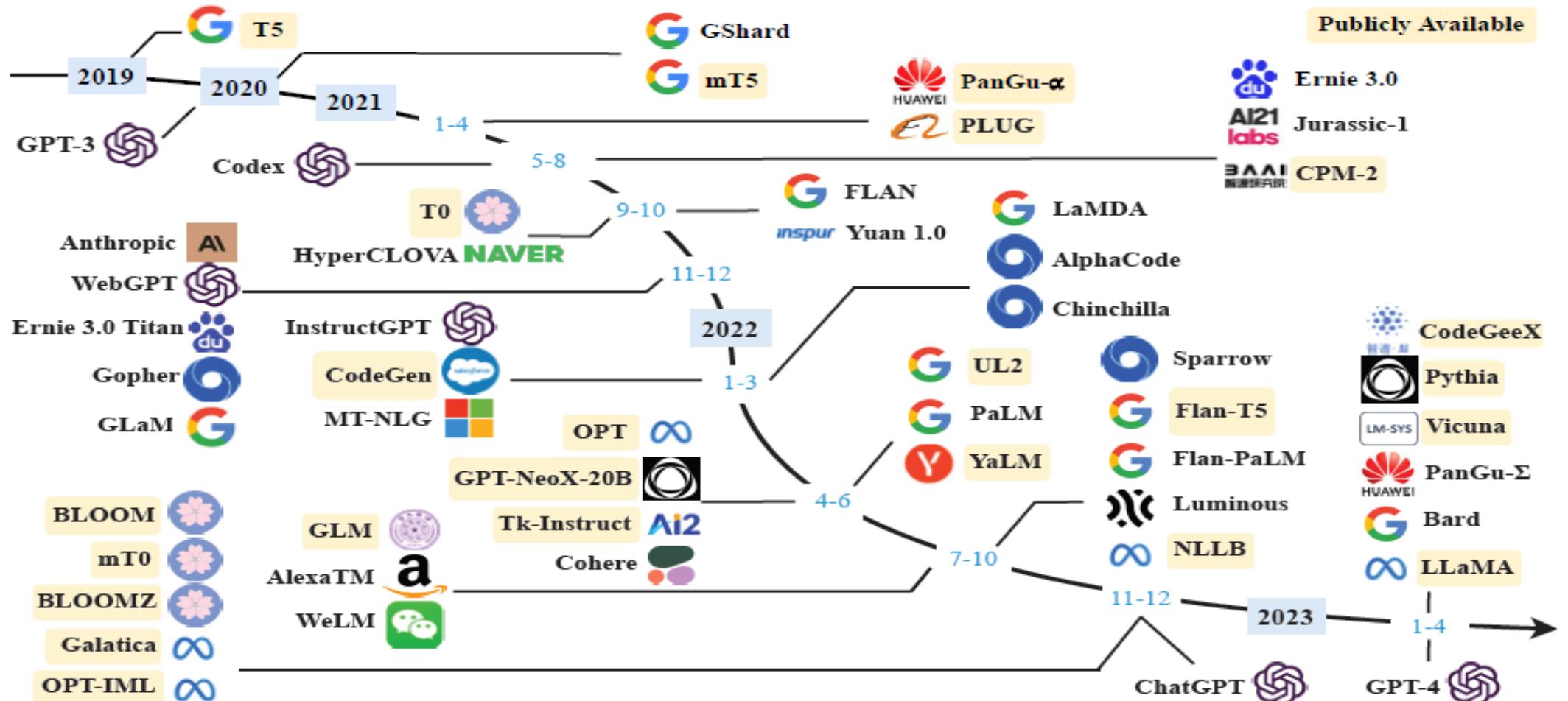
Ilya Sutskever
(OpenAI首席科学家)

将语言模型在学习预测下一个字符的过程中来学习对整个宇宙的规律。



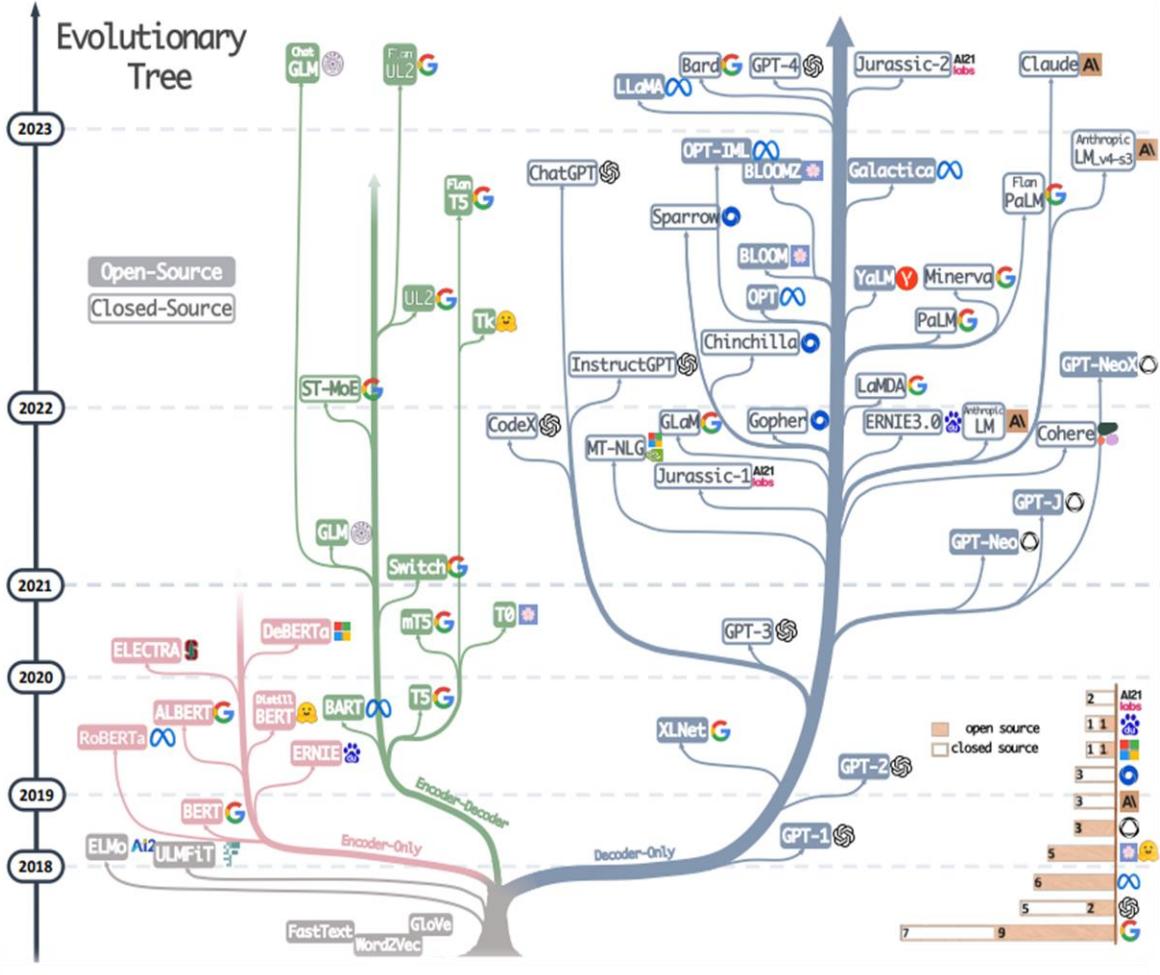
Ding, CCL2023

The timeline of existing LLM



(Xin Zhao et al., 2023)

The Evolutionary of LLM



- **GPT** "Improving Language Understanding by Generative Pre-Training". 2018.
- **GPT-2** "Language Models are Unsupervised Multitask Learners". 2018.
- **GPT-3** "Language Models are Few-Shot Learners". NeurIPS 2020.
- **OPT** "OPT: Open Pre-trained Transformer Language Models". 2022.
- **PaLM** "PaLM: Scaling Language Modeling with Pathways". Aakanksha Chowdhery et al arXiv 2022.
- **BLOOM** "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model". 2022.
- **MT-NLG** "Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model". 2021.
- **GLaM** "GLaM: Efficient Scaling of Language Models with Mixture-of-Experts". ICML 2022.
- **Gopher** "Scaling Language Models: Methods, Analysis & Insights from Training Gopher", 2021.
- **chinchilla** "Training Compute-Optimal Large Language Models". 2022.
- **LaMDA** "LaMDA: Language Models for Dialog Applications". 2021.
- **LLaMA** "LLaMA: Open and Efficient Foundation Language Models". 2023.
- **GPT-4** "GPT-4 Technical Report". 2023.
- **BloombergGPT** BloombergGPT: A Large Language Model for Finance, 2023,
- **GPT-NeoX-20B**: "GPT-NeoX-20B: An Open-Source Autoregressive Language Model". 2022.

The Development History of GPT (Generative Pre-Trained Transformer)



Existing Models

- 近年来大规模语言模型（在本次调查中规模超过100亿的模型）的统计数据，包括容量评估、预训练数据规模（以令牌数量或存储大小计算）以及硬件资源成本。
- 在这张表格中，我们仅包括了那些有关技术细节的公开论文的大规模语言模型。

Model	Release Time	Size (B)	Base Model	Adaptation IT	RLHF	Pre-train Data Scale	Latest Data Timestamp	Hardware (GPUs / TPUs)	Training Time	Evaluation ICL	Evaluation CoT
Publicly Available	T5 [73]	Oct-2019	11	-	-	1T tokens	Apr-2019	1024 TPU v3	-	✓	-
	mT5 [74]	Oct-2020	13	-	-	1T tokens	-	-	-	✓	-
	PanGu- α [75]	Apr-2021	13*	-	-	1.1TB	-	2048 Ascend 910	-	✓	-
	CPM-2 [76]	Jun-2021	198	-	-	2.6TB	-	-	-	-	-
	T0 [28]	Oct-2021	11	T5	✓	-	-	512 TPU v3	27 h	✓	-
	CodeGen [77]	Mar-2022	16	-	-	577B tokens	-	-	-	✓	-
	GPT-NeoX-20B [78]	Apr-2022	20	-	-	825GB	-	96 40G A100	-	✓	-
	Tk-Instruct [79]	Apr-2022	11	T5	✓	-	-	256 TPU v3	4 h	✓	-
	UL2 [80]	May-2022	20	-	-	1T tokens	Apr-2019	512 TPU v4	-	✓	✓
	OPT [81]	May-2022	175	-	-	180B tokens	-	992 80G A100	-	✓	-
	NLLB [82]	Jul-2022	54.5	-	-	-	-	-	-	✓	-
	GLM [83]	Oct-2022	130	-	-	400B tokens	-	768 40G A100	60 d	✓	-
	Flan-T5 [64]	Oct-2022	11	T5	✓	-	-	-	-	✓	✓
	BLOOM [69]	Nov-2022	176	-	-	366B tokens	-	384 80G A100	105 d	✓	-
	mT0 [84]	Nov-2022	13	mT5	✓	-	-	-	-	✓	-
	Galactica [85]	Nov-2022	120	-	-	106B tokens	-	-	-	✓	✓
	BLOOMZ [84]	Nov-2022	176	BLOOM	✓	-	-	-	-	✓	-
	OPT-IML [85]	Dec-2022	175	OPT	✓	-	-	128 40G A100	-	✓	✓
	LLaMA [57]	Feb-2023	65	-	-	1.4T tokens	-	2048 80G A100	21 d	✓	-
	CodeGeeX [86]	Sep-2022	13	-	-	850B tokens	-	1536 Ascend 910	60 d	✓	-
	Pythia [87]	Apr-2023	12	-	-	300B tokens	-	256 40G A100	-	✓	-
Closed Source	GPT-3 [85]	May-2020	175	-	-	300B tokens	-	-	-	✓	-
	GShard [88]	Jun-2020	600	-	-	1T tokens	-	2048 TPU v3	4 d	-	-
	Codex [89]	Jul-2021	12	GPT-3	-	100B tokens	May-2020	-	-	✓	-
	ERNIE 3.0 [90]	Jul-2021	10	-	-	375B tokens	-	384 V100	-	✓	-
	Jurassic-1 [91]	Aug-2021	178	-	-	300B tokens	-	800 GPU	-	✓	-
	HyperCLOVA [92]	Sep-2021	82	-	-	300B tokens	-	1024 A100	13.4 d	✓	-
	FLAN [62]	Sep-2021	137	LaMDA-PT	✓	-	-	128 TPU v3	60 h	✓	-
	Yuan 1.0 [93]	Oct-2021	245	-	-	180B tokens	-	2128 GPU	-	✓	-
	Anthropic [94]	Dec-2021	52	-	-	400B tokens	-	-	-	✓	-
	WebGPT [72]	Dec-2021	175	GPT-3	-	✓	-	-	-	✓	-
	Gopher [59]	Dec-2021	280	-	-	300B tokens	-	4096 TPU v3	920 h	✓	-
	ERNIE 3.0 Titan [95]	Dec-2021	260	-	-	-	-	-	-	✓	-
	GLaM [96]	Dec-2021	1200	-	-	280B tokens	-	1024 TPU v4	574 h	✓	-
	LaMDA [63]	Jan-2022	137	-	-	768B tokens	-	1024 TPU v3	57.7 d	-	-
	MT-NLG [97]	Jan-2022	530	-	-	270B tokens	-	4480 80G A100	-	✓	-
	AlphaCode [98]	Feb-2022	41	-	-	967B tokens	Jul-2021	-	-	-	-
	InstructGPT [61]	Mar-2022	175	GPT-3	✓	✓	-	-	-	✓	-
	Chinchilla [34]	Mar-2022	70	-	-	1.4T tokens	-	-	-	✓	-
	PaLM [56]	Apr-2022	540	-	-	780B tokens	-	6144 TPU v4	-	✓	✓
	AlexaTM [99]	Aug-2022	20	-	-	1.3T tokens	-	128 A100	120 d	✓	✓
	Sparrow [100]	Sep-2022	70	-	-	✓	-	64 TPU v3	-	✓	-
	WeLM [101]	Sep-2022	10	-	-	300B tokens	-	128 A100 40G	24 d	✓	-
	U-PaLM [102]	Oct-2022	540	PaLM	-	-	-	512 TPU v4	5 d	✓	✓
	Flan-PaLM [64]	Oct-2022	540	PaLM	✓	-	-	512 TPU v4	37 h	✓	✓
	Flan-U-PaLM [64]	Oct-2022	540	U-PaLM	✓	-	-	-	-	✓	✓
	GPT-4 [46]	Mar-2023	-	-	✓	✓	-	-	-	✓	✓
	PanGu- Σ [103]	Mar-2023	1085	PanGu- α	-	329B tokens	-	512 Ascend 910	100 d	✓	-

(Xin Zhao et al., 2023)

Dataset Statistics

- 信息量估计：
 - 人一生要说的话 5TB
 - 所有语言的wiki百科 5PB (1PB=1024TB)
 - 所有记录在纸上的内容 200PB
 - YouTube所有视频 400PB, Facebook总信息量 500PB
 - 2000年互联网上的信息量 1000PB, 2010年互联网信息量 450EB; 目前互联网总信息量 3600EB (1EB=1024PB)
 - 世界上所有的电子数据约 100ZB (1ZB=1024EB)
- LLM训练 (ChatGPT) 用了320TB文字信息, 约3000亿单词/字
- $320T/5T=64$ 个人

Statistics of commonly-used data sources.

Corpora	Size	Source	Latest Update Time
BookCorpus [122]	5GB	Books	Dec-2015
Gutenberg [123]	-	Books	Dec-2021
C4 [73]	800GB	CommonCrawl	Apr-2019
CC-Stories-R [124]	31GB	CommonCrawl	Sep-2019
CC-NEWS [27]	78GB	CommonCrawl	Feb-2019
REALNEWS [125]	120GB	CommonCrawl	Apr-2019
OpenWebText [126]	38GB	Reddit links	Mar-2023
Pushift.io [127]	2TB	Reddit links	Mar-2023
Wikipedia [128]	21GB	Wikipedia	Mar-2023
BigQuery [129]	-	Codes	Mar-2023
the Pile [130]	800GB	Other	Dec-2020
ROOTS [131]	1.6TB	Other	Jun-2022

(Xin Zhao et al., 2023)

Dataset Distribution

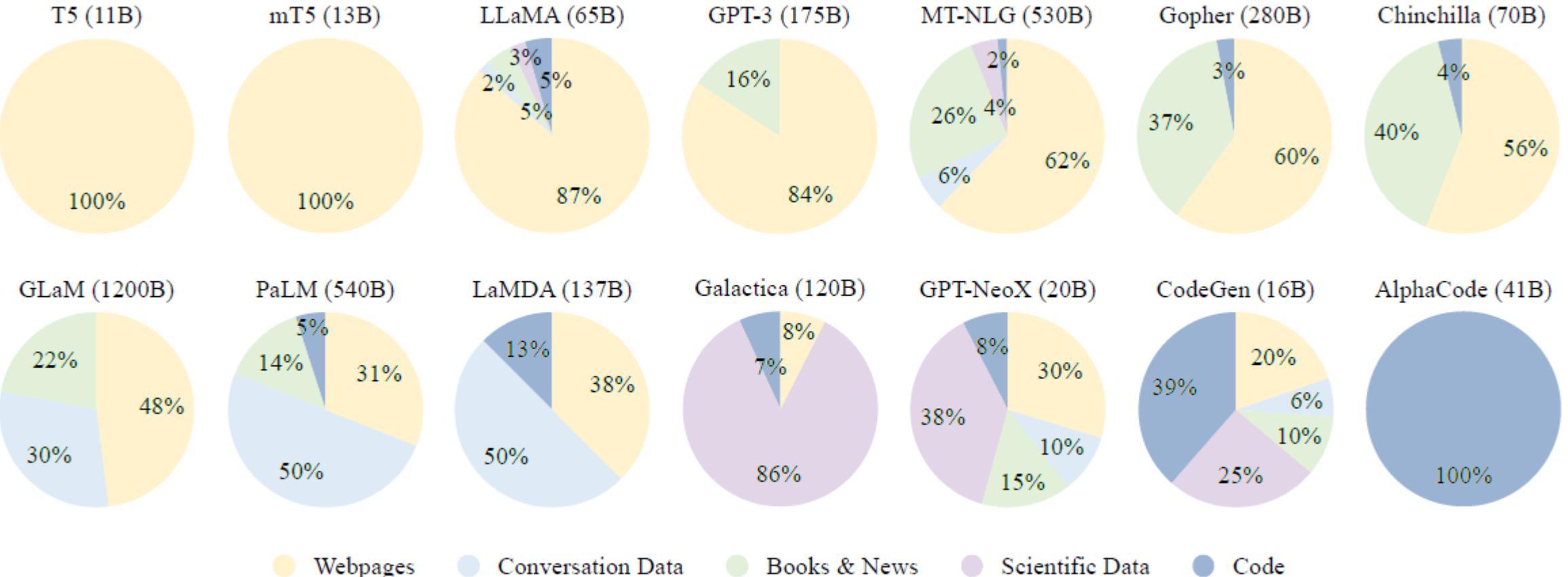
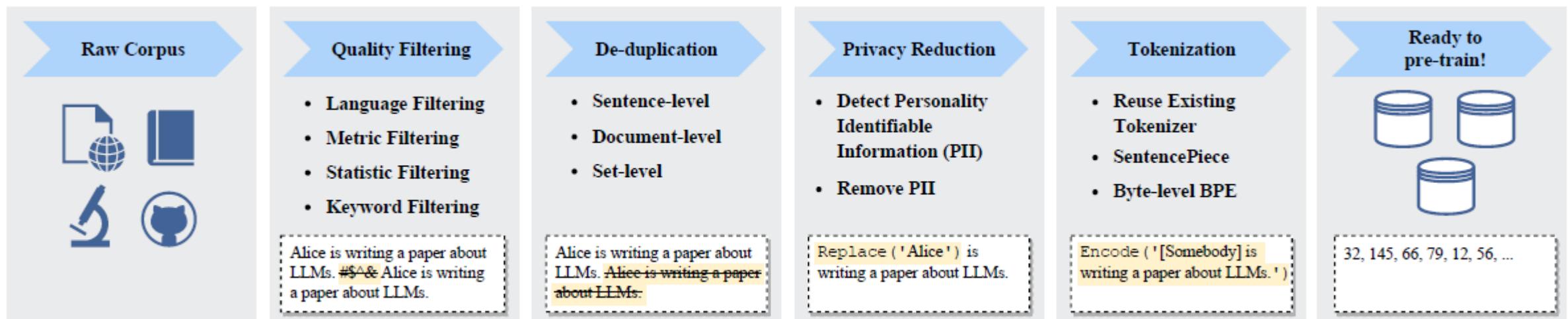


Fig. 2. Ratios of various data sources in the pre-training data for existing LLMs.

(Xin Zhao et al., 2023)

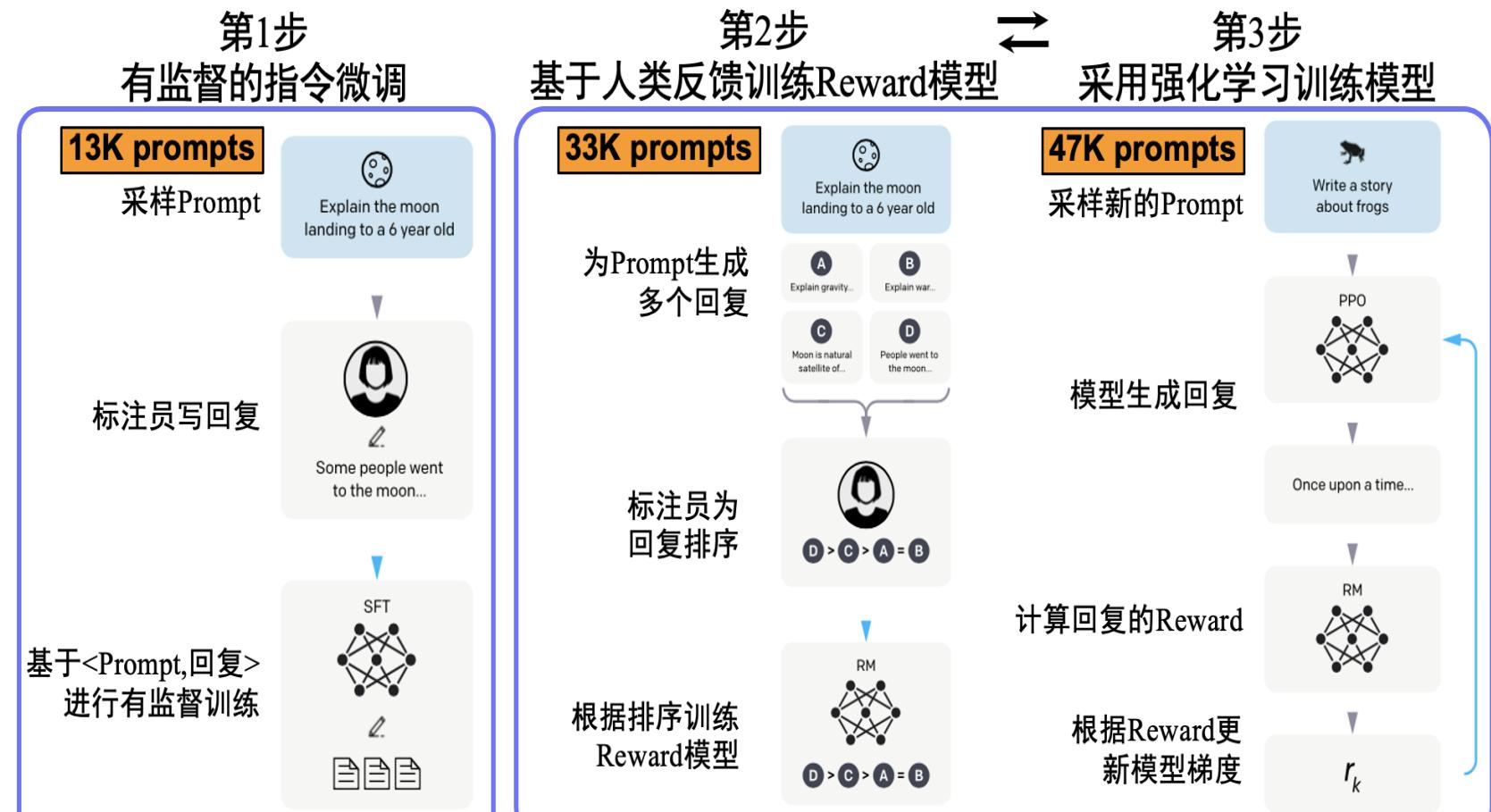
Dataset Preprocessing

- InstructGPT: 这是GPT-3的一个变体，专门针对用户指令进行了优化。它在**2021年底**左右被公开介绍。
- WebGPT: 截至**2022年1月**，WebGPT还处于研究和开发阶段。具体的发布时间可能在2022年之后。
- ChatGPT 3.5: **2022年年底发布**



(Xin Zhao et al., 2023)

The Main Architecture of GPT



有监督指令微调

基于人工反馈的强化学习 (RLHF)

米尔阿迪力江

<http://www.miradeljan.com>

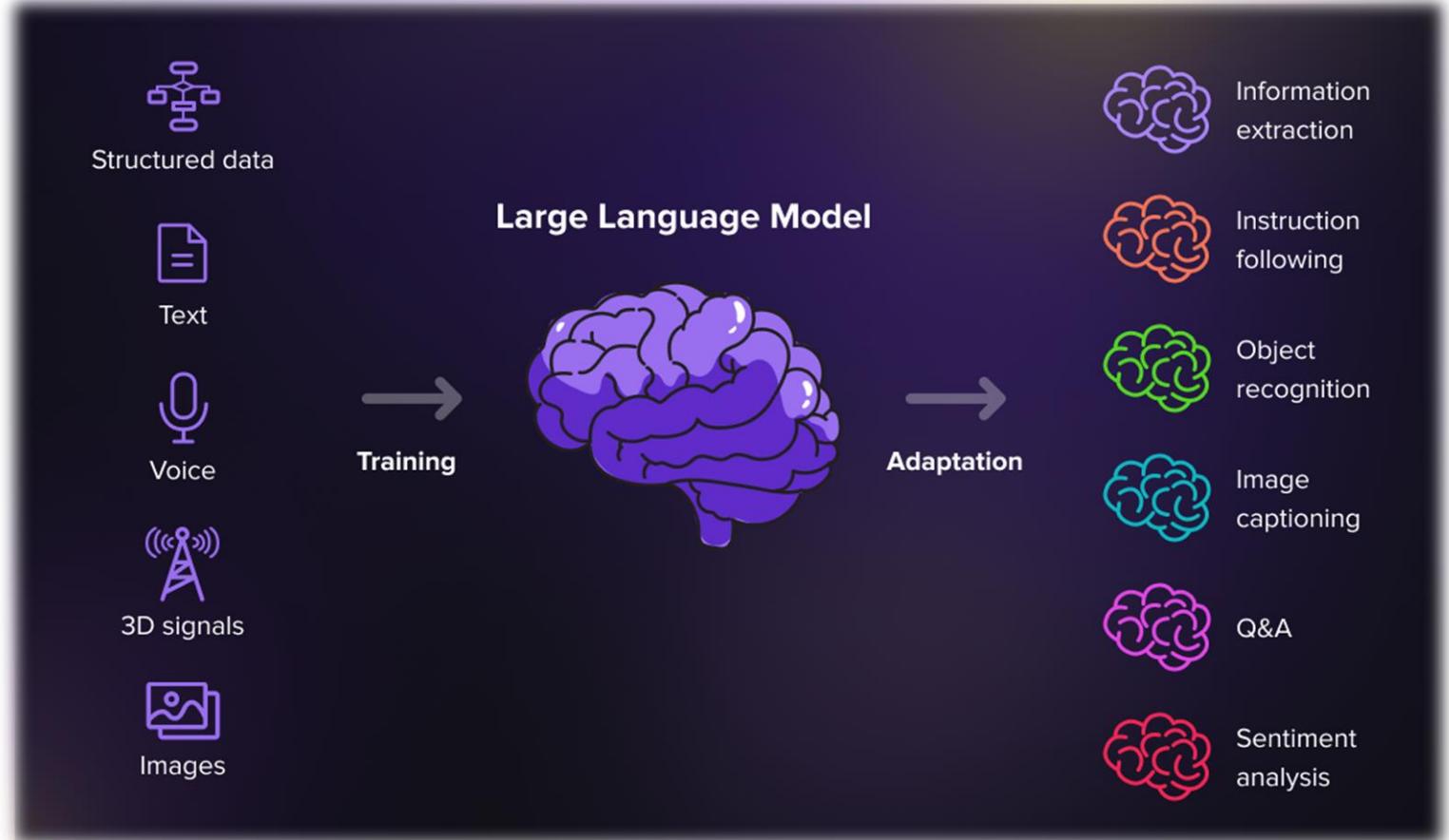


Outline

- Introduction of LLM
- LLM Performance in NLP
- Fine-Tuning Strategies & Challenges for LLM
- Machine Translation with LLMs
- LRLs Machine Translation with LLM
- Future of NLP with LLMs

LLM performance in NLP

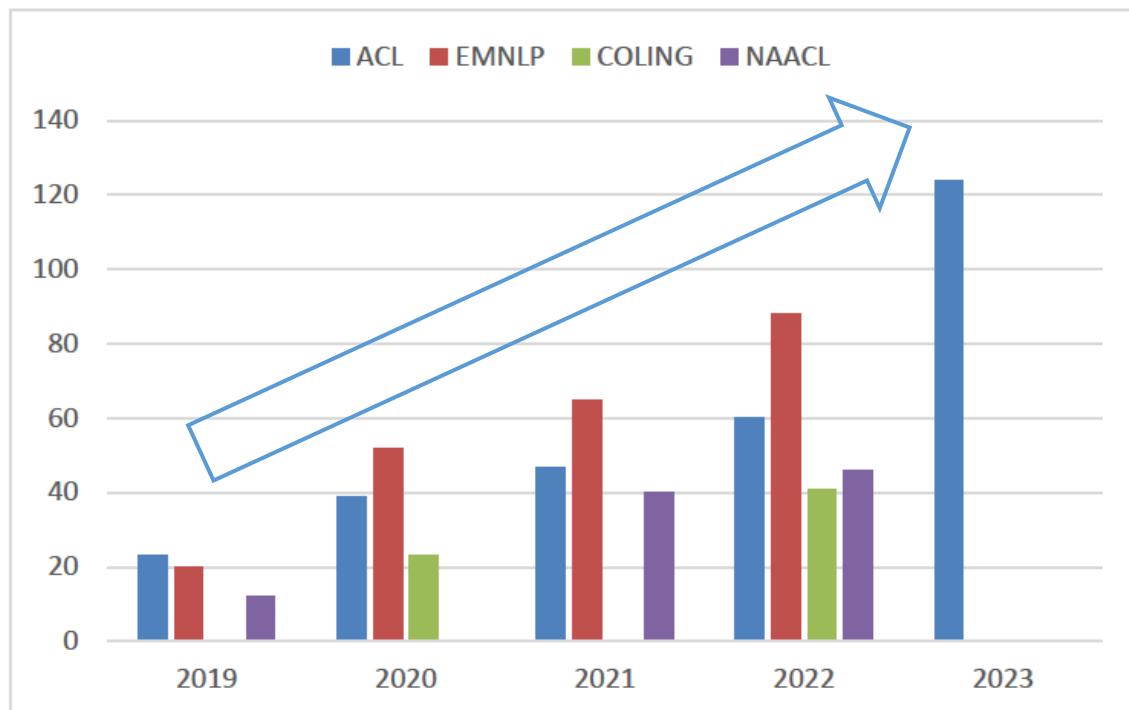
LLM performance on NLP down-stream tasks



<https://blog.gopenai.com/>

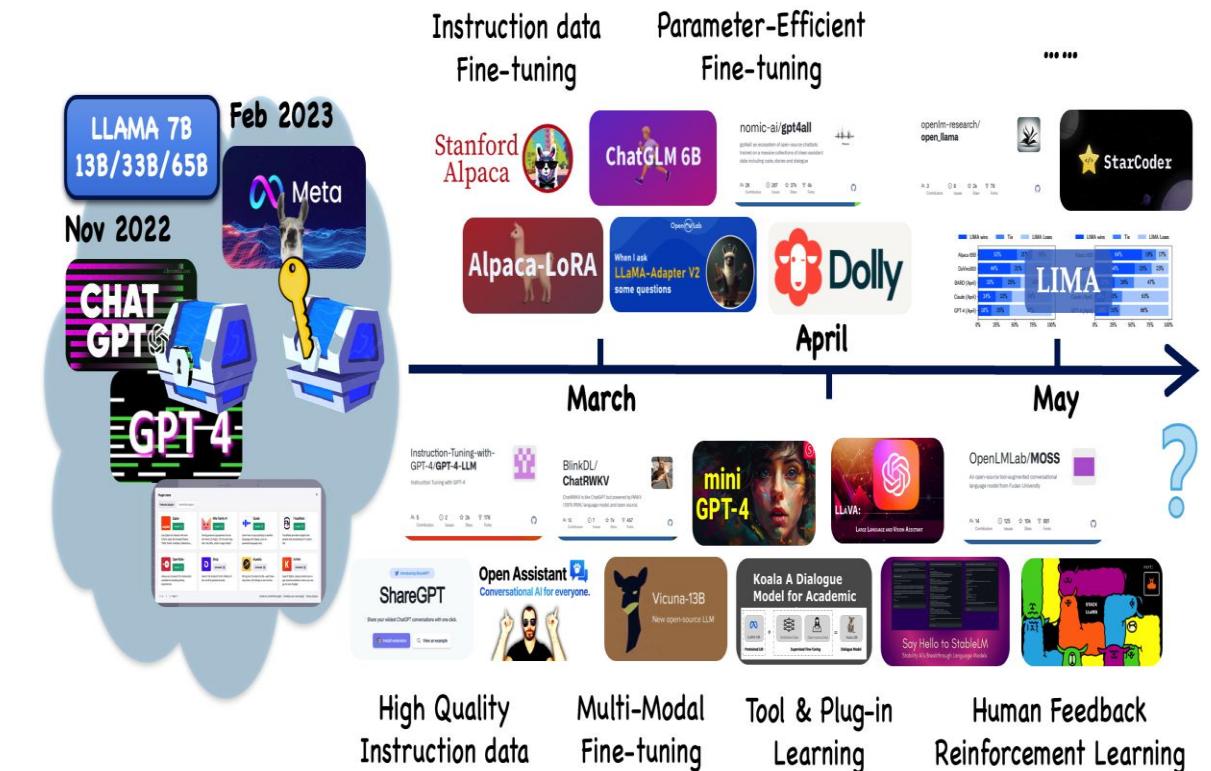
Basic Fact - foundational models is rapidly evolving

语言模型论文粗略统计



相关信息统计自ACL Anthology: <https://www.aclweb.org/anthology/>

一堆指令微调模型层出不穷



(Ding et al., CCL2023)

LLM Everywhere --- Conversation



You

能跟我聊一会儿吗?



ChatGPT

当然可以，很高兴与您聊天。您有什么想谈论的话题吗？无论是科技、文化、旅行还是日常生活的点滴，我都乐于倾听和参与讨论。您可以随时开始。



You

你问我一个问题吧



ChatGPT

好的，让我来问一个问题：您最近读过的一本书是什么？您对它的看法如何？



You

创业有关但是名字忘了



ChatGPT

关于创业的书籍通常包含很多有价值的见解和经验。即使忘记了书名，您能分享一下从这本书中学到的最重要的一两个点吗？比如关于创业的策略、创新思维或者是成功和失败的教训等。这些内容往往能够启发人们对于创业或职业生涯有更深的思考。



ChatGPT 4o ▾

共享

Li

新疆昌吉学院如何哇？你怎么看这所学校啊？



新疆昌吉学院 (Changji University) 位于新疆昌吉回族自治州，是一所以本科教育为主的地方性高校。以下是我对这所学校的一些客观评价，希望能帮助你了解：

优点：

1. 区域优势：

昌吉地处新疆的核心区域，距离乌鲁木齐较近，交通较为便利。学校也因地处民族区域，具有多民族文化的融合特点，对学生的包容性较强。

2. 专业设置：

学校的重点专业集中在师范教育、农业科技、机械工程等领域，尤其注重培养地方所需的实用型人才。

- 对于立志从事教师职业的学生，学校的师范类专业有一定优势，很多毕业生在新疆本地中小学就业。

3. 校园环境：

学校位于昌吉州，周边自然风光优美，环境相对安静，适合学习与生活。

4. 政府扶持：

新疆近年来在教育方面有较大投入，昌吉学院作为地方高校，得到了政策和资金的支持，尤其是面向少数民族和当地经济发展的专业。

LLM Everywhere --- Conversation, QA

新疆昌吉学院什么时候成立的?

新疆昌吉学院是什么时候成立的?



已完成搜索

昌吉学院的前身是成立于1959年的昌吉师范学校。随后，学校经历了几次重要的发展阶段：

- 1985年，昌吉师范学校升格为昌吉师范专科学校。
- 2001年5月，经国家教育部和自治区人民政府批准，昌吉师范专科学校升格为本科院校，正式命名为昌吉学院。

因此，昌吉学院作为一所本科院校是在2001年成立的，但其教育历史可以追溯到1959年。

相关链接 · 8

深度搜索



阿里
通义



已搜索 5 个网站 ▾

新疆昌吉学院的前身是1959年成立的昌吉师范学校。随后，1985年升级为昌吉师范专科学校，2001年5月正式升格为本科院校昌吉学院【12】【13】【14】。

GPT 4o

这所学校位于昌吉市，紧邻乌鲁木齐，享有便利的地理位置和优美的自然环境。经过多年发展，昌吉学院已从以师范教育为主的专科院校，逐步成长为涵盖多个学科的综合性高校。学校注重地方经济与教育的结合，为当地社会发展培养了大量优秀人才【15】。



新疆昌吉学院是啥时候成立的?



Kimi



新疆昌吉学院是什么时候成立的?



新疆昌吉学院成立于2001年5月，其前身是成立于1959年的昌吉师范学校。2001年5月，昌吉师范学校升格为昌吉学院，成为区属高等本科院校，由教育部和新疆维吾尔自治区人民政府批准设立。学院位于新疆天山北坡经济带、乌昌石城市群核心发展区、丝绸之路经济带新疆核心区的昌吉市，拥有园丁校区和新城校区两个校区。

参考4条信息源

1. 昌吉学院

百度

2. 昌吉学院

掌上高考

3. 昌吉学院

百度百科

由文心大模型 3.5 生成

百度
文心

LLM Everywhere --- Conversation, QA



SparkDesk 阅书万卷，我是你的知识宝典

新疆昌吉学院是什么时候成立的?

使用工具：联网搜索

新疆昌吉学院成立于1959年。

昌吉学院的前身是昌吉师范学校，这所学校在1985年升格为昌吉师范专科学校。到了2001年5月，经过教育部批准，学校升格为普通本科院校，并更名为昌吉学院。作为一所省属普通本科院校，昌吉学院坐落在新疆维吾尔自治区昌吉市，由新疆维吾尔自治区人民政府举办，新疆维吾尔自治区教育厅主管。

基于 10 个搜索来源

昌吉学院的前身是成立于 1959 年的昌吉师范学校 ① ② ③ 。1985 年升格为昌吉师范专科学校，2001 年 5 月经国家教育部和自治区人民政府批准升格为普通本科院校——昌吉学院 ① ② ③ 。

相关视频

- 考上海昌吉学院学哪个专业最好？-抖音
- 昌吉学院——新疆一流应用型大学-抖音
- 昌吉学院介绍-抖音



由星火生成

ChatGLM 清华大学

讯飞星火

+ 创建新对话

新疆昌吉学院是什么时候成立的?

2024-10-22 20:36:57

很抱歉，根据您提供的参考资料，没有包含关于新疆昌吉学院成立时间的信息。因此我无法回答这个问题。如果您需要了解新疆昌吉学院的成立时间，建议您查询该学院的官方网站或相关历史资料。

以上内容为 AI 生成，不代表开发者立场



腾讯元宝

元宝

引用 8 篇资料作为参考

新疆昌吉学院是什么时候成立的?

新疆昌吉学院成立于2001年，其前身为成立于1959年的昌吉师范学校 ① 。以下是关于昌吉学院的相关信息：

昌吉学院的历史沿革

- 1959年：昌吉师范学校成立。
- 1985年：升格为昌吉师范专科学校。
- 2001年5月：经国家教育部和自治区人民政府批准升格为普通本科院校——昌吉学院 ① 。

LLM Everywhere --- Conversation, QA, Outline



You

给我提供一个大模型概要有关的提纲



ChatGPT

当然，这里是一个关于大型机器学习模型（如GPT、BERT等）的概要提纲。这个提纲将帮助您深入了解大型模型的基本概念、架构、应用场景以及它们在未来发展中可能面临的挑战和机遇。

1. 引言

- 大型模型在人工智能领域的兴起背景
- 大型模型的定义和重要性

2. 大型模型的基本原理和架构

- 深度学习和神经网络基础
- 大型模型的核心组件（如变换器架构）
- 数据和算法在模型训练中的作用

3. 主要大型模型的发展历程

- 不同代的大型模型（如GPT系列、BERT、T5等）的发展和比较
- 关键技术突破和创新点

4. 大型模型的应用场景

- 自然语言处理（NLP）
- 图像识别和处理
- 其他领域的应用（如医疗、金融、游戏等）

5. 大型模型的优势和挑战

- 处理复杂问题和大数据集的能力
- 模型偏见和伦理问题
- 计算资源和环境影响
- 数据隐私和安全性问题

6. 未来发展趋势

- 技术创新和模型优化



You

帮我写一个slides提纲，有关机器学习基础与实践。



ChatGPT

当然，以下是一个关于“机器学习基础与实践”的幻灯片提纲。这个提纲将帮助您设计一个包含机器学习的基本概念、关键技术方法、实际应用案例，以及未来发展趋势的完整演示。

1. 幻灯片 1: 标题页

- 标题：机器学习基础与实践
- 副标题：从理论到应用
- 演讲者姓名和机构

2. 幻灯片 2: 介绍

- 机器学习的定义和重要性
- 本次演讲的目的和结构

3. 幻灯片 3: 机器学习的历史简介

- 从最初的概念到现代的发展
- 关键技术突破和里程碑事件

4. 幻灯片 4: 机器学习的类型

- 监督学习、非监督学习、半监督学习、强化学习
- 每种类型的基本概念和应用示例

5. 幻灯片 5: 关键算法和模型

- 决策树、支持向量机、神经网络、深度学习等
- 每种算法的基本原理和适用场景

6. 幻灯片 6: 数据处理和特征工程

- 数据预处理的重要性
- 特征选择和特征提取
- 实例展示数据处理流程

7. 幻灯片 7: 机器学习的实际应用案例



LLM Everywhere --- Conversation, QA, Outline, Recommend



You

如果想从事人工智能开发行业的话，对一个新人你有什么建议呀？



ChatGPT

从事人工智能（AI）开发是一个充满挑战和机遇的领域。对于新人来说，以下是一些建议：

1. **基础教育**: 学习计算机科学、数学和统计学的基础知识。对于人工智能特别是机器学习而言，线性代数、概率论、微积分和算法是非常重要的。
2. **编程技能**: 精通至少一种编程语言，Python 是最受欢迎的语言之一，因为它拥有大量的AI和机器学习库，例如TensorFlow和PyTorch。
3. **在线课程和资源**: 利用在线平台如Coursera、edX、Udacity等，参加AI和机器学习的课程。
4. **实际项目**: 通过实际项目来应用你的知识。GitHub上有很多开源项目，你可以参与其中来提高你的技能。
5. **理解业务**: 了解AI技术如何解决实际问题，并对其可能带来的业务和社会影响有所认识。
6. **读原始论文**: 阅读领域内的经典和最新论文，如在arXiv上发布的论文，以保持对最新研究动态的了解。
7. **网络和社区**: 参加Meetups、Conferences、Workshops等，建立职业网络，与同行交流。
8. **批判性思维**: 培养批判性思维能力，对数据、模型和结果进行仔细分析。
9. **伦理和法律知识**: 理解与人工智能相关的伦理和法律问题，例如数据隐私、算法偏见等。
10. **持续学习**: AI领域在快速发展，持续学习新技术、新算法是非常重要的。

记住，建立一个坚实的基础，然后不断地学习和实践是关键。每个人的学习路径都是独特的，找到最适合你的方式去探索和成长。



You

是非这个行业的，想转行做人工智能，有啥建议的学习方向吗？



ChatGPT

如果您想从其他行业转行进入人工智能（AI），以下是一些建议的学习方向和步骤：

1. **学习基础**: 首先，您需要建立在数学（特别是线性代数、概率论和统计学）和编程（推荐Python）方面的坚实基础。这些是深入理解AI的前提。
2. **了解AI的子领域**: 人工智能包含了机器学习、深度学习、自然语言处理、计算机视觉等子领域。初步了解这些子领域，确定您最感兴趣的领域。
3. **参加在线课程**: 参加Coursera、edX、Udacity等平台上的入门课程，例如Andrew Ng的《机器学习》课程，或者Google的机器学习速成课程。
4. **项目实践**: 通过小项目开始实践，比如使用Python编写一个简单的数据分析脚本，或者使用机器学习库如scikit-learn来构建简单模型。
5. **深化专业知识**: 在有了基础之后，可以通过专门的书籍、课程和项目来深化特定领域的知识，例如深度学习、强化学习或者自然语言处理。
6. **参与开源项目**: GitHub上有许多AI相关的开源项目，参与这些项目能够帮助您理解实际应用中的问题，并提高解决问题的能力。
7. **建立网络**: 加入AI相关的论坛和社区，如Reddit的r/MachineLearning、Kaggle等，与其他从业者交流心得。
8. **了解伦理和社会影响**: AI技术的使用会带来伦理和社会问题，理解这些问题对于负责任地开发和应用AI技术很重要。
9. **求知若渴**: AI是一个不断发展的领域，保持好奇心和学习的渴望，不断更新您的知识和技能。
10. **思考如何结合自己的背景**: 思考您之前的行业经验和知识如何与AI结合。许多AI应用是跨领域的，您之前的专业知识可能会成为宝贵的资产。

转行到AI领域可能既充满挑战也富有成就感。重要的是保持耐心，不断学习，并乐于解决问题。



Message ChatGPT...



ChatGPT can make mistakes. Consider checking important information.

LLM Everywhere --- Conversation, QA, Outline, Recommend, MT

You

"我喜欢新疆北疆，景色非常迷人，特别美丽，不过从没去过。"\n麻烦把以上这几句话翻译成英语，德语，法语，阿拉伯语，土耳其语，泰语等几个多语种。

ChatGPT

当然，以下是您提供的句子翻译成不同语言的版本：

英语：

"I really like the northern part of Xinjiang. The scenery is very enchanting and incredibly beautiful, though I've never been there."

德语：

"Ich mag den Norden von Xinjiang sehr. Die Landschaft ist sehr bezaubernd und unglaublich schön, obwohl ich noch nie dort war."

法语：

"J'aime beaucoup le nord du Xinjiang. Le paysage est très charmant et incroyablement beau, bien que je n'y sois jamais allé."

阿拉伯语：

"أُحِبُّ شَمَالْ شِينْجِيَانْغْ كَثِيرًا. الْمَنَاظِرُ طَبِيعِيَّةٌ سَاحِرَةٌ وَجَمِيلَةٌ لِلْغاِيَةِ، عَلَى الرَّغْمِ مِنْ أَنِّي لَمْ أَذْهَبْ إِلَى هَنَاكَ قَطْ."

土耳其语：

"Xinjiang'in kuzeyini çok seviyorum. Manzara çok büyüleyici ve inanılmaz güzel, ama hiç oraya gitmedim."

泰语：

"ฉันชอบภาคเหนือของชิงเจียงมาก ทิวทัศน์ที่นั่นมีเสน่ห์และสวยงามมาก แม้ว่าฉันจะไม่เคยไปที่นั่น."

这些翻译应该准确传达了原文的意思和情感。↓

You

把以上句子可以翻译成东南亚地区几个国家的语言吗？

ChatGPT

当然，我可以将您提供的句子翻译成几种东南亚地区国家的语言：

印度尼西亚语：

"Saya sangat suka bagian utara Xinjiang. Pemandangannya sangat memukau dan sangat indah, meskipun saya belum pernah ke sana."

马来语：

"Saya sangat suka bahagian utara Xinjiang. Pemandangannya sangat menarik dan sangat cantik, walaupun saya belum pernah ke sana."

越南语：

"Tôi rất thích phần phía bắc của Tân Cương. Cảnh quan ở đó rất hấp dẫn và cực kỳ đẹp, mặc dù tôi chưa từng đến đó."

泰语 (重复)：

"ฉันชอบภาคเหนือของชิงเจียงมาก ทิวทัศน์ที่นั่นมีเสน่ห์และสวยงามมาก แม้ว่าฉันจะไม่เคยไปที่นั่น."

菲律宾语：

"Gusto ko talaga ang hilagang bahagi ng Xinjiang. Ang tanawin ay napaka-kaakit-akit at lubos na maganda, bagaman hindi pa ako nakapunta doon."

这些翻译尽可能准确地保持了原文的含义和情感。



LLM Everywhere --- Conversation, QA, Outline, Recommend, MT, Coding



You

麻烦给我写一个python脚本，需要把tsv格式的文件转换成json格式的文件。



ChatGPT

当然，我可以为您提供一个Python脚本来将TSV格式的文件转换成JSON格式的文件。以下是脚本的一个基本示例：

```
python

import csv
import json

# 将TSV文件转换为JSON文件的函数
def tsv_to_json.tsv_file_path, json_file_path):
    # 读取TSV文件
    with open.tsv_file_path, 'r', encoding='utf-8') as tsvfile:
        reader = csv.DictReader(tsvfile, delimiter='\t')

    # 创建一个列表来保存所有行的数据
    data = []
    for row in reader:
        data.append(row)

    # 写入JSON文件
    with open(json_file_path, 'w', encoding='utf-8') as jsonfile:
        json.dump(data, jsonfile, indent=4, ensure_ascii=False)

    # 使用示例
    tsv_file = 'example.tsv' # TSV文件的路径
    json_file = 'output.json' # 输出JSON文件的路径
    tsv_to_json.tsv_file, json_file)
```



You

我不喜欢用with open格式来文件操作，重写一下！



ChatGPT

当然，我可以按照您的要求修改脚本，不使用`with open`来进行文件操作。请注意，虽然`with open`是推荐的方式，因为它可以自动处理文件的关闭，但是如果您不想使用它，可以直接使用`open`和`close`来手动管理文件的打开和关闭。下面是修改后的脚本：

```
python

import csv
import json

def tsv_to_json.tsv_file_path, json_file_path):
    # 打开TSV文件
    tsvfile = open.tsv_file_path, 'r', encoding='utf-8')
    reader = csv.DictReader(tsvfile, delimiter='\t')

    # 创建一个列表来保存所有行的数据
    data = []
    for row in reader:
        data.append(row)

    # 关闭TSV文件
    tsvfile.close()

    # 打开JSON文件
    jsonfile = open(json_file_path, 'w', encoding='utf-8')
    json.dump(data, jsonfile, indent=4, ensure_ascii=False)

    # 关闭JSON文件
    jsonfile.close()

    # 使用示例
    tsv_file = 'example.tsv' # TSV文件的路径
    json_file = 'output.json' # 输出JSON文件的路径
    tsv_to_json.tsv_file, json_file)
```

LLM Everywhere --- Conversation, QA, ... , MT, Coding, Reading Paper

ChatGPT 4o ▾

模型

GPT-4o

适用于大多数任务

GPT-4o with canvas 测试版

在写作和代码方面开展协作

o1-preview

使用高级推理

o1-mini

推理速度更快

模型更多

临时聊天

GPT-4o mini

适合处理日常任务

GPT-4

传统模型

ChatGPT 4o ▾

共享

Li



TooILM.pdf
PDF

what is the main idea of this paper? BTW, Could you please summarize the pros and cons of this paper?

Main Idea of the Paper:

The paper "ToolLLM: Facilitating Large Language Models to Master 16000+ Real-World APIs" introduces a novel framework aimed at enhancing the tool-use capabilities of open-source large language models (LLMs), particularly focusing on API integration. The authors propose ToolLLaMA, an enhanced version of LLaMA, trained using their ToolBench dataset. ToolBench includes instructions and real-world APIs, generated and annotated with the help of ChatGPT. They also propose a **decision tree-based reasoning strategy (DFSDT)** to improve complex decision-making and reasoning. A neural API retriever is integrated to recommend relevant APIs for each task, making the model effective in multi-tool scenarios.

Pros:

1. Extensive API Collection:

- The paper collects over 16,000 real-world APIs across 49 categories, enabling LLMs to generalize to diverse, practical applications.

2. Multi-Tool Scenarios:

- ToolLLaMA can manage complex tasks involving multiple tools, surpassing existing solutions that only support single-tool operations.

3. Depth-First Search Decision Tree (DFSDT):

LLM Everywhere --- Conversation, QA, ... , MT, Coding, Reading Paper

Cons:

1. Dependence on ChatGPT for Dataset Construction:

- The reliance on ChatGPT for instruction generation and annotation raises concerns about bias or limitations inherent in ChatGPT's outputs.

2. High Computational Costs:

- Using DFSDT and multiple API calls increases computational overhead, especially compared to simpler baseline methods like ReACT.

3. Context Length Limitation:

- Despite extending context length to 8192 tokens, some long API responses may still exceed the limit, requiring further truncation or compression.

4. Complex Implementation:

- The multi-step reasoning process and decision-tree-based strategy may complicate practical implementation and increase latency.

5. Limited Evaluation Metrics:

- While pass and win rates provide useful insights, other aspects like user experience or system efficiency are not thoroughly evaluated.

This summary captures the core contributions and trade-offs of the ToolLLM framework. Let me know if you need more detailed insights or further analysis on specific sections!

please summarize shortly

Main Idea:

The paper introduces **ToolLLM**, a framework to improve large language models' (LLMs) ability to use over 16,000 real-world APIs. It develops **ToolLLaMA**, trained on a new dataset called **ToolBench**. The framework integrates a **neural API retriever** and introduces a **DFSDT reasoning strategy** to handle complex, multi-tool tasks efficiently.

Pros:

- Supports complex multi-tool scenarios.
- Strong generalization to unseen APIs.
- DFSDT improves decision-making compared to ReACT.
- API retriever automates tool selection.
- Publicly available dataset and code for community use.

Cons:

- Relies heavily on ChatGPT for dataset creation.
- High computational cost due to DFSDT and API calls.
- Some responses may exceed the model's context length.
- Implementation complexity increases latency.
- Limited focus on efficiency and user experience metrics.

Let me know if you need further clarifications!

LLM Everywhere --- Conversation, QA, … , MT, Coding, Reading Paper

please translate the shortly summarization into the
chinese!

主要内容:

该论文介绍了**ToolLLM**框架，用于提升大型语言模型（LLM）使用超过16,000种真实世界API的能力。通过新数据集**ToolBench**训练出**ToolLaMA**模型。该框架集成了神经API检索器，并引入了**DFSDT**推理策略，以有效处理复杂的多工具任务。

优点:

- 支持复杂的多工具场景。
- 对未见过的API具有强大的泛化能力。
- DFSDT比ReACT在决策上表现更佳。
- API检索器实现自动化工具选择。
- 数据集和代码公开，有助于社区创新。

缺点:

- 数据集构建过于依赖ChatGPT。
- DFSDT和API调用增加了计算成本。
- 部分响应可能超过模型的上下文长度。
- 实现复杂度高，增加了延迟。
- 缺乏对效率和用户体验的评估。

如需进一步说明，请告诉我！

给我推荐一下跟大模型tool learning 有关的最新的研究工
作（学术论文或者slides）~

已搜索 4 个网站 ~

Here are some of the latest research papers and resources on **tool learning with large language models (LLMs)**:

- "Tool Learning with Large Language Models: A Survey"**

This paper, published in May 2024, provides a comprehensive overview of how LLMs are augmented with external tools to handle complex tasks. It covers the full lifecycle of tool integration—task planning, selection, calling, and response generation. The paper also highlights several benchmarks and evaluation strategies and discusses the challenges and future directions in this field.

[Link to the survey](#) PAPERS WITH CODE.
- "Large Language Models as Tool Makers (LATM)"**

This research introduces a framework where powerful models like GPT-4 generate reusable tools (Python functions) from task demonstrations, which smaller models (e.g., GPT-3.5) can later utilize efficiently. The tool-making process involves proposing, verifying, and wrapping tools for reuse across various tasks. This setup demonstrates how tool creation and usage can improve efficiency and adaptability.

[Explore the LATM paper](#) ARXIV.

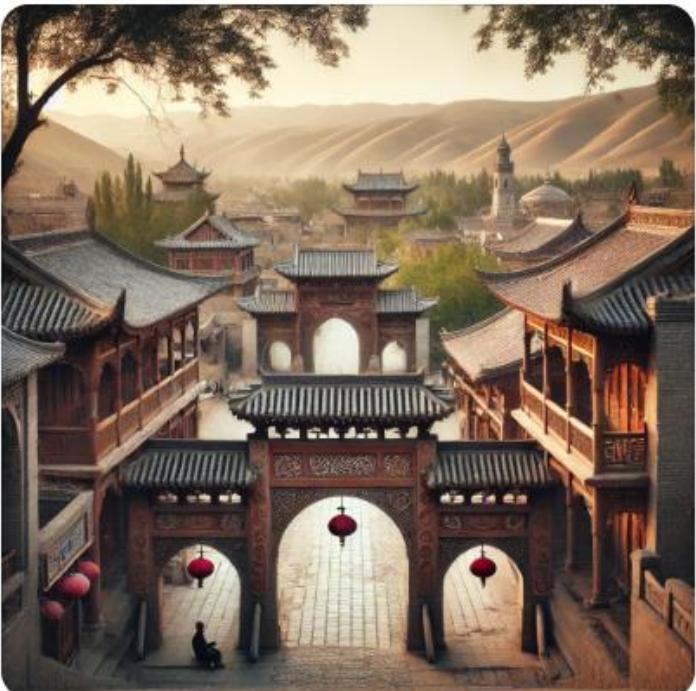
These resources represent cutting-edge directions in **tool learning**, focusing on creating reusable tools and improving reasoning abilities by augmenting LLMs with external functions or frameworks. This research demonstrates promising steps toward enhancing efficiency and flexibility for complex problem-solving.

LLM --- Image Generation (DALL-E)

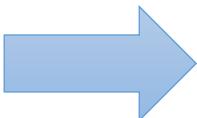


请给我生成一个昌吉古城的照片?

我觉得不太像古城，应该路边有一些小店铺吧！



已根据描述生成了一张反映新疆昌吉古城风貌的图片。请查看上方图像，如果还有其他需求或修改，请告诉我！



已根据您的描述更新了昌吉古城的场景，加入了街边小店和市集氛围。请查看上方图像，如有任何进一步的修改需求，请告诉我！

LLM --- Image Generation (DALL-E)



Currently, I cannot directly edit or combine the uploaded image with a new gate. However, you could take the following steps to achieve the desired result:

1. Use a Photo Editing Tool:

- Tools like **Photoshop**, **GIMP** (free), or **Canva** allow you to add a gate image as an overlay on top of the original image.
- You can select and position the gate at the entrance to align it perfectly with the street layout.

2. AI-based Image Editors:

- You could use **DALL-E's inpainting** feature to introduce a gate directly into the existing image. Inpainting allows you to modify parts of an image seamlessly by masking an area where you want the gate to appear.

3. Send Back to Me:

- If you'd like, you can use a masking tool to mark where you want the gate, upload that version, and I can generate a gate-focused image overlay using DALL-E to fit that space.

This way, you can have the entrance gate integrated professionally and aligned with the lively street view of Changji. Let me know if you'd like specific steps for one of the tools or more help with inpainting instructions!



我想在这张图的基础上加一扇大门，当作古城的入口，把这张图给我改一下~

我想要一张新疆昌吉著名建筑“昌吉古城”傍晚的样子！



这是一幅展现新疆昌吉古城傍晚景色的图像。夕阳的余晖洒在传统的中式建筑上，瓦顶街道旁悬挂着红色灯笼，温暖的光线为街景增添了一种怀旧且舒适的氛围。小商铺仍在营业，人们悠闲地在街上散步。背景中的山峦与城市景观自然融合，天色在橙色与紫色之间过渡，捕捉到了傍晚时分的美妙时光。

如有更多调整或进一步需求，请告诉我！

Training Procedure of LLM (GPT)

	预训练	有监督微调	奖励建模	强化学习
数据集合	原始数据 数千亿 单词：图书、百科、网页等	标注用户指令 数万 用户指令和对应的答案	标注对比对 百万量级 标注对比对	用户指令 十万量级 用户指令
算法	语言模型训练	语言模型训练	二分类模型	强化学习方法
模型	基础模型	SFT 模型	RM 模型	RL 模型
资源需求	1000+GPU 月级别训练时间	1-100GPU 天级别训练时间	1-100GPU 天级别训练时间	1-100GPU 天级别训练时间

(Prof. Huang., CCL2023)

Knowledge Representation

	信息存储方式	表示方式的精确度	知识调用方式	调用方式的自然度	研究领域	代表应用	代表公司
数据库	结构化	高	SQL语句	低	数据库	DBMS	Oracle Microsoft
互联网	非结构化	中	关键词	中	信息检索	搜索引擎	Google Microsoft
大模型	参数化	低	自然语言	高	自然语言处理	ChatGPT	OpenAI Microsoft Google

(Prof. Liu., CCL 2023)

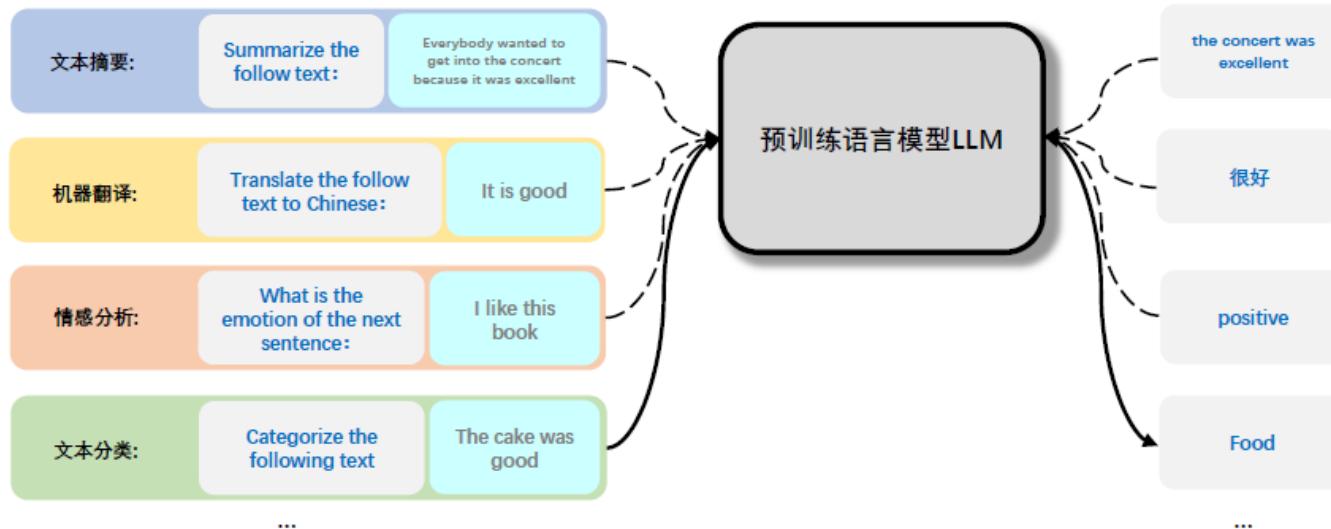
Outline

- Introduction of LLM
- LLM Performance in NLP
- Fine-Tuning Strategies & Challenges for LLM
- Categorization of Tools
- Benchmarking Tool Use
- Open-source Tool-Kit

Fine-Tuning Strategies & Challenges for LLM

What Is the Instruction Tuning?

- Instruct tuning指在**多个任务上**，使用自然语言的形式构建指令数据集，并对模型进行**微调**
- Instruct通常由**人类的自然语言描述与任务输入拼接而成**
- 根据人工标注的指令形式，可分为学术性指令和人工自然语言指令



上图为Instruct tuning示例，训练阶段（虚线）在机器翻译、情感分析等多个任务上构造指令微调LLM，推断阶段（实线）可泛化到文本分类任务上

学术性指令

Template 1

<premise>
Based on the paragraph above, can we conclude that <hypothesis>?
<options>

Template 2

<premise>
Can we infer the following?
<hypothesis>
<options>

Template 3

Read the following and determine if the hypothesis can be inferred from the premise:
Premise: <premise>
Hypothesis: <hypothesis>
<options>

Template 4, ...

人工自然语言指令

Use-case Prompt

Brainstorming	List five ideas for how to regain enthusiasm for my career
Generation	Write a short story where a bear goes to the beach, makes friends with a seal, and then returns home.

Rewrite	This is the summary of a Broadway play:
---------	---

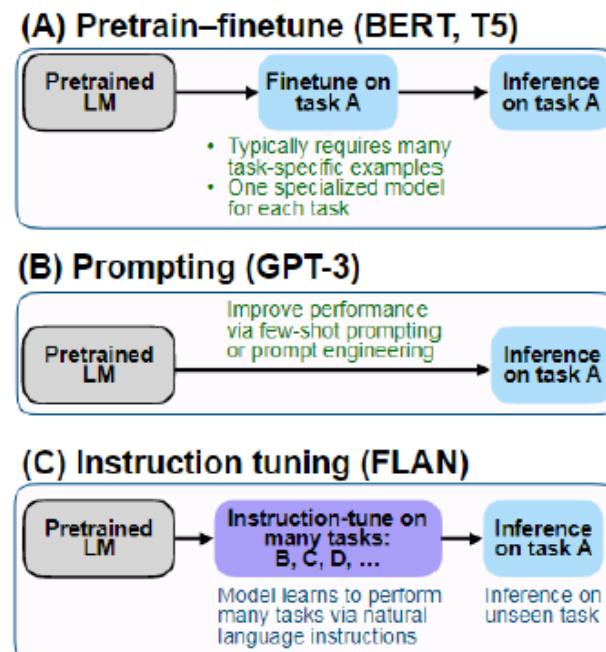
[summary]

This is the outline of the commercial for that play:
....

内容来源：小牛翻译-生成式LLM系列直播, 5期, 刘新宇

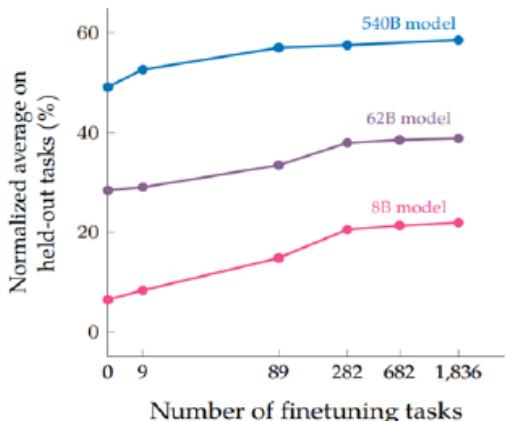
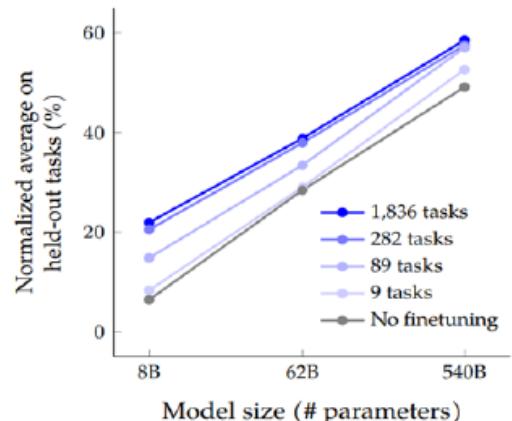
Why Instruction Tuning?

- 使大模型的输入与人类习惯“对齐”
- 指令微调可以提高模型的指令理解能力，使得模型能根据输入的Instruction自动识别任务类型，并执行对应的操作
- Instruct tuning区别于fine-tuning和prompting，需要标注数据更少，可在未见任务上指令泛化



Instruction tuning 与 fine-tuning 和 prompt 对比

研究表明，当模型规模越大，微调时见过的指令任务种类越多，其泛化能力越强。

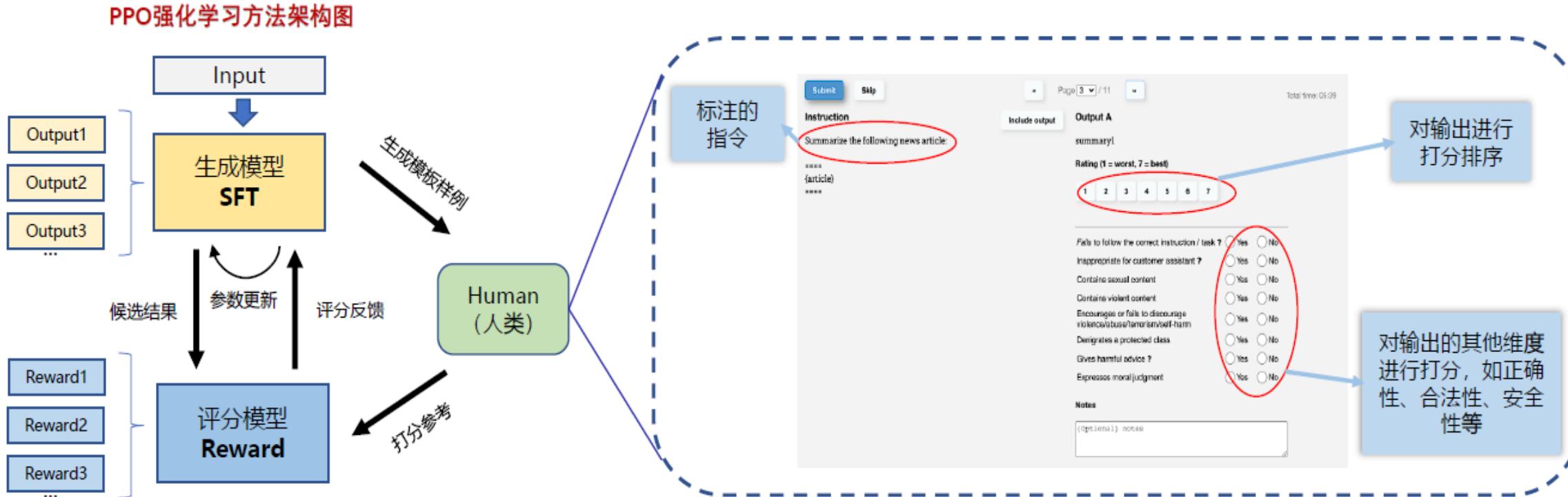


Flan-PaLM模型在不同scale下进行instruct tuning性能对比

内容来源：小牛翻译-生成式LLM系列直播, 5期, 刘新宇

What Is the RLHF (Reinforcement Learning with Human Feedback)?

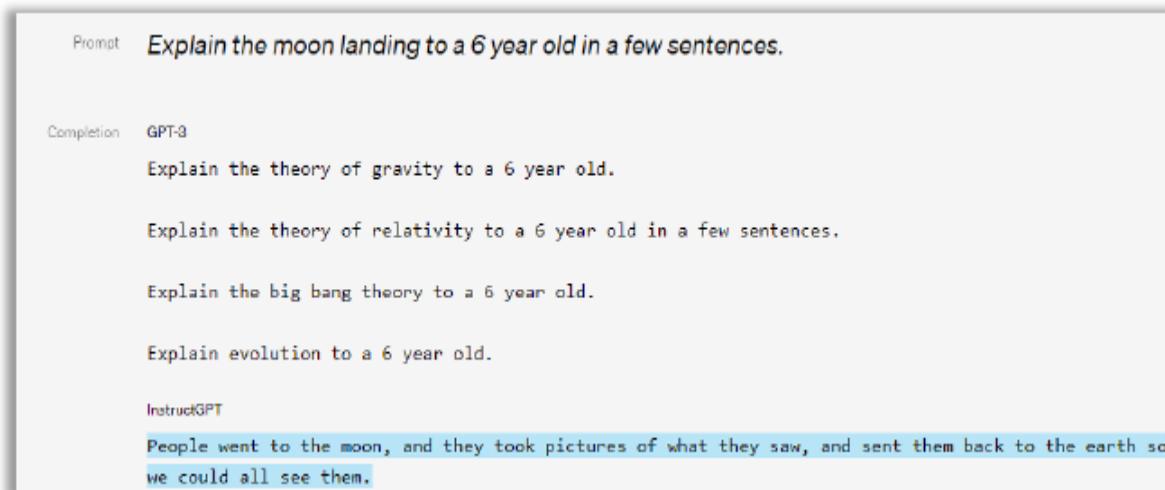
- RLHF是一种**基于人类反馈的强化学习方法**，使用基于Agent的强化学习框架进行模型参数的更新
- 通过交互界面进行打分，使得模型可以利用人类反馈进行学习
- **评分模型**训好后，可与**生成模型**形成迭代闭环，完成LLM的训练



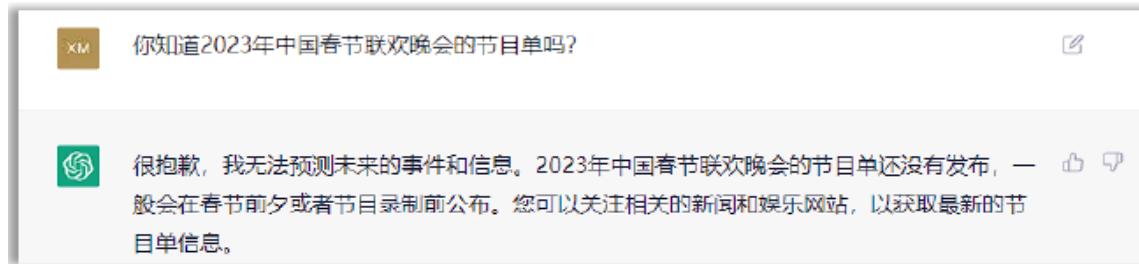
内容来源：小牛翻译-生成式LLM系列直播, 5期, 刘新宇

Why RLHF?

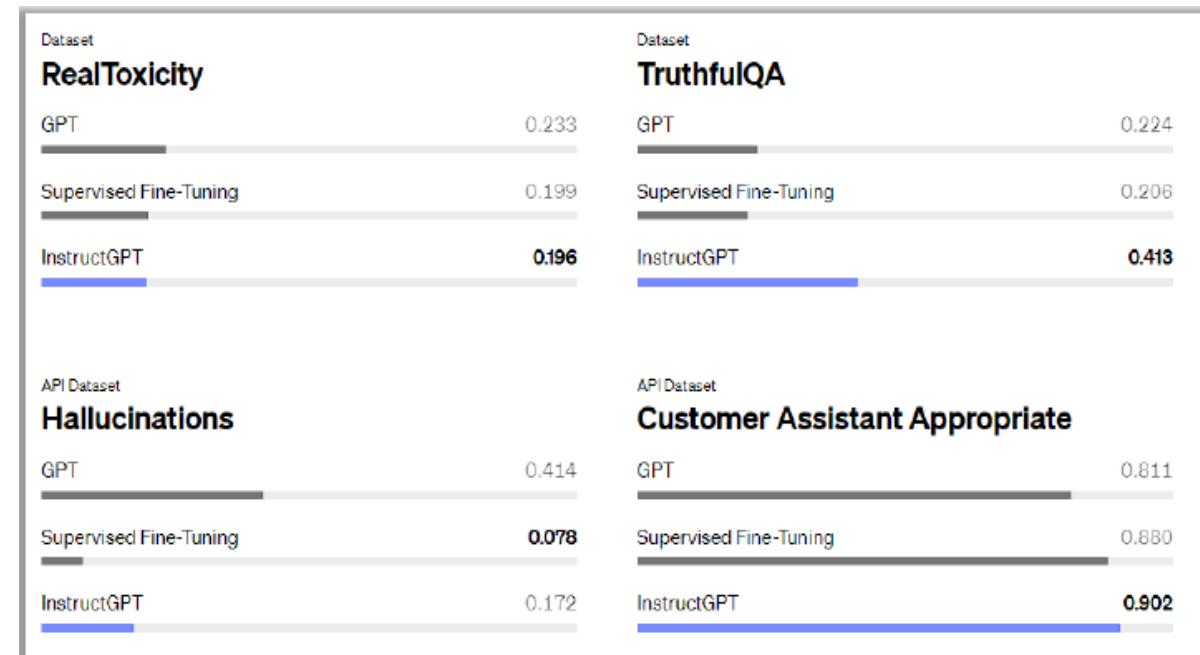
- 使大模型的输出与人类期望“对齐”
- RLHF可以提高模型的鲁棒性与安全性，使得生成结果更加遵循指令和事实，并大幅度减少有毒输出
- RLHF使得模型具备了Say no的能力



使用了RLHF的InstructGPT比GPT-3更加遵循指令



可以拒绝回答



使用了RLHF的InstructGPT相对于GPT-3产生的谎言更少(TruthfulQA)，毒性更少(RealToxicity)，同时不会频繁的编造事实(hallucinates)

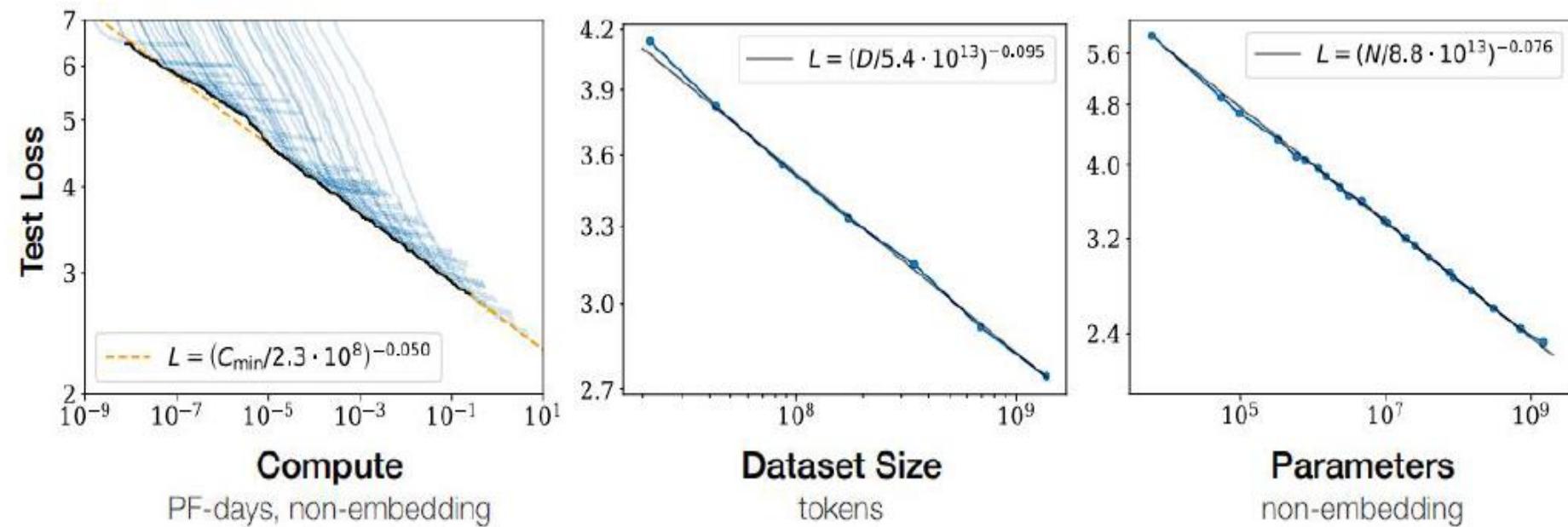
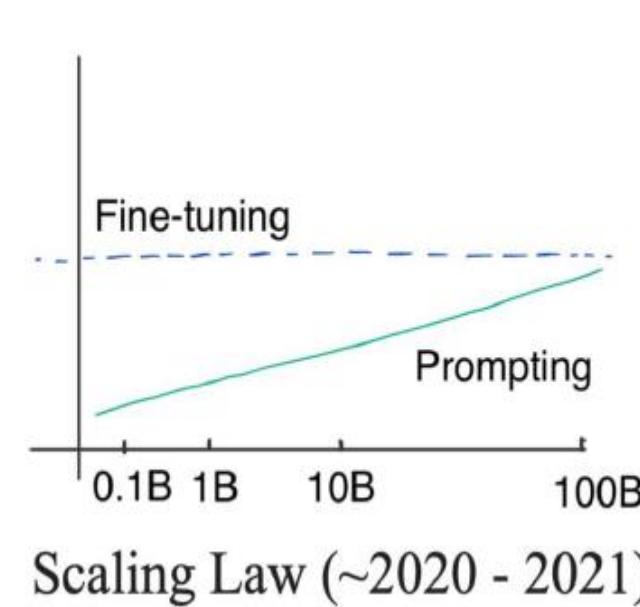
内容来源：小牛翻译-生成式LLM系列直播, 5期, 刘新宇

What Is the Scaling Law?

- 语言模型的Scaling law是一条log-linear曲线，指模型性能随着**模型规模呈指数级增长而线性增加**
- 模型规模包括**计算规模、数据规模以及参数规模**

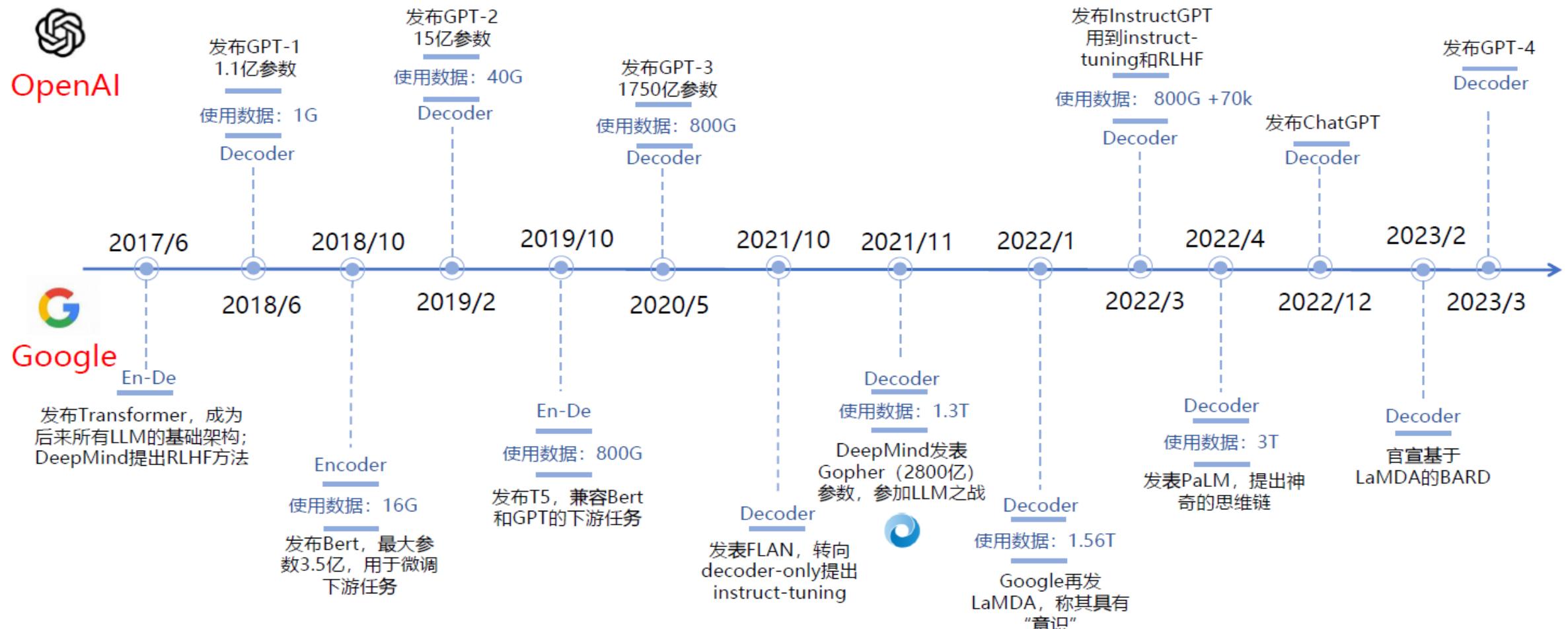
Scaling law的恐怖之处在于，模型的性能似乎存在一个与规模有关的上界。

模型被“锁定”在相应的标度内，无法明显的突破。



内容来源：小牛翻译-生成式LLM系列直播, 5期, 刘新

Making the LLM getting Larger

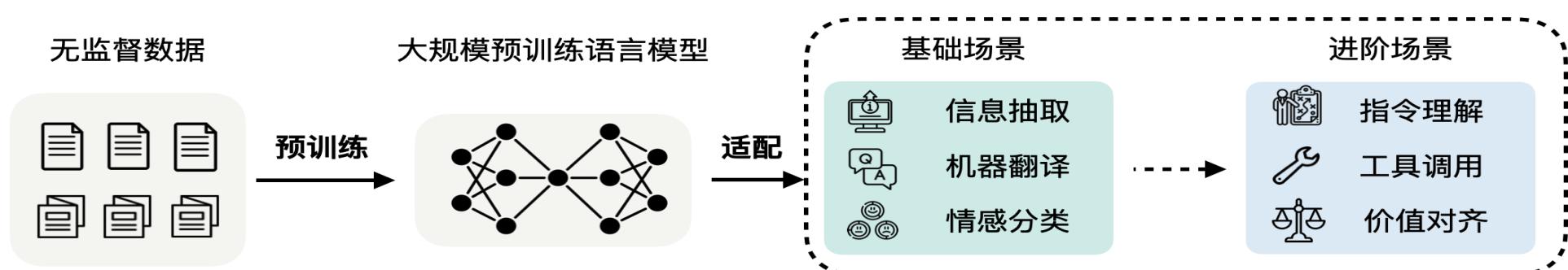


Scaling Law

- **Scaling Law** 昭示了更大的计算量一定会带来更好的效果（更低的**预训练损失**）



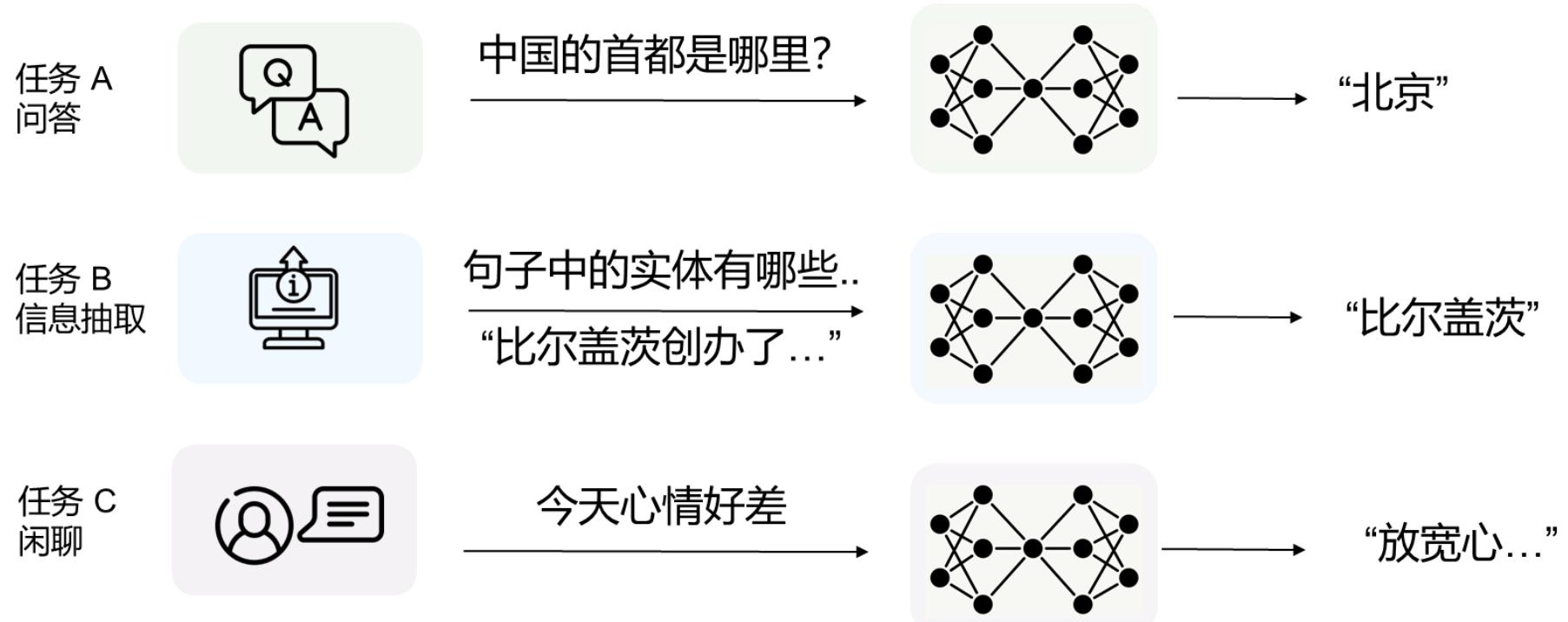
- 更大的计算量也会带来更好的**下游任务表现**



Ph.D. Ding CCL2023

Big Model Fine-tuning Task

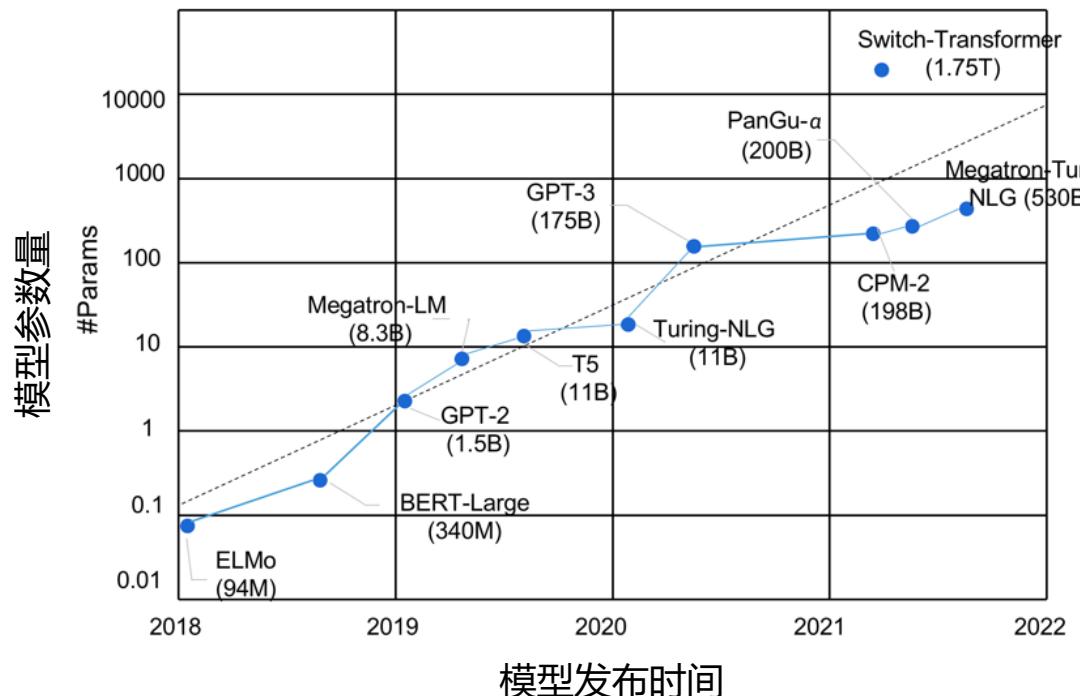
- BERT时代的模型微调：对预训练模型在**特定任务的全量数据上进行全参数微调**
- **缺少多任务之间的泛化能力**



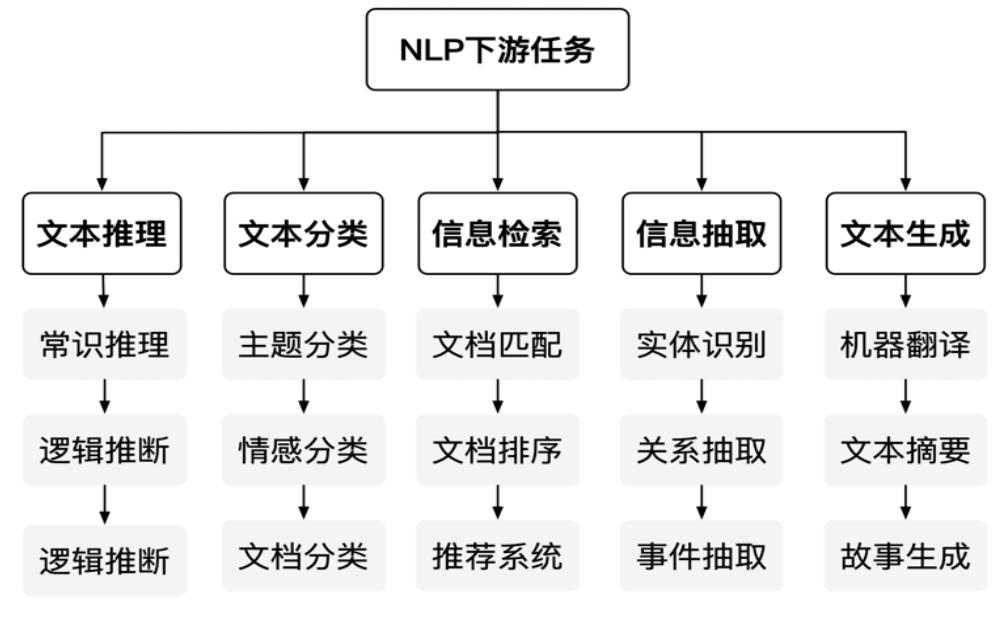
Prof. Liu CCL2023

Main Challenge --- Difficult to fine-tune LLM

- 大模型在微调中更新所有参数，需要在**计算**中占用庞大的显存
- 大模型在微调中对每一个任务存储一个大模型，需要占用庞大的**存储空间**



模型的参数规模越来越大，使之更加难以微调

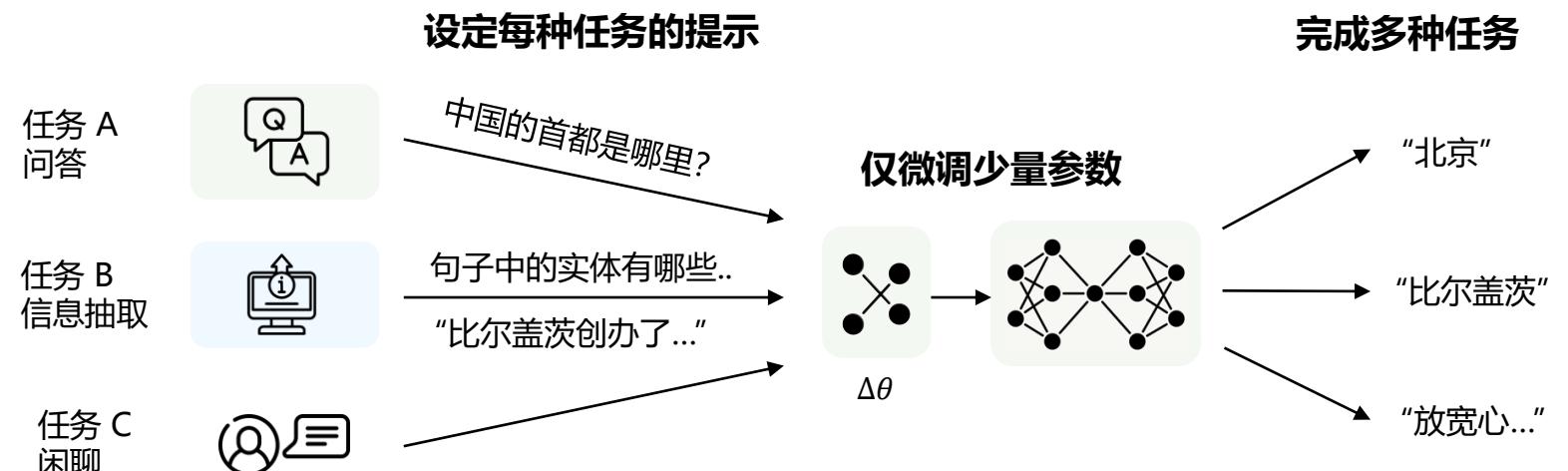


任务种类繁多，难以对所有任务都进行全参数微调

Prof. Liu CCL2023

Key Points for Fine-tuning the LLM

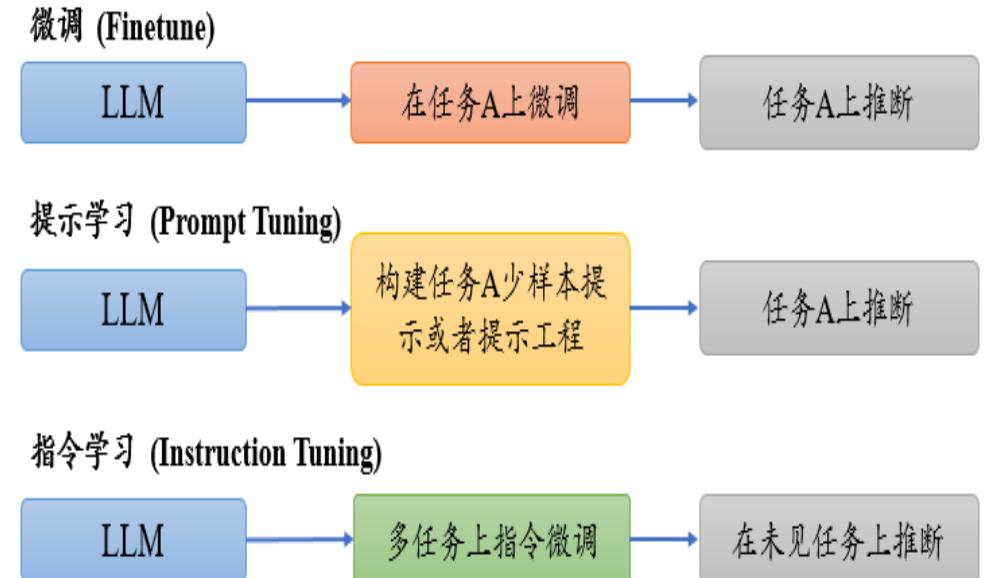
- 大模型时代的模型微调：具备少数据、少参数、强任务的泛化能力
 - 少量数据上的**提示微调**
 - 少量参数上的**增量微调**
 - 多种任务上的**指令微调**



Prof. Liu CCL2023

Fine-tuning the LLM

- 微调是大模型适配到各种场景的关键技术
- **数据高效：**使用**提示微调**组织训练和任务格式
 - 统一预训练与下游任务形式
 - 仅需少量样本就可适配新任务
- **参数高效：**使用**增量微调**激活大模型能力
 - 计算 & 存储高效
 - 仅需少量参数就可与全参数微调相当
- **任务泛化：**使用**指令微调**打破任务壁垒
 - 极大增强了大模型的意图理解能力
 - 构建ChatGPT的关键一步



微调、提示学习以及指令学习对比



Outline

- Introduction of LLM
- LLM Performance in NLP
- Fine-Tuning Strategies & Challenges for LLM
- Categorization of Tools
- Benchmarking Tool Use
- Open-source Tool-Kit

Categorization of Tools

Tools and Intelligence

- Tools are **extensions** of human **capabilities** designed to enhance productivity, efficiency, and problem-solving
- Throughout history, humans have been the **primary agents** in the **invention** and **manipulation** of tools
- Question: can **artificial intelligence** be as capable as humans in tool use?



Image from Internet

Tools and Intelligence

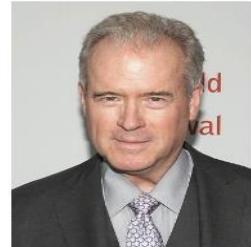
- The answer is **yes** with **foundation models**
 - Strong semantic understanding
 - Extensive **world knowledge**
 - Powerful reasoning and planning capabilities...



Noam Chomsky



Edward Feigenbaum



Robert Mercer



Yoshua Bengio

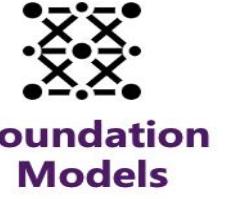
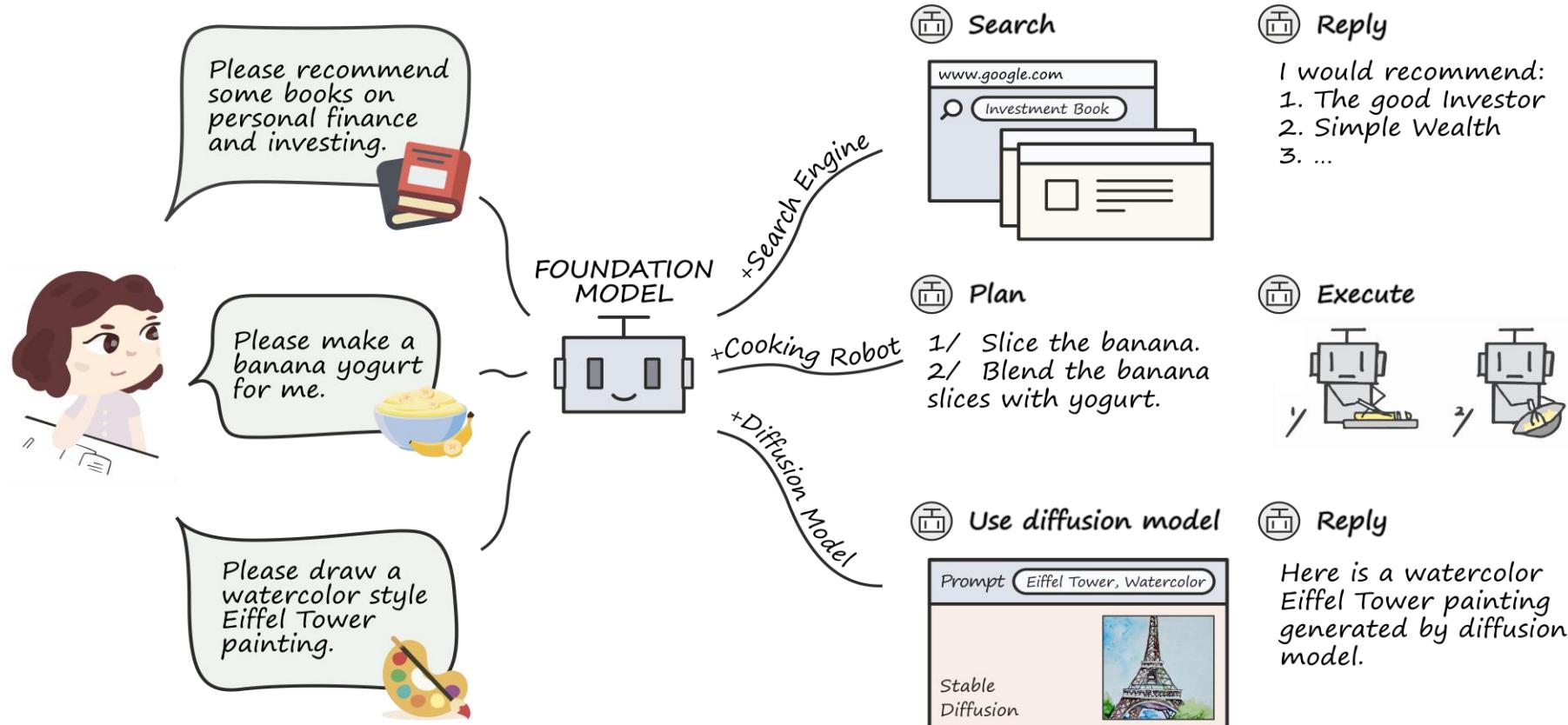


Image from Internet

Tools and Intelligence

- Tool Learning [1]: foundation models can follow human instructions and manipulate tools for task-solving

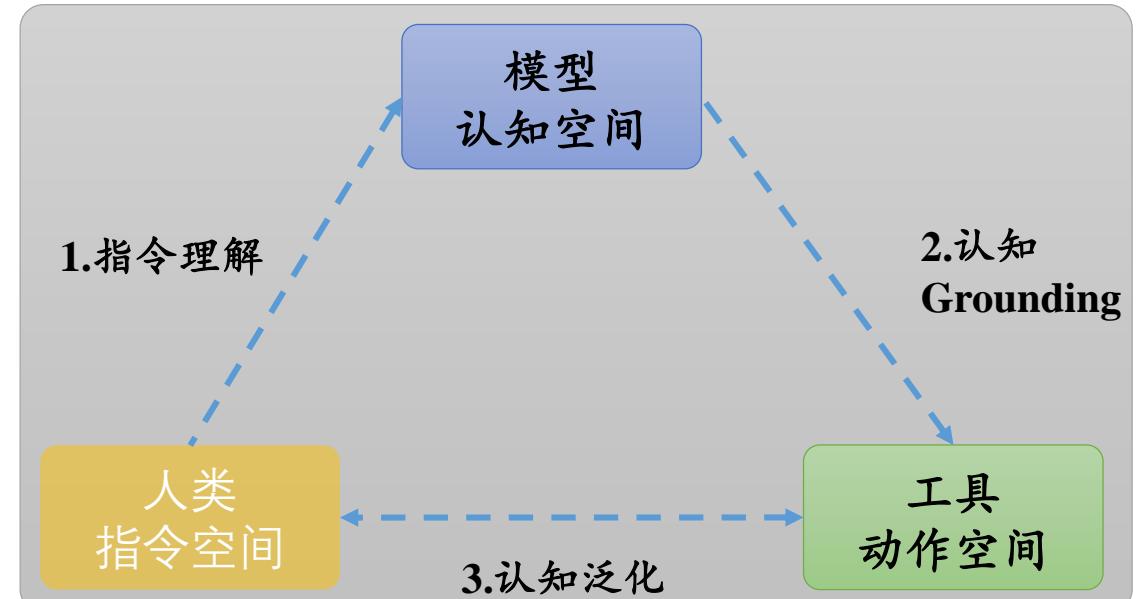
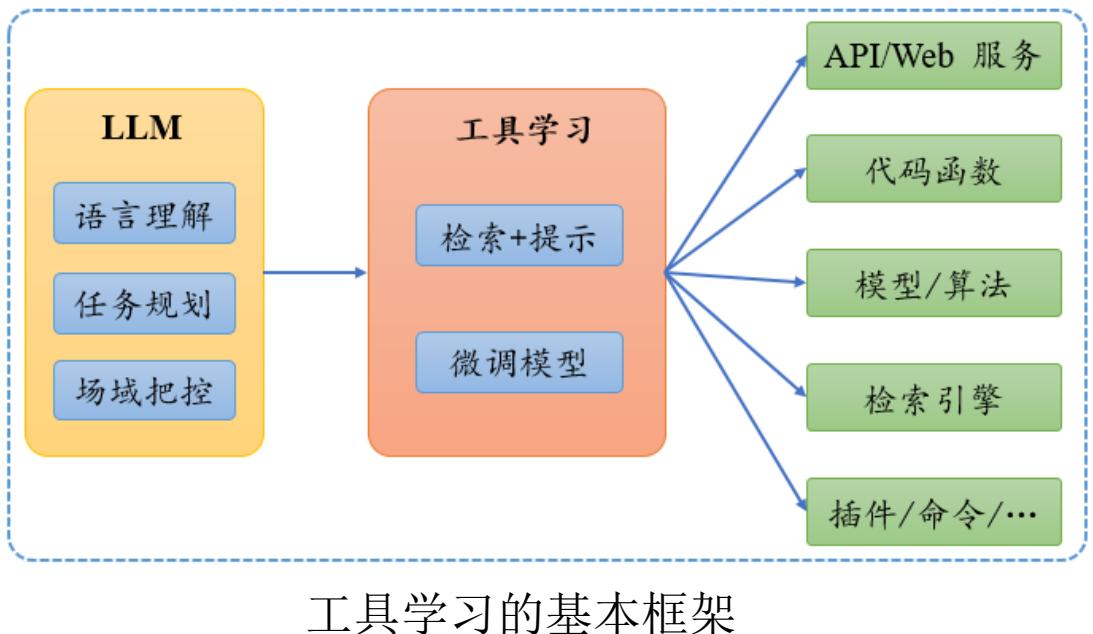


[1] Qin, Yujia, et al. "Tool Learning with Foundation Models." arXiv preprint arXiv:2304.08354 (2023).

Categorization of Tools

- Deterministic tools:  
 - Tools with fixed functionalities and stable input-output mappings
 - e.g., calculator, clock...
- Black-box API:  
 - Tools with their internal workings undisclosed
 - e.g., Bing search API, Google Translation API...
- Neural models: 
 - Neural networks optimized for specific tasks
- Robots:  
 - Tools that interact with the physical world through actuators, sensors, and control systems

Categorization of Tools



Categorization of Tools

- Tool-augmented learning
 - Augment foundation models with the **execution results** from **tools**
 - Tools are viewed as **complementary resources** that **aid in** the generation of high-quality outputs

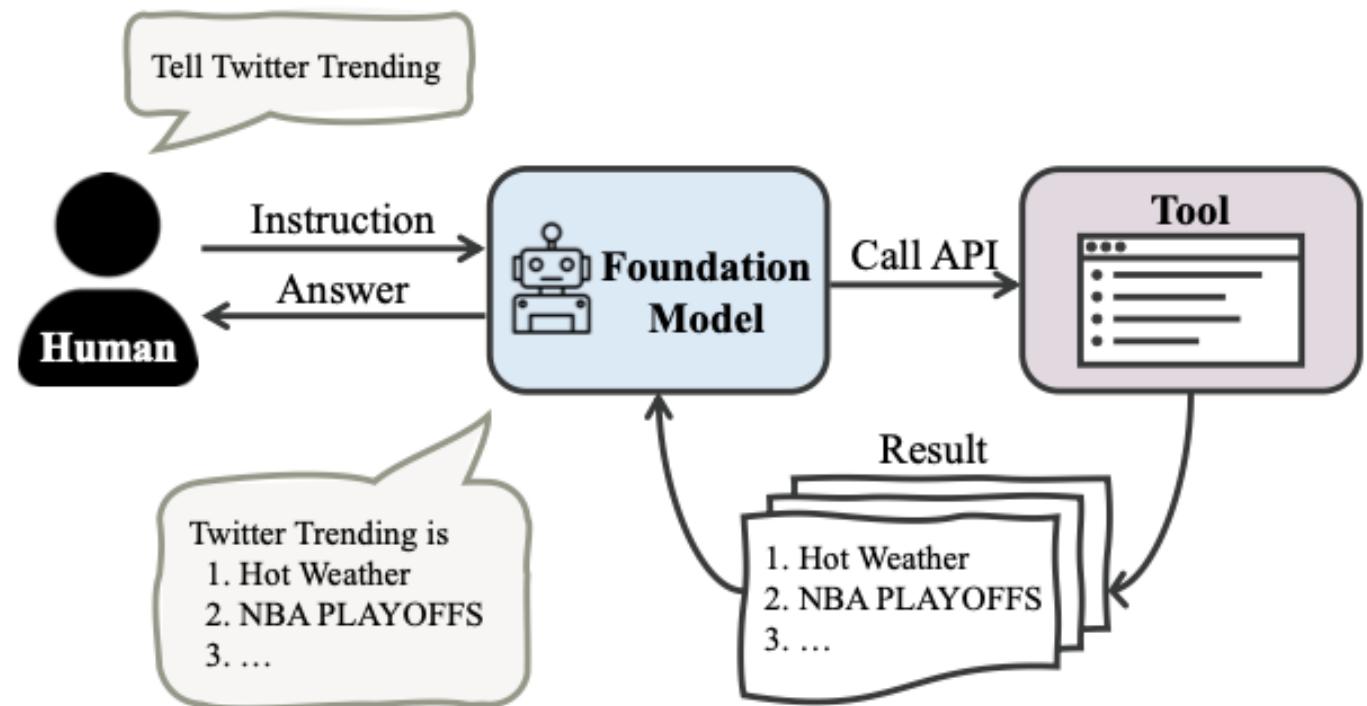


Image from Internet

Tool-augmented learning

Tool Documentation Enables Zero-Shot Tool-Usage with Large Language Models Preprint 2023.
Cheng-Yu Hsieh, Si-An Chen, Chun-Liang Li, Yasuhisa Fujii, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, Tomas Pfister [[pdf](#)]

ToolQA: A Dataset for LLM Question Answering with External Tools Preprint 2023.
Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, Chao Zhang [[pdf](#)]

Toolformer: Language Models Can Teach Themselves to Use Tools Preprint 2023.
Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, Thomas Scialom [[pdf](#)]

Generalization through Memorization: Nearest Neighbor Language Models ICLR 2020.
Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, Mike Lewis [[pdf](#)]

Improving Language Models by Retrieving from Trillions of Tokens PMLR 2022.

.....

<https://github.com/thunlp/ToolLearningPapers>

Categorization of Tools

- **Tool-oriented learning**

- Utilize models to govern tools and **make sequential decisions** in place of humans
- **Exploiting** foundation models' vast **world knowledge** and **reasoning ability** for complex reasoning and planning

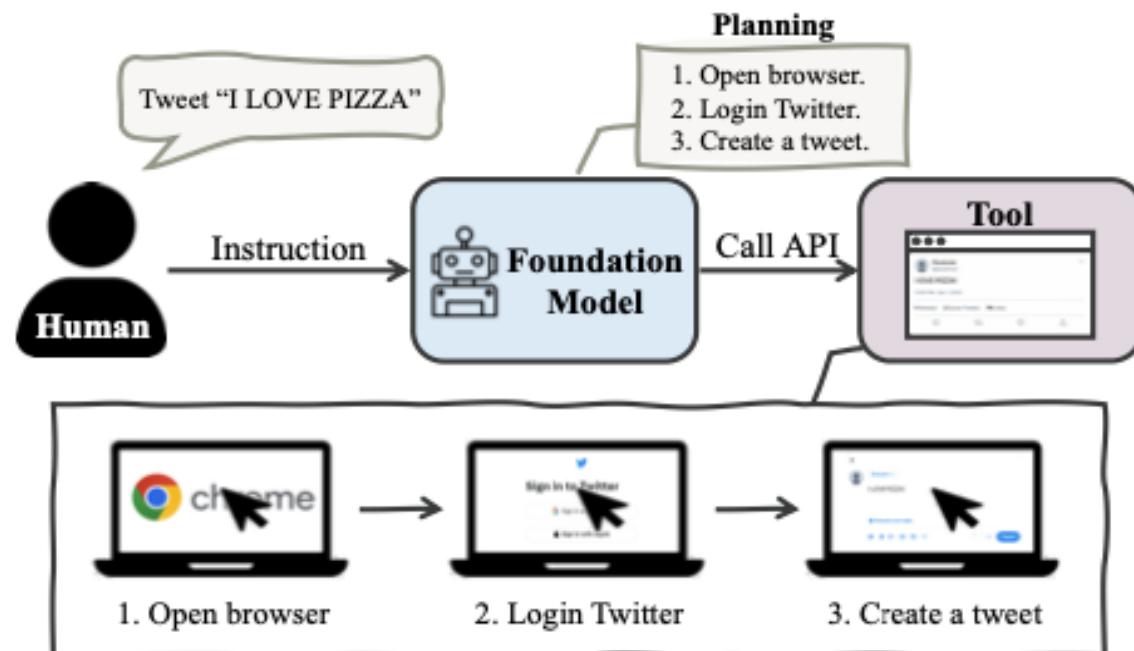


Image from Internet

Tool-oriented learning

ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs Preprint 2023.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, Maosong Sun [[pdf](#)]

TPTU: Task Planning and Tool Usage of Large Language Model-based AI Agents Preprint 2023.

Jingqing Ruan, Yihong Chen, Bin Zhang, Zhiwei Xu, Tianpeng Bao, Guoqing Du, Shiwei Shi, Hangyu Mao, Xingyu Zeng, Rui Zhao [[pdf](#)]

WebCPM: Interactive Web Search for Chinese Long-form Question Answering Preprint 2023.

Yujia Qin, Zihan Cai, Dian Jin, Lan Yan, Shihao Liang, Kunlun Zhu, Yankai Lin, Xu Han, Ning Ding, Huadong Wang, Ruobing Xie, Fanchao Qi, Zhiyuan Liu, Maosong Sun and Jie Zhou [[pdf](#)]

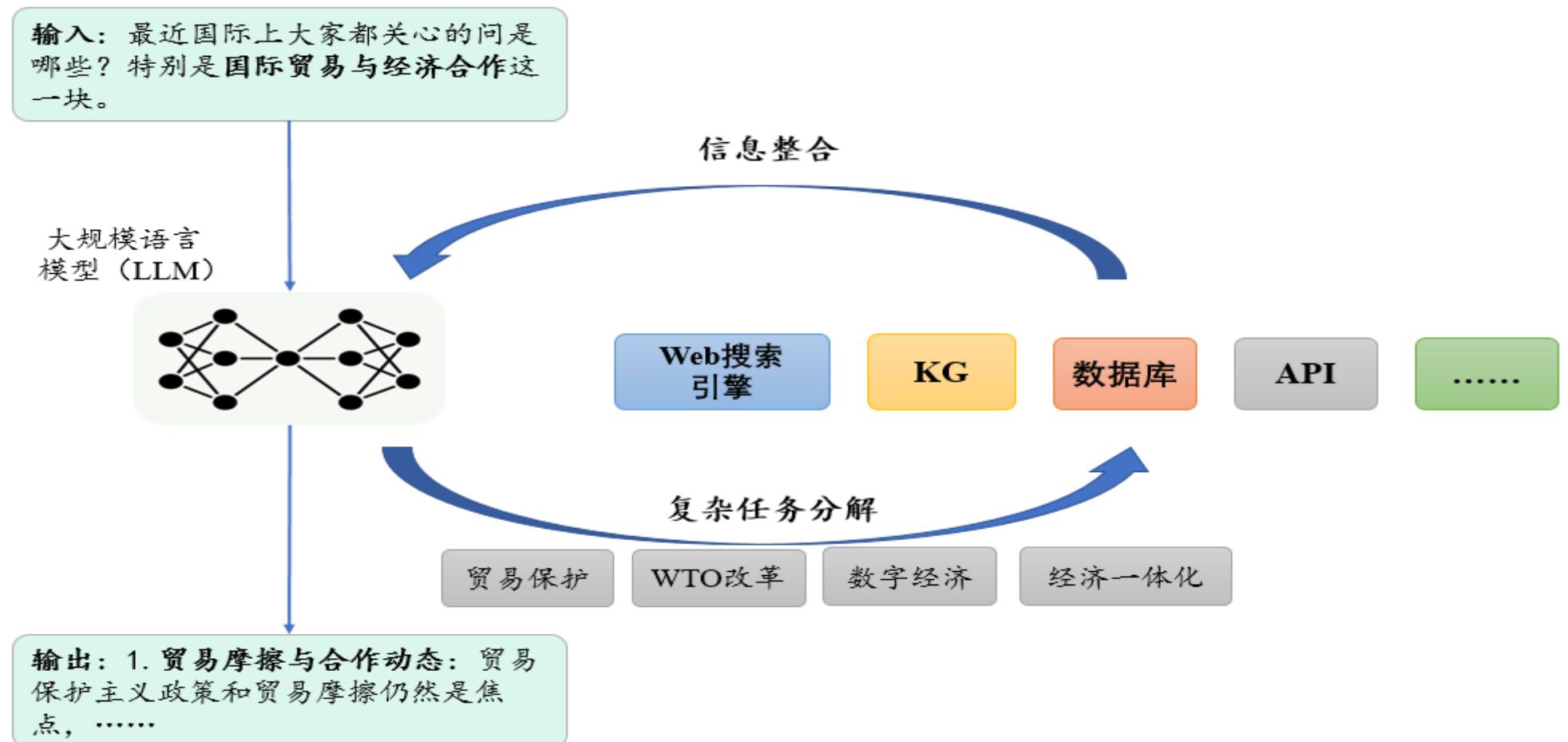
CREATOR: Disentangling Abstract and Concrete Reasonings of Large Language Models through Tool Creation Preprint 2023. Cheng Qian, Chi Han, Yi R. Fung, Yujia Qin, Zhiyuan Liu, Heng Ji [[pdf](#)]

Describe, Explain, Plan and Select: Interactive Planning with Large Language Models Enables Open-World Multi-Task Agents Preprint 2023.

.....

Inability of Foundation Models

- Models **cannot** provide **real-time** information

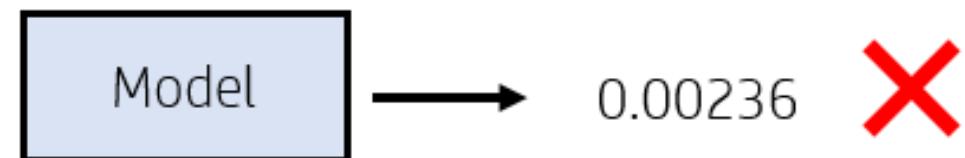


Inability of Foundation Models

- Models cannot handle **specialized tasks**

Input: Please calculate $12321 / 23432$

without tools



with tools



Tool Creation

- From Tool User to **Tool Creator**
 - Humans are the primary agents that created and used tools from Stone Age to 21st century
 - Most tools are created for humans, not AI
- Tools Made for **Models**
 - Modularized: compose tools into smaller units
 - New input and output formats: more computable and suitable for AI

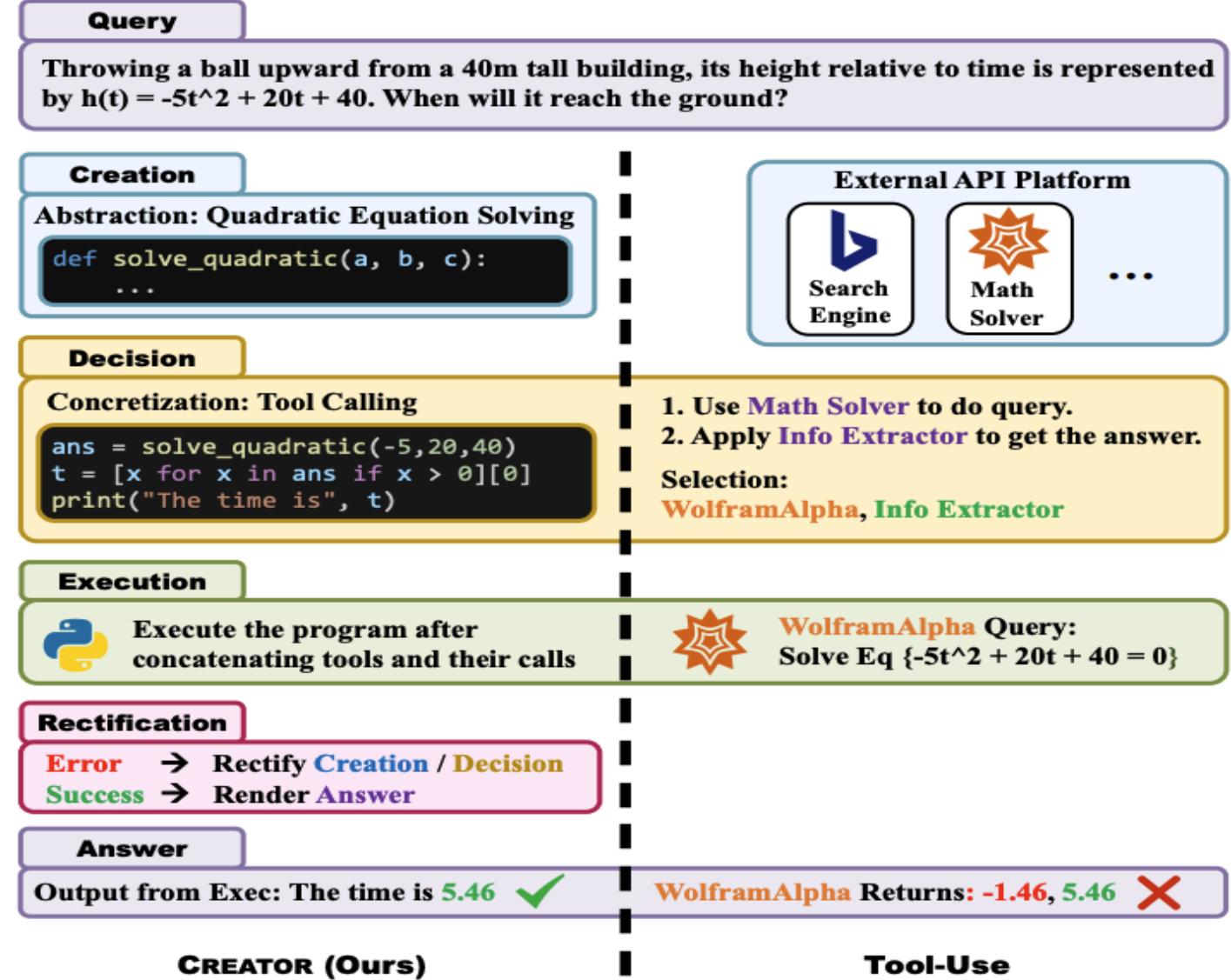
Tool Creation

- **Limitations** of Existing Works
 - Most existing work tends to concentrate on a **limited number** of tools
 - The reasoning process employed by models for determining the optimal utilization of tools **is inherently complex**
 - The current pipelines **lack an error-handling mechanism** after retrieving execution results
- Instead of letting LLMs act as the users of tools, we enable them to be the **creators** [1]

[1] Qian, Cheng, et al. "CREATOR: Disentangling Abstract and Concrete Reasonings of Large Language Models through Tool Creation."

Tool Creation

- Four Procedures
 - Creation
 - Decision
 - Execution
 - Rectification



Qian, Cheng, et al. "CREATOR: Disentangling Abstract and Concrete Reasonings of Large Language Models through Tool Creation."

Tool Creation

- Experiments
 - Datasets: MATH, TabMWP
 - Significant improvements over PoT and pure CoT

Method	Setting	Algebra	Counting & Probability	Geometry	Itmd. Algebra	Number Theory	Pre-Algebra	Pre-Calculus	Average (weighted)
Standard	w/o CoT	25.7	25.8	22.4	13.9	18.5	40.9	21.8	25.3
	w/ CoT	50.9	36.1	24.5	17.5	23.2	58.6	16.7	37.9
PoT (w/o Rectify)	w/o CoT	58.2	48.5	35.4	25.8	53.1	66.8	25.0	49.8
	w/ CoT	54.0	47.8	32.5	22.3	48.9	64.5	19.9	46.5
PoT (w/ Rectify)	w/o CoT	63.8	51.9	35.9	28.6	59.2	70.0	28.2	53.9
	w/ CoT	61.4	48.8	34.6	23.7	54.5	67.6	34.6	51.2
Tool Use	w/o CoT	47.3	35.1	27.0	20.5	30.8	56.8	31.4	39.0
	w/ CoT	55.3	37.8	28.7	20.5	34.8	61.8	26.9	43.0
Tool Create (whole)	w/o Demo	58.0	53.3	34.2	21.8	55.7	63.4	33.3	49.6
	w/o CoT	64.1	55.7	35.9	42.7	61.6	69.0	37.2	57.2
	w/ CoT	62.7	50.9	33.8	31.4	61.4	68.7	31.4	54.0
CREATOR (ours)	w/o Demo	66.6	53.6	33.8	29.4	59.8	68.7	34.6	54.9
	w/o CoT	71.5	55.3	41.4	41.9	60.4	71.7	35.3	59.7
	w/ CoT	63.1	58.1	34.6	35.0	61.8	69.7	32.1	55.7

[1] Qian, Cheng, et al. "CREATOR: Disentangling Abstract and Concrete Reasonings of Large Language Models through Tool Creation."

WebCPM

- Motivation
 - WebGPT is **not public**, and its inner workings remain **opaque**
- WebCPM's **Efforts**
 - Open-source interactive web search interface
 - The **first public QA dataset** that involves interactive web search, and also the **first Chinese LFQA dataset**
 - Framework and Model Implementation

WEBCPM:
Interactive Web Search for Chinese Long-form Question Answering

Yujia Qin¹, Zihan Cai¹, Dian Jin¹, Lan Yan¹, Shihao Liang³, Kunlun Zhu³,
Yankai Lin^{2*}, Xu Han¹, Ning Ding¹, Huadong Wang¹, Ruobing Xie⁴, Fanchao Qi¹,
Zhiyuan Liu^{1*}, Maosong Sun^{1*}, Jie Zhou⁴

¹NLP Group, DCST, IAI, BNRIST, Tsinghua University, Beijing

²Gaoling School of Artificial Intelligence, Renmin University of China, Beijing

³ModelBest Inc. ⁴Pattern Recognition Center, WeChat AI, Tencent Inc.

qyj20@mails.tsinghua.edu.cn

WebCPM

- Interface (search mode) and pre-defined actions

Question | 麦田怪圈是什么？它们是如何形成的？ *What are crop circles? How are they made?*

Query | 麦田怪圈如何形成？ *How do crop circles form?*

Action Buttons: Undo, Reset

Window (search mode)

难解谜团：麦田怪圈究竟是如何形成的？ *Page <1>*
Unsolved mysteries: How did crop circles form?

麦田怪圈出现最多的季节是在春天和夏天，有人认为，夏季天气变化无常，龙卷风是造成怪圈的主要原因...
The crop circles appear most often in spring and summer.
Some people think that the weather in summer is erratic, and tornadoes are the main cause of the strange circles...

Title of page <2> *Page <2>*
A snapshot of the page content

Title of page <3> *Page <3>*
A snapshot of the page content

Go Back Number of remaining actions (86/100) **Finish**

Action Name	Functionality
Search <query>	Call Bing search with <query>
← Go Back	Return to the previous window
Load Page <1>	Load the details of page <1>
Load Page <2>	Load the details of page <2>
Load Page <3>	Load the details of page <3>
↑ Scroll Up	Scroll up for a pre-set stride
↓ Scroll Down	Scroll down for a pre-set stride
„ Quote <content>	Extract <content> from the current page as a supporting fact
↑ Merge	Merge two facts into a single fact
Finish	End the search process

<https://github.com/thunlp/WebCPM>

WebCPM

- WebCPM contains **5,500** high-quality question-answer pairs, together with 14,315 supporting facts and **121,330** web search actions
- Data format: <question, web search actions, supporting facts, answer>

Resource	IWSQA (this work)	DuReader (He et al., 2018)	CMRC (Cui et al., 2019)	C ³ (Sun et al., 2020)	WebGPT (Nakano et al., 2021)	GopherCite (Menick et al., 2022)
Language?	ZH	ZH	ZH	ZH	EN	EN
Is Public?	✓	✓	✓	✓	✗	✗
Targets long-form QA?	✓	✗	✗	✗	✓	✓
Has free-form answer?	✓	✓	✗	✗	✓	✓
Has web search behavior?	✓	✗	✗	✗	✓	✗
Avg. question length	29.0	9.6	16.3	12.2	-	-
Avg. supporting fact length	512.5	187.3	495.5	116.9	-	-
Avg. answer length	244.4	104.9	17.0	5.5	-	-

<https://github.com/thunlp/WebCPM>

Question:

麦田怪圈是什么？它们是如何形成的？

Action Sequence:

Search → Load Page <1> → Scroll Down × 5 → Scroll Up → Scroll Down × 11 → Go Back → Search → Load Page <1> → Go Back → Load Page <3> → Scroll Down × 4 → Scroll Up × 3 → Quote → Scroll Down → Quote → Scroll Down → Quote → Scroll Down → Quote → Finish

Supporting Facts:

1. 麦田怪圈 (Crop Circle)，是指在麦田或其它田地上，通过某种未知力量（大多数怪圈是人类所为）把农作物压平而产生出来的几何图案。这个神秘现象有时被人们称之为“Crop Formation”。麦田怪圈的出现给了对支持外星人存在论的人们多种看法。
2. 人为说：人为说一般认为，麦田圈是用木板压成的。木板两头系上绳子形成圈套，在制作时，一脚踩在木板上拖动木板压倒麦子，并拉着细绳与圆心保持固定的距离，逐渐就可以形成一个圆圈。为了便于制造，主要形状所有圆圈的直径都可以被6除尽。以前曾经出现过制作麦田圈被当场抓获的事情，制作者使用的就是这种工具。
3. 自然形成说：也有人认为，麦田圈只是一种，成因还未被人类发现。就像雷电，古时候人类也是以为是雷神电母做的，对于麦田圈中经常出现人文信息的现象，他们认为这只是人们“先入为主”造成的错觉。
4. 磁场说：有专家认为，磁场中有一种神奇的移动力，可产生一股电流，使农作物“平躺”在地面上。美国专家杰弗里·威尔逊研究了130多个麦田怪圈，发现90%的怪圈附近都有连接高压电线的变压器，方圆270米内都有一个水池。由于接受灌溉，麦田底部的土壤释放出的离子会产生负电，与高压电线相连的变压器则产生正电，负电和正电碰撞后会产生电磁能，从而击倒小麦形成怪圈。

Answer:

麦田怪圈是指在麦田或其它田地上，通过某种未知力量（大多数怪圈是人类所为）把农作物压平而产生出来的几何图案。**【1】**而麦田怪圈的形成有三个说法，人为说、自然说与磁场说。人为说一般认为，麦田圈是用木板压成的。木板两头系上绳子形成圈套，在制作时，一脚踩在木板上拖动木板压倒麦子，并拉着细绳与圆心保持固定的距离，逐渐就可以形成一个圆圈。为了便于制造，主要形状所有圆圈的直径都可以被6除尽。**【2】**自然形成说认为麦田怪圈的成因还未被人类发现。就像雷电，古时候人类也是以为是雷神电母做的，对于麦田圈中经常出现人文信息的现象，他们认为这只是人们“先入为主”造成的错觉。**【3】**磁场说认为，磁场中有一种神奇的移动力，可产生一股电流，使农作物“平躺”在地面上。**【4】**

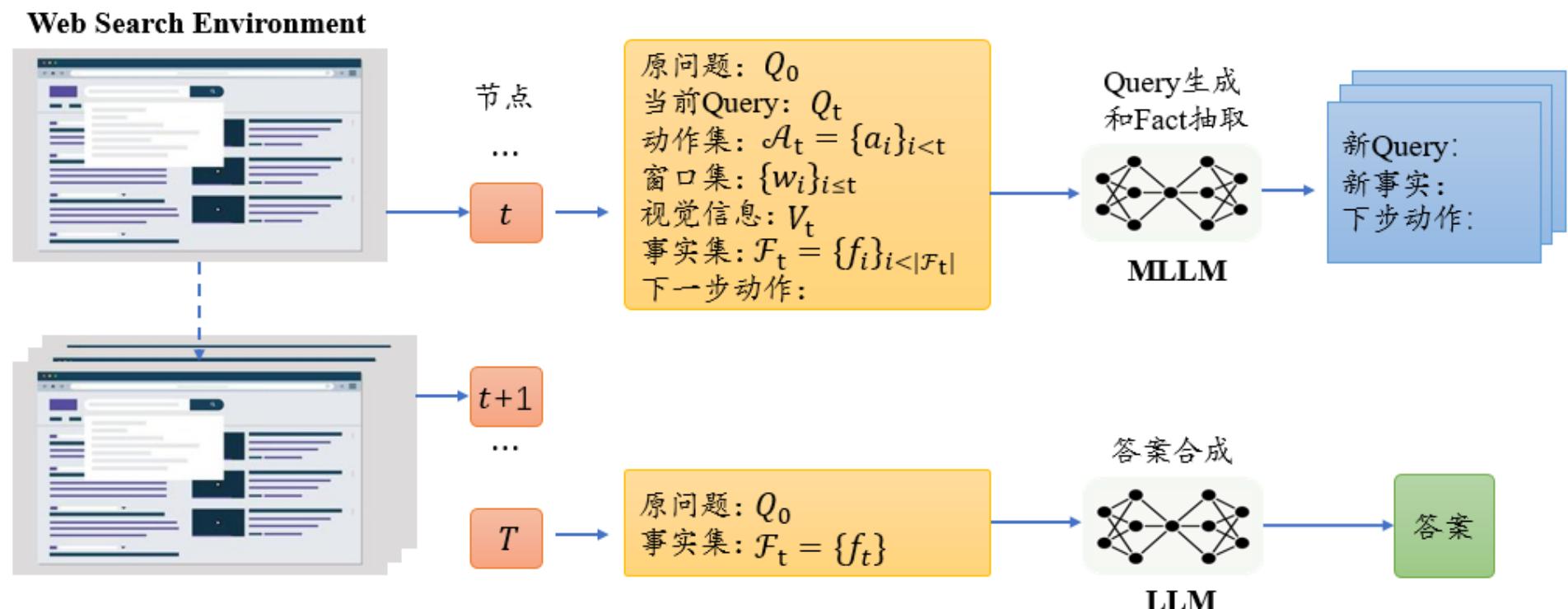
WebCPM

- WebCPM's framework consists of **two models**:
- **Search model**, consisting of:
 - **Action** prediction module
 - Search query generation module
 - Supporting fact extraction module
- Information synthesis model

<https://github.com/thunlp/WebCPM>

WebCPM

For an action sequence of T steps, the **search model** executes actions to collect supporting facts, which are sent to the **synthesis model** for answer generation.



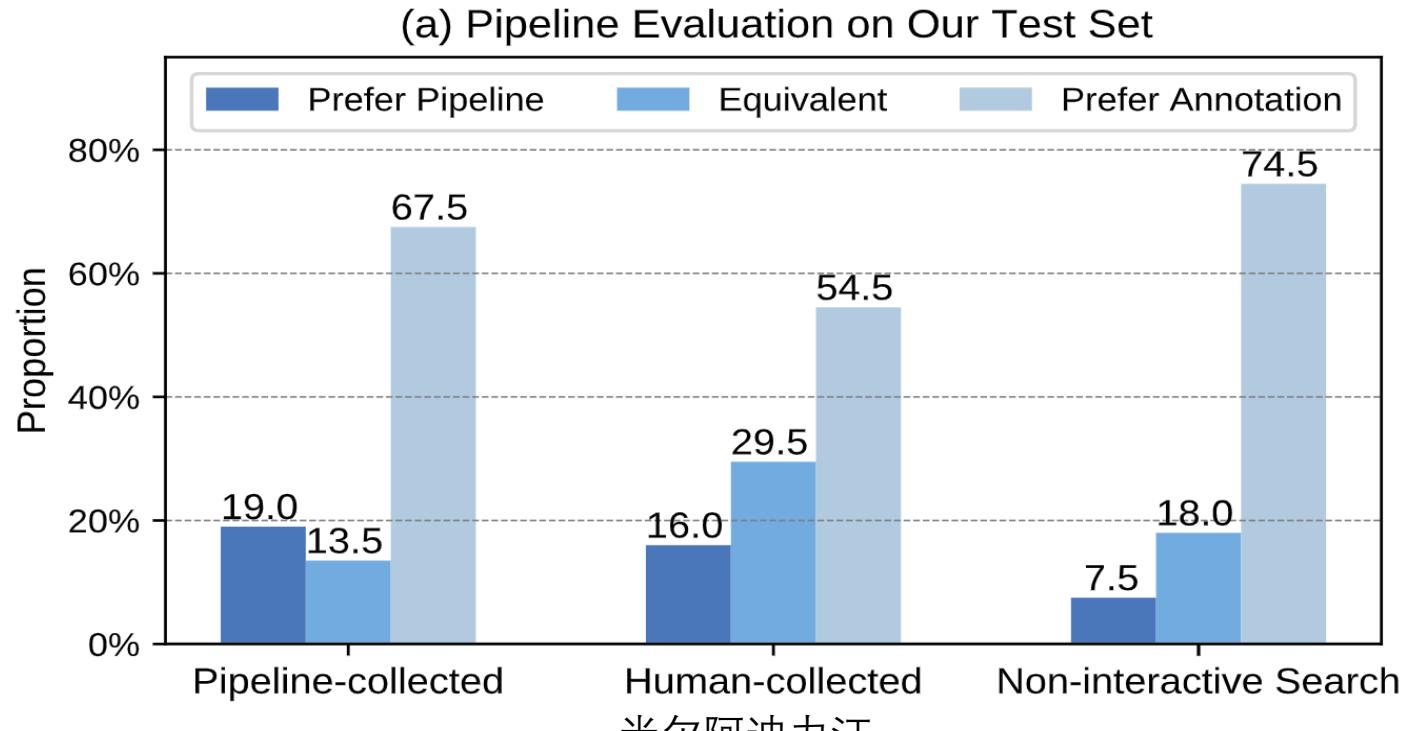
WebCPM

Holistic Pipeline Evaluation (based on human preference)

Model-generated Answer v.s. Human Annotation

Three sources of supporting facts are sent to the synthesis model

(1) pipeline-collected, (2) human-collected, (3) non-interactive search (TF-IDF)



Outline

- Introduction of LLM
- LLM Performance in NLP
- Fine-Tuning Strategies & Challenges for LLM
- Categorization of Tools
- Benchmarking Tool Use
- Open-source Tool-Kit

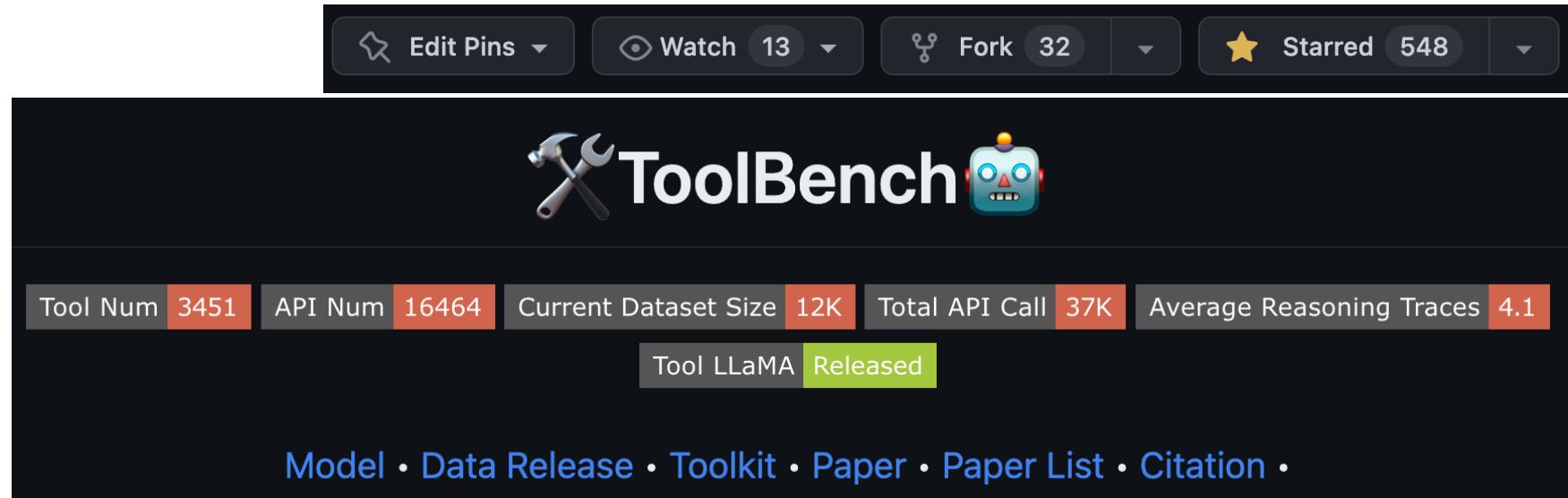
Benchmarking Tool Use

Tool Bench

- Open-source, e.g., LLaMA & Vicuna, remain **significantly limited** in performing high-level tasks, such as **tool use**
- Reason: current instruction tuning largely focuses on **basic language tasks** instead of the **tool-use domain**
- This is **in contrast** to SOTA LLMs, e.g., ChatGPT, which have demonstrated excellent tool-use capabilities
- How to bridge the gap?

Tool Bench

- Introducing ToolBench, an instruction-tuning dataset for tool use
- Elicit tool-use capabilities within **open-source** LLMs



<https://github.com/OpenBMB/ToolBench>

Tool Bench

- Introducing ToolBench, an instruction-tuning dataset for tool use
- Elicit tool-use capabilities within open-source LLMs



TOOLLLM: FACILITATING LARGE LANGUAGE MODELS TO MASTER 16000+ REAL-WORLD APIs

Yujia Qin^{1*}, Shihao Liang^{1,2*}, Yining Ye¹, Kunlun Zhu^{1,2}, Lan Yan^{1,2}, Yaxi Lu¹, Yankai Lin^{3†}, Xin Cong¹, Xiangru Tang⁴, Bill Qian⁴, Sihan Zhao¹, Runchu Tian¹, Ruobing Xie⁵, Jie Zhou⁵, Mark Gerstein⁴, Dahai Li^{2,6}, Zhiyuan Liu^{1†}, Maosong Sun^{1†}

¹Tsinghua University ²ModelBest Inc. ³Renmin University of China

⁴Yale University ⁵WeChat AI, Tencent Inc. ⁶Zhihu Inc.

yujiaqin16@gmail.com

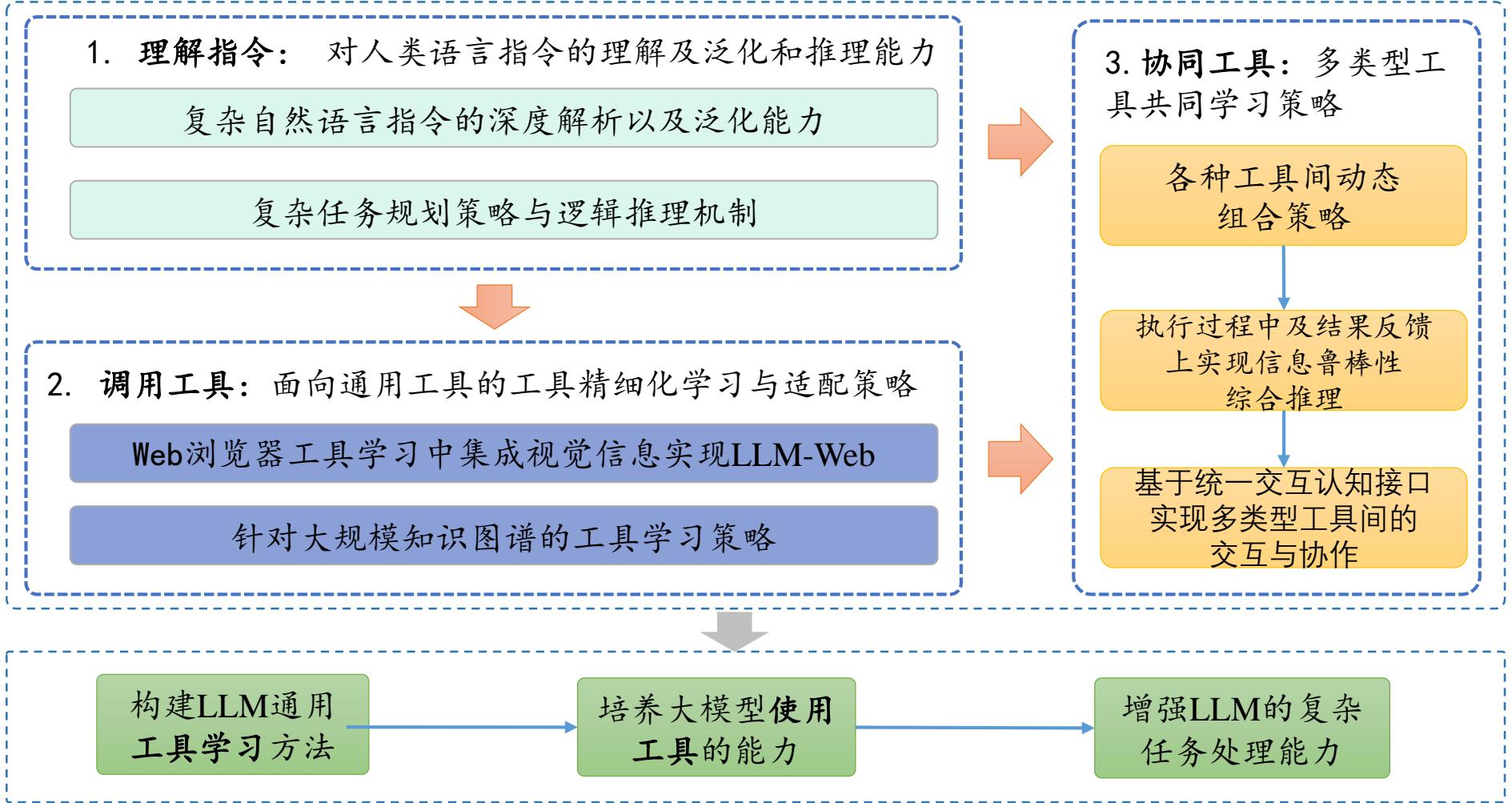
<https://github.com/OpenBMB/ToolBench>

Tool Bench

Steps



Goal



Tool Bench

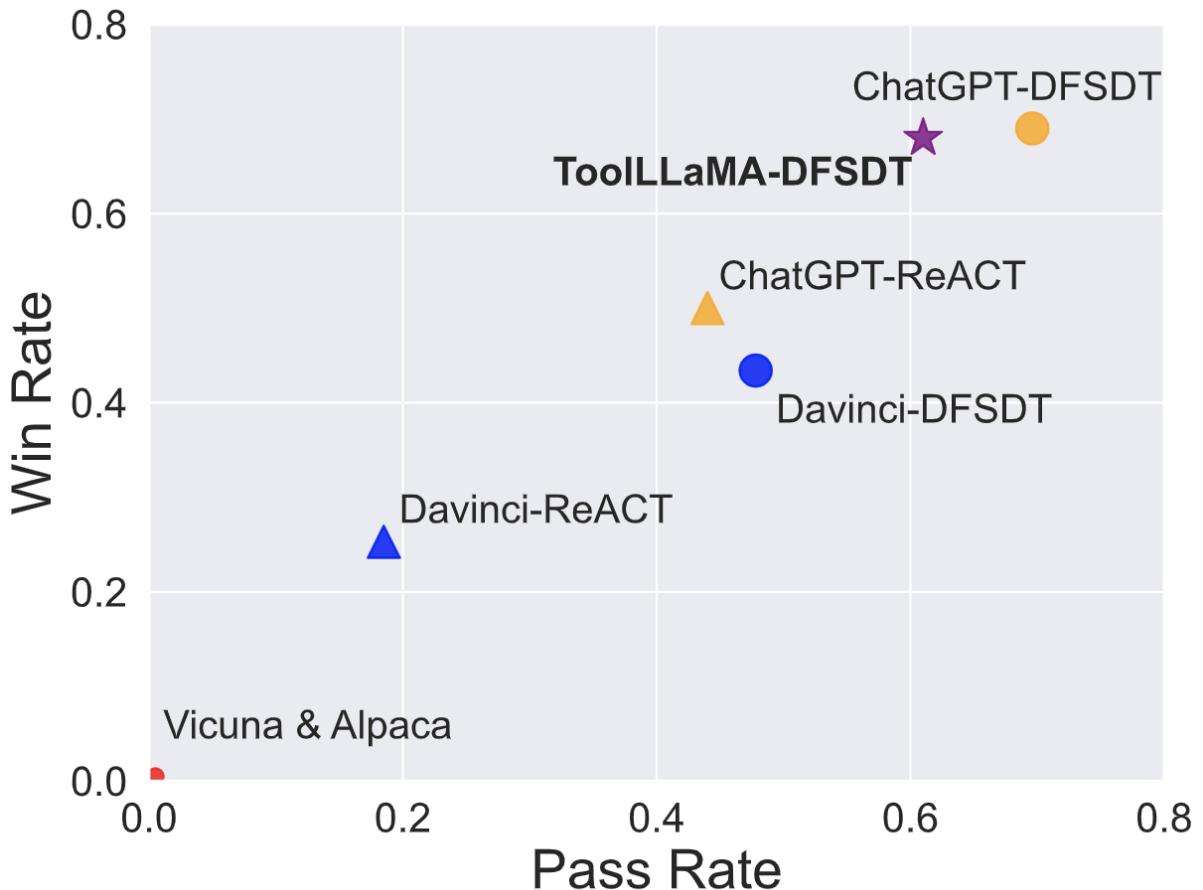
- Highlights:
 - **16000+** real-world APIs gathered from RapidAPI
 - **Real** API call for building this dataset
 - Support single-tool and **multi-tool** scenarios
 - API retriever
 - Complex multi-step reasoning

Resource	ToolBench (this work)	APIBench (Patil et al., 2023)	API-Bank (Li et al., 2023a)	ToolAlpaca (Tang et al., 2023)	T-Bench (Xu et al., 2023b)
Real-world API?	✓	✗	✓	✗	✓
Real API Response?	✓	✗	✓	✗	✓
Multi-tool Scenario?	✓	✗	✗	✗	✗
API Retrieval?	✓	✓	✗	✗	✗
Multi-step Reasoning?	✓	✗	✓	✓	✓
Number of tools	3451	3	53	400	8
Number of APIs	16464	1645	53	400	232
Number of Instances	12657	17002	274	3938	2746
Number of Real API Calls	37204	0	568	0	0
Avg. Reasoning Traces	4.1	1.0	2.1	1.0	5.9

<https://github.com/OpenBMB/ToolBench>

Tool Bench

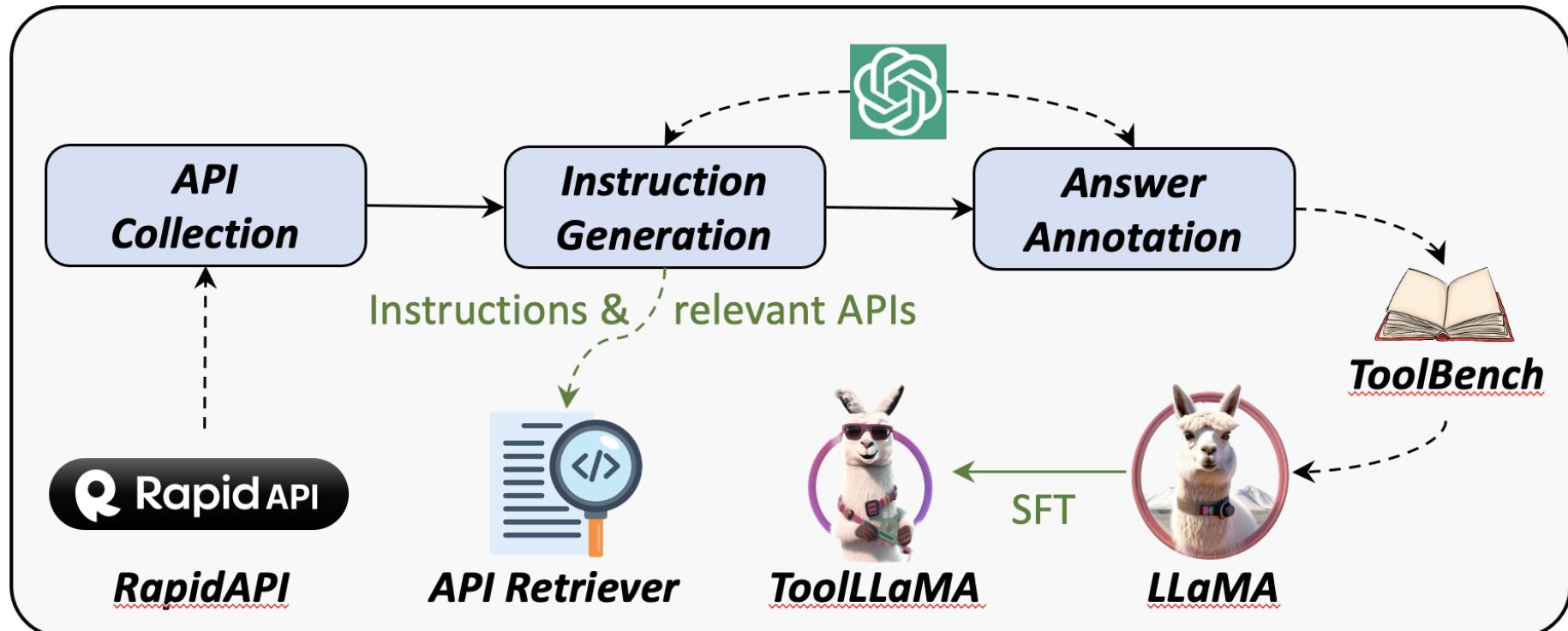
- Highlights:
 - Enhanced reasoning strategy: DFSDT (ours) >> ReACT
 - Automatic Evaluator **ToolEval**
 - ToolLLaMA performs on par with ChatGPT (turbo-16k)
 - Robust **generalization** to unseen APIs



<https://github.com/OpenBMB/ToolBench>

Tool Bench Construction --- Overview

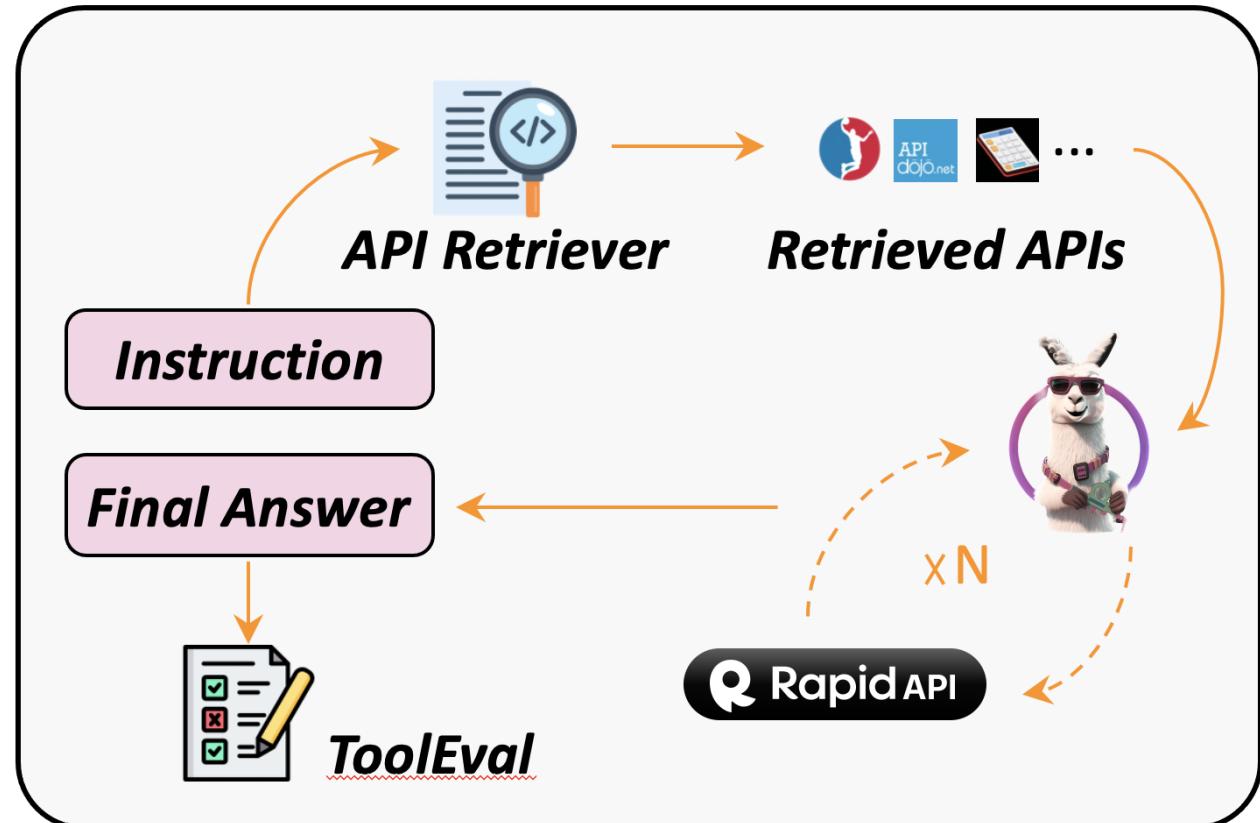
- Three phases for constructing ToolBench



<https://github.com/OpenBMB/ToolBench>

Tool Bench Construction --- Overview

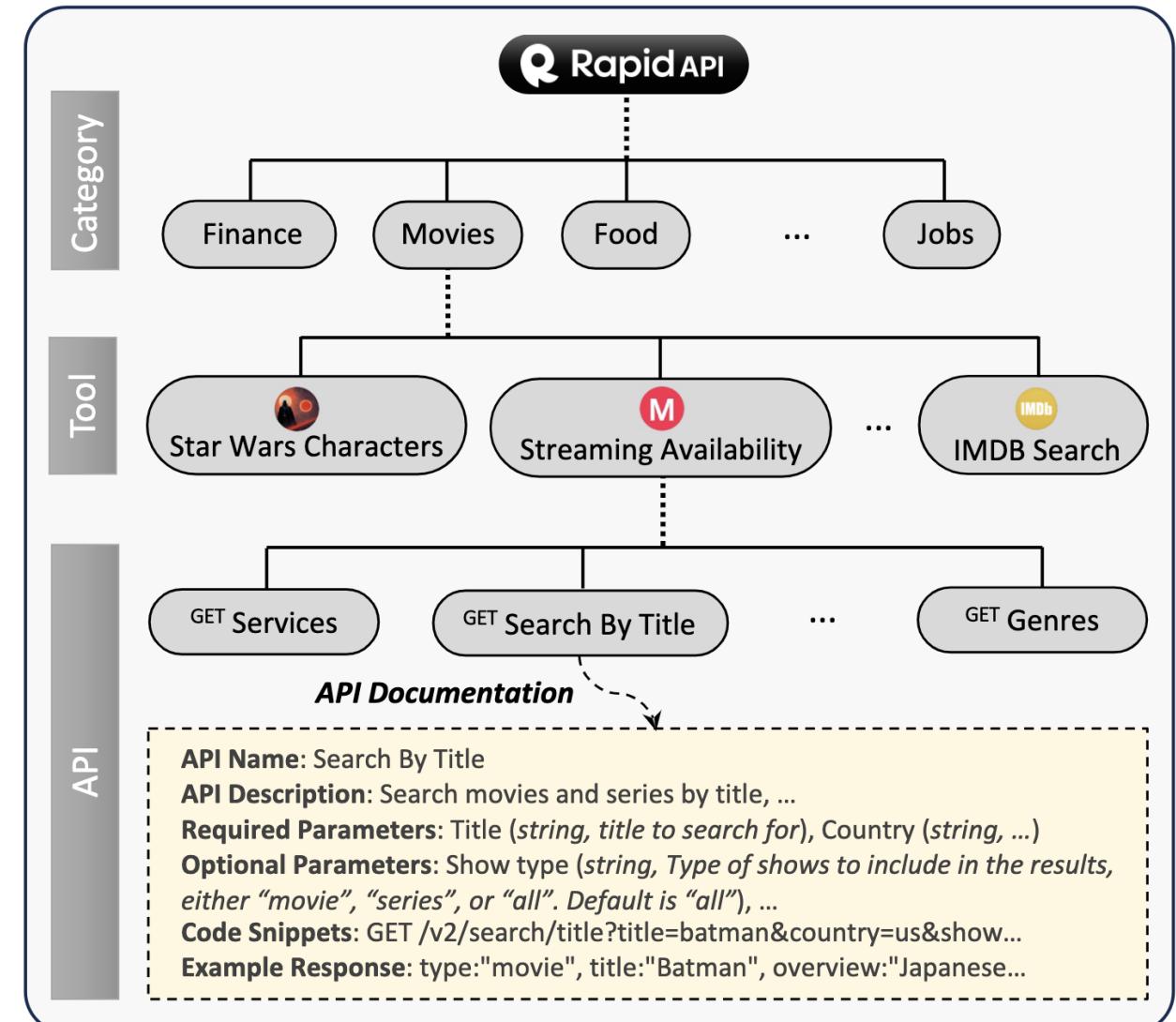
- Inference and evaluation of the pipeline



<https://github.com/OpenBMB/ToolBench>

Constructing Tool Bench

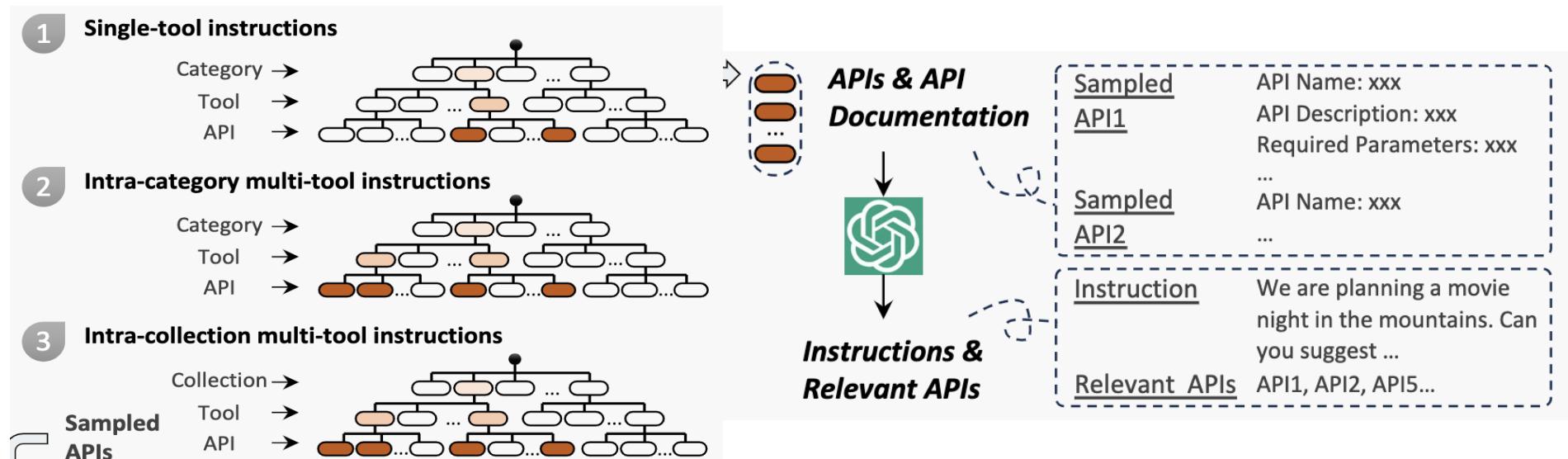
- API Collection
 - [RapidAPI Hub](#)
 - <https://rapidapi.com/hub>
- Filter from 50000+ APIs
- Retain 16000+ APIs
- 49 categories



Constructing Tool Bench

- Instruction Generation
 - Single-tool + Multi-tool
 - (1) sample a few APIs
 - (2) generating instructions and their relevant APIs

ChatGPT ($\{[\mathbb{S}_1^{\text{rel}}, \text{Inst}_1], \dots, [\mathbb{S}_{N'}^{\text{rel}}, \text{Inst}_{N'}]\} | \text{API}_1, \dots, \text{API}_N, \text{seed}_1, \dots, \text{seed}_3$).
 $\{\text{API}_1, \dots, \text{API}_N\} \in \mathbb{S}_{\text{API}}$,
 $\{\text{seed}_1, \dots, \text{seed}_3\} \in \mathbb{S}_{\text{seed}}$



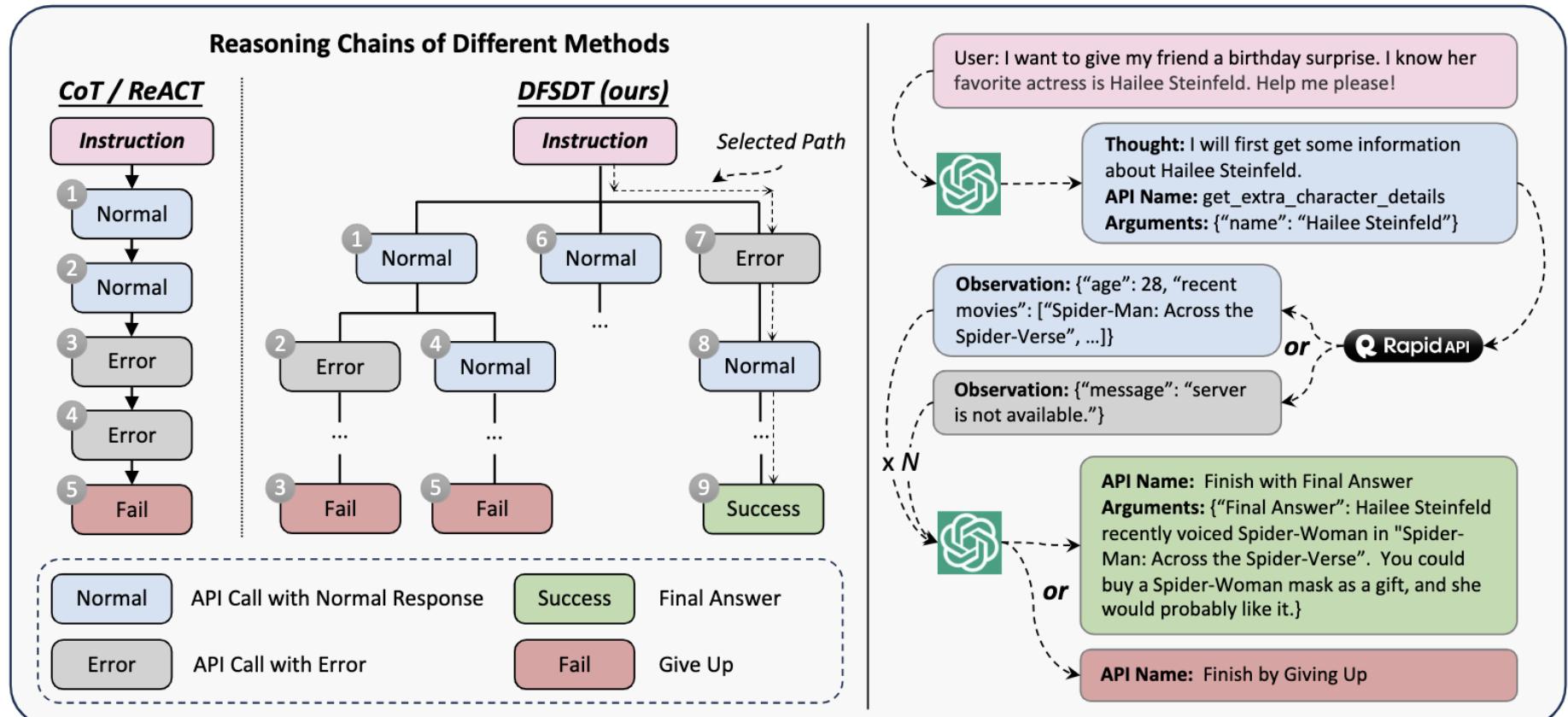
Constructing Tool Bench

- Answer Annotation
 - gpt-3.5-turbo-16k: function call feature
- Failure of ReACT / CoT
 - Error Propagation: a mistaken action may propagate the errors further
 - Limited Exploration: although the action space is infinite, CoT or ReACT only explores one possible path

Method	Single-tool (I1)	Category (I2)	Collection (I3)	Average
ReACT	43.98	23.62	20.42	29.34
ReACT@N	50.80	36.14	32.87	39.94
DFSDT	54.10	47.35	44.80	48.75

Constructing Tool Bench

- Answer Annotation

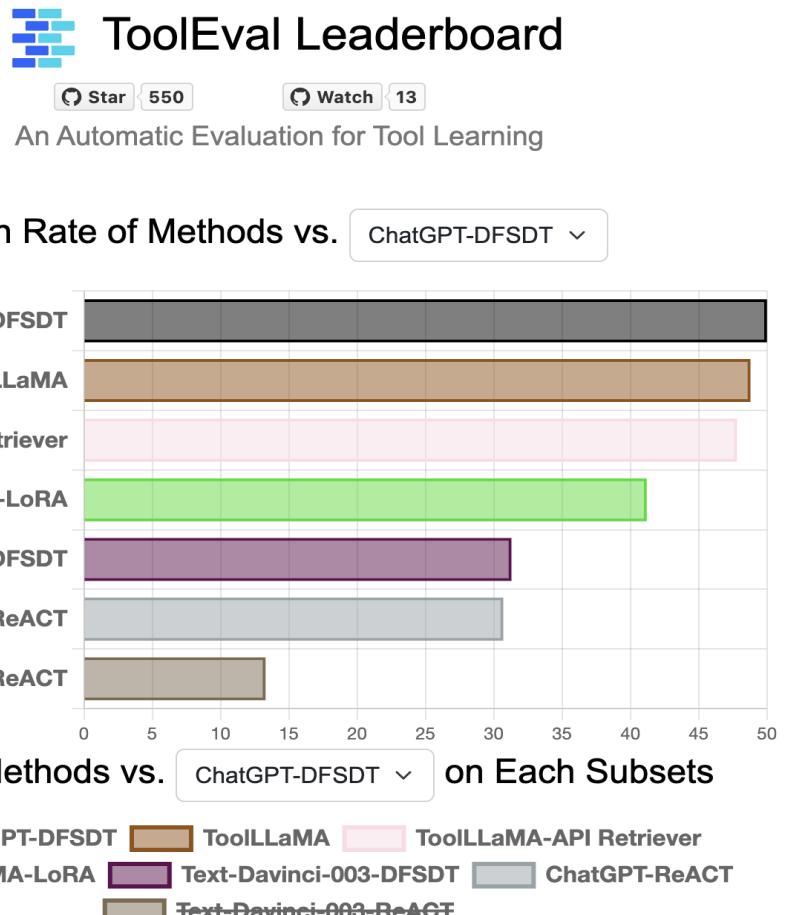


Tool Eval

- An automatic evaluator based on ChatGPT
- Two metrics:
 - Pass Rate: the proportion of completing an instruction within limited OpenAI API calls
 - Preference: the **quality/usefulness** comparison between two answers, i.e., which is better?
- High-correlation with human experts (~80%)

Tool Eval

- An automatic evaluator based on ChatGPT



Tool LLaMA

- Excellent generalization to unseen instructions and APIs
- Significantly better than ChatGPT+ReACT

Model	I1-Inst.		I1-Tool		I1-Cat.		I2-Inst.		I2-Cat.		I3-Inst.		Average	
	Pass	Win												
ChatGPT-ReACT	56.0	-	62.0	-	66.0	-	28.0	-	22.0	-	30.0	-	44.0	-
Vicuna (ReACT & DFSDT)	0.0	-	0.0	-	0.0	-	0.0	-	0.0	-	0.0	-	0.0	-
Alpaca (ReACT & DFSDT)	0.0	-	0.0	-	0.0	-	0.0	-	0.0	-	0.0	-	0.0	-
Text-Davinci-003-DFSDT	53.0	46.0	58.0	38.0	61.0	39.0	38.0	46.0	38.0	45.0	39.0	48.0	47.8	43.7
ChatGPT-DFSDT	78.0	68.0	84.0	59.0	89.0	57.0	51.0	78.0	58.0	77.0	57.0	77.0	69.6	69.3
ToolLLaMA-DFSDT	<u>68.0</u>	<u>68.0</u>	<u>80.0</u>	<u>59.0</u>	<u>75.0</u>	<u>56.0</u>	<u>47.0</u>	<u>75.0</u>	<u>56.0</u>	<u>80.0</u>	<u>40.0</u>	<u>72.0</u>	<u>61.0</u>	<u>68.3</u>

- DFSDT >> ReACT

Method	Single-tool (I1)	Category (I2)	Collection (I3)	Average
ReACT	43.98	23.62	20.42	29.34
ReACT@N	50.80	36.14	32.87	39.94
DFSDT	54.10	47.35	44.80	48.75

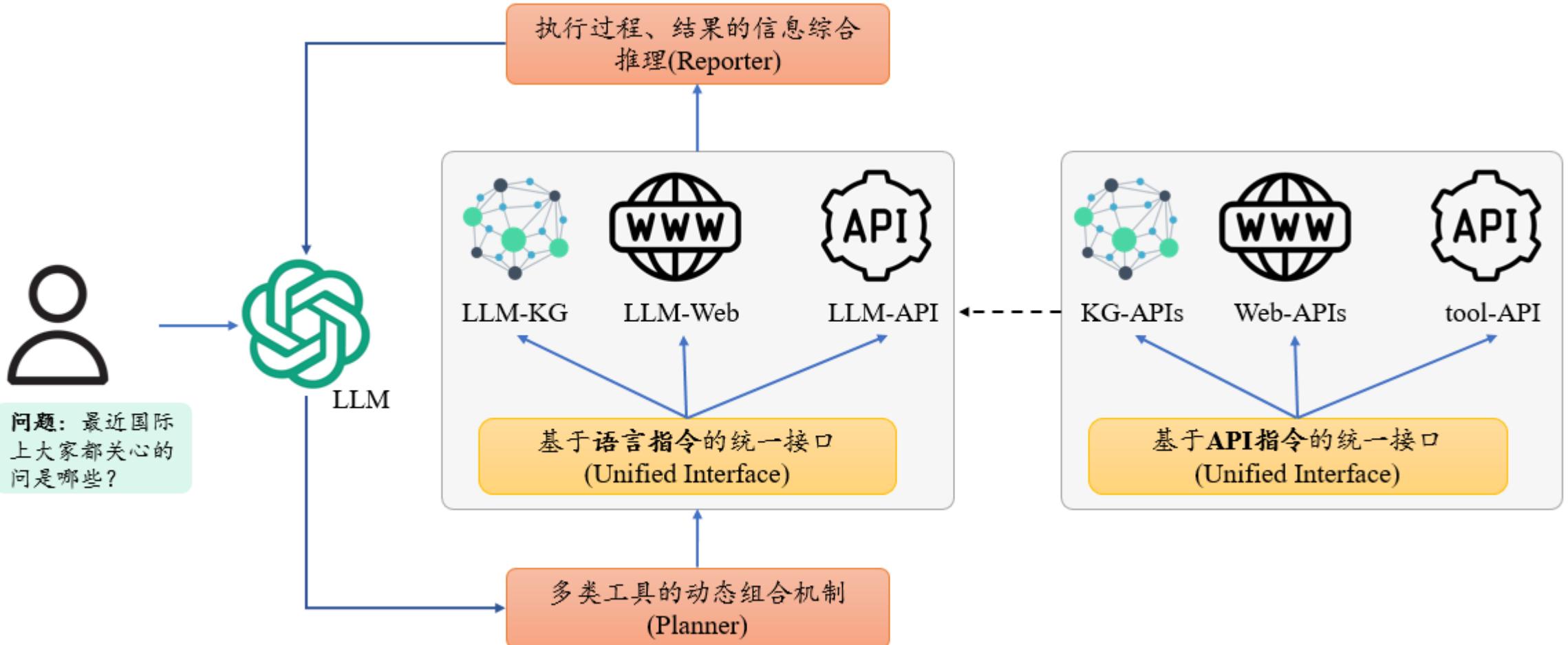
API Retriever

- Despite sifting through a large pool of APIs, the retriever exhibits remarkable retrieval precision, returning APIs closely aligned with the ground truth

Instruction	API Retriever (ours)			BM25			Ada Embedding		
	NDCG1	NDCG3	NDCG5	NDCG1	NDCG3	NDCG5	NDCG1	NDCG3	NDCG5
Single-tool (I1)	84.20	89.59	89.65	18.37	17.97	19.65	57.52	54.90	58.83
Category (I2)	68.24	77.43	77.90	11.97	9.85	10.95	36.82	28.83	30.68
Collection (I3)	81.65	87.24	87.13	25.23	18.95	20.37	54.59	42.55	46.83
All	75.73	83.19	83.06	15.84	13.98	15.63	46.59	41.06	43.95

Model	I1-Inst.		I1-Tool		I1-Cat.		I2-Inst.		I2-Cat.		I3-Inst.		Average	
	Pass	Win												
ToolLLaMA	68.0	-	80.0	-	75.0	-	47.0	-	56.0	-	40.0	-	61.0	-
→API Retriever	62.0	54.0	62.0	39.0	72.0	49.0	45.0	54.0	55.0	51.0	47.0	52.0	57.2	49.8
→ReACT	19.0	21.0	21.0	13.0	24.0	16.0	6.0	9.0	14.0	9.0	6.0	10.0	15.0	13.0
→LoRA	51.0	34.0	63.0	44.0	61.0	39.0	38.0	38.0	42.0	42.0	45.0	54.0	50.0	41.8

Multi-type Tool Collaborative Learning Framework





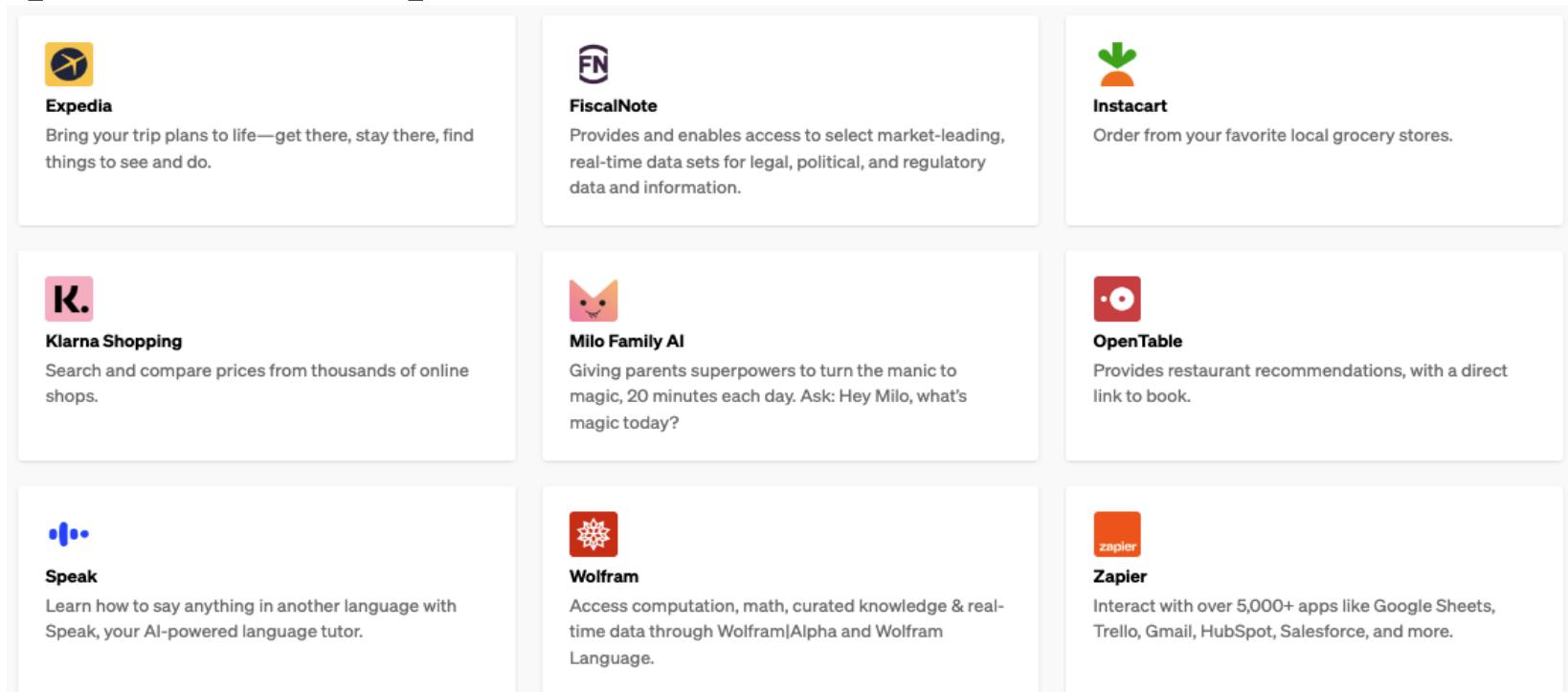
Outline

- Introduction of LLM
- LLM Performance in NLP
- Fine-Tuning Strategies & Challenges for LLM
- Categorization of Tools
- Benchmarking Tool Use
- Open-source Tool-Kit

Open-source Tool-Kit

ChatGPT Plugins

- OpenAI's official tool library
 - Empower ChatGPT with **broader applications**
 - By simply providing APIs with descriptions, ChatGPT is enabled to call applications and complete more complex tasks



 Expedia Bring your trip plans to life—get there, stay there, find things to see and do.	 FiscalNote Provides and enables access to select market-leading, real-time data sets for legal, political, and regulatory data and information.	 Instacart Order from your favorite local grocery stores.
 Klarna Shopping Search and compare prices from thousands of online shops.	 Milo Family AI Giving parents superpowers to turn the manic to magic, 20 minutes each day. Ask: Hey Milo, what's magic today?	 OpenTable Provides restaurant recommendations, with a direct link to book.
 Speak Learn how to say anything in another language with Speak, your AI-powered language tutor.	 Wolfram Access computation, math, curated knowledge & real-time data through Wolfram Alpha and Wolfram Language.	 Zapier Interact with over 5,000+ apps like Google Sheets, Trello, Gmail, HubSpot, Salesforce, and more.

ChatGPT Plugins

- Plugin: web browsing

TD Can you tell me what person/movie won the oscar for these categories?

- Best Actor
- Best soundtrack
- Best picture
- Best supporting actor

And then come up with a poem linking them together

Finished browsing ^

🔍 Searched: **"Oscar winners Best Actor Best Soundtrack Best Picture Best Supporting Actor"**

🔗 Clicked on: [CNN Oscar winners 2023: See the full list | CNN](#)

⬅ Going back to last page

🔗 Clicked on: [CNN Oscar winners 2023: See the full list | CNN](#)

⬅ Going back to last page

🔗 Clicked on: [ABC News List of 2023 Oscar Winners - ABC News](#)

≡ Reading content

☑ Finished browsing

ChatGPT Plugins

- Plugin: code interpreter

JW

a brick weighs 30% of a brick plus five GPUs
there are 100 bricks and 200 GPUs in a metric ton
how much does a brick weigh?



Let's call the weight of a brick "B" and the weight of a GPU "G". We can set up an equation based on the information given:

```
from sympy import symbols, Eq, solve
B, G = symbols('B G')
eq1 = Eq(B, 0.3 * B + 5 * G)
eq2 = Eq(100 * B + 200 * G, 1000)
solution = solve((eq1, eq2), (B, G))
solution[B]
```

```
# Result
7.812500000000000
```

Finished calculating ^

in the total
ual to 1

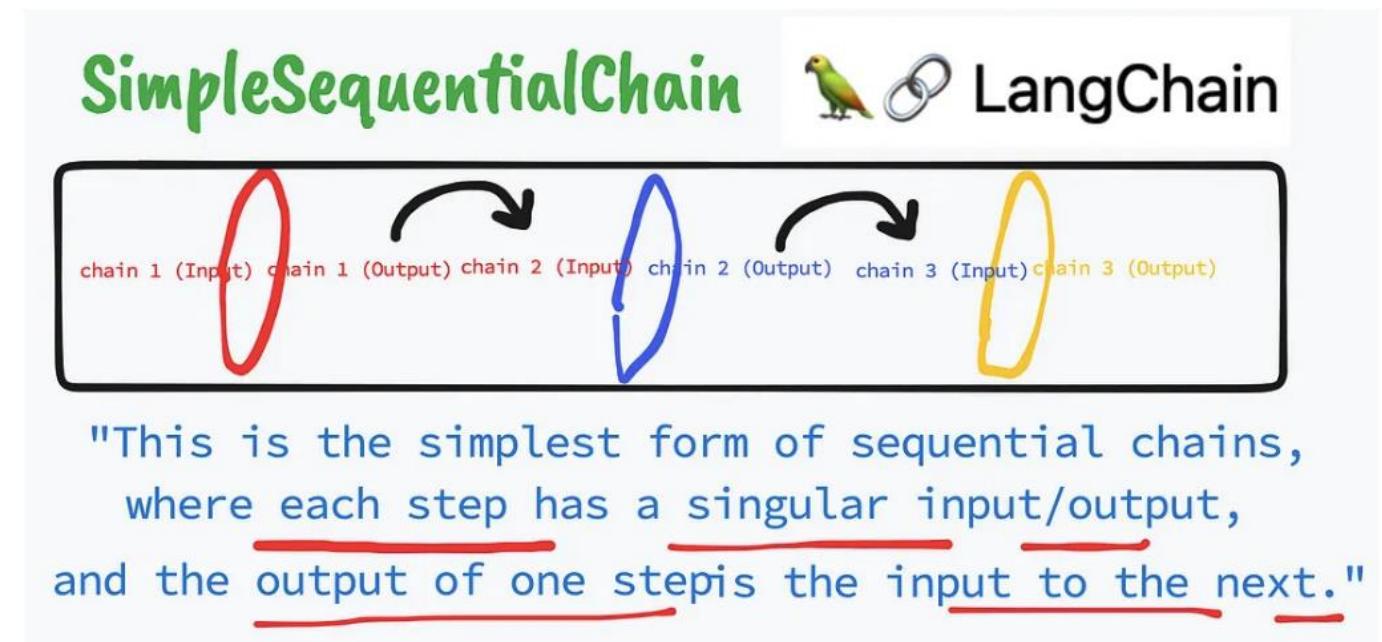
I the
(G).

The weight of a brick is 7.8125 kg.

Open-Source Solution

- **LangChain** 

- Integrated library to “chain” language models
- Also the first library to leverage tools



<https://langchain.readthedocs.io/>

Open-Source Solution

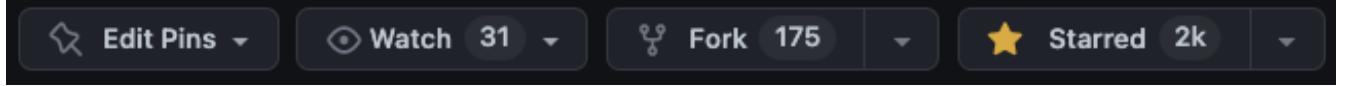
- BabyAGI

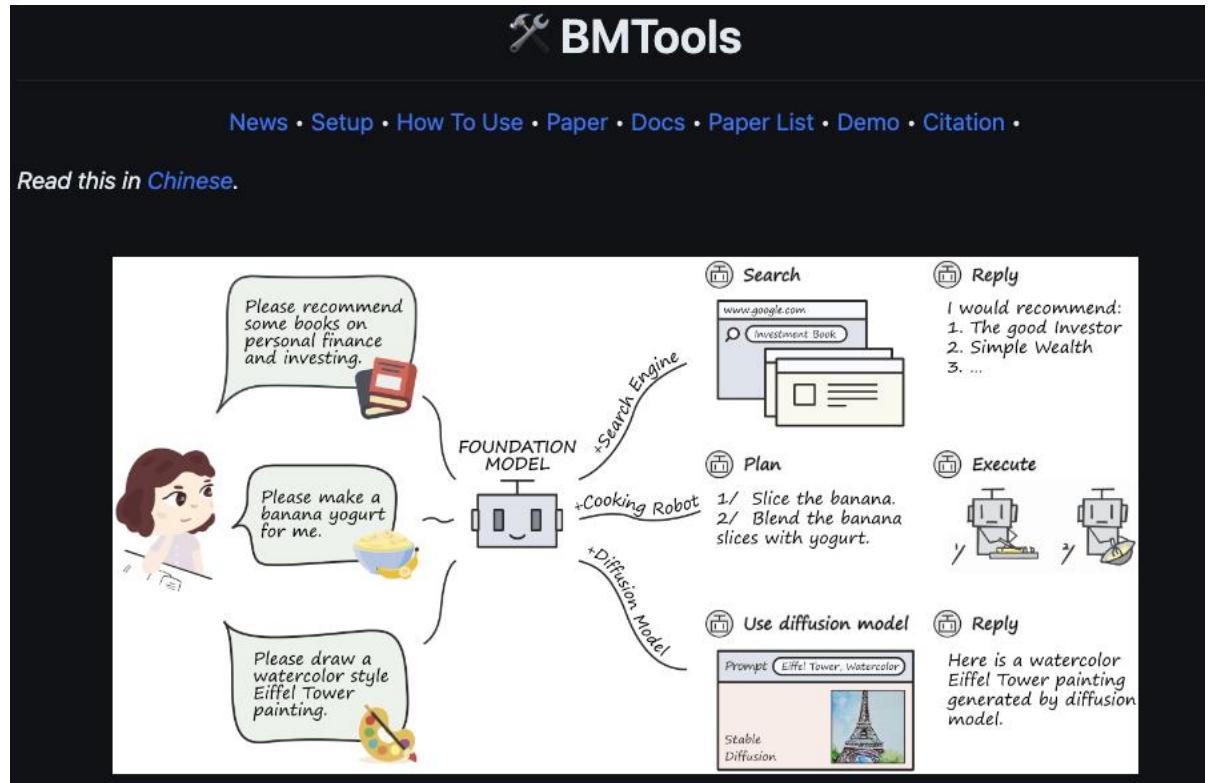
- Pull the first task from the task list
- Send the task to the execution agent, which uses ChatGPT to complete the task based on the context
- Enrich the result and stores it in Pinecone
- Create new tasks and reprioritizes the task list based on the objective and the result of the previous task



<https://github.com/yoheinakajima/babyagi>

Open-Source Solution

- **BMTools**  
- An open-source repository that extends language models to use tools and serves as a platform for the community to build and share tools



<https://github.com/OpenBMB/BMTools/tree/main>

- Features:
 - Users can easily build a **new** plugin by writing python functions and use external ChatGPT-Plugins
 - Users can host their **local models** (e.g., LLaMA, CPM) to use tools

Stanford Alpaca



CPM–Bee

双语大模型，集
多种语言能力于
一体



- Features:
 - 30+ tools supported, welcome contributing!



database



Weather API



PPT



Google Scholar



Hugging Face

Huggingface Models

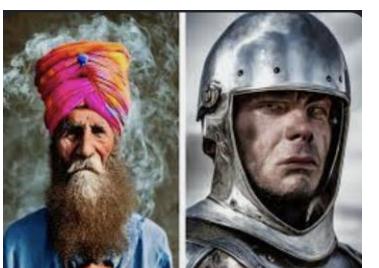


Image Generation

- AgentVerse
- From **individual** intelligence to **collective** intelligence
- Provide a flexible framework that simplifies the process of building custom multi-agent environments for LLMs
- Will support **multi-agent collaboration** for tool learning



<https://github.com/OpenBMB/AgentVerse>

- AgentVerse
 - From **individual** intelligence to **collective** intelligence
-

AgentVerse: Facilitating Multi-Agent Collaboration and Exploring Emergent Behaviors in Agents

Weize Chen^{1*}, Yusheng Su^{1*}, Jingwei Zuo¹, Cheng Yang^{2✉}, Chenfei Yuan¹,
Chen Qian¹, Chi-Min Chan¹, Yujia Qin¹, Yaxi Lu¹, Ruobing Xie³,
Zhiyuan Liu^{1✉}, Maosong Sun¹, Jie Zhou³

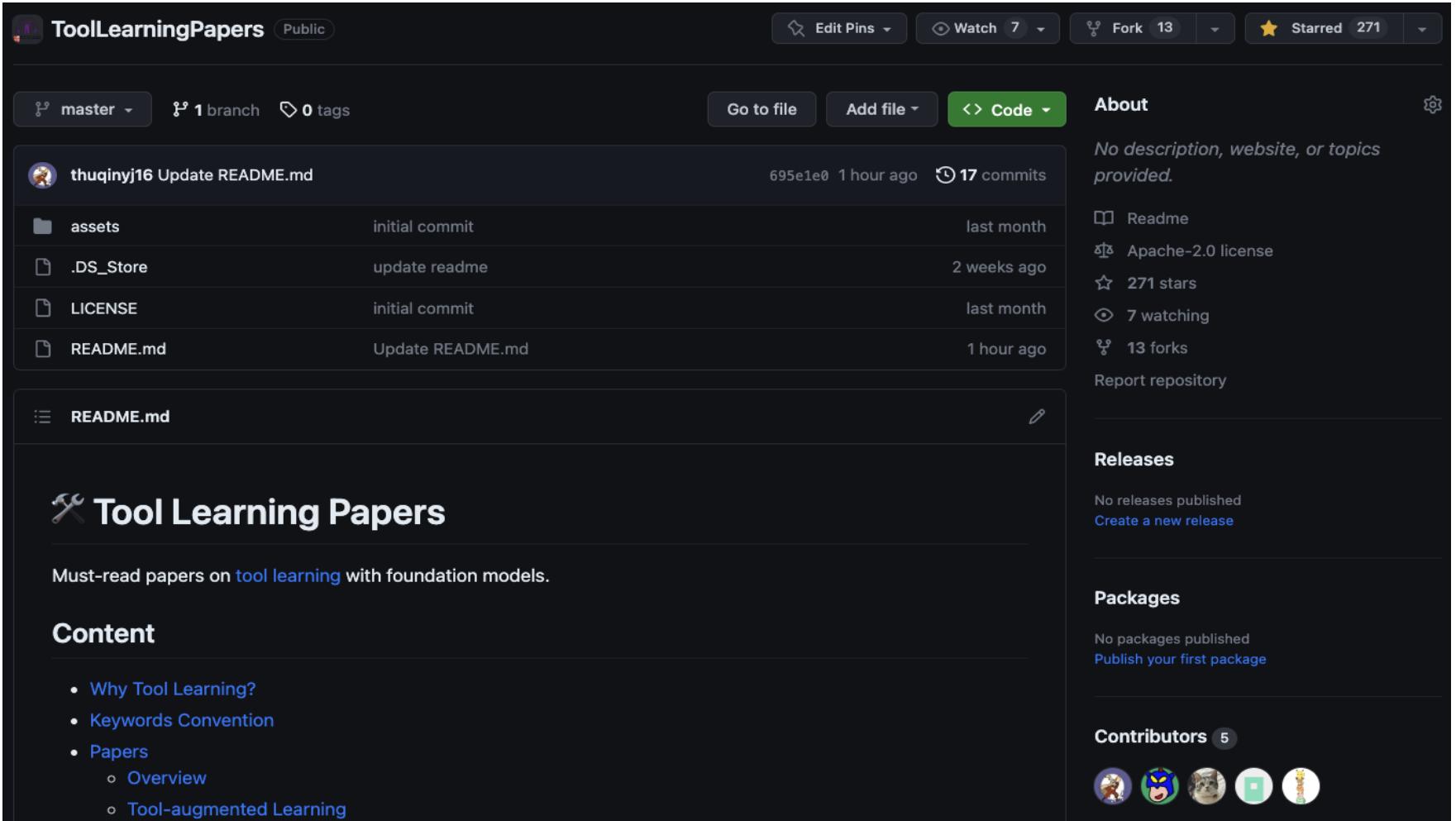
¹ Department of Computer Science and Technology, Tsinghua University

² School of Computer Science, Beijing University of Posts and Telecommunications

³ Pattern Recognition Center, WeChat AI, Tencent Inc.

chenwz21@mails.tsinghua.edu.cn, yushengsu.thu@gmail.com

• Tool Learning paper list



The screenshot shows the GitHub repository page for 'ToolLearningPapers'. The repository is public and has 1 branch and 0 tags. The commit history shows 17 commits from user 'thuqinyj16' over the last month. The repository has 271 stars, 13 forks, and 7 watchers. It includes sections for About, Releases, Packages, and Contributors.

About
No description, website, or topics provided.

Code

Commits

File	Commit Message	Time
assets	initial commit	last month
.DS_Store	update readme	2 weeks ago
LICENSE	initial commit	last month
README.md	Update README.md	1 hour ago

README

Tool Learning Papers

Must-read papers on [tool learning](#) with foundation models.

Content

- Why Tool Learning?
- Keywords Convention
- Papers
 - Overview
 - Tool-augmented Learning

About

No description, website, or topics provided.

Code

Commits

File	Commit Message	Time
assets	initial commit	last month
.DS_Store	update readme	2 weeks ago
LICENSE	initial commit	last month
README.md	Update README.md	1 hour ago

Readme

Apache-2.0 license

271 stars

7 watching

13 forks

[Report repository](#)

Releases

No releases published

[Create a new release](#)

Packages

No packages published

[Publish your first package](#)

Contributors 5



<https://github.com/thunlp/ToolLearningPapers>

Future Research Topic for LLM

NLP领域未来科研工作的发展方向掀起广泛的探讨

中国科学: 信息科学

SCIENTIA SINICA Informationis

评述

大模型时代的自然语言处理：挑战、机遇与发展

车万翔⁴, 窦志成¹¹, 冯岩松¹, 桂韬³, 韩先培¹⁰, 户保田⁵, 黄民烈⁶, 黄萱菁^{2*},
刘康⁹, 刘挺⁴, 刘知远^{6*}, 秦兵⁴, 邱锡鹏², 万小军¹, 王宇轩⁸, 文继荣¹¹, 严睿¹¹,
张家俊⁹, 张民^{5,7*}, 张奇², 赵军⁹, 赵鑫¹¹, 赵妍妍⁴

1. 北京大学王选计算机研究所, 北京 100080
2. 复旦大学计算机科学技术学院, 上海 200438
3. 复旦大学现代语言学研究院, 上海 200433
4. 哈尔滨工业大学计算学部, 哈尔滨 150001
5. 哈尔滨工业大学(深圳), 深圳 518055
6. 清华大学计算机科学与技术系, 北京 100084
7. 苏州大学计算机科学与技术学院, 苏州 215006
8. 之江实验室, 杭州 310000
9. 中国科学院自动化研究所, 北京 100190
10. 中国科学院软件研究所, 北京 100190
11. 中国人民大学高瓴人工智能学院, 北京 100872

* 通信作者. E-mail: xjhuang@fudan.edu.cn, liuzzy@tsinghua.edu.cn, zhangmin2021@hit.edu.cn

† 作者和单位排名不分先后, 同等贡献, 按拼音排序

Towards a comprehensive understanding of the impact of large language models on natural language processing: challenges, opportunities and future directions[†] (sciengine.com)

- LLM
 - Challenges
 - Opportunities
 - Future Directions
- 43 Pages
- NLP down-stream tasks
 - 文本分类
 - 结构化预测
 - 语义分析
 - 知识图谱与文本信息抽取
 - 情感计算
 - 文本生成
 - 自动文摘
 - **机器翻译**
 - 对话系统
 - 信息检索
 - 自动问答

Future Research Topic for LLM

A PhD Student's Perspective on Research in NLP in the Era of Very Large Language Models

Oana Ignat*, Zhijing Jin*, Artem Abzaliev, Laura Biester, Santiago Castro, Naihao Deng, Xinyi Gao, Aylin Gunal, Jacky He, Ashkan Kazemi, Muhammad Khalifa, Namho Koh, Andrew Lee, Siyang Liu, Do June Min, Shinka Mori, Joan Nwatu, Veronica Perez-Rosas, Siqi Shen, Zekun Wang, Winston Wu, Rada Mihalcea

LANGUAGE AND INFORMATION TECHNOLOGIES (LIT)

UNIVERSITY OF MICHIGAN

{oignat, jinzhi, mihalcea}@umich.edu

Abstract

Recent progress in large language models has enabled the deployment of many generative NLP applications. At the same time, it has also led to a misleading public discourse that “it’s all been solved.” Not surprisingly, this has in turn made many NLP researchers – especially those at the beginning of their career – wonder about what NLP research area they should focus on. This document is a compilation of NLP research directions that are rich for exploration, reflecting the views of a diverse group of PhD students in an academic research lab. While we identify many research areas, many others exist; we do not cover those areas that are currently addressed by LLMs but where LLMs lag behind in performance, or those focused on LLM development. We welcome suggestions for other research directions to include: <https://bit.ly/nlp-era-llm>

1 Background

Language models represent one of the fundamental building blocks in NLP, with their roots traced back to 1948 when Claude Shannon introduced Markov chains to model sequences of letters in English text (Shannon, 1948). They were then heavily used in connection to the early research on statistical ma-

idea: given a series of previous words or characters, we can predict what will come next. The new large language models (LLMs) benefit from two main developments: (1) the proliferation of Web 2.0 and user-generated data, which has led to a sharp increase in the availability of data; and (2) the growth in computational capabilities through the introduction of Graphics Processing Units (GPUs). Together, these developments have facilitated the resurgence of neural networks (or deep learning) and the availability of very large training datasets for these models.

Current LLMs have output quality comparable to human performance, with the added benefit of integrating information from enormous data sources, far surpassing what one individual can accumulate in their lifetime. The number of applications that benefit from using LLMs is continuously growing, with many cases where the LLMs are used to replace entire complex pipelines. LLMs becoming “lucrative” has led to a surge in industry interest and funding, alongside a sharp increase in the number of research publications on LLMs. For instance, a search on Google Scholar for “language models” leads to 50,000 publications over the past five years, a third of the roughly 150,000 papers published during the past 25 years.

- 博士研究生关于大语言模型时代自然语言处理研究的视角
- [2305.12544.pdf \(arxiv.org\)](https://arxiv.org/pdf/2305.12544.pdf)
- Multilinguality and Low-Resource Languages

Future Research Topic for LLM

Choose Your Weapon: Survival Strategies for Depressed AI Academics

Julian Togelius and Georgios N. Yannakakis*

April 14, 2023

Abstract

Are you an AI researcher at an academic institution? Are you anxious you are not coping with the current pace of AI advancements? Do you feel you have no (or very limited) access to the computational and human resources required for an AI research breakthrough? You are not alone; we feel the same way. A growing number of AI academics can no longer find the means and resources to compete at a global scale. This is a somewhat recent phenomenon, but an accelerating one, with private actors investing enormous compute resources into cutting edge AI research. Here, we discuss what you can do to stay competitive while remaining an academic. We also briefly discuss what universities and the private sector could do improve the situation, if they are so inclined. This is not an exhaustive list of strategies, and you may not agree with all of them, but it serves to start a discussion.

1 Introduction

As someone who does AI research in a university, you develop a complicated relationship to the corporate AI research powerhouses, such as DeepMind, Open AI, Google Brain and Meta AI. Whenever you see one of these papers that train some kind of gigantic neural net model to do something you weren't even sure a neural network could do, unquestionably pushing the state of the art and reconfiguring your ideas of what is possible, you get conflicting emotions. On the one hand: it's very impressive. Good on you for pushing AI forward. On the other hand: how could we possibly keep up? As an AI academic, leading a lab with a few PhD students and (if you're lucky enough) some postdocs, perhaps with a few dozen GPUs in your lab, this kind of research is simply not possible to do.

To be clear, this was not always the case. As recently as ten years ago, if you had a decent desktop computer and an internet connection you had everything you needed to compete with the best of researchers out there. Ground-breaking papers were often written by one or two people who ran all the experiments on their regular workstations. It is useful to point this out particularly for those who've come into the research field within the last decade, and for which the need for giant compute resources is a given.

- Try Scaling Anyway
- Scale Down
- Reuse and Remaster
- Analysis Instead of Synthesis
- **RL! No Data!**
- **Small Models! No Compute!**
- Work on Specialized Application Areas or Domains
- Solve Problems Few Care About (For Now!)
- Try Things that Shouldn't Work
- Do Things That Have Bad Optics
- Start it Up; Spin it Out!
- Collaborate or Jump Ship!
-

[2304.06035.pdf \(arxiv.org\)](https://arxiv.org/abs/2304.06035)

Future Research Topic for LLM

国内外一些优秀的 Survey 文章不能不重视

- [A Survey of Large Language Models](#)
- [Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond](#)
- [Tool Learning with Foundation Models](#)
- [Evaluating Large Language Models: A Comprehensive Survey](#)
- [Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models](#)
- [Interactive Natural Language Processing](#)
- [A Survey on In-context Learning](#)
- [Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and future Prospects](#)
- [Aligning Large Language Models with Human: A Survey](#)
- [A Comprehensive Survey of AI-Generated Content \(AIGC\)](#)
- [Explanation-Based Human Debugging of NLP Models: A Survey](#)
- [Augmented Language Models: a Survey](#)
- [Training language models to follow instructions with human feedback](#)
- [Deep reinforcement learning from human preferences](#)
- [A REVIEW ON LARGE LANGUAGE MODELS: ARCHITECTURES, APPLICATIONS, TAXONOMIES, OPEN ISSUES AND CHALLENGES](#)

Thank You!