2024 年 AI 大模型技术变迁情况回溯

中移智库

中国移动智慧家庭运营中心

2024 年 12 月

「水木人工智能学堂」

水木AI知识荟 & 交流社群📢

📚每日分享行业报告、行业资讯等！

🔗链接海量AI行业精英！

🎉不定时进行名校名企行活动！

🚀足不出门，尽在水木AI知识荟！

🔥扫码添加小编微信，免费进水木AI交流群

交流社群

去噪星球

去噪星球 每日仅需0.5元

公众号：水木人工智能学堂

摘要

2024 年大模型技术快速演进，相较于 2023 年的大小模型之争，技术的进步方向逐步向应用落地方向倾斜，降低端侧模型部署门槛，缩短模型推理时延，提升模型交互能力，大模型的发展迎来了新的变化。本文将从技术视角梳理 2024 年以来大模型各个领域发生的变化，以行业领先实践为佐证，提出大模型技术演进方向。语言大模型发展迎来新范式，通过强化学习优化内部思维链策略，以提升复杂逻辑推理能力。多模态大模型架构正向端到端演进，决策准确性和灵敏度提升推动机器人场景应用落地。在视频生成领域，DiTs 架构的可扩展性优势显现，推动 AI 应用商业化。在硬件部署方面，模型压缩、安全控制等技术正降低部署门槛。在智能体实践方面，垂类大模型开始在智能终端环境应用。在合成数据策略方面，自我奖励语言模型生成合成数据，试图打破数据瓶颈。

中移智库

一、 语言大模型领域：后训练阶段规模定律显现，以强化学习优化内部思维链策略或成大模型发展新范式

在2024年之前，语言类大模型的参数量快速扩展带来的“规模定律”获得业界普遍共识，因模型参数规模扩展、数据集质量提升以及人工微调为语言模型展现出前所未有的泛化能力和通用能力。而今年9月OpenAI公开发布o1推理大模型后，使得语言类大模型在解决专业科学、代码和数学模型等复杂逻辑推理问题的能力上更进一步。通过研究o1的技术原理发现，其以强化学习优化模型内部思维链推理逻辑步骤，模拟人的思考过程，以加深对问题的理解程度从而提升处理复杂推理任务能力。o1推理大模型的发布标志著语言大模型的“规模定律”正延展至后训练阶段，OpenAI首席技术官米拉穆拉迪称通过强化学习优化思考策略或将是未来大模型发展新范式。

通过强化学习学会了精炼其思维链并优化所用的策略，学会识别并纠正错误，将复杂的步骤分解为更简单的部分，并在

中移智库

当前方法无效时尝试不同的途径。通过这个过程显著提升了模型的推理能力。在多个高难度推理基准测试中，o1 的表现出色，超越人类专家和 GPT-4o，展示了其强大的推理能力和在某些领域的专业知识。

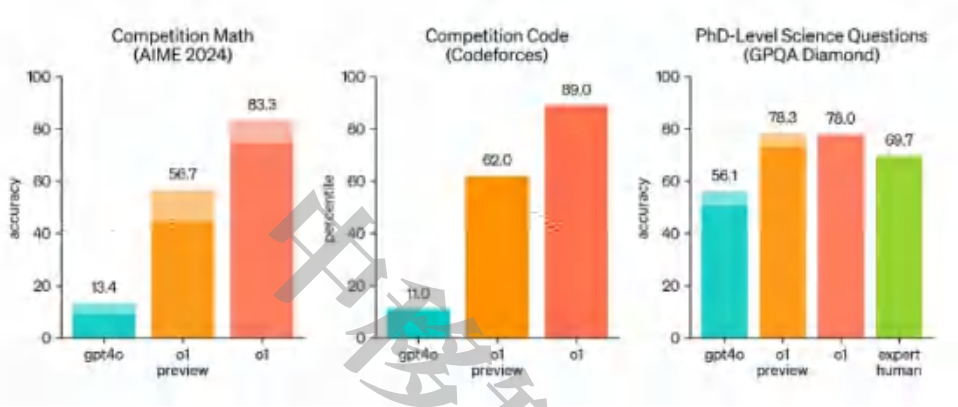


图 1：以图文匹配任务多模态大模型实现架构

无独有偶，斯坦福团队今年 3 月发布论文提出的 Q-STaR 语言模型也有类似的原理，利用强化学习优化中间图例过程，实现并行原理生产、混合原理预测、优化原理生成的能力。其测试在 7B 参数规模的语言类模型上表现优异，经调整后的语言模型在零样本测试准确率大幅提升。

二、 多模态大模型领域：主流模型架构从跨模态向端到端演进，提高决策准确性的同时提升模型灵敏度，以

满足无人驾驶、人形机器人应用场景下的需求

过去业界多模态大模型多采用基于语言模型为主干的跨模态架构，其往往通过模态特定的编码器（RNN、CNN）转化为统一的向量表示后再输入语言模型，依靠语言模型来处理模态融合后的特征交互。但是这样带来的问题是任务响应时间长、损失模态间交互细节。

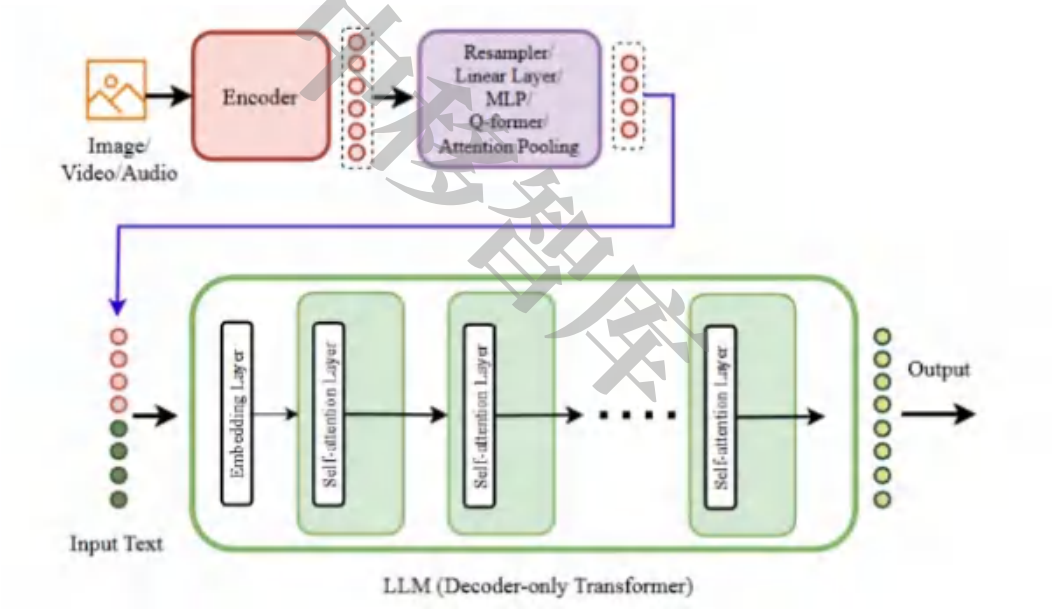


图 2：以图文匹配任务多模态大模型实现架构

2024 年以来以 GPT-4o、Gemini 为代表的多模态大模型纷纷开始使用端到端支持多种模态统一输入输出的模型架构。在该架构下引入分词器，通过将图像、音频等连续信号转换为离散

的 token 序列，然后与文本模态做统一表示，共同输入到基于自注意力的 Transformer 等模型中，实现端到端的学习。通过简化了模型的输入接口，减少模态间的信息损失，提升了模型处理即时任务的响应时间。

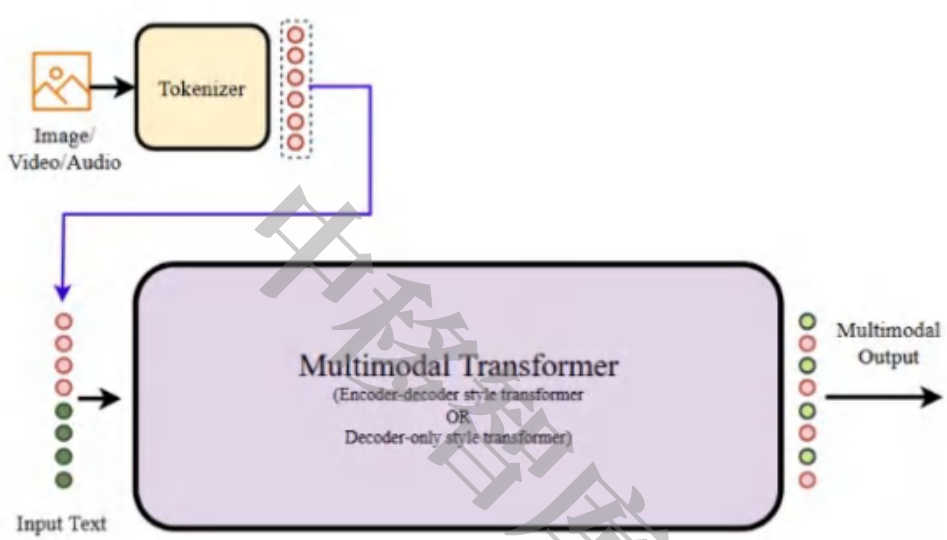


图 3：端到端多模态模型架构图

2024GTC 大会上，英伟达发布了人形机器人项目 GR00T，旨在开发人形机器人的通用基础模型，该模型便是基于控制、执行和决策三个层级分层实现模型的端到端训练学习，最终通过合并反向反馈来得到精准输出结果，相比于直接通过语言大模型来让机器人实现自主决策可大幅提升机器人处理复杂任务的

精度、高效性以及灵活性。相比于人形机器人，端到端架构更早应用于无人驾驶领域，特斯拉早在2023年便发布了史上第一个端到端AI无人驾驶系统——FSD Beta V12,打破了传统无人驾驶模块化、人为预设规则解决方案的定式，只需通过摄像头、激光雷达等传感器数据输入，无需任何预设规则，便可直接输出控制车辆方向和速度的操作指令，使得无人驾驶方案变得更高效、成本更低。

三、 视频生成领域：DiTs核心架构的可拓展性优势显现，基于数据处理、视频标注、音频模型的微创新推动视频生成应用更加平价高效，为AI应用带来更大商业化空间

2024年以来，国内外科技大厂发布的视频生成模型多以DiTs为基础，基于Transformer架构的扩散模型在视频生成任务中可扩展性优势凸显，即相较于原先的U-Net卷积网络架构，Transformer骨干架构可以提供基于参数规模和训练数据量提升而带来更优越的性能。同时通过Transformer的窗口注意力机制

可有效降低高维视频信号对算力的需求，解决 Transformer 输入序列长度增加带来内存巨额开销的问题。

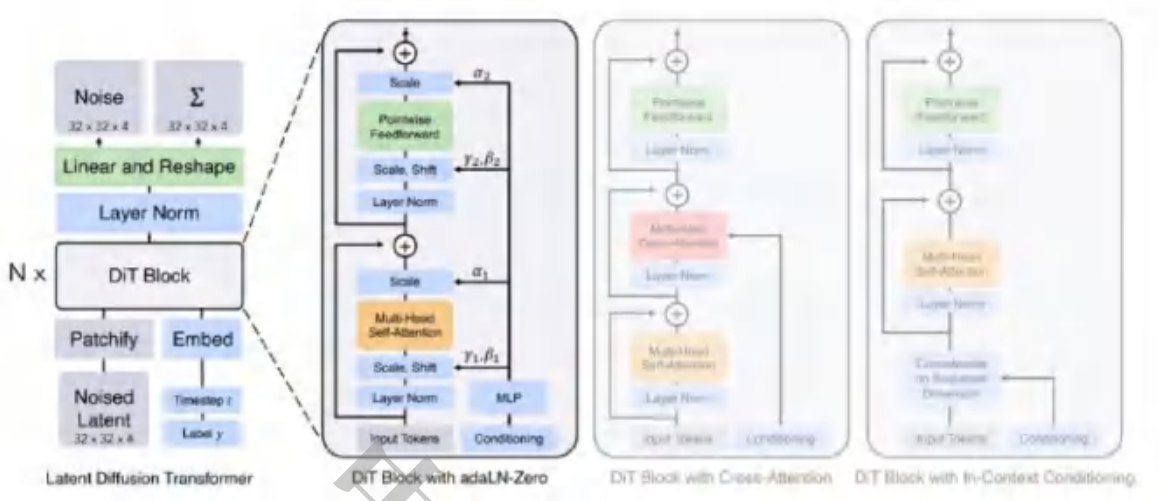


图 4：DiTs 模型架构中的注意力机制

2024 年 2 月，OpenAI 发布视频生成模型 Sora 便是基于 DiTs 架构，在生成视频的像素稳定性、前后逻辑连续性以及信息丢失等方面有大幅提升。Sora 除了采用 DiTs 架构以外，还在数据处理和视频标注领域做了创新。在数据方面，基于视频编码器将样本空间数据进行时间空间维度压缩和 Patch 化处理，再通过相应解码器实现隐空间向视频像素空间的映射，以训练新的视频压缩网络实现长视频生成的能力。在视频标注方面，复用

DALL-E3 的重标注技术，对视频数据生成高质量文字标注，借助 GPT 对提示词进行扩展从而提升视频生成效果。

2024 年 10 月，Meta Movie Gen 视频生成模型发布，其延续了原先视频生成模型架构的基础上，叠加了一个 13B 参数转换器模型 Meta Gen Audio，通过数百万个小时的音频参考数据的对比总结，Meta Gen Audio 可精准匹配声音和画面之间的对应关系，根据不同情绪和环境的提示词，找到与画面完美契合的音乐。

虽然视频生成模型的商业化仍处于早期，以国内公司为例如 Vidu、PixVerse、可灵等视频生成模型目前已经开面向 C 端采用订阅模式收费，年费标准版会员 4-5s 视频生成价格折合为 0.025-0.1 美元左右，面向 B 端的 API 调用价格暂未确定，但伴随着架构持续成熟以及各类创新技术推动下，视频生成有望更加平价高效，为 AI 应用带来更大商业化空间。

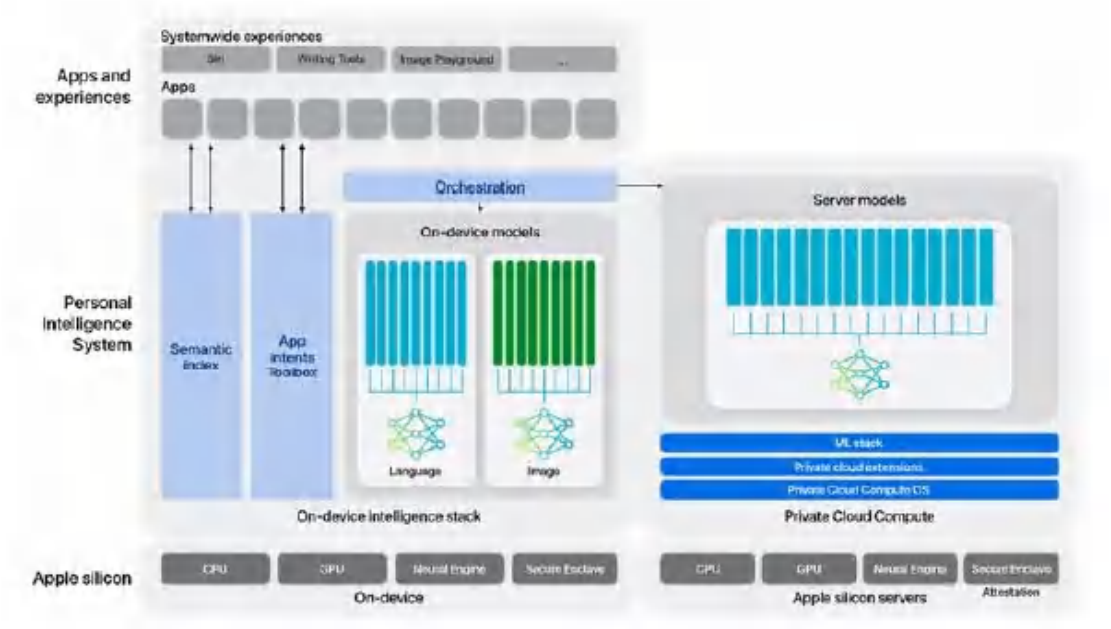
四、 硬件部署实践方面：在端云结合架构下，模型

中移智库

压缩、安全控制、闪存运行以及推理优化降低大模型硬件部署门槛，为AI+硬件赋能筑基

苹果作为智能硬件全球领先的科技公司，在软硬件、操作系统领域拥有强势地位，其在2024年6月发布的Apple Intelligence便为大模型硬件部署实践提供了很好的指引。

Apple Intelligence采用端云结合方案，即分别在设备端和服务器端部署大小语言模型（AFM-on-device和AFM-server），不同应用可以通过统一的语义索引、意图检测等工具调用AFM模型，当遇到复杂任务超出端侧模型处理能力时，任务会被发送至服务器端模型进行处理。



中移智库

图 5：Apple Intelligence 端云结合架构

为保障在端侧设备上运行模型同时避免精度上损失，苹果创新推出量化压缩叠加适配器的架构，一方面采用量化压缩的方法降低模型大小，同时通过 LoRA 适配器来恢复量化模型的精度。该适配器本身也是由特定任务上精度恢复训练得到，训练与开发成本较低，能够较好平衡模型良好性能和模型轻便化需求。

为保障模型安全可控，苹果制定了 Responsible AI 原则——用户赋能，即工具仅提供智能工具，尊重用户选择及隐私；代表全球，服务全球用户避免种族歧视；谨慎设计，开发设计过程设施保护措施，防止 AI 工具滥用或产生风险；隐私保护，通过端侧离线处理和云基础设施创新实现隐私保护，例如使用私有云计算来保护用户数据及隐私。该四大原则被整合到基础模型开发的每一个环节中，包括数据的收集与处理、模型训练、模型评估、用户反馈等。

中移智库

为提升大模型端侧运行流畅度，苹果研发了 LLM-in-a-flash 技术，让大模型可以不受限于 DRAM 的限制，在推理时将参数加载至闪存中来辅助完成计算，分担存储压力，从而降低端侧设备部署大模型门槛。具体来说，苹果以“滑动窗口”和“低秩预测器”方式精简加载的参数数据量，以大幅缩短数据从闪存加载至 DRAM 因速度慢导致的时延。通过“行列捆绑”技术借助大模型矩阵运算数据连续存放的特性，发挥闪存顺序读取速度的优势，从而加速闪存的读取传输。另外还通过 LazyLLM 预填充动态剪枝技术提升模型的推理速度，苹果研究人员发现模型在推理预填充阶段，仅有少部分的 token 是有用的，若对每一层生成的 token 进行剪枝，后续层无需对这些无效 token 进行计算，从而很大程度的节省计算量。在不损失模型精度的前提下大幅降低模型推理时的预填充计算量，缓解预填充阶段出现的计算瓶颈问题。

经测试，在 LLM-in-a-flash 和 LazyLLM 预填充动态剪枝技

中移智库

术的优化下，大模型推理的时延大幅降低，在端侧可运行的模型规模也得到大幅提升，为 AI+硬件赋能筑基。

Configuration				Performance Metrics			
Hybrid	Predictor	Windowing	Bundling	DRAM (GB)	Flash → DRAM(GB)	Throughput (GB/s)	I/O Latency (ms)
✗	✗	✗	✗	0	13.4 GB	6.10 GB/s	2130 ms
✓	✗	✗	✗	6.7	6.7 GB	6.10 GB/s	1090 ms
✓	✓	✗	✗	4.8	0.9 GB	1.25 GB/s	738 ms
✓	✓	✓	✗	6.5	0.2 GB	1.25 GB/s	164 ms
✓	✓	✓	✓	6.5	0.2 GB	2.25 GB/s	87 ms

Tasks	Method	Llama 2		XGen	
		Score	TTFT Speedup (×)	Score	TTFT Speedup (×)
Single-Document QA	Baseline	**25.79**	1.00	**25.19**	1.00
	Random Token Drop	20.05	1.20	18.32	1.58
	Static Token Pruning	21.89	1.18	19.30	1.61
	Prompt Compression	22.88	0.12	15.31	0.20
	LazyLLM (Ours)	25.59	**1.36**	25.00	**1.96**
Multi-Document QA	Baseline	**22.43**	1.00	**20.71**	1.00
	Random Token Drop	16.77	1.19	14.86	1.37
	Static Token Pruning	19.93	2.16	17.23	2.11
	Prompt Compression	8.42	0.13	11.56	0.19
	LazyLLM (Ours)	22.31	**2.34**	20.68	**2.65**

图 6：采用 LLM-in-a-flash 和 LazyLLM 提出的优化方法模型推理时延显著降低

五、 智能体实践方面：以面向 UI 交互与操作的垂类大模型核心，结合用户意图理解及应用逻辑推演，开始在 Android、Windows 等智能终端环境应用

2024 年以来，AI Agent 领域出现诸多进展。围绕 UI 交互与

操作的模型相比传统大语言模型、多模态模型在基于手机、平板等智能终端实现UI界面理解、数字推理任务领域具备更好的表现能力，更适应智能体在智能终端复杂UI环境场景下落地应用。

早在2023年10月，Adept公司（Adept由前谷歌大脑主管和OpenAI工程副总裁David LUAN创立，公司成立初衷便是打造AI teammate类通用操作工具来帮助人完成工作。）就正式发布并开源80亿参数多模态大模型Fuyu-8B，其具备图表、图形和文本理解能力之外，能够厘清复杂图形中元素的相互关系，类似手机内各类APP中button的意义，并能够根据用户指令准确归纳图表信息。2024年1月，基于Fuyu-8B发布了Fuyu-Heavy多模态模型，进一步加强模型在UI界面理解和数学推理能力，以及适配多平台的可扩展性。在规模仅为传统多模态模型5%-10%的基础下，在多项基准测试以及标准文本测试中不输GPT-4V和Gemini Ultra。

中移智库

	MMMU	VQAv2	AI2D	ChartQA
Adept Fuyu-Heavy	48.3	76.2	81.2	75.4
Gemini Pro	47.9	71.2	73.9	74.1
Gemini Ultra	59.4	77.8	79.5	80.8

	MMLU	GSM8K	MATH	HumanEval
Adept Fuyu-Heavy	72.1	82.9	29.5	58.0
Gemini Pro	71.8	86.5 (Maj1@32)	32.6	67.7
Inflection-2	79.6	81.4	34.8	44.5
Grok-1	73	62.9	23.9	63.2

图 7：Fuyu-Heavy 具备出色多模态推理和文本生成能力

苹果也在今年发布了自己首个手机端 Agent 的多模态大模型——Ferret-UI，在理解屏幕整体功能基础上，能够基于人机对话自主推断任务并提出相应可行操作，从而帮助用户完成界面导航等开放式任务的能力在这个模型上得到了加强。

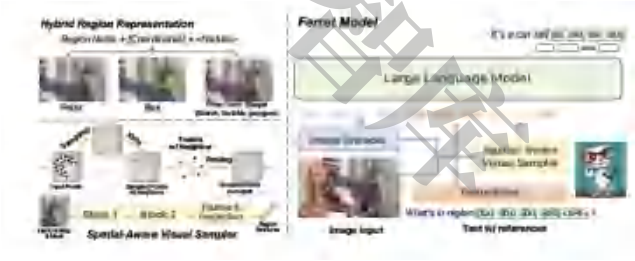


图 8：：Ferrert 模型核心架构

其通过图形编码器和空间感知视觉采样器来处理 UI 屏幕界面里图像嵌入以及混合区表示的输入问题，混合区域表示输入技术（Hybrid Region Representation）是在语言模型下提升引用、定位能力以及二者间紧密程度的创新技术，提升语言模

型理解和描述图形元素的能力。Ferret-UI 不仅在架构上做了特殊调整，其训练数据集也包含大量的UI 任务训练样本，可以有效的加强模型对 UI 任务的理解和执行能力。在 iPhone 和 Andriod 测试承接中，Ferret-UI 在处理大多数 UI 基础任务时，准确性高于 GPT-4v。

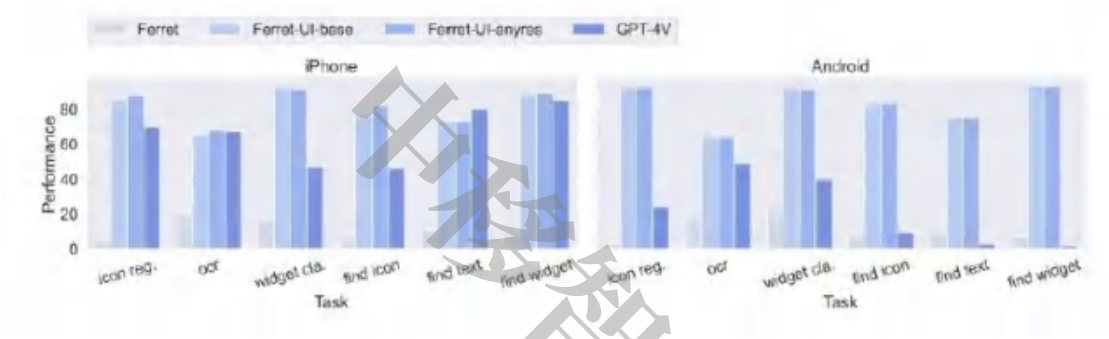


图 9：：Ferrert-UI 处理能力强于 GPT-4v

六、 大模型合成数据策略方面：以自我奖励语言模型生成合成数据，打破 Scaling Law“数据墙”瓶颈

Scaling Law 的“数据墙”问题正成为当前大模型迈向通用人工智能道路上的瓶颈，有机构预言（巴克莱投资银行在《AI 的下一步是什么》中提及，随着 GPT5 向 GPT6 迈进的时刻，是合成数据技术需要发挥的时刻，否则缩放定律会崩溃，阻碍模型

的改进），互联网上所有文本数据可能在 GPT6 推出之时消耗殆尽，若想进一步提升大模型性能，拓展数据集扩展的能力将会成为大模型大厂的核心竞争力。

2024 年 7 月，Meta 发布的 Llama 3.1 405B 便是完全基于合成数据方式，使用自我奖励语言模型进行训练，其在训练的过程中并没有依赖任何人类编写的答案，而是完全基于 Llama 2 语言模型生成的合成数据。具体做法是先基于少量人工标注数据预训练一个初始模型 A，基于问题生成多个候选回复，然后让大语言模型 B 对自己生成的回复打分，并根据打分形成新的训练数据，从而继续训练模型。该过程是迭代进行，在每次迭代中模型的遵循指令能力和打分能力都会提升。

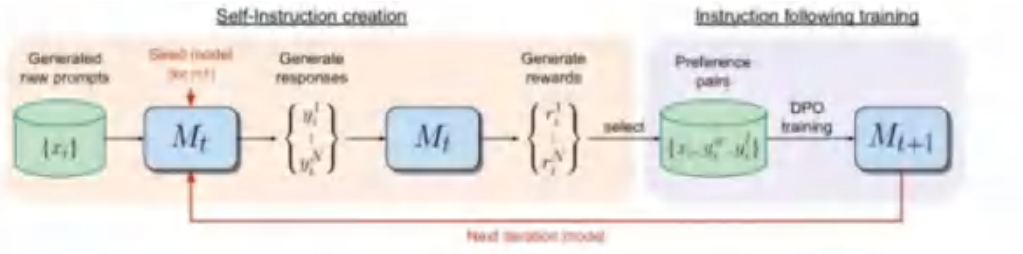


Figure 1: **Self-Rewarding Language Models.** Our self-alignment method consists of two steps: (i) *Self-Instruction creation*: newly created prompts are used to generate candidate responses from model M_t, which also predicts its own rewards via LLM-as-a-Judge prompting. (ii) *Instruction following training*: preference pairs are selected from the generated data, which are used for training via DPO, resulting in model M_{t+1}. This whole procedure can then be iterated resulting in both improved instruction following and reward modeling ability.

图 10：：基于合成数据的自我奖励模型训练机制

实验评估结果表明，Llama 3.1 405B 在常识、可操作性、数学、工具使用和多语言翻译等一系列任务中，都能与 GPT-4、GPT-4o 和 Claude 3.5 Sonnet 相媲美。在现实场景中，Llama 3.1 405B 进行了与人工评估的比较，其总体表现优于 GPT-4o 和 Claude 3.5 Sonnet。

中移智库

审核：王颢、马春山 | 智慧家庭运营中心

作者：赵石轩 | 智慧家庭运营中心