芯跑科技
iRUNtech

芯跑科技丨月报资讯

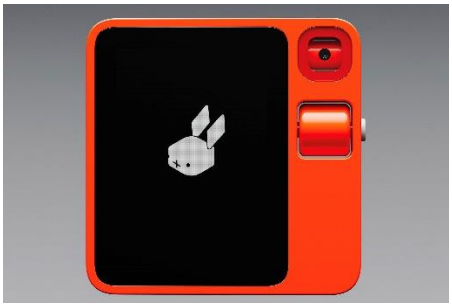
2024 年 2 月，第 050 期：AI Agent，重塑人机交互方式

AI Agent，重塑人机交互方式

前言：

比尔·盖茨在其 GatesNotes 上写到：“要在计算机上执行任何任务，您必须告诉设备要使用哪个应用程序。您可以使用 Microsoft Word 和 Google Docs 起草商业提案，但它们无法帮助您发送电子邮件、分享自拍照、分析数据、安排聚会或购买电影票。即使是最好的网站也无法完全了解您的工作、个人生活、兴趣和关系，并且使用这些信息为您做事的能力也有限。这种事情只有在今天与另一个人（例如亲密的朋友或私人助理）合作时才有可能实现。未来五年，这种情况将彻底改变。您不必为不同的任务使用不同的应用程序。您只需用日常语言告诉您的设备您想做什么。根据您选择与之分享的信息量，该软件将能够做出个性化响应，因为它将对您的生活有丰富的了解。在不久的将来，任何上网的人都将能够拥有一个由远远超出当今技术的人工智能驱动的个人助理。”【01】

比尔·盖茨还强调：“代理不仅会改变每个人与计算机交互的方式。它们还将颠覆软件行业，带来自我们从键入命令到点击图标以来最大的计算革命。”



图：Rabbit R1，新一代 AI Agent 交互设备

那 AI Agent 是如何做到的呢？有什么新的产品和模式？目前技术发展到哪一步？接下来，我们就试着回答这些问题。

一、关于 AI Agent 的定义、原理、分类及常见应用模式

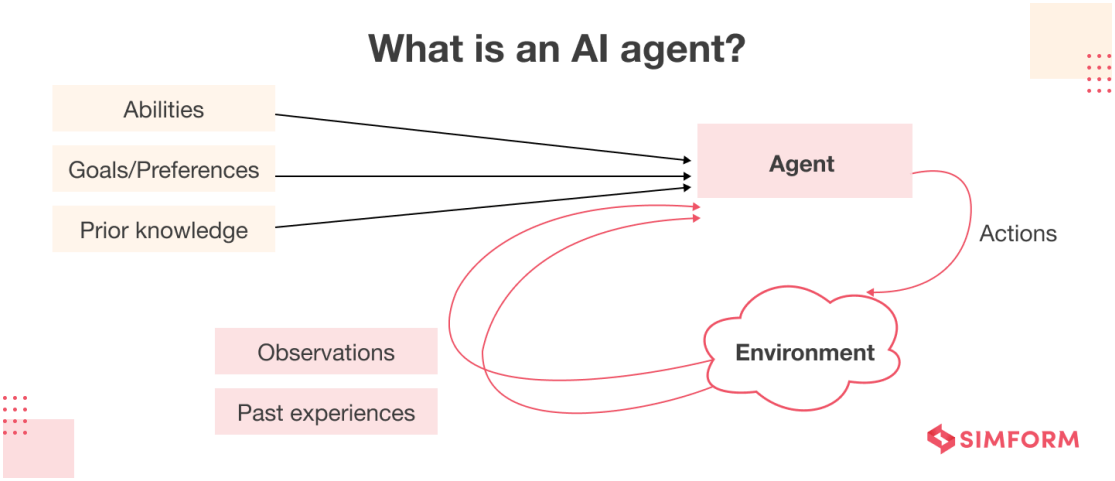
1、AI Agent 的定义

自 20 世纪 80 年代计算机科学家开始探索如何开发可以像人类一样交互的智能软件以来，人工智能代理就已出现。从那时起，这个概念已经发展到包括可以独立做出决策和完成任务的人工智能代理。

人工智能代理是一种软件程序，旨在与其环境交互，感知接收到的数据，并根据该数据采取行动以实现特定目标。人工智能代理模拟智能行为，它们可以像基于规则的系统一样简单，也可以像高级机器学习模型一样复杂。他们使用预先确定的规则或经过训练的模型来做出决策，并且可能需要外部控制或监督。

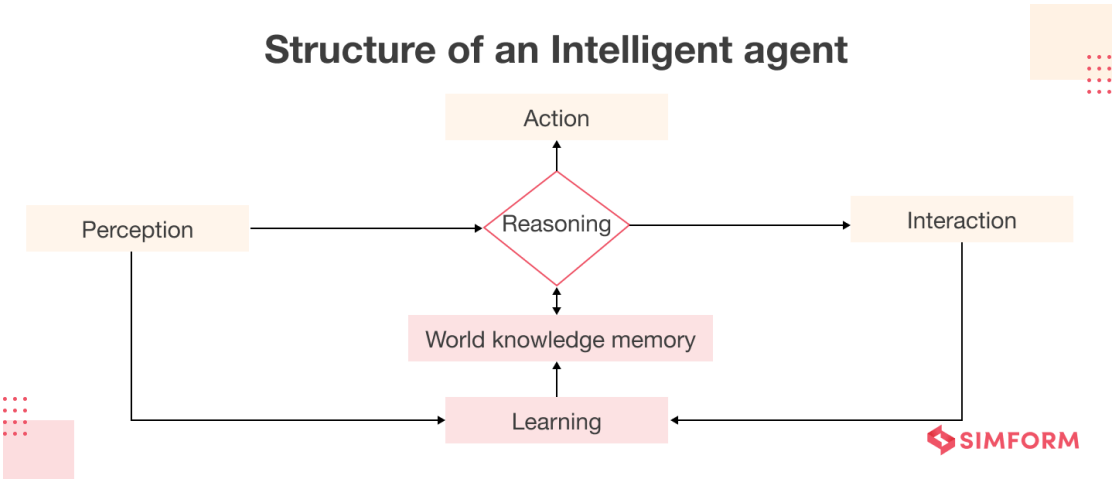
自主人工智能代理是一种先进的软件程序，可以在没有人类控制的情况下独立运行。它可以自行思考、行动和学习，无需人类不断输入。这些代理广泛应用于医疗保健、金融和银行等不同行业，使事情运行得

更顺畅、更高效。他们可以适应新情况，从经验中学习，并利用自己的内部系统做出决策。



图片来自：Simform

目前，与人工智能的交互遵循一种熟悉的仪式。您输入提示，AI 模型会根据输入计算响应。每次您想要新的输出时，您都必须提供提示。总是有人来启动这个过程。人工智能代理以不同的方式工作。他们被设计为独立思考和行动。您唯一需要提供的就是一个目标——研究竞争对手或购买披萨。**他们将根据环境的反馈和自己的内心独白生成一个任务列表并开始工作。就好像人工智能代理可以自我提示，不断发展和适应，以尽可能最好的方式实现他们的目标。**



图片来自：Simform

与主流自动化相比（您根据数据或系统状态设置一系列触发器并配置接下来发生的事情），人工智能代理可以在存在大量新信息的不可预测的环境中工作。这是人工智能自动化。人工智能代理也可以非常好地使用计算机。**他们可以浏览网页、使用应用程序、读写文件、使用信用卡付款，甚至可以控制您的笔记本电脑。**

这标志着向 AGI（通用人工智能）又迈进了一步。我们越来越接近这样的时刻：机器能够在任何主题或专业领域执行与人类相同的任务，并且具有完全的灵活性和卓越的性能。

2、AI Agent 如何工作?

当您输入目标时，AI Agent 会进行目标初始化。它将你的提示传递给核心 LLM（现在使用的是 GPT-3.5 和 GPT-4），并返回其内部独白的第一个输出，表明它理解它需要做什么。

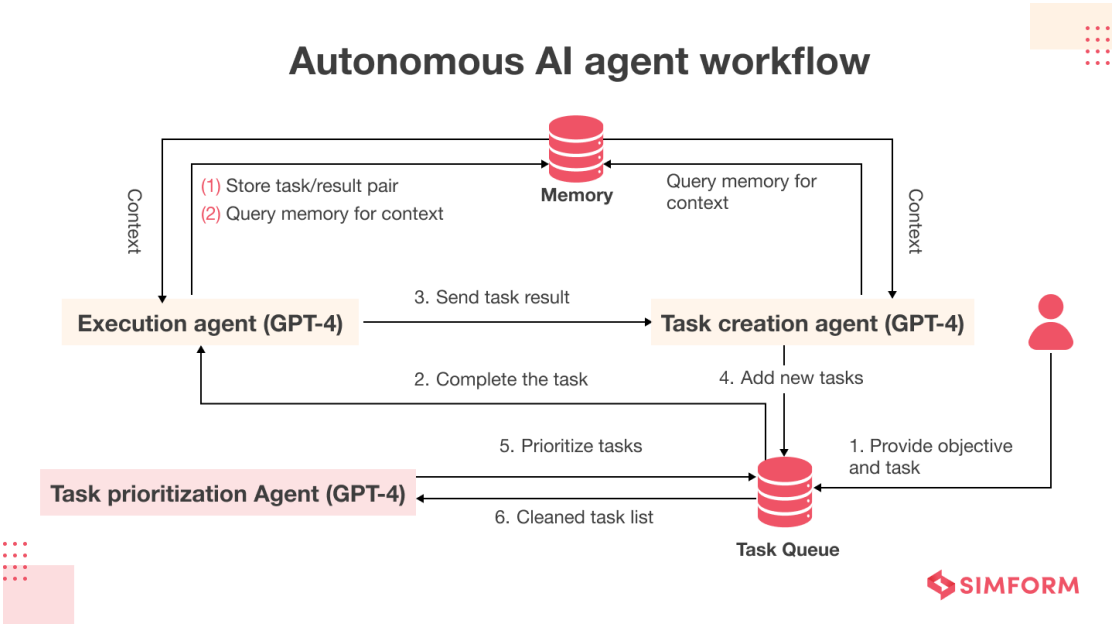
下一步是创建任务列表。根据目标，它将生成一组任务并了解应按什么顺序完成这些任务。一旦它决

定有一个可行的计划，它就会开始搜索信息。

由于 AI Agent 可以像您一样使用计算机，因此它可以从互联网收集信息。我们还看到一些代理能够与其他人工智能模型连接来外包任务和决策，让它们访问图像生成、地理数据处理或计算机视觉功能。

所有数据均由代理存储和管理，这样它就可以将其转发回给您，并在前进过程中改进其策略。

当任务从列表中划掉时，智能体通过从外部来源和内部独白收集反馈来评估距离目标还有多远。并且在实现目标之前，代理将不断迭代，创建更多任务，收集更多信息和反馈，并不断前进。



图片来自：Simform

一般来说，一个典型的 AI Agent 的算法框架由分析模块、内存模块、规划模块和操作模块组成。分析模块的目的是识别代理的角色。记忆和规划模块将代理置于动态环境中，使其能够回忆过去的行为并计划未来的行动。动作模块负责将代理的决策转化为具体的输出。在这些模块中，分析模块影响记忆和规划模块，这三个模块共同影响行动模块。

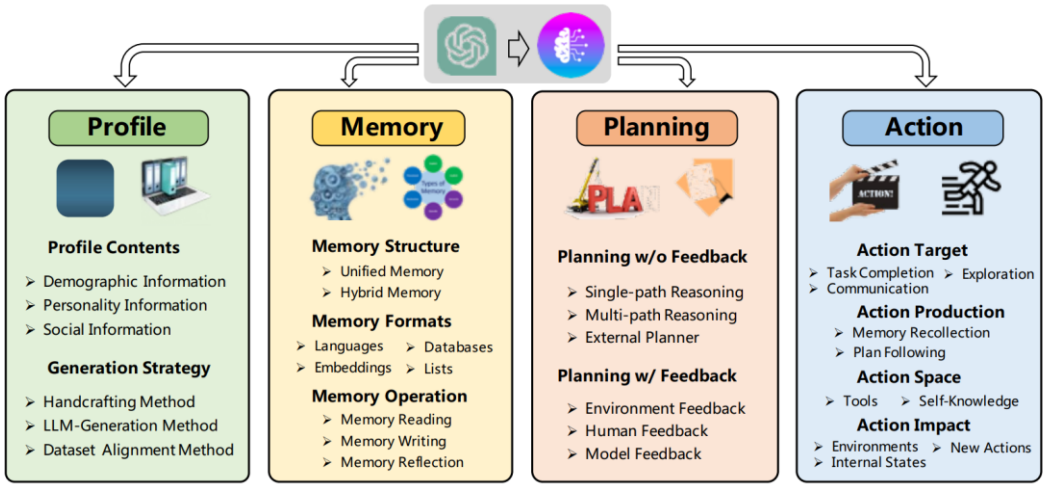


Figure 2: A unified framework for the architecture design of LLM-based autonomous agent.

图：一个典型的 AI Agent 的算法框架，来源：A Survey on Large Language Model based Autonomous Agents[02]

一般而言，在开发 AI Agent 之前，需要考虑以下几方面的信息：

- 定义机器人角色：

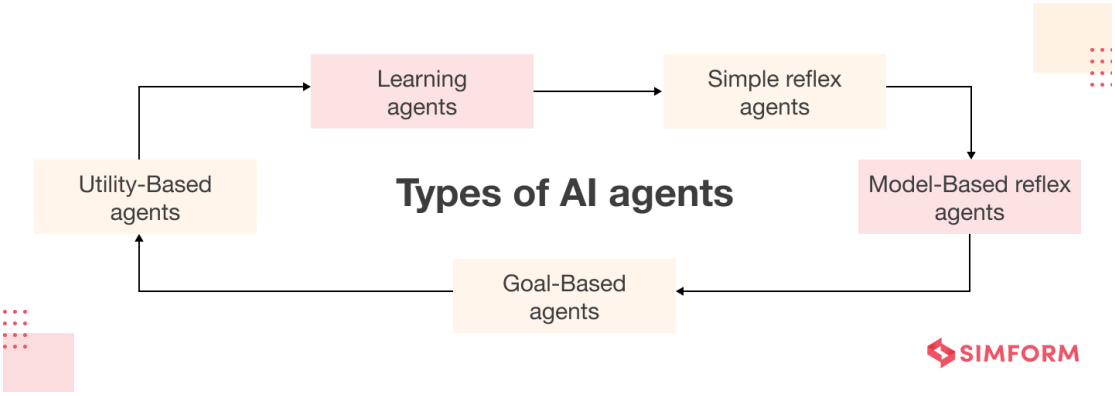
- 您的机器人应该是什么样子？
- 应该叫什么名字呢？
- 您的机器人有什么性格？它有性别吗？
- 您的机器人应该如何处理困难的情况和问题？

- 设计对话流程：
  - 您的用例可以期待什么类型的对话？
- 定义评估计划：
  - 您如何衡量成功？
  - 您希望使用哪些衡量标准来改进您的服务？

3、AI 代理有哪些类型？

以下是人工智能代理的五种主要类型：

- 简单的反射代理被编程为根据预定义的规则响应特定的环境刺激。
- 基于模型的反射代理是反应代理，它们维护环境的内部模型并使用它来做出决策。
- 基于目标的代理执行程序以实现特定目标，并根据评估当前环境状态采取行动。
- 基于效用的代理会考虑其行为的潜在结果，并选择能够最大化预期效用的结果。
- 学习代理执行机器学习技术来随着时间的推移改进他们的决策。



图片来自：Simform

4、AI Agent 的常见应用模式

对话式人工智能应用程序的实现有一些常见的模式：

知识机器人：知识机器人可以被设计为提供几乎任何主题的信息。例如，一个知识机器人可能会回答有关事件的问题，例如“本次会议有哪些机器人活动？”或“下一场雷鬼表演什么时候？”另一个机器人可能会回答与 IT 相关的问题，例如“如何更新我的操作系统？”还有一个机器人可能会回答有关联系人的问题，例如“约翰·多伊是谁？”或“Jane Doe 的电子邮件地址是什么？”

客服机器人：无论机器人拥有多少人工智能，有时它仍然可能需要将对话交给人类。在这种情况下，机器人应该识别何时需要切换并为用户提供平滑的过渡。

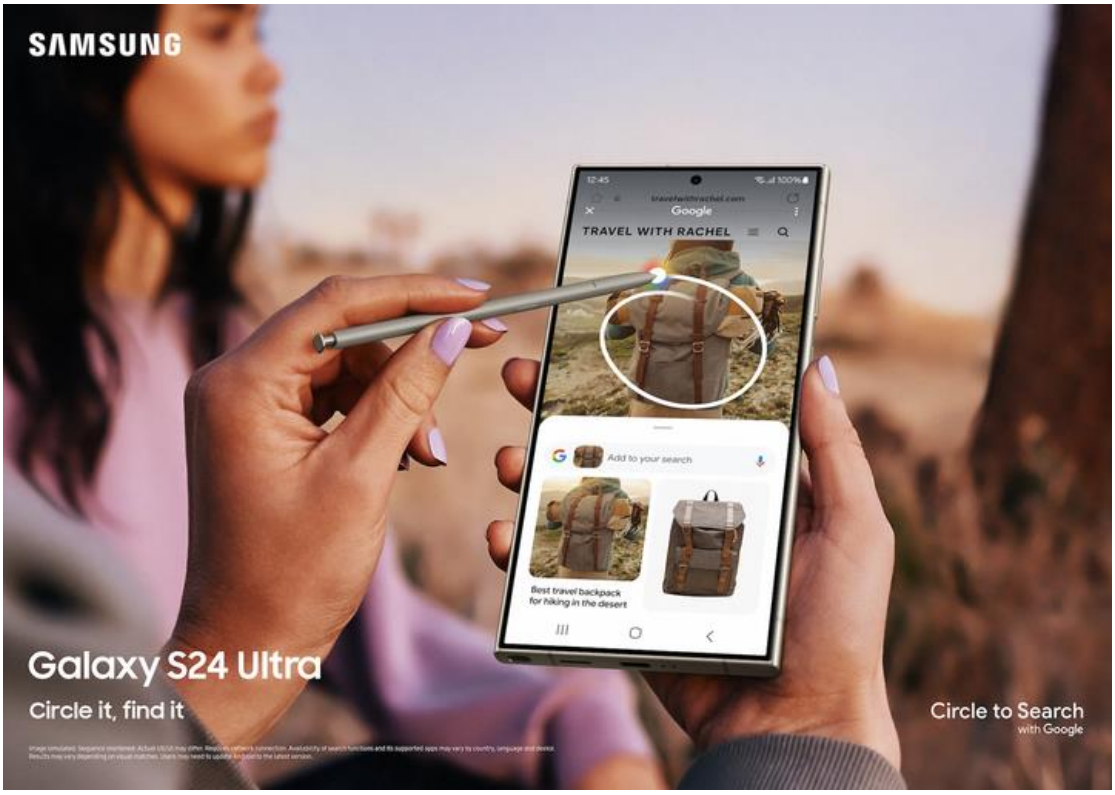
将机器人嵌入到应用程序中：虽然机器人通常存在于应用程序之外，但它们也可以与应用程序集成。例如，您可以在应用程序中嵌入知识机器人来帮助用户查找信息。您还可以在帮助台应用程序中嵌入机器人，作为传入用户请求的第一响应者。机器人可以独立解决简单的问题，并将更复杂的问题交给人工代理。

将机器人嵌入到网站中：与将机器人嵌入到应用程序中一样，机器人也可以嵌入到网站中，以实现跨渠道的多种通信模式。

二、AI Agent 产品及模式

2024 年既是 AI 爆发的元年，也是一个摸着石头过河的时期，没有人知晓最理想的 AI 终端该是什么形态。但毋容置疑的是，只有当 AI 终端无缝融入日常生活时，它才算得上是真正的「AI 神器」。

1、Samsung AI 手机 Galaxy S24



三星电子 2024 年 1 月份推出了 Galaxy S24 Ultra、Galaxy S24+ 和 Galaxy S24，利用 Galaxy AI 带来全新移动体验。人工智能增强了 Galaxy S24 系列上的几乎所有体验，从通过智能文本和通话翻译实现无障碍通信，到利用 Galaxy ProVisual Engine 最大限度地提高创作自由度，再到设定新的搜索标准，从而改变 Galaxy 用户发现周围世界的方式。

Galaxy AI 引入了有意义的智能，旨在增强生活的各个方面，尤其是手机最基本的作用：沟通。当您需要克服语言障碍时，Galaxy S24 可以让您变得比以往更容易。与来自国外的其他学生或同事聊天。在另一个国家度假时预订。借助 Live Translate，这一切都可以实现，即在本机应用程序中对电话进行 2 种 双向实时语音和文本翻译。不需要第三方应用程序，设备上的人工智能可以使对话完全私密。

借助 Interpreter，实时对话可以在分屏视图上即时翻译，这样站在对方对面的人就可以阅读对方所说内容的文本转录。它甚至无需蜂窝数据或 Wi-Fi 即可工作。

对于消息和其他应用程序，聊天助手可以帮助完善对话语气，以确保沟通听起来符合预期：就像给同事的礼貌消息或社交媒体标题的简短朗朗上口的短语。三星键盘内置的 AI 还可以实时翻译 13 种语言的消息。在车内，Android Auto 将自动汇总收到的消息并建议相关回复和操作，例如向某人发送您的预计到达时间，以便您可以在专注于道路的同时保持联系。

三星笔记中的 Note Assist 也让组织工作得到了极大的提升，它具有人工智能生成的摘要、使用预制格式简化笔记的模板创建以及封面创建功能，使笔记可以通过简短的预览轻松识别。对于录音，当有多个发言者时，Transcript Assist 使用 AI 和语音转文本技术来转录、总结甚至翻译录音。

通信并不是 Galaxy S24 系列在未来发挥手机基本优势的唯一方式。在线搜索几乎改变了生活的方方面面。Galaxy S24 标志着搜索史上的一个里程碑，它是第一款与 Google 一起推出直观、手势驱动的 Circle to

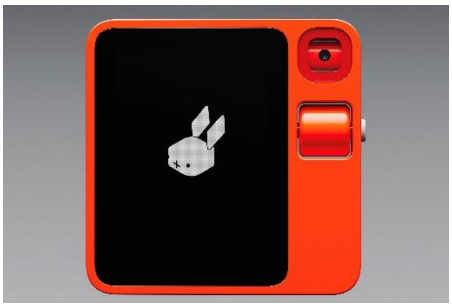
Search 的手机。

在三星手机的刺激下，OPPO 创始人陈明永在一封内部信中称：“2024 年是 AI 手机元年。未来 5 年，AI 对手机行业的影响，完全可以比肩当年智能手机替代功能机。从行业发展阶段来看，AI 手机也将成为继功能机、智能手机之后，手机行业的第三阶段。为此，公司已做好充分准备，并专门成立了 AI 中心，我们的资源将向 AI 集中”。除此之外，魅族也宣布 All in AI，将停止传统“智能手机”新项目，全力投入“明日设备” AI For New Generations。

产品成熟度高、应用生态完善、算力资源充足，都让 AI 手机、AI 电脑成为承载 AI Agent 的第一设备。受益于 AI 带来的新生产力的提升，手机、电脑等消费电子将迎来一波新的增长。

2、Rabbit R1

近期，科技初创公司 Rabbit 推出新一代 AI 互动硬件设备，一种小型橙色对讲机式设备。



图：R1

在拉斯维加斯消费电子展上预先录制的主题演讲中，Rabbit 的创始人 Jesse Lyu 要求该设备为他计划去伦敦度假；主题演讲展示了该设备为他设计行程并预订行程。他订购了披萨，预订了 Uber，并教设备如何使用 Midjourney 生成图像。（视频链接：https://www.youtube.com/watch?v=uJnhh7YSr5Q）

这款名为 Rabbit r1 的小工具是日益活跃的新硬件类别中的最新产品：便携式人工智能优先设备，可以使用自然语言与用户交互，避开屏幕和基于应用程序的操作系统。r1 的零售价为 199 美元，与 Humane Ai Pin（一款于 11 月推出的售价 699 美元的可穿戴设备，提供类似功能套件）以及售价 299 美元的 Meta 和 Rayban 智能眼镜（配有人工智能助手）相比，它是更便宜的竞争对手。

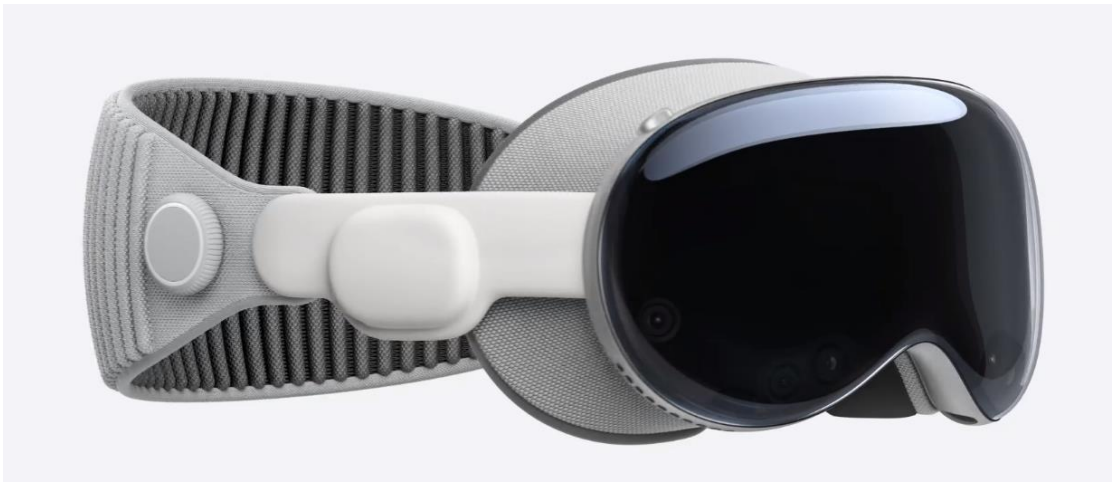
随着 Rabbit R1 的推出，科技行业掀起了一场风暴，这是一款挑战数字交互规范的突破性设备。Rabbit R1 旨在简化我们与智能手机和应用程序交互的方式。通过整合 Rabbit OS 的大型动作模型 (LAM)，该设备可以跨各种应用程序界面执行任务，模仿人类交互。此功能涵盖从发送消息到控制音乐，甚至购买杂货，所有这些都通过用户友好的界面进行。内部组件包括 MediaTek Helio P35 处理器、4GB RAM 和 128GB 存储空间，与其时尚的外观相得益彰。

Rabbit R1 不仅仅是一个 AI 小工具；它还是一个 AI 设备。这是创新的声明。它学习和自主复制动作的能力标志着人工智能技术的重大进步。尽管存在电池寿命和用户可训练性方面的问题，Rabbit Inc. 仍致力于持续改进，使该设备始终处于技术发展的前沿。

据该公司推文称，截至 1 月 18 日，第五批 10,000 台 r1 设备已售罄。第六批 50,000 台设备的预订现已开放 rabbit.tech，交货日期为六月和七月。微软首席执行官萨蒂亚·纳德拉 (Satya Nadella) 就是其中一位粉丝。“我认为兔子操作系统和设备的演示非常棒，”他告诉彭博社。“在（史蒂夫）乔布斯推出 iPhone 后，（这）可能是我见过的最令人印象深刻的演示之一，它捕捉到了未来的愿景。以代理为中心的操作系统和界面是有可能向前发展的。

3、Apple vision pro

Vision Pro 是苹果推出的 MR 头戴式显示设备，严格来说，vision Pro 不算是 AI Agent 的硬件，但从最近的 1-2 年智能硬件的发展脉络的视野来看，它算是一个里程碑式的产品。



苹果总裁库克称，Vison Pro 开创了一类新的计算设备，能将数字世界融入真实世界，从而实现增强现实（AR）。这款设备兼容 iOS 和 iPadOS 的各种软件，可以办公、娱乐，拍摄空间视频，并且只需要手、眼和语音就能交互。苹果称这种新的计算范式为「空间计算」。

Apple Vision Pro 的发布标志着模拟现实技术向前迈出了有趣的一步。该耳机利用先进的 AR 和 VR 功能，致力于创建现实和数字世界的无缝融合。

借用法国社会学家让·鲍德里亚 (Jean Baudrillard) 的拟像和模拟理论，我们能更好的理解 Vision Pro 对这个时代产生的影响，在鲍德里亚看来，现代社会已经进入了一个模拟阶段，符号和符号与现实世界的参考脱节，创造了一种超现实的状态。当对象或概念的表示取代了它应该表示的现实时，就会发生这种情况。通过创建一个让用户可以像真实一样进行交互的沉浸式虚拟环境，Apple 的 Vision Pro 可能会带领我们进一步进入这个超现实时代。[03]

增强现实： Vision Pro 能够将数字信息叠加到我们的物理环境中，体现了鲍德里亚拟像的第二阶段：现实的扭曲。它通过创建现实世界的数字孪生来实现这一点，该孪生并不完全是准确的表示，而是一个风格化的、用户友好的版本，其中包含附加信息和交互元素。这代表着我们对周围世界的体验可能发生重大转变，进一步模糊了真实与模拟之间的界限。

现实的幻觉： Vision Pro 还可以切换到完全虚拟现实，可以说概括了鲍德里亚拟像的第三阶段：现实的替代。在这个阶段，虚拟环境不仅扭曲了现实，而且完全取代了现实。这种新的“现实”并不是对任何真实事物的参考，而是一种独立的模拟。有人可能会说，在 Vision Pro 的超现实世界中，现实不再是先决条件。

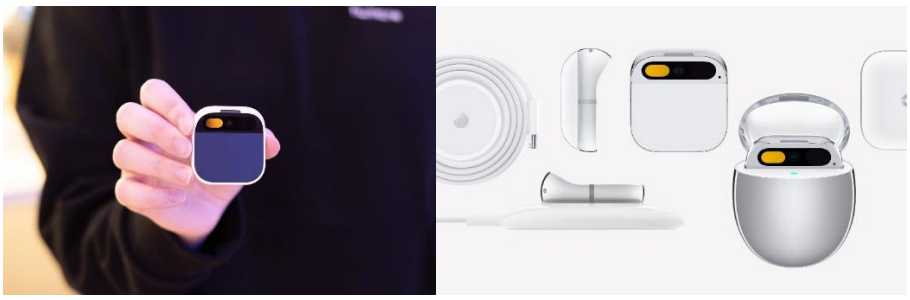
无处不在的模拟： Vision Pro 的出现将我们带入了一个数字模拟无处不在的世界。通过允许用户在这个虚拟环境中与熟悉的 iPhone 和 iPad 应用程序进行交互，Apple 将数字标准化为我们日常生活的一部分。这继续模糊了真实与模拟之间的界限，并加强了我们对数字地图和模型来导航世界的依赖。

音频协同：空间音频是 Apple 推出的一项流行功能，为 Vision Pro 的超真实体验带来了新的维度。通过提供根据用户的动作而变化的三维声场，空间音频加深了数字环境中的沉浸感，增强了模拟现实中的临场感。这项技术不仅增强了娱乐价值，还进一步侵蚀了真实与模拟之间的界限。这是 Vision Pro 的设计旨

在吸引多种感官，增强替代现实的幻觉并进一步挑战我们对现实世界构成的看法的另一个例子。

近期也有苹果内部宣布放弃研发 10 年以上的汽车计划，转向 AIGC 的消息。Apple Vision Pro 为娱乐、教育和工作提供了令人兴奋的新可能性。随着近期 AGI 和 AIGC 技术的发展，可以预见的是在 Apple Vision Pro 上会有让人惊喜的 AI Agent 的出现。

4、AI Pin



说起智能硬件，AI Pin 也是热度很高的产品。AI Pin 是一款以大语言模型为基底的 AI 装置，重量不到 100 克，没有屏幕，可直接别胸上，以投影显示结果，还可用语音与手势与它互动。许多人惊呼宛如科幻片里的个人助理，将颠覆 iPhone 时代。

Humane 的创办人查得利（Imran Chaudhri）与班吉欧茉（Bethany Bongiorno）皆是前 Apple 员工，试图将 Pin 与过往市场上的装置区分开来。

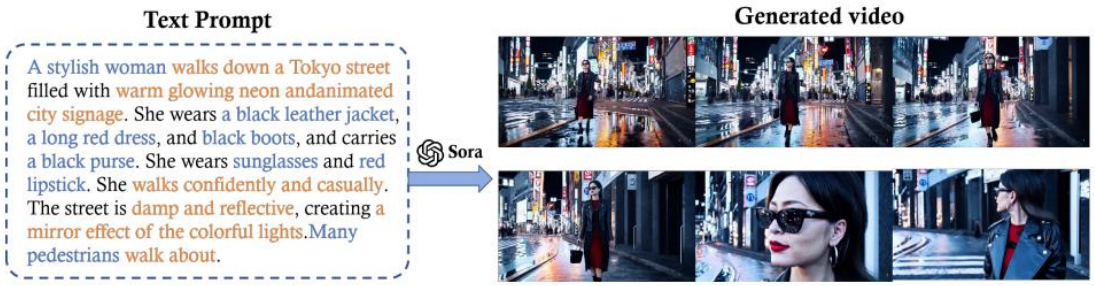
Pin 具备手机应该有的功能，上网、语音通话，同时也配备麦克风与相机，因此可以在动态中摄影，或者是边走边对话，但它强调比手机更好，不用屏幕、可以用投影呈现内容，人们则能够透过声音、手势、手部接触等方式与 Pin 互动。

对比手机内建智慧助理，与 Pin 的对话显得更为自然，不用特别唤醒（wake）装置，也不用非得以下命令的方式才能互动；对比其他企图替人类创造新体验的 VR/MR 头盔，Pin 的尺寸轻薄，可以配戴在衣服上，不会干扰工作与生活。

5、Sora

Sora 不是 AI Agent 的承载硬件，而更像一个“动画生成的 AI Agent”。Sora 是一种文本到视频生成 AI 模型，由 OpenAI 于 2024 年 2 月发布。该模型经过训练，可以根据文本指令生成现实或想象场景的视频，并显示出模拟物理世界的潜力。Sora 标志着 AI 视频生成技术的关键时刻。这不仅仅是制作视频；它是关于打造身临其境且无可否认的真实体验。该工具能够生成具有多个角色、动态运动和复杂背景的复杂场景，这不仅展示了对用户提示的深入理解，还展示了对现实世界物理的深入理解。

有评论指出：“在生成式人工智能的短暂历史中，所有令人震惊、令人不安的时刻中，很少有能比 OpenAI 上周推出的 Sora 更胜一筹。该公司的第一个根据书面提示制作视频的工具，它仍然是一个研究项目，仅供评估其潜在危险的团队和选定的创意专业人士使用。OpenAI 和其他人分享的剪辑都在不到一分钟的时间内，而且是无声的。但他们所证明的人工智能能够模拟相机拍摄的视频的能力表明，我们最终将面临成功伪造我们周围世界的镜头的冲击，OpenAI 的 Sora 很棒，很可怕，离杀死现实又近了一步”



有研究表明，Sora 是一个具有灵活采样维度的扩散转换器，如下图所示。它由三个部分组成：(1) 时空压缩器首先将原始视频映射到潜在空间。(2) ViT 然后处理 (3) 类似 CLIP 的调节机制接收 LLM 增强的用户指令和潜在的视觉提示，以指导扩散模型生成风格或主题视频。 经过许多去噪步骤后，获得生成视频的潜在表示，然后使用相应的解码器将其映射回像素空间。[04]

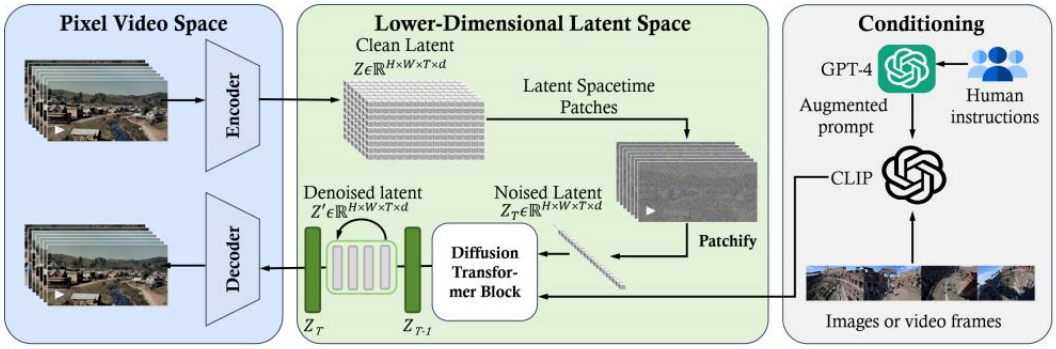


Figure ：**Reverse Engineering**: Overview of Sora framework

Sora 不仅仅是一个视频生成模型，它为能够理解和模拟现实世界的更先进模型奠定了基础。OpenAI 将 Sora 视为迈向通用人工智能 (AGI) 的垫脚石，未来人工智能系统将拥有广泛的、类似人类的理解和能力。通过掌握视频生成和解释的复杂性，Sora 为实现 AGI 的更广泛目标做出了贡献，标志着创造造福全人类的人工智能之旅的一个重要里程碑。

在 Sora 推出之后，国内不少团队也开始跟进并迅速实现突破。纵观国内玩家，已有超 15 家企业推出了视频生成工具，既包括百度、阿里、腾讯、字节等 6 家巨头，也包括爱诗科技、生数科技、智象未来等 9 家创企。

国内大厂发布的类Sora产品/模型（截至3月1日）

序号	公司	产品/模型	发布/更新时间	能否试用	文生视频	图生视频
1	字节跳动	CapCut AI Video	2024年2月	✓*	✓	
		Boximator	2024年2月			✓
		MagicVideo-V2	2024年1月		✓	
		PixelDance	2023年11月			✓
2	腾讯	DynamiCrafter	2024年2月	✓		✓
		VideoCrafter2	2024年1月	✓	✓	✓
		AnimateZero	2023年12月			✓
3	阿里巴巴	ModelScope T2V	2023年12月	✓	✓	
		I2VGen-XL	2023年12月	✓		✓
4	百度	UniVG	2024年1月		✓	✓
5	万兴科技	天幕多媒体大模型	2024年1月		✓	
6	美图	MiracleVision 奇想智能大模型	2023年12月		✓	✓

注：仅统计视频生成类产品/模型，如有遗漏欢迎补充

* 试用入口暂时关闭

▲国内大厂发布的Sora类产品/模型（智东西统计制表，数据截至2月27日）

国内创业公司发布的类Sora产品/模型（截至3月1日）

序号	公司	产品/模型	发布时间	能否试用	文生视频	图生视频
1	Morph Studio	Morph Studio	2023年6月	✓	✓	✓
2	HiDream.ai (智象未来)	Pixeling	2023年8月	✓	✓	✓
3	爱诗科技	PixVerse	2023年11月	✓	✓	✓
4	MewXAI	艺映AI	2023年11月	✓	✓	✓
5	NeverEnds	NeverEnds	2023年12月	✓	✓	✓
6	右脑科技	Vega AI	2023年12月	✓	✓	✓
7	李白AI实验室	神采 PromeAI	2023年12月	✓		✓
8	Möbius	Möbius	2024年1月	✓	✓	
9	生数科技	PixWeaver	2024年1月	✓*	✓	

注：仅统计视频生成类产品/模型，如有遗漏欢迎补充

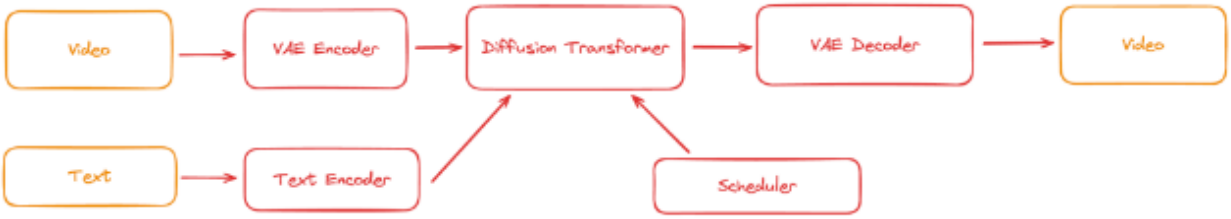
* 试用入口暂时关闭

▲国内创业公司发布的Sora类产品/模型（智东西统计制表，数据截至3月1日）

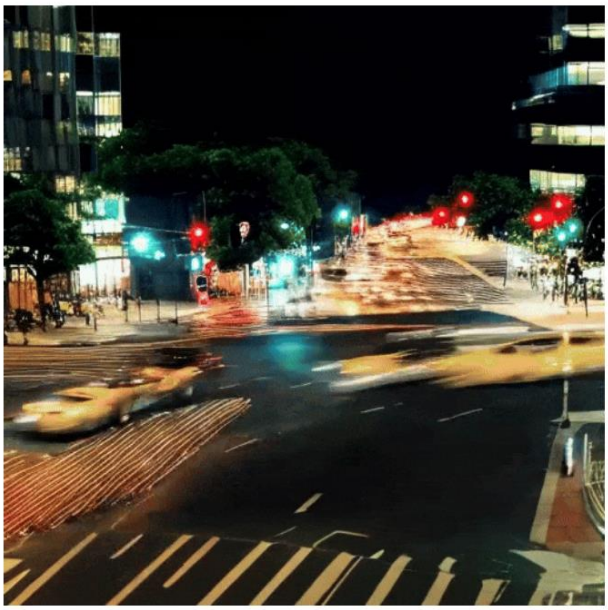
根据智东西的测试报告[05]，字节短暂上线的 CapCut AI Video 功能最佳，尤其胜在运动平滑度和成像质量上。Morph Studio、NeverEnds 在创企中领跑，且稳定性较高，在体验过程中没有出现大翻车的情况。VideoCrafter2、Pixeling 生成质量不够稳定，出现了几次大翻车的情况；PixVerse、Vega AI 也出现了不同程度的翻车。 对于未来中国的 LLM 及“中国 Sora”，我们仍可以抱有一个乐观的态度去关注它们后续的发展。

在 Open-Sora 路径上，国内的潞晨科技做出了让人印象深刻的产品。潞晨科技曾经获得过创新工场和真格基金的千万种子轮融资，又获数亿元 A 轮融资。核心成员来自美国加州伯克利、斯坦福、清北、新加坡国立、南洋理工等世界一流高校，在国际顶级学术刊物或会议共发表论文近百篇，曾创造 ImageNet、BERT、AlphaFold、ViT 训练速度的世界纪录，在谷歌、微软、英伟达、IBM、英特尔等头部科技公司拥有丰富的任职经验，团队在高性能计算，人工智能，分布式系统方面已有十余年的技术积累。

潞晨科技自研的 Colossal-AI，又第一时间快速开源完整的 Sora 复现架构方案 Open-Sora，还可降低 46%复现成本，并将模型训练输入序列长度扩充至 819K patches，并将 Sora 可能使用的训练 pipeline 归纳为下图。



继推出成本直降 46%的 Sora 训练推理复现流程后，Colossal-AI 团队全面开源全球首个类 Sora 架构视频生成模型「Open-Sora 1.0」——涵盖了整个训练流程，包括数据处理、所有训练细节和模型权重。



Open-Sora 1.0生成的都市繁华掠影

关于以上文生视频的模型架构、训练好的模型权重、复现的所有训练细节、数据预处理过程、demo 展示和详细的上手教程，Colossal-AI 团队已经全面免费开源在 GitHub。团队在采访中也提及他们将会继续维护和优化 Open-Sora 项目，预计将使用更多的视频训练数据，以生成更高质量、更长时长的视频内容，并支持多分辨率特性，切实推进 AI 技术在电影、游戏、广告等领域的落地。

三、AI Agent 技术前沿

目前有代表性的 AI Agent 都基于 LLM（通用大模型），除了底层 LLM 技术外，为了能够更好的实现帮助使用者解决问题和完成任务，有两个核心的问题是 AI Agent 公司要解决的。一是设计良好的算法架构，充分使用 LLM 的能力；二是给出设计的架构，如何使 Agent 获得完成特定任务的能力。这涉及到两个核心：拆分问题以及 prompt engineering [06]。

拆分问题：需要构思好怎么将当前的任务进行更好的拆分成一个个子任务，以确保这些子任务足够简单、完成的准确率足够高；当这些子任务都确保能被很好的完成，那么最终的任务就能很好的完成了。

所以拆分问题很关键，拆分的粒度如果太粗，子任务难度系数就高，就很容易失败；拆分的粒度如果太西，调用逻辑就很繁琐，整个链路就会很臃肿，所以对需求和业务的理解越深刻，拆解才会越相对合理。同时拆分的是否合理还会影响另外一个关键问题：后期自研模型的开发。当我们想利用大模型开发 Agent 的时候，一般来说会先去使用 GPT4 去试一试（毕竟其目前是大模型的天花板），如果它都完成的不好，那要么自己当前 Agent 的设计框架需要进一步完善，要么就是这个事目前大模型还真的是很难完成。当 GPT4 完成的还不错后，出于安全和成本等考虑我们必定是想走自研模型这条路的，做到自主可控，那么我们就可以前期使用 GPT4 去积累数据，然后用这部分数据去蒸馏训练出自己的大模型。那么如果你的任务拆分的粒度太细，假设有 100 个子任务（这里指要最终调用大模型能力），那么如何将这 100 个任务同时进行大模型训练，平衡住所有任务的能力，这是很难的（每个子任务训练一个大模型也太不现实了），当然如果拆分的粒度太粗，单个子任务本身就很难，那单训练好这个子任务可能都是问题，就更别提要融合所有子任务了；

prompt engineering：当我们把当前这个 Agent 需要完成的任务多步拆解后以及理顺子任务之间的联动调用链路后，那么完成这些子任务就需要调用大模型了（当然有时候是调用一些其他插件比如计算器、搜索引擎等等；即使是调用插件其实上一步也一般是需要调用大模型来分析出要调用哪个插件以及插件需要的参数）。

既然是调用大模型，那么如何写好 prompt 让大模型完全 get 到你的需求，这是非常关键的，如果没写好 prompt，那么子任务就失败了，整个链路就 run 不起来。这块工作也是最繁琐和最需要经验的，要不断的去试进而润色出一个很棒的 prompt。甚至笔者觉得在某些场景下，当你要做一个 agent 项目时，prompt engineering 是你第一步要去做的事情，先去试着写几个 prompt 看看大模型能完成的怎么样？自己感受感受摸个底，这样也才可以有更多灵感看看怎么将任务进行更好的拆解，通过多步调用大模型（也即上面说的第一个核心点）来合作完成。

1、底层 LLM 发展前沿

目前，LLM 领域领先的，主要是 OpenAI、Google、亚马逊及 Facebook，其中，OpenAI 的 ChatGPT 以及 Google 的 Gemini 及 Gemma 一时瑜亮。

GPT-4 [07]，这是 OpenAI 扩大深度学习努力的最新里程碑。GPT-4 是一个大型多模态模型（接受图像和文本输入，发出文本输出），虽然在许多现实场景中能力不如人类，但在各种专业和学术基准上表现出人类水平的表现。例如，它通过了模拟律师考试，分数在考生中排名前 10%；相比之下，GPT-3.5 的得分在底部 10% 左右。OpenAI 花了 6 个月的时间，利用对抗性测试项目和 ChatGPT 的经验教训，迭代 调整 GPT-4，在真实性、可操纵性和拒绝超出护栏方面取得了有史以来最好的结果（尽管远非完美）。

Simulated exams	GPT-4 estimated percentile	GPT-4 (no vision) estimated percentile	GPT-3.5 estimated percentile
Uniform Bar Exam (MBE+MEE+MPT)[1]	298/400 ~90th	298/400 ~90th	213/400 ~10th
LSAT	163 ~88th	161 ~83rd	149 ~40th
SAT Evidence-Based Reading & Writing	710/800 ~93rd	710/800 ~93rd	670/800 ~87th
SAT Math	700/800 ~89th	690/800 ~89th	590/800 ~70th
Graduate Record Examination (GRE) Quantitative	163/170 ~80th	157/170 ~62nd	147/170 ~25th
Graduate Record Examination (GRE) Verbal	169/170 ~99th	165/170 ~96th	154/170 ~63rd
Graduate Record Examination (GRE) Writing	4/6 ~54th	4/6 ~54th	4/6 ~54th
USABO Semifinal Exam 2020	87/150 99th–100th	87/150 99th–100th	43/150 31st–33rd
USNCO Local Section Exam 2022	36/60	38/60	24/60
Medical Knowledge Self-Assessment Program	75%	75%	53%
Codeforces Rating	392 below 5th	392 below 5th	260 below 5th
AP Art History	5 86th–100th	5 86th–100th	5 86th–100th
AP Biology	5 85th–100th	5 85th–100th	4 62nd–85th
AP Calculus BC	4 43rd–59th	4 43rd–59th	1 0th–7th

Gemini【08]，Gemini 是整个 Google 团队大规模协作努力的成果。它是从头开始构建的多模式，这意味着它可以概括和无缝地理解、操作和组合不同类型的信息，包括文本、代码、音频、图像和视频。从自然图像、音频和视频理解到数学推理，Gemini Ultra 的性能在大型语言模型 (LLM) 研发中使用的 32 个广泛使用的学术基准中的 30 个上超过了当前最先进的结果。Gemini Ultra 的得分高达 90.0%，是第一个在 MMLU（大规模多任务语言理解）上超越人类专家的模型，该模型结合了数学、物理、历史、法律、医学和伦理学等 57 个科目来测试知识和解决问题的能力。

Gemma【09] 是一系列轻量级、最先进的开放式模型，采用与创建 Gemini 模型相同的研究和技术而构建。Gemma 由 Google DeepMind 和 Google 的其他团队开发，其灵感来自 Gemini，其名称反映了拉丁语 gemma，意思是“宝石”。

有消息称 Openai 即将推出 GPT-5，GPT-5 预计将具有更好的自然语言理解（NLU）能力，这意味着它将能够掌握复杂句子的含义并提供更准确的响应。这将消除重新表述查询的需要，并使 ChatGPT 5 在提供相关信息方面更加用户友好和高效。GPT-5 相对于 GPT-4 进步的另一个关键方面是其扩展的信息处理能力。这意味着 GPT-5 将能够处理和生成更大的文本块，从而获得更好的准确性和更连贯的输出。这一改进至关重要，因为它将允许 ChatGPT 5 无缝处理更长的对话和更复杂的任务。以前的 GPT 模型（包括 GPT-4）面临的挑战之一是在其输出中产生幻觉或虚假信息。有了 GPT-5，预计这些问题将显着减少。这一改进对于确保 ChatGPT 5 提供可靠和准确的信息、提高其对用户的价值并降低传播错误信息的风险至关重要。

总而言之，虽然 ChatGPT 5 的具体发布日期仍然未知，但很明显，语言理解方面的进步、令牌容量的扩展以及输出中幻觉或虚假信息的减少将使其成为强大的变革性人工智能工具。这些改进将塑造对话式人工智能的未来，并有可能让我们更接近实现通用人工智能 (AGI)。

TEXT

Capability	Benchmark Higher is better	Description	Gemini Ultra	GPT-4 API numbers calculated where reported numbers were missing
General	MMLU	Representation of questions in 57 subjects (incl. STEM, humanities, and others)	**90.0%** CoT@32*	86.4% 5-shot** (reported)
Reasoning	Big-Bench Hard	Diverse set of challenging tasks requiring multi-step reasoning	**83.6%** 3-shot	83.1% 3-shot (API)
	DROP	Reading comprehension (F1 Score)	**82.4** Variable shots	80.9 3-shot (reported)
	HellaSwag	Commonsense reasoning for everyday tasks	87.8% 10-shot*	**95.3%** 10-shot* (reported)
Math	GSM8K	Basic arithmetic manipulations (incl. Grade School math problems)	**94.4%** maj1@32	92.0% 5-shot CoT (reported)
	MATH	Challenging math problems (incl. algebra, geometry, pre-calculus, and others)	**53.2%** 4-shot	52.9% 4-shot (API)
Code	HumanEval	Python code generation	**74.4%** 0-shot (IT)*	67.0% 0-shot* (reported)
	Natural2Code	Python code generation. New held out dataset HumanEval-like, not leaked on the web	**74.9%** 0-shot	73.9% 0-shot (API)

* See the technical report for details on performance with other methodologies

** GPT-4 scores 87.29% with CoT@32 - see the technical report for full comparison

2、国内 AI 大模型发展近况

国内已有发布 200 个左右的大模型，也有许多有前途的 AI 大模型。其中，字节跳动和百度被认为是最有前途的。他们拥有高质量的中文私有数据、专业的团队、丰富的 GPU 资源，以及强大的技术实力。此外，智谱和 Moonshot 也是国内大模型创业公司的代表，他们拥有独特的商业模式和融资优势。其他公司也各有特点，但目前尚未公开表示具体的技术进展和优势。同时，也需要注意一些创业公司可能存在的问题，如创始人争议、技术不够成熟、缺少技术护城河等。

序号	公司	大模型	省市	类别	官网	说明
1	百度	文心一言	北京	通用	✔	有APP，衍生灵医Bot
2	智谱华章	清言	北京	通用	✔	有APP，开源小模型ChatGLM-6B和ChatGLM2-6B
3	百川智能	百川	北京	通用	✔	开源小模型baichuan-7B和Baichuan-13B，baichuan-2
4	达观数据	曹植	上海	工业	✔	试用需账号
5	上海人工智能实验室	书生	上海	通用	✔	开源小模型书生·浦语，OpenMEDLab浦医
6	科大讯飞	星火	安徽合肥	通用	✔	试用需账号,有APP
7	深度求索	Deepseek Coder	浙江杭州	代码	✔	
8	商汤科技	日日新	上海	通用	✔	
9	春田知韵（抖音）	豆包	北京	通用	✔	开源多模态7B小模型BuboGPT，豆包是云雀的聊天机器人
10	中国科学院自动化研究所	紫东·太初	北京	通用	✔	紫东太初2.0号称100B参数，全模态
11	阿里云	通义千问	浙江杭州	通用	✔	试用需账号,开源小模型Qwen-7B和Qwen-7B-Chat
12	华为	盘古,盘古气象,盘古-Σ	广东深圳	工业	✔	华为+鹏城,华为云盘古
13	复旦大学	MOSS	上海	科研	✔	试用需账号
14	智源人工智能研究院	悟道·天鹰,悟道·EMU	北京	通用	✔	悟道3.0,视界视觉，AQUILA天鹰座，Aquila-7B,AquilaChat-7B,AquilaCode-7B-NV,AquilaCode-7B-TS,HuggingFace,EMU基于LLaMA
15	浙江大学	启真,TableGPT,智海-录问,智海-三乐,PromptProtein	浙江杭州	垂直	✔	医学大模型提供基于LLaMA-7B、CaMA-13B和ChatGLM-6B 三个版本,用于PromptProtein的模型，法律大模型智海-录问基于Baichuan-7B，智海-三乐基于Qwen-7B
16	面壁智能	CPM,CPM-Bee	北京	通用	✔	面壁智能,CPM-Bee-10B,OpenBMB
17	元象科技	XVERSE-13B	广东深圳	通用	✔	模型下载
18	腾讯	混元	广东深圳	通用	✔	
19	云知声	山海	北京	医学	✔	
20	东北大学	TechGPT,PICA	辽宁沈阳	科研	✔	TechGPT->BELLE->LLaMA，图谱构建和阅读理解问答;PICA->ChatGLM2-6B情感大模型

国内主要大模型厂商

总之，人工智能大模型是未来的发展趋势，国内有许多有前途的公司和团队，但需要不断提高技术实力和创新能力才能在这个领域中保持领先地位。

四、展望

目前，AI 底层技术的迭代快的惊人。承载 AI 应用技术的硬件、软件平台及应用软件也处于快速生长期，不断有耳目一新的产品推出市场。

关于共识的部分是，AI 将重塑人机交互方式，带来新的一波硬件升级的机会。同时，我们所熟悉的软件平台、应用软件都值得被 AI 重新做一遍。产业界要抓住这一波机会，用 AI 更新自己的硬件产品，提升自身的产品竞争力，努力跟上这一波时代机会。

对于新的竞争者，这也许是一次“鲤鱼跃龙门”的机会。而对于软件平台、应用软件而言，这是一个重塑产品形态、平台规则的机会。底层技术的突破带来重新洗牌的机会，对于电商、游戏、品牌宣传等领域，AI 技术的应用更快速，抢先一步，也许就是制胜之机。

作为 AI 市场的参与者，我们也将持续关注 AI 底层技术的迭代、新硬件产品及其上下游投资机会，时刻关注新的应用形态及产品，努力在 AI 时代，在我们的投资“甜区”，寻找优质的投资机会。

参考材料：

【01】AI is about to completely change how you use computers, By Bill Gates

【02】A Survey on Large Language Model based Autonomous Agents, Lei Wang, Chen Ma* , Xueyang Feng* , Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, Ji-Rong Wen Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China.

【03】Apple's Vision Pro: The Precarious New Age of Hyperreality, By John Nosta;

【04】Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models, by Yixin Liu1* Kai Zhang1* Yuan Li1* Zhiling Yan1* Chujie Gao1*Ruoxi Chen1* Zhengqing Yuan1* Yue Huang1* Hanchi Sun1* Jianfeng Gao2 Lifang He1 Lichao Sun1, 1Lehigh University 2Microsoft Research

【05】全网首个“中国版 Sora”横评！15 家企业对决，字节领跑，智东西;

【06】大模型 AI Agent 前沿调研，小小梦想，https://zhuanlan.zhihu.com/p/668363633;

【07】https://openai.com/research/gpt-4

【08】https://blog.google/technology/ai/google-gemini-ai/#performance

【09】https://blog.google/technology/developers/gemma-open-models/