

AI时代新起点，寻新投资方向（三）

AI Agent，大模型时代重要落地方向

证券分析师：张良卫
执业证书编号：S0600516070001
联系邮箱：zhanglw@dwzq.com.cn

研究助理：郭若娜
执业证书：S0600122080017
guorn@dwzq.com.cn

2023年10月12日

并购家 免费行业报告网

IPOIPO.CN 每日更新 不用注册会员 无广告 永久免费

- **大模型时代的AI AGENT = LLM + 规划能力 + 记忆 + 工具。**AI领域AGENT概念由来已久，这一轮LLM给AI AGENT提供了突破性技术方案，对AI AGENT性能表现至关重要；同时需借助外部工具在实际应用中保持长期一致性和准确性。目前AI AGENT的探索大致分为自主智能体和智能体模拟。
- **方向一：自主智能体，力图实现复杂流程自动化。**自主智能体有望带来软件行业交互方式和商业模式变革。基座大模型能力解决下限问题，在实际企业应用场景中自主智能体的架构设计、工程能力、垂类数据质量等也至关重要，垂类/中间件玩家亦有机会。其中，单智能体相对更适用于较简单的任务，在C端应用上有一定潜力，代表性玩家包括中心化应用的ChatGPT、去中心化应用的adept AI、可定制和平台化的Cortex、MindOS等；但其在B端场景上略显乏力，基本无法完成较为复杂的工作，多智能体优势相对更加突出。代表性玩家包括MetaGPT、ChatDev、Showrunner。
- **方向二：智能体模拟，力图更加拟人可信。**1) **陪伴类智能体**强调情感情商等人类特征，具有“人格”，且能够记住与用户的历史交流，代表应用如PI、Character ai、replica、glow等。我们认为国内情绪消费市场仍有较大想象空间，陪伴类智能体或受益于情绪消费趋势红利，成为LLM时代重要的AI原生应用。我们预计陪伴类智能体大部分商业价值集中在供给方而非平台，我们更加看好具备丰富IP储备或者能让用户定制智能体的玩家。2) **交互智能体**：强调与环境交互的能力，智能体之间/与虚拟世界内事物之间可互动，可能涌现出超越设计者规划的场景和能力，大模型不确定性反而成为优势，有望成为AIGC重要部分。特别是对开放世界游戏等行业，可增强玩家沉浸感，解决开放世界内容消耗快的问题；多可信agent技术成熟后可能会孵化出新的游戏品类。代表项目如斯坦福大学开源的Smallville 小镇，应用如网易《逆水寒》、昆仑万维《Club Koala》。
- **投资建议：**1) 推荐在AI AGENT方向有直接布局的昆仑万维、中文在线、盛天网络，建议关注天地在线等。2) 游戏板块推荐研运能力出色且产品储备充沛的厂商：恺英网络、吉比特、三七互娱、巨人网络、宝通科技等，港股的创梦天地等，建议关注神州泰岳、世纪华通等。3) IP资源建议关注艺人IP（华策影视等）、影视IP（光线传媒、奥飞娱乐、博纳影业、上海电影、中国电影等）。4) 教育推荐南方传媒、凤凰传媒、皖新传媒、新东方，建议关注世纪天鸿、好未来、高途集团、科大讯飞、佳发教育、盛通股份、传智教育、鸿合科技等；电商及企业服务推荐焦点科技、华凯易佰、吉宏股份；创作类自主智能体发展有助于提高内容行业生产效率，推荐动画电影龙头光线传媒。
- **风险提示：**产业进展不及预期，监管风险，市场竞争加剧风险



- 1. AI AGENT: LLM提供新基座，自动化/拟人化是两大方向

 - 1.1 AI AGENT: 感知并反应，具备自主性/可迭代/规划性
 - 1.2 新时代AI AGENT: LLM+规划+记忆+工具
 - 1.3 两大方向：自动化（自主智能体）、拟人化（智能体模拟）
- 2. 自主智能体：自动化，新一轮生产力革命

 - 2.1 自主智能体：软件新范式，非大模型玩家亦有机会
 - 2.2 自主智能体：实验性VS实操性，单智能体VS多智能体
- 3. 智能体模拟：拟人化，新的精神消费品

 - 3.1 陪伴类，提供情绪价值
 - 3.2 重交互，提高用户体验
- 4. 投资建议

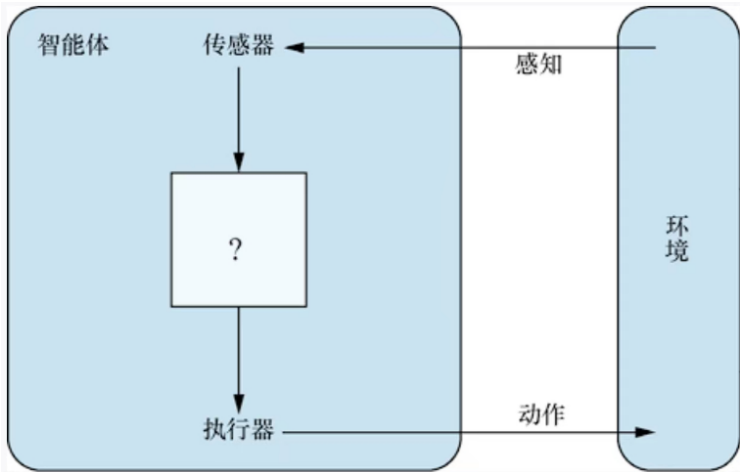
- 5. 风险提示

1、AI AGENT: LLM提供新基座，自动化/拟人化是两大方向

1.1 AI AGENT：感知并反应，具备自主性/可迭代/规划性

- **智能体agent**：通过传感器感知环境（收集信息）并通过执行器作用于该环境（采取行动）的事物。 Agent概念最早起源于M. Minsky (AI之父之一) 于1986年出版的《Society of Mind》（也是神经网络热潮开始的年份），1994年AI学术界内便已开始较多对agent的讨论，本篇报告我们采用22年出版的《人工智能：现代方法》（第四版）中的定义。
- **理性智能体概念**是研究人工智能方法的核心。区别于普通计算机程序，理性智能体具备自主性、可学习迭代、可制定并实现目标等特点。人工智能专注于研究和构建做正确的事情的智能体——理性智能体（Rational Agent/Intelligent Agent，也是本报告“AI AGENT”定义），即设计理性智能体程序实现智能体函数，完成从感知到动作的映射。对于每个可能的感知序列，给定感知序列提供的证据和智能体所拥有的任何先验知识，理性智能体应该选择一个期望最大化其性能度量的动作。

图：智能体通过传感器和执行器与环境交互



图：智能体结构拆解

智能体 = 架构 + 程序

具备物理传感器和执行器的**计算装置**，使程序可以使用来自传感器的感知，运行程序，并将程序生成的动作选择反馈给执行器

将当前感知作为传感器输入，将动作返回给执行器，是**智能体函数**的具体实现。
• 智能体函数：抽象的数学描述，描述智能体行为，将任意给定的感知序列映射到一个动作；可能依赖整个感知历史

1.2 新时代AI AGENT：LLM+规划+记忆+工具，大模型重要落地方向

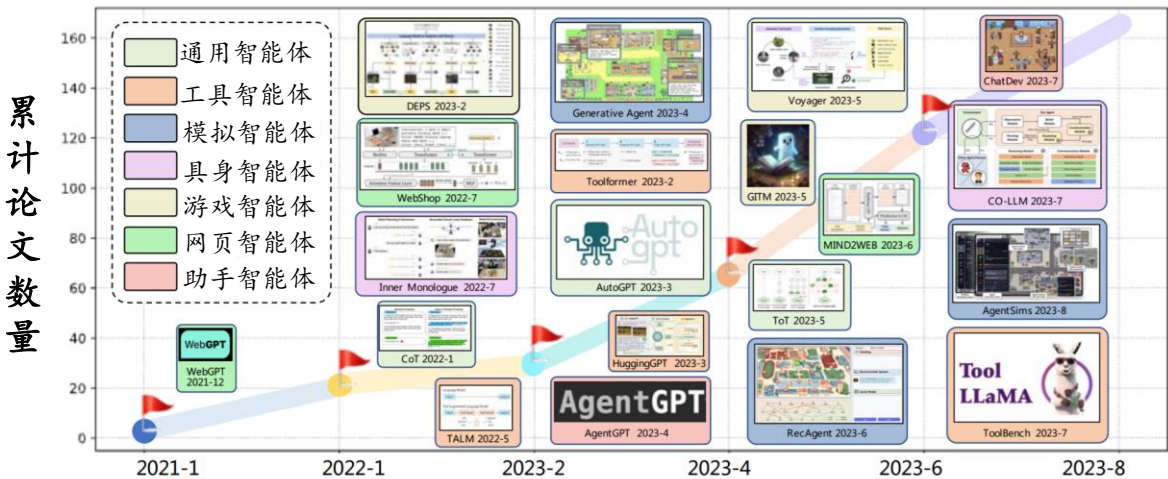
- 大模型时代的AI AGENT = LLM（核心控制器，构建核心能力）+ 规划能力 + 记忆 + 工具。其中基座模型能力至关重要。
- 我们认为LLM给AI AGENT底层提供了一个突破性技术方案：过去强化学习基于深度学习框架可让agent学到技能，但agent本身并没有真正理解问题和技能，泛化性也较差，只能用于特定领域，主要用在游戏和用来制作低维控制/计划，代表性应用是围棋领域的AlphaGo；LLM带来了深度学习新范式，思维链和强大的自然语言理解能力有望让agent具备强大的学习能力和迁移能力，从而让创建广泛应用且实用的agent成为可能。
- 由于生成式LLM存在幻觉问题，记忆力短，在实际应用中难以保持长期一致性和准确性，且agent间合作也是重要趋势，除了等待基座模型自身迭代之外，借助外部力量（向量存储、检索、代码等）是重要方法，完整的AGENT框架应该具备这些能力。我们认为补齐了大模型短板的AI AGENT更具备实用性，将是**大模型重要落地方向**。前特斯拉总监、OpenAI科学家Karpathy公开表示“如今AI智能体才是未来最前沿的方向”“相比大模型训练，OpenAI内部目前更关注Agent领域”。



1.3 两大方向：自动化（自主智能体）、拟人化（智能体模拟）

- 结合目前学术界和产业界基于LLM开发的AI AGENT应用情况，我们将目前AI AGENT划分为两大类：
- 自主智能体，力图实现复杂流程自动化。当给定自主智能体一个目标时，它们能自行创建任务、完成任务、创建新任务、重新确定任务列表的优先级、完成新的首要任务，并不断重复这个过程，直到完成目标。准确度要求高，因而更需要外部工具辅助减少大模型不确定性的负面影响。
- 智能体模拟，力图更加拟人可信。分为强调情感情商智能体以及强调交互的智能体，后者往往是在多智能体环境中，可能涌现出超越设计者规划的场景和能力，大模型生成的不确定性反而成为优势，多样性使其有望成为AIGC重要组成部分。
- 我们认为两大方向并不是完全割裂的，相反，自动化与拟人化将作为ai agent两大核心能力并行发展，随着底层模型成熟以及行业探索更加深入，有望进一步扩大ai agent适用范围，提升其实用性。

图：基于LLM的AGENT领域产品增长情况



图：当前AI AGENT两大方向——自主智能体和智能体模拟对比

		准确度要求	记忆	外部工具	当前应用	
初始动力	用户给定	自主智能体	高	短期记忆为主	较多	AutoGPT、chatgpt+插件、adept、metagpt等
	开发者设定的内部目标	智能体模拟	一般，更多要求拟人程度	短期记忆+长期记忆	较少	Pi、Smallville小镇、Voyager、GITM等

2、自主智能体：自动化，新一轮生产力革命

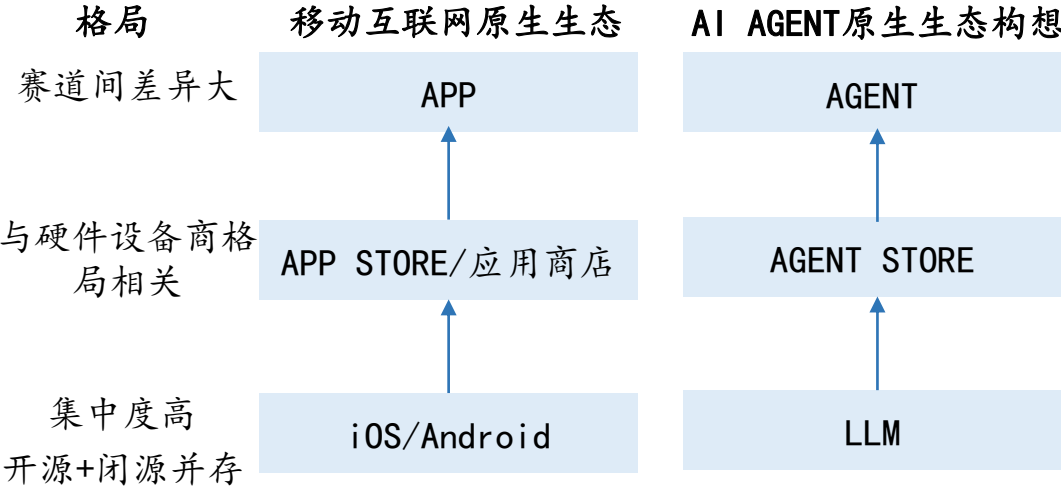
2.1 自主智能体：软件新范式，非大模型玩家亦有机会

- 自主智能体，力图实现复杂流程自动化。真格基金管理合伙人戴雨森将AI和人类协作的程度类比为自动驾驶的不同阶段，AI Agent约为自动驾驶的L4阶段，Agent完成任务，人进行外部辅助和监督。
- 自主智能体有望带来软件行业交互方式和商业模式变革：
 - 交互方式变革：相比过去的APP/软件，从人适应应用变成应用适应人，Agent的决策/规划/执行等环节需要更深的用户需求理解以及更强的工程细节打磨。如目前Agent运行中常常遇见无休止的扩展、误解输出格式等问题，这类问题不单单依靠大模型能力提升，对Agent架构的设计和垂类数据的学习也有要求。
 - 商业模式变革：按服务内容收费转换成按token收费，对Agent功能实用性要求更高。
- 基座大模型能力固然重要，但其只能解决下限问题，在实际企业应用场景中自主智能体的架构设计、工程能力、垂类数据质量等也至关重要，垂类/中间件玩家亦有机会。准确度和效率是自主智能体重要指标（决策式AI更擅长做的事，也意味着更低的容错度），企业对于低门槛定制AGENT也存在需求，专注具体领域/提供AGENT框架的玩家仍有可为。

表：AI发展阶段对比自动驾驶不同阶段

等级	名称	自动化程度	含义	示例
L1	Tool	无	人类完成所有工作，没有任何显性的AI辅助	目前绝大多数软件产品
L2	Chatbot	助理	人类完成绝大部分工作，类似向AI询问意见，了解信息，AI提供信息和建议但不直接处理工作	初代ChatGPT和Chatbot
L3	Copilot	部分自动化	人类和AI进行协作，工作量相当。AI根据人类prompt完成工作初稿，人类进行目标设定，修改调整，最后确认	Copilot、Jasper
L4	Agent	条件自动化	AI完成绝大部分工作，人类负责设定目标、提供资源和监督结果，AI完成任务拆分，工具选择，进度控制，现目标后自主结束工作	AutoGPT
L5	species	完全自动化	完全无需人类监督，AI自主拆解目标，寻找资源，选择并使用工具，完成全部工作，人类只需给出目标	类似冯·诺什曼机器人

图：AI AGENT原生生态构想



2.2 自主智能体：实验性VS实操性，单智能体VS多智能体

- 目前行业内对自主智能体的探索主要分为四大类：
- 实验性项目VS实操性应用：前者虽然实际运行中出错概率高，但其创意、思路和开发经验对开发者仍有不少启发和贡献，如掀起这一波AGENT开发热潮的AutoGPT。实操性应用更加强调与实际场景的适配。
- 单智能体VS多智能体：单智能体相对更适用于较简单的任务（如比价等），在C端应用上有一定潜力（如chatgpt+插件、adept ai等）；但其在B端场景上略显乏力，由于缺乏完整性评估、任务队列过长、大模型幻觉等原因，基本无法完成较为复杂的工作（如软件开发等），多智能体优势相对更加突出（如metagpt、chatdev等）。

表：主要自主智能体一览

项目	实用性	单智能体/多智能体	配置模块	记忆		规划	动作	能力获取	发布时间
				操作	结构				
WebGPT	实验性	单	-	-	-	-	使用工具	有微调	2021/12
ReAct	实验性	单	-	-	-	有反馈规划	使用工具	有微调	2022/10
HuggingGPT	实验性	单	-	-	统一存储器 (没有短期和长期记忆的区别)	无反馈规划	使用工具	-	2023/03
AutoGPT	实验性	单	-	只有读/写操作	混合存储器	有反馈规划	使用工具	无微调	2023/04
ChatDev	实操性	多	手工输入	-	混合存储器	有反馈规划	不使用工具	无微调	2023/07
MetaGPT	实操性	多	手工输入	有读/写/反射操作	混合存储器	有反馈规划	使用工具	-	2023/08

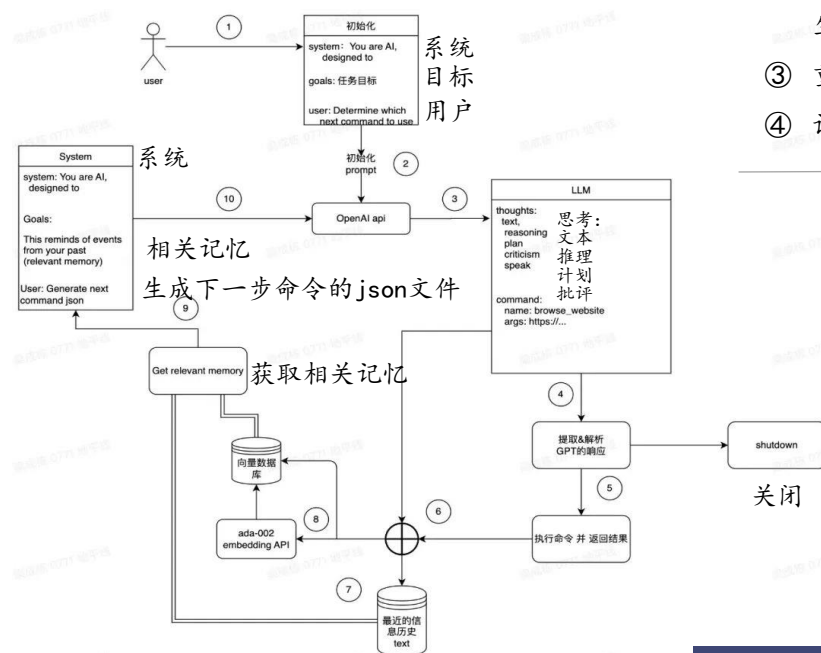
注：“-”表示相关应用论文中没有相关表述

资料来源：A Survey on Large Language Mod

2.2.1 单智能体（1）：实验性项目，如AutoGPT

- **AutoGPT**：2023年3月开源的第一款AI AGENT实验性项目，由游戏开发者Toran Bruce Richards开发。AutoGPT可根据用户给定的目标，自动生成所需提示，并利用GPT4和各种工具API执行多步骤的项目，无需人类的干预和指导。上线5个月在github上的STAR数量已超过149K，具有一定代表性，对后来各种AGENT发展具有启发意义。
- **使用多个外部工具**：1) 能克隆 github 仓库、启动其他agent、发言、发送推文和生成图片，支持所有矢量数据库和 LLM 提供商/文本到图片模型/浏览器。2) 通过集成 Pinecone数据库，它可进行长期内存存储，从而保存上下文并基于此进行决策改进。3) 通过python内部List结构，保存text文本，每次会获取最近一次的历史信息。
- **应用场景偏向于办公/开发类**（自动化流程工作流、进行深入市场研究、写代码/程序、开发网站或APP），但实操效果一般，容易陷入死循环。
- **其他相似项目**：BabyAGI（4月开源的实验性项目，由小型风投Untapped Capital合伙人Yohei Nakajima开发，类似简单版本的AUTOGPT）、AGENTGPT等

图：AUTOGPT运行流程图



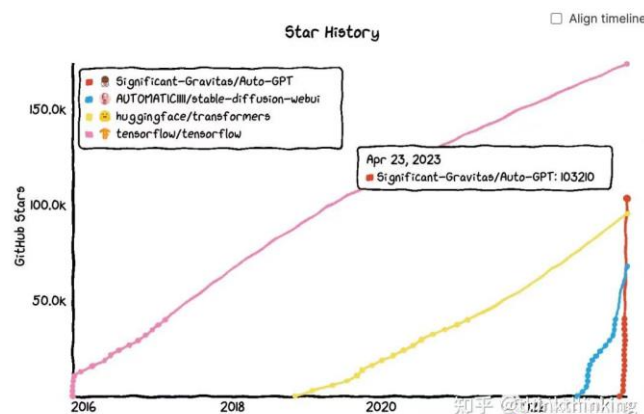
案例1：借助GPT4+PINECONE，“写一篇关于太空的文章”

- ① PINECONE里查找关于太空的文章，有则直接展示文章，无则下一步
- ② 用GPT-4生成提示，如“太空是什么”，用GPT4回答这个提示，并把生成的提示和回答存储在PINECONE里，并用它们作为文章第一段。
- ③ 重复直到文章足够长/足够完整/达到一定字数限制/时间限制。
- ④ 询问用户是否满意，是则结束运行，否会根据用户反馈修改补充。

案例2：借助多个外部工具，“实现多语言的自动翻译功能”

- ① 在聊天对话中，用户输入一段文本。
- ② AutoGPT可以与外部的机器翻译工具（如Google Translate）结合，利用外部翻译工具将其翻译为另一种语言。
- ③ 输出翻译结果。

图：AUTOGPT是github历史上星数增长最快的开源项目



图：AUTOGPT在github上的问答区热门问题多为实操效果不佳

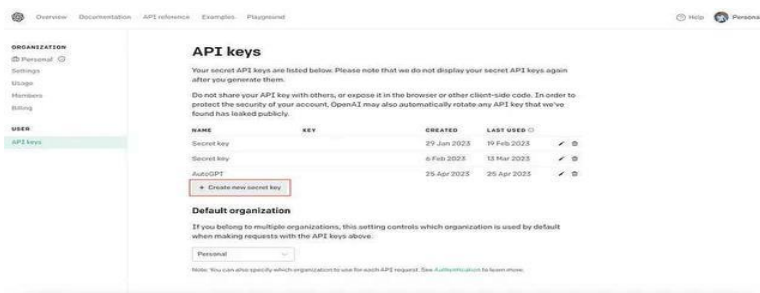


2.2.1 实验性项目，如代码开发类GPT Engineer

- **GPT Engineer**：主要作者Anton Osika在6月11日推出的开源代码生成工具，基于GPT模型，能根据用户的指示和需求生成高质量的代码，功能包括创建新函数、修复现有代码错误等等，支持多种编程语言。截至2023年9月，github星星数量接近44k。
- **亮点一：可定制性**。用户可根据自己的编码风格、项目需求和编程习惯进行设置，让GPT Engineer生成符合自身要求的代码。
- **亮点二：上下文感知**。可理解代码上下文，并生成与之相适应的代码片段，用户无需为适应其生成方式而做出额外调整，从而提高工作效率。
- **不足**：1) 无法进行迭代，在创建完应用程序后不能要求它更改任何内容。2) 构建复杂的代码时仍需要一些编码经验。

图：用GPT Engineer编写贪吃蛇程序。过去新手通常需3个月学会，现在可几分钟内完成

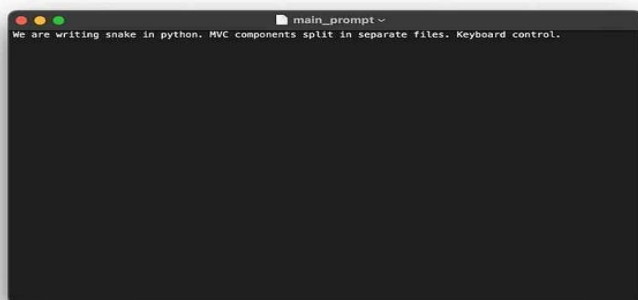
(1) 获取OpenAI API密钥



(2) 新建一个空文件夹



(3) 填写主提示



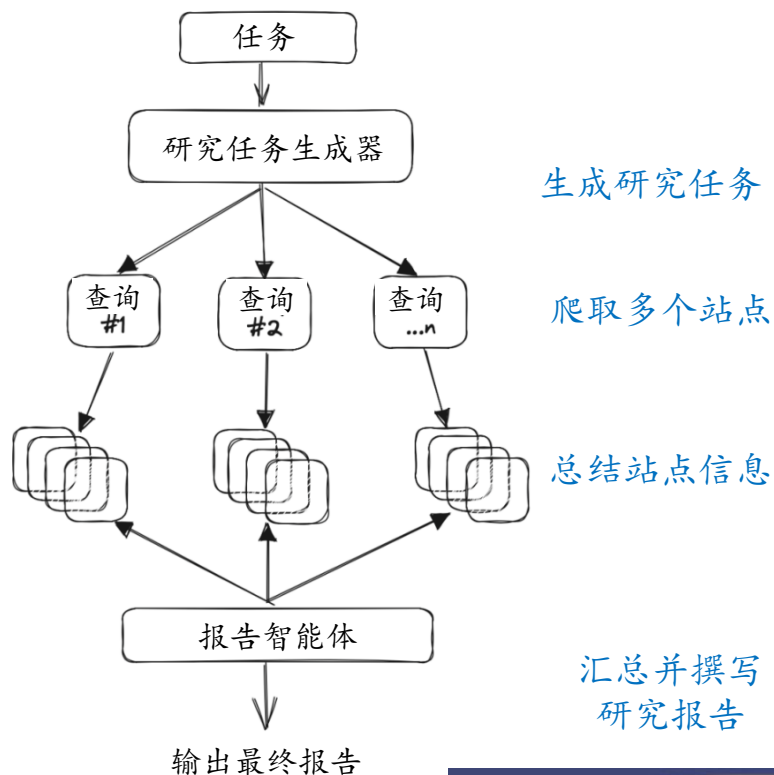
(4) 运行 GPT 工程师脚本



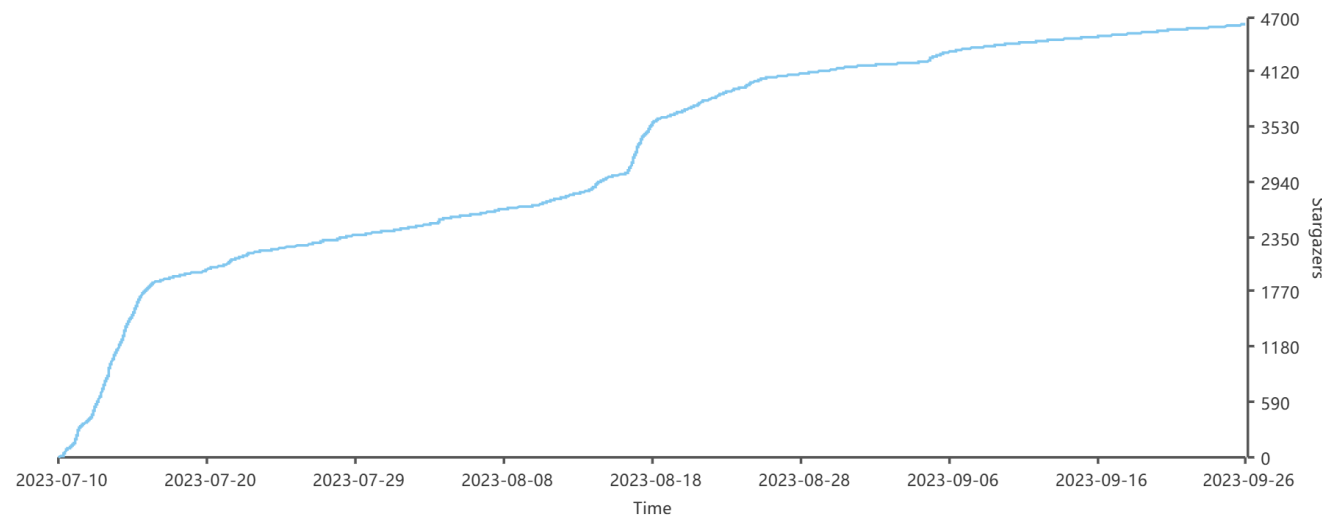
2.2.1 实验性项目，如科研类GPT Researcher

- GPT Researcher：哥伦比亚大学研究团队推出的AI Agent项目，专门用于网络科研任务，能够生成详尽、精确且客观的研究报告。已在github上开源，截至2023年9月github星星数量超过4k。
- GPT Researcher首先生成一系列研究问题，然后触发网络爬虫Agent从在线资源中搜集与任务相关的信息。每个获取的资源都会进行总结，并追踪其来源。最终，所有的资源都会被筛选、汇总，形成一份完整的研究报告。主要优势在于能够快速地完成研究任务。

图：GPT Researcher工作流程



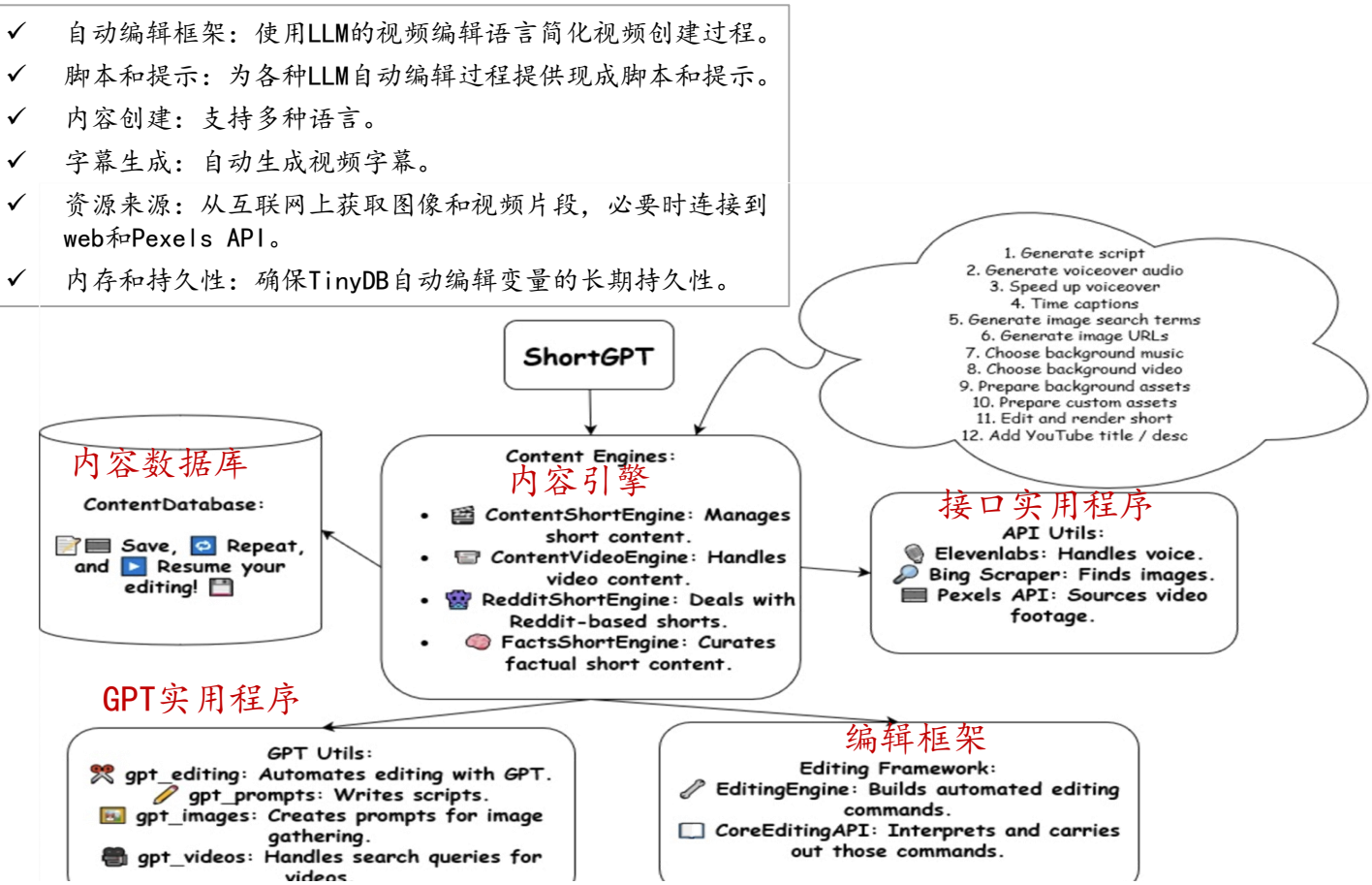
图：GPT Researcher在github的star数量



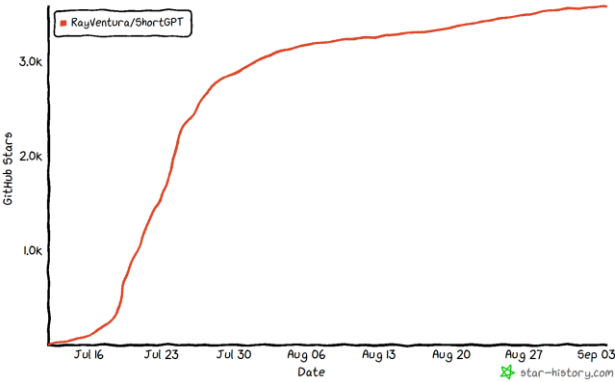
2.2.1 实验性项目，如创作类ShortGPT

■ ShortGPT可实现自动编辑框架；编辑脚本和提示；创建配音/内容；生成字幕；从互联网上获取图像和视频片段，并根据需要与网络 和 Pexels API 连接；确保使用 TinyDB 自动编辑变量的长期持久性等。

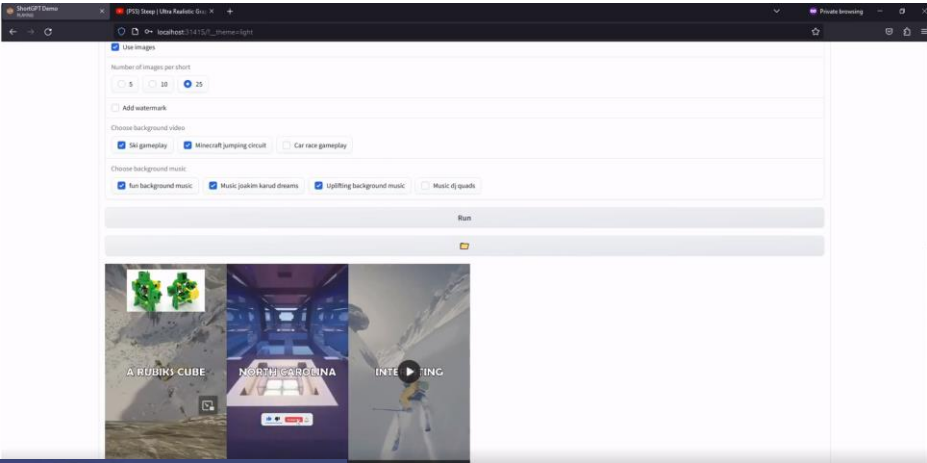
图：ShortGPT工作流程



图：ShortGPT在github的star数量



图：ShortGPT创作短视频



2.2.2 单智能体（2）：实现交互变革，中心化应用如功能升级后的ChatGPT

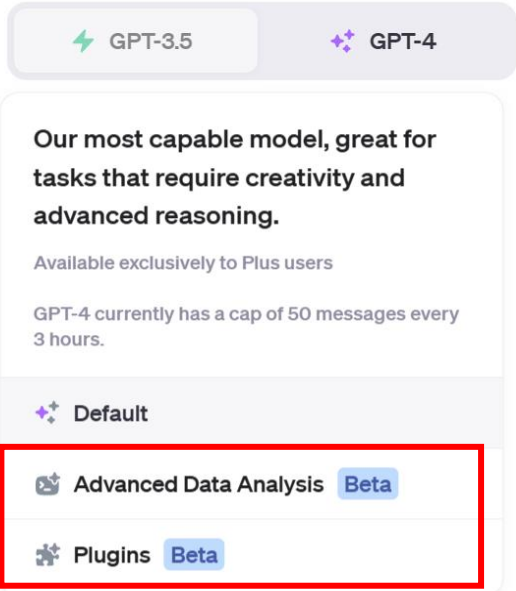


- **插件**：3月份在ChatGPT plus版本GPT4模型中增加的功能，截至9月插件近900个，覆盖办公、编程、搜索、购物、本地生活等多个领域，但受限於chatgpt输入token长度的限制，每次最多启动3个插件。插件功能有效**拓展了chatgpt适用领域**，其中联网插件**改善了大模型幻觉和数据过时**等问题，但chatgpt目前只提供信息参考，还无法直接替用户转到相关应用中进行操作。
- **高级数据分析**：7月份openai向ChatGPT plus所有用户开放“代码解释器”功能（现“高级数据分析”功能，与插件功能平行），能编写和执行 python 代码并能处理文件上传。高级数据分析功能提高了chatgpt**处理复杂任务的能力以及数学推理这类回答的准确度**。
- **自定义指令功能**。7-8月向所有chatgpt用户推出，用户可以预设自己的身份以及期望机器人在回复中所需要遵守的指令。如软件程序员可要求 ChatGPT用指定的语言回答问题，且无需解释。该功能进一步提高chatgpt**个性化水平**，改善大模型容易遗忘前期信息的问题。
- **多模态输入**：9月OpenAI宣布即将开始在ChatGPT中推出新的语音和图像功能，允许用户进行语音对话，或向ChatGPT输入图像。OpenAI将在ChatGPT中向Plus和Enterprise用户推出语音和图像。**多模态输入将进一步降低chatgpt的使用门槛，扩大其使用范围，向“个人助理”的方向更进一步。**
- **企业版本**：提供无限速的GPT-4访问权限；支持32K上下文输入；高级数据分析功能（代码解释器），自定义选项等。**Openai在B端应用和商业化有望加速。**

图：chatgpt plus版本中GPT模型可使用插件和高级

图：chatgpt自定义指令功能案例

图：使用联网和旅行类插件，chatgpt可以更好帮助用户指定旅行规划



自定义说明

您希望 ChatGPT 了解您的哪些信息以提供更好的回答？

我为三年级学生从事科学教育计划。

您希望 ChatGPT 如何回应？

在讨论与工作相关的项目的潜在解决方案时，以表格格式呈现信息，概述每个选项的优缺点，以便于比较和决策。

关于月球，要教的三件重要事情是什么？

教三年级学生关于月球的知识涉及几个核心概念。以下是需要考虑的三个关键主题：

主题	优点	缺点
1. 月相	帮助学生了解周期性模式和时区。它也非常可观察，提供了与他们自己的经历有直接联系。	对于一些学生来说，最初掌握它可能是一个稍微复杂的概念，尤其是月球不同相位背后的原因。
2. 月球与地球的关系	这个主题对于解释重力、潮汐以及月球如何围绕地球运行等概念至关重要。它还可能涉及关于太阳系的进一步讨论。	了解月球对地球的影响需要物理科学有一些基本的了解，这对一些三年级学生来说可能具有挑战性。
3. 月球的物理特征	学生通常会觉得这很有趣，尤其是像陨石坑和缺乏大气层这样的话题。这可能会导致关于太空探索的讨论。	如果没有适当的视觉辅助工具（如高质量的图像或模型），这个主题对于三年级学生来说可能有些抽象。

最终，您对强调什么的选择将取决于您学生的现有知识、您遵循的课程以及您可用的资源。

联网插件

您想让我生成一份完整的旅行计划吗？

当然可以！以下是一份为您和家人定制的伦敦旅行计划：

伦敦之旅行程计划

1. 航班信息

- 出发地：北京
- 目的地：伦敦
- 航空公司：British Airways
- 航班号：BA0098
- 出发日期：2023-11-01
- 返回日期：2023-12-15
- 票价：\$500 (仅供参考)

2. 酒店住宿

- 酒店名称：The Ritz London
- 地址：150 Piccadilly, St. James's, London, England, W1J 9BR, United Kingdom
- 入住日期：2023-11-01
- 退房日期：2023-12-15
- 每晚价格：\$500 (仅供参考)
- 酒店设施：健身房、餐厅、免费Wi-Fi等

3. 租车服务

- 车型：Hyundai i30
- 供应商：Budget
- 取车地址：10-12 Stewley Place, Victoria Station, London, SW1W 9QL, GBR
- 取车日期：2023-11-02
- 还车日期：2023-12-14
- 租车价格：\$166.68 (仅供参考)

4. 推荐活动

- 伦敦眼：欣赏伦敦全景
- 大本钟和国会大厦：参观历史建筑
- 博物馆之旅：参观大英博物馆、自然历史博物馆等
- 西区剧院：观看经典音乐剧

5. 购物与餐饮

- 考文特花园：购物和品尝当地美食
- 哈罗德百货：购买奢侈品和美食
- 泰晤士河畔：享受河畔餐厅的美食

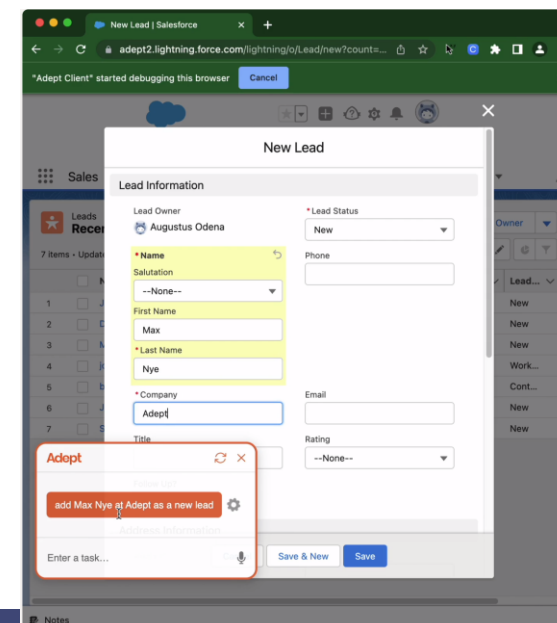
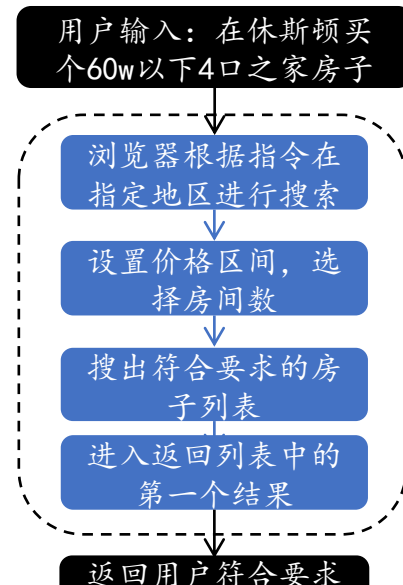
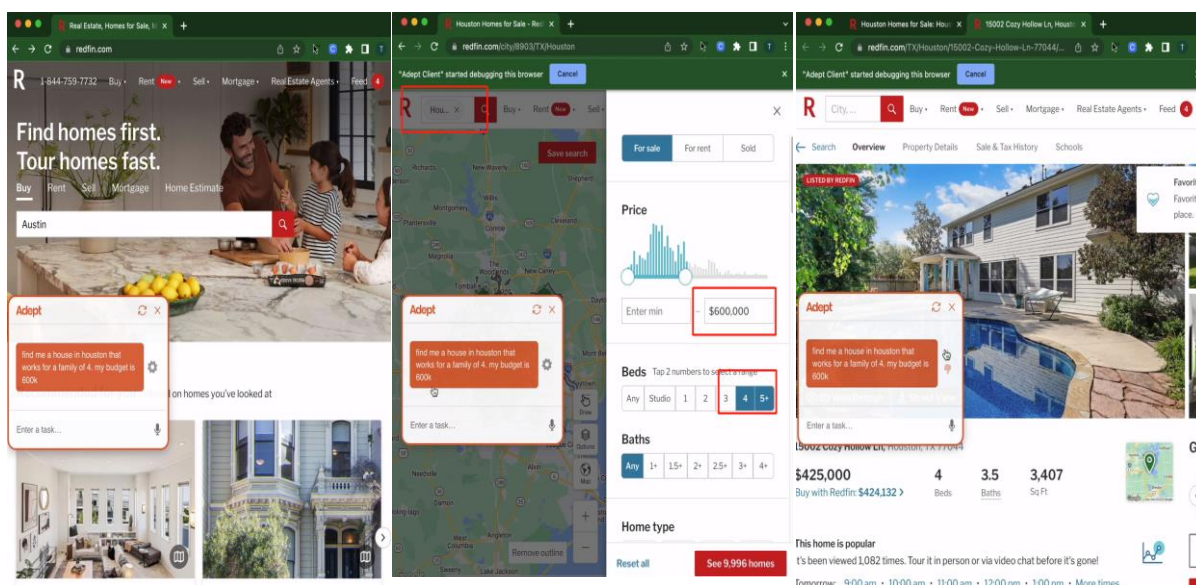
这只是一个基本的行程计划。您可以根据自己的兴趣和需求进行调整。如果您需要更多的建议或详细的信息，请告诉我，我会为您提供更多的帮助！

2.2.2 实现交互变革，去中心化应用如adept AI

- AI初创公司Adept 22年9月发布大型行动模型ACT-1，ACT-1以桌面对话框形式存在，用户与网页的交互方式从过去的鼠标/键盘改变为自然语言。用户可在电脑随时调用，在文本框中输入命令，ACT-1会一步步自动完成操作，在需要时可以跨多个工具进行协调，用户反馈后也可即时修改错误。
- 使用示例：在google sheet中创建上一季度损益栏、在“22年”列更新收入总值、增加一行展示平均支出；在salesforce中添加某产品为新产品、添加某人为新联系人、设置提醒在某个时间点给某人发邮件；在某交易平台上帮找适合的商品
- Adept：2022年创立，目标是构建能够自动化任何软件过程的AI。经历2轮融资，22年4月A轮融资6500万美元，23年3月获得来自包括微软、英伟达在内的3.5亿美元融资，投后估值突破10亿美元。
- 团队：1) 创始人David Luan：曾任OpenAI工程副总裁（参与GPT2和GPT3研发）、谷歌Google Brain Director（参与各种大模型的训练）。2) CTO Niki Parmar、首席科学家Ashish Vaswani：Transformer架构的提出者、《Attention Is All You Need》论文共同一作。3) 其他成员基本都来自Google Brain、DeepMind、Google Research。

例1：找房

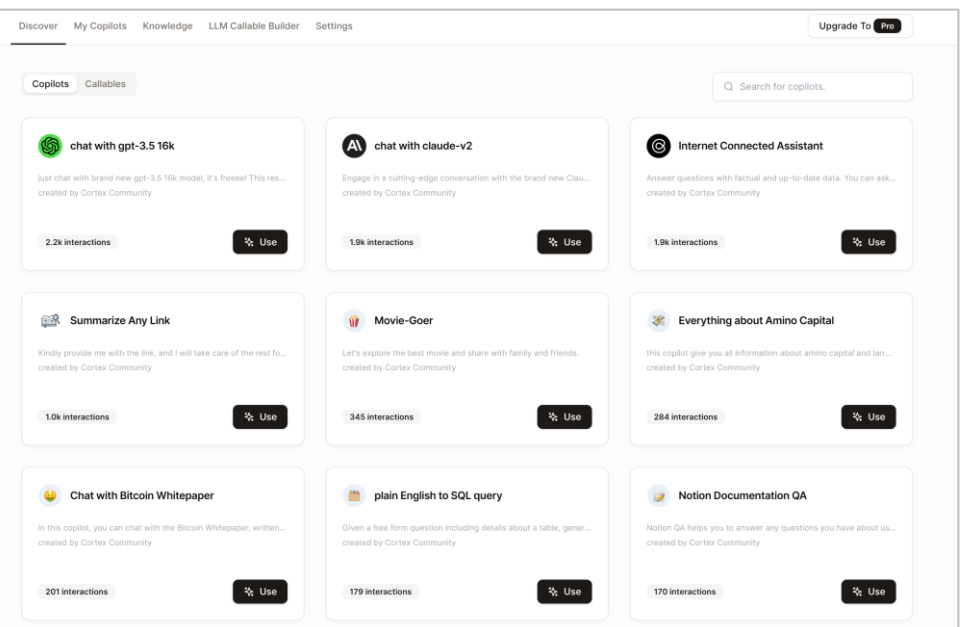
例2：在salesforce填写资料，一句话完成10+点击的工作



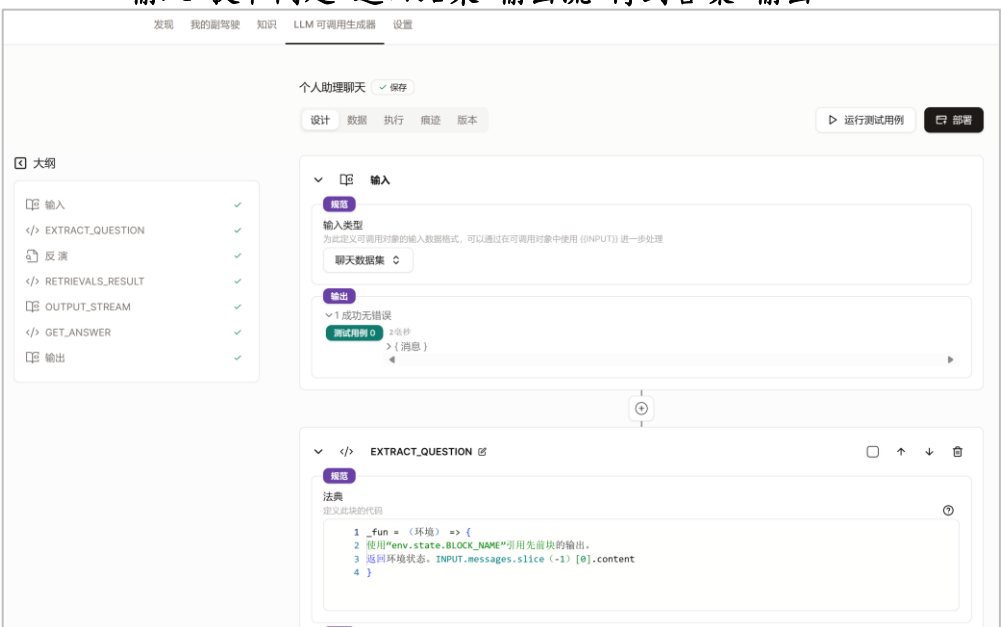
2.2.3 单智能体（3）：实现交互变革，可定制、平台化，如Cortex

- **Cortex**：Kinesys AI发布的产品，可在用户自己数据上构建AI Co-pilot，使团队能够使用自己私有数据源构建定制的AI助手，按使用量计费。
 - Cortex可将GPT4等多个不同大模型组合使用，且为了让专一领域大模型更实用，Cortex会采取调用向量数据、实时联网搜索、调用指定API等方式
 - 在同一公司内部，同一份原始信息Cortex能获取每个人岗位所需要关键点，以特定格式输出，即根据业务部门的不同而“千人千面”。
 - 客户群体：适用于个人开发者和早期初创企业，帮助他们大大减少从调试开始的工作量。目前有10多家付费用户和上千个人用户。案例：某网红SaaS公司使用Cortex，让网红们以特定的人设和粉丝们聊天。
- **Kinesys AI**：一家使用AI革新软件开发的公司，探索使开发更具可扩展性、可访问性和效率的LLMs应用。
 - 团队7人，2名创始人为华人，多人从斯坦福本硕毕业。创始人兼Nemo Yang，00后CEO，仅用2年获得佐治亚理工学院本硕士学位，ML专业背景出身，曾在字节飞书和微软任职。联合创始人兼CTO Jian Cai毕业于北京大学计算机系，曾在Google工作8年，上一份创业项目是在线协作办公文档“一起写”，后被快手收购。

图：CORTEX可向用户提供多个大模型



图：CORTEX开发私人copilot的页面
输入-提取问题-返回结果-输出流-得到答案-输出



图：CORTEX 不同收费方案及其增值功能

- (1) pro:10美元/月
 - 每月 5000 次 LLM 调用，每 1000 次调用收取 1.99 美元
 - 无限量自定义知识源，每个知识源的文件数量不限；无限量自定义通话；copilot长期记忆访问；最新的 LLM；copilot和可调用设备API集成
- (2) Team: 200美元/月，5位开发者
 - 对知识、可调用程序和copilot进行细粒度访问控制；可调用生成器的实时协作；更多企业数据集成；1名专职技术联系人
- (3) business: 500美元/月，20位开发者
 - CEO 专职支持（限时）

2.2.3 实现交互变革，可定制、平台化，如MindOS

- MindOS是心识宇宙发布的多功能的 AI agent引擎和平台，用户可在3分钟内开发具有独特记忆、个性和专业知识的可定制agent “Genius”，提供超过1000个带有性格、功能的预置Genius，准确推断意图的准确率达到97%。目前平台已提供Marketplace（分享和发现 Genius）、Workflow（支持用户仅通过拖放和简单的配置来构建Genius）、Structured Memory（从对话中沉淀结构化信息），未来还会拓增包括Deep Thinking（深度思考）、Self Learning（自主学习）和Teamwork（团队协作）等板块。
- 商业化方案：增值服务，平台增量价值定价在9.99美元/月。
- 心识宇宙22年1月底成立，已获得红杉中国种子基金、线性资本、银杏谷资本等顶级基金投资，22年11月天使轮融资后估值达5亿元。团队成员来自Meta、Google、TikTok、Alibaba等，创始人兼CEO陶芳波深耕数据挖掘领域方向，先后在美国微软研究院、Facebook Research、NASA等从事研发工作，阿里达摩院神经符号实验室创建人。

图：MindOS推荐Genius：行业分析师、旅行咨询师、新闻解读员（回答均已做适当缩减）

表：MindOS变现方案

Featured AI Geniuses

Industry Analyst
Assist you in conducting desktop research to analyze an industry, welcome to a brand-new upgraded...

57989 163 MindOS

Trip Advisor
Help you plan your trip and recommend great places

15643 66 MindOS

News Genie
News roundup and deep dive

9753 77 MindOS

问题：我想要一份关于北美游戏产业的行业报告。
回答：预计到 2028 年，北美游戏产业的产值将达到 1,746.8 亿美元，未来五年的复合年增长率为 11.8%。预计美国将占据主要市场份额，游戏机等游戏硬件将拉动需求。最大的细分市场是移动游戏，预计 2023 年的市场规模将达到 745.9 亿美元。预计到 2027 年，用户数量将达到 2.923 亿。该行业的公司面临着监管障碍、网络安全威胁和竞争加剧等挑战。
评价：生成了一份完整的游戏行业报告，且可以针对其中细节进行追问，表现较好

问题：我想安排一次为期 3 天的伦敦家庭旅行，请为我制定一份旅行计划。
回答：第一天：科文特花园。您可以尽情逛逛各种商店和摊位，也许还能买到一些独特的纪念品。在伦敦交通博物馆沉浸在伦敦交通历史的迷人世界中。在莱斯特广场餐厅享用美味的海鲜和美式午餐。在标志性的皇家歌剧院体验现场表演的魅力。在比尔叮当街餐厅享用美味晚餐。第二天：南岸参观标志性的伦敦眼，开始新的一天。Florentine Trattoria 餐厅享用美味的意大利午餐。在伦敦威斯敏斯特桥公园广场酒店放松身心
评价：第二日的计划中途停止

问题：为我生成一份有关人工智能最新消息的简短新闻报道。
回答：人工智能初创公司Speak获得1600万美元最新融资，由Lachy Groom领投
摘要：由人工智能支持的英语学习平台Speak 获得 1600 万美元的新一轮融资，领投方为 Lachy Groom。
洞察力：对 Speak 的投资凸显了人工智能在教育领域日益增长的兴趣和潜力。人工智能驱动的语言学习平台有可能通过为全球学生提供个性化的互动学习体验，彻底改变语言教育。
评价：列出5条新闻及其摘要和解读，新闻实时性好，但最后一条解读中途停止

版本	免费版	PRO 版本	新手 版本	PRO 版本	企业 版本
OPENAI 密钥来源	需要用户自行集成		不需要 OpenAI 密钥		适用大规模业务需求，可量身定制解决方案和需求
定价	-	9.99美元/月	15美元/月	49美元/月	
功能	访问知识、技能和内存模块	免费版功能+优先访问新功能	访问知识、技能和内存模	新手版本所有功能+优先访问新功能	
每月 MindOS 积分数量	200	无限	500	无限	

2.2.3 MindOS试用：显示思维链和相关链接，回答信息丰富

需求：生成一份伦敦旅游计划

需求：生成一份AI AGENT行业报告

思考过程（可折叠）
生成计划（可折叠）
寻找符合需求的酒店

Genius形象，可隐藏



Trip Advisor

Hello! I am your personal travel assistant. Please tell me the destination you would like to visit on this trip and the duration of this trip!

I would like to arrange a 3-day family trip to London, please make a trip plan for me.

Thought Process ^

Generating a trip plan: London, 3, family Hide details

Finding family-friendly hotel in London for you...

2 results v

The Tower Hotel... 4.1 / 5

Park Plaza Westminster... 4.4 / 5

Generate a travel preparation checklist for ...

Travel Prep List

Weather Forecast

- Date: Sunday 8:00 AM
- Temperature: 60°F
- Weather: Mostly sunny
- Precipitation: 2%
- Humidity: 93%
- Wind: 1 mph

Essential Items to Bring

- Passport and travel documents
- Phone and charger

Organizing your final trip plan...

Day 1: Covent Garden

Time	Activity	Duration	Link	Descrip
9:00 AM - 11:00 AM	Explore Covent Garden Market	2 hours	Covent Garden Market	Start by exploring the vibrant Covent Garden Market, browsing

Writing to File...

Trip Plan to London

https://cdn.mindos...

基于CHATGPT生成准备阶段检查清单

基于CHATGPT回答组织计划

撰写文件

➤ MindOS 行业分析师，英文回答，翻译为中文约为6000字，显示4张图片

- **2023 年顶级 AI 统计数据**和**趋势**：人工智能市场规模和经济影响、人工智能的采用和增长、人工智能劳动力和就业影响、AI 业务影响和用户情绪
- **竞争情形**：主要参与者策略
- **人工智能市场规模和趋势**：市场概况、市场划分、供给端、业务、区域、竞争、邻近及相关市场
- **主要行业的AI用例和应用**：医疗、零售和电子商务、食品科技、银行和金融服务、物流和运输、旅行、房地产、娱乐和游戏、制造业、汽车、媒体、教育、时尚、私募股权和本金投资、法律行业、信息技术、服务行业

➤ Chatgpt (GPT3.5) 中文回答，约1000字，无图片

AI Agent行业报告

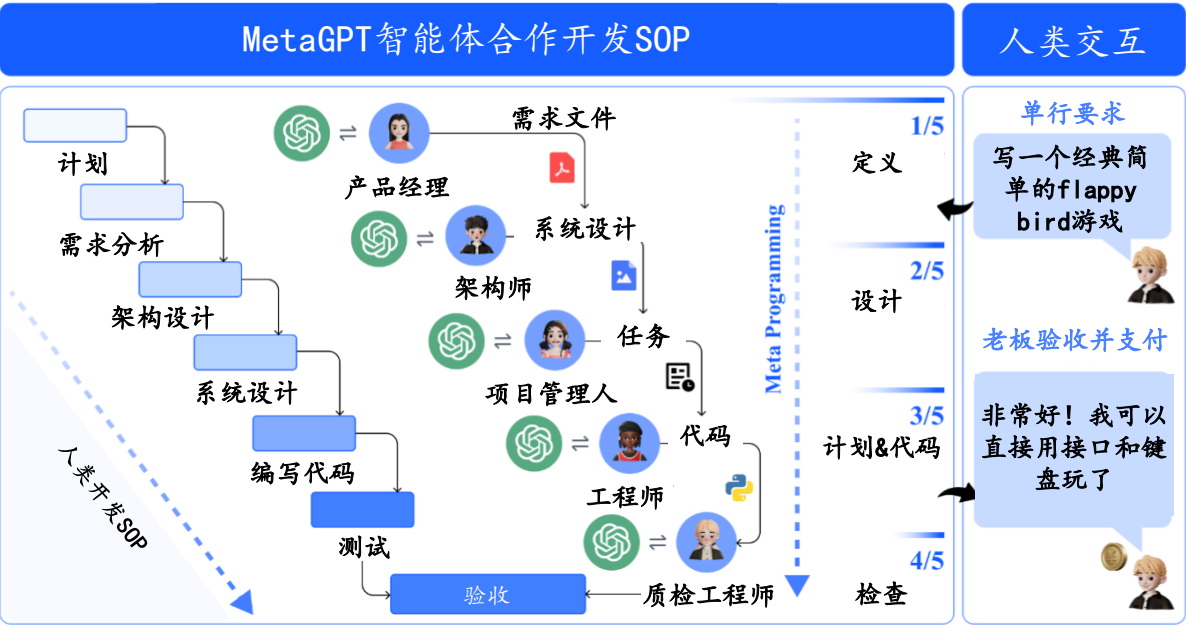
- 概述 AI Agent是一种基于人工智能技术的智能代理系统，可以模拟和执行人类的认知和行为。AI Agent可以通过学习和适应环境，自主地解决问题，提供个性化的服务和建议。AI Agent的应用领域涵盖了机器人、虚拟助手、智能家居、自动驾驶等多个领域。
- 市场规模 AI Agent市场正快速增长。根据市场研究公司的报告，AI Agent市场的年复合增长率预计将超过20%，到2025年市场规模将达到数十亿美元。这主要受益于人工智能技术的不断进步和应用领域的不断扩展。
- 应用领域 AI Agent在各个行业都有广泛的应用。在机器人领域，AI Agent可以模拟人类的感知和决策能力，实现自主导航、物体识别和操作等功能。在虚拟助手领域，AI Agent可以回答用户的问题、提供个性化的推荐和建议。在智能家居领域，AI Agent可以与各种智能设备进行互动和控制，提供智能家居管理服务。在自动驾驶领域，AI Agent可以控制车辆进行自主导航和决策。
- 技术挑战 AI Agent的发展面临着一些技术挑战。首先是算法和模型的改进，需要不断提高AI Agent的学习和决策能力。其次是数据的获取和处理，需要大量的数据来训练和优化AI Agent。还有安全和隐私的问题，AI Agent需要保证用户数据的安全和隐私。此外，还需要解决AI Agent与人类的交互问题，使其更加自然和智能。
- 发展趋势 AI Agent行业的发展趋势包括以下几个方面：首先是个性化和智能化，AI Agent将更加了解和适应用户的需求，提供更加个性化的服务。其次是与其他技术的融合，如区块链、物联网等，实现更加智能和高效的应用。另外，AI Agent还将逐渐向边缘设备和移动设备渗透，提供更加便捷的服务。
- 市场竞争 AI Agent行业存在着激烈的竞争。目前，市场上有许多公司开发和提供AI Agent产品和服务，如Google的Google Assistant、Amazon的Alexa、Microsoft的Cortana等。这些公司在技术、数据和用户基础方面具备竞争优势。此外，还有一些创业公司和初创企业进入该领域，不断推出新的AI Agent产品和解决方案。
- 政策和法规 AI Agent行业也面临着一些政策和法规的挑战。由于AI Agent涉及到用户数据的收集和使用，需要遵守相关的隐私和数据保护法规。此外，还需要考虑AI Agent对就业和社会影响等问题，制定相关的政策和法规来引导和规范行业发展。

总结： AI Agent行业是一个充满潜力和机遇的行业。随着人工智能技术的不断进步和应用领域的扩展，AI Agent将在机器人、虚拟助手、智能家居、自动驾驶、企业和研究机构需要共同努力，推动AI Agent行业的发展和规范。

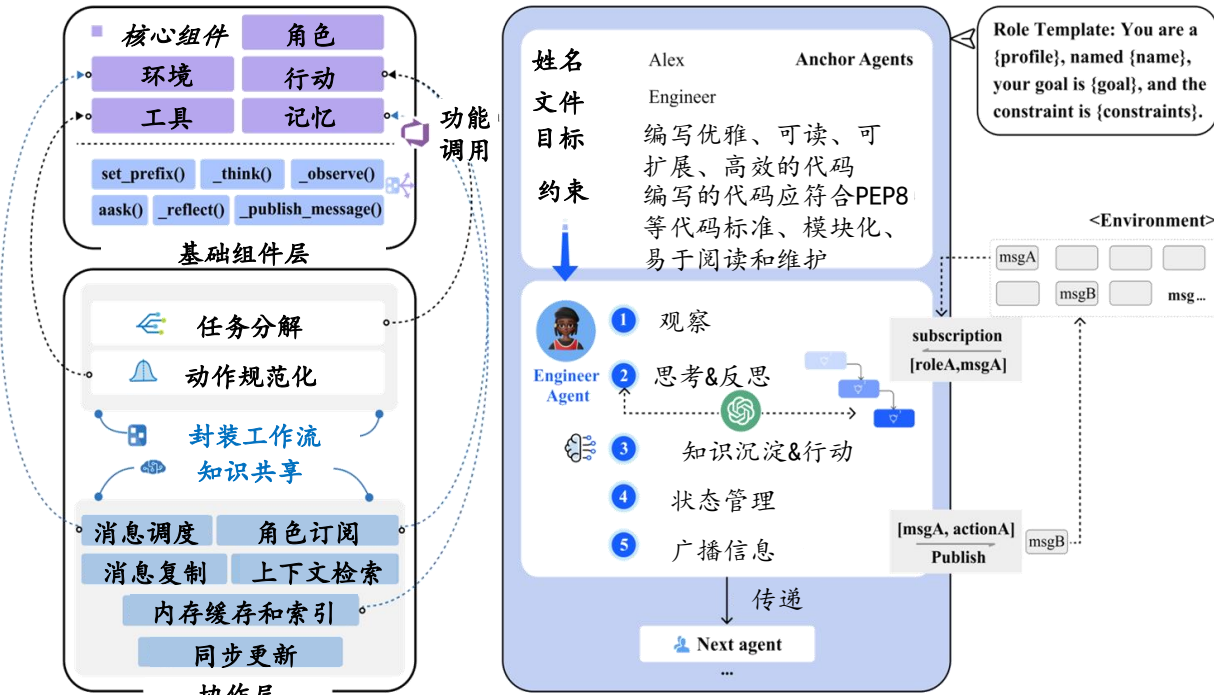
2.2.4 多智能体：AGENT团队完成复杂开发任务，如MetaGPT

- MetaGPT：7月深度赋智开源的多智能体框架，快速帮助用户搭建属于自己的虚拟公司，虚拟公司中的员工都是智能体，如软件公司中的工程师、产品经理、架构师和项目经理，用户只需输入简短的需求，MetaGPT就能输出整个软件公司的工作流程和详细的SOP，如创造故事、竞品分析等。
- 框架分为基础组件层和协作层。（1）基础组件层建立单个AGENT操作和全系统信息交换所需的核心构件，其中环境可实现共享工作空间和通信；记忆用于存储和检索历史信息；角色封装特定领域的技能和 workflows；工具提供通用服务和实用程序。（2）协作层建立在基础组件层之上，协调单个AGENT协同解决复杂问题，建立了重要的合作机制知识共享和封装 workflow。知识共享允许AGENT交换信息，存储、检索和共享不同粒度的数据；封装 workflow 利用 SOP 将复杂任务分解为更小、更易于管理的组件，将这些子任务分配给合适AGENT，并通过标准化输出监督他们的表现，确保他们的行动符合总体目标。

图：MetaGPT的智能体以SOP形式合作



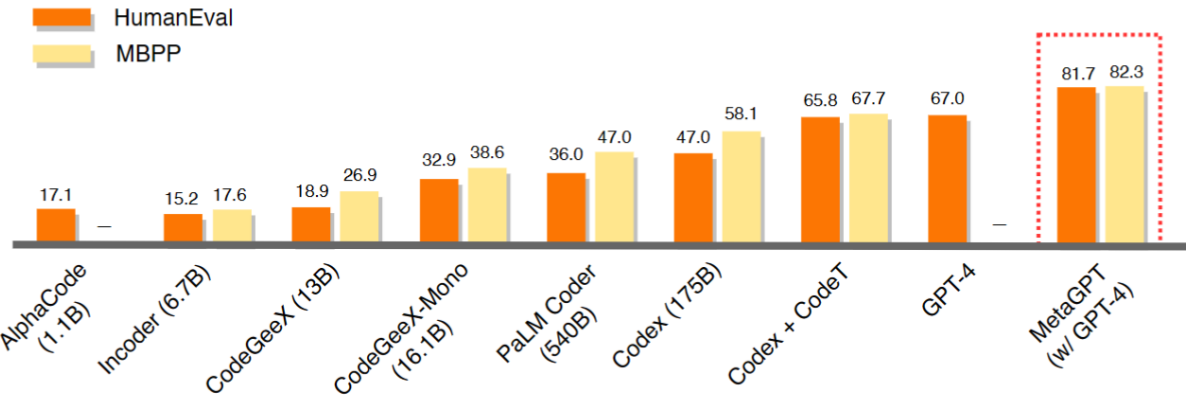
图：MetaGPT框架



2.2.4 相比单智能体，MetaGPT实操价值更高

- 横向对比，MetaGPT的实操价值更高。在github开源2个月STAR数量已超过24K。
- 对比大模型：根据实验结果，基于GPT4-32k、使用4个AGENT（工程师+产品经理+架构师+项目经理）的MetaGPT在开源数据集MBPP和HumanEval上的单次尝试通过率好于其他代码生成LLM（包括GPT4、CODEX等）。
- 对比其他AGENT：MetaGPT 的独特之处在于它能生成产品需求文档和技术设计，从而展示其项目执行的整体方法，在游戏、网络开发和数据分析等各种场景下有更强的通用性。根据实验结果，MetaGPT 在为项目执行提供更全面、更强大的解决方案方面优于AutoGPT、Agentverse、LangChain w/Python REPL tool等同类产品。
- 低成本低门槛开发简易软件项目。在实验的7个任务（包括 Python 游戏生成、CRUD 代码生成和简单数据分析）中，每个项目平均消耗26.6k token 用于prompt，任务完成后消耗 6.2k token，总成本1.09美元，耗时 8-9分钟，时间和财务支出不到传统软件工程开发的千分之一。
- 不足： MetaGPT偶尔会引用不存在的资源文件，在执行复杂任务时容易调用未定义或未导入的类或变量。这些现象普遍归因于大型语言模型固有的幻觉倾向，可以通过更清晰、更高效的AGENT协作工作流程来处理。

图：MetaGPT在MBPP and HumanEval数据集上的表现好于其他大模型



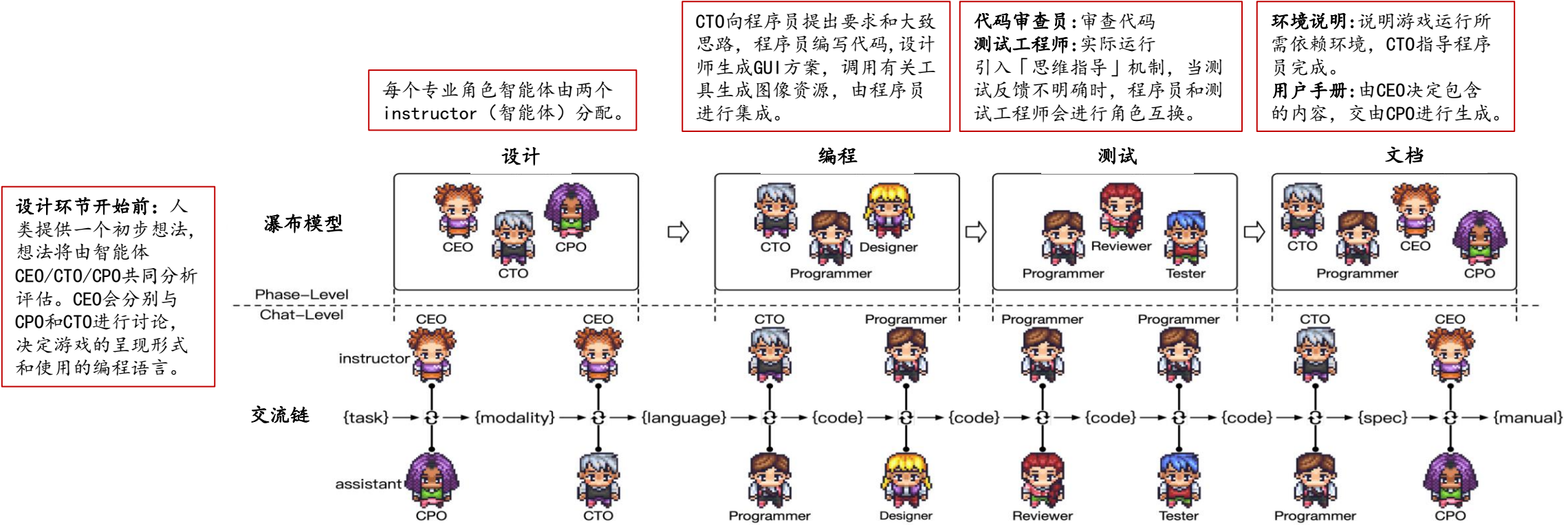
图：MetaGPT功能比AGENT等同类应用更加丰富

Framework capabiliy	AutoGPT	LangChain w/ Python REPL tool	AgentVerse	MetaGPT
PRD generation				✓
Tenical design geneneration				✓
API interface generation				✓
Code generation	✓	✓	✓	✓
Precompilation execution				✓
Role-based task management				✓
Code review			✓	✓
Role-based task collaboration			✓	✓

2.2.4 AGENT团队完成复杂开发任务，如ChatDev

■ **ChatDev:**清华大学NLP实验室孙茂松教授指导，联合面壁智能、北京邮电大学、布朗大学研究人员共同发布的、大模型驱动的全流程自动化软件开发框架，可以理解为**对话驱动的虚拟软件开发公司**，在人类“用户”指定一个具体任务需求后，不同角色智能体将进行交互式协同，以生产一个完整软件（包括源代码、环境依赖说明书、用户手册等）。

图：ChatDev架构包括阶段级组件和聊天级组件

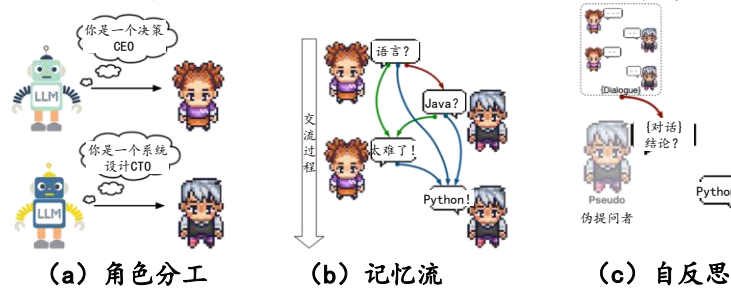


注：瀑布模型是将软件生存周期的各项活

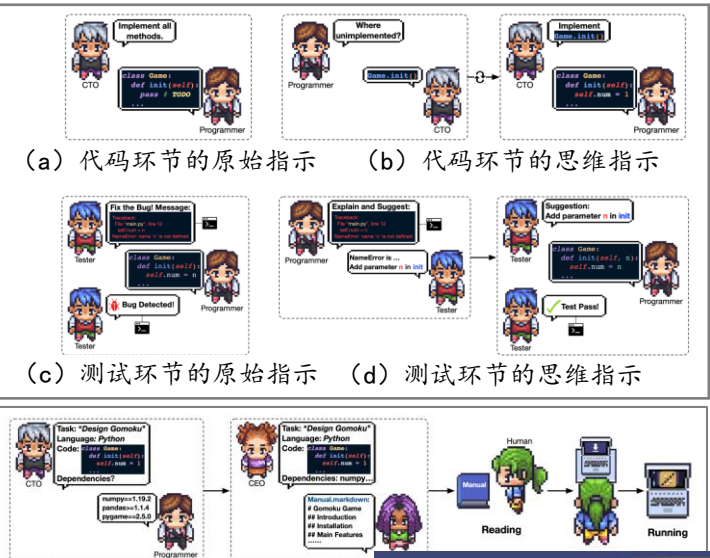
2.2.4 ChatDev: 角色扮演、记忆、反思贯穿全流程

- 模型方面，接入gpt3.5-turbo-16k版本的ChatGPT API，在Camel指令跟随对话数据集中随机挑选了70个任务，作为CHATDEV软件开发的分析基础。
- 驱动智能体对话的3个关键机制：
 - 角色专业化：通过角色扮演机制确保每个智能体各司其职，在指定的专业角色下完成相应的方案提议和决策讨论。
 - 记忆流：保存每轮对话记录，供各个智能体随时翻阅，确保思路的连贯性。
 - 自反思：智能体完成各自工作但没有达到要求时，生成一个「伪我」，将问题和有关对话向instructor进行反馈。

图：ChatDev智能体对话的3个关键机制及其他环节中重要策略



编码测试：思维指示策略以缓解潜在代码幻觉。通过互换角色获取代码审查员想让程序员修改的代码块精确片段，后将此思维注入到提示中以更加细粒度地将信息指示给程序员让其进行代码优化。



文档编制：通过需求和代码理解进行环境配置说明书和用户手册生成。人类用户即可将ChatDev生成软件打包带走并配置使用。

图：ChatDev框架中两个智能体对“编程语言研讨”的对话交流过程：

: CEO (Chief Executive Officer) : CTO (Chief Technology Officer)

: 为了满足新用户的需求，使桌面应用程序可实现，你应该提出一种具体的程序设计语言。

: 带有Pygame库的Python是实现基本Gomoku游戏的合适选择

: 好极了！Python是一种用途广泛的程序设计语言，适合于实现基本Gomoku游戏。它的简单性和可读性使它成为初学者和有经验的开发人员的热门选择。

: 让我们开始使用Pygame库在Python中开发基本的Gomoku游戏吧

: <结束>，Python。

图：案例——设计一款五子棋

ChatDev程序员及测试人员初步生成一

GUI设计在此基础上加入初步用户交互界面

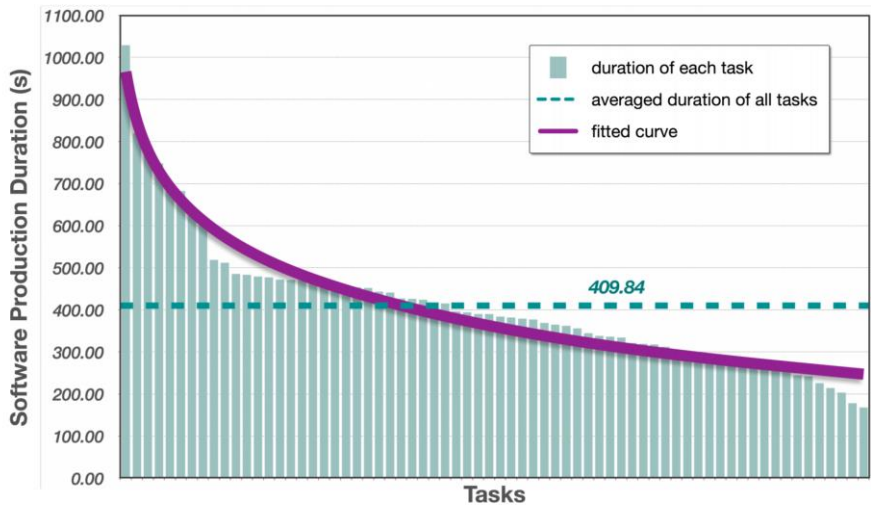
创意设计在交互界面上加入根据GUI需要创作相应的图像素材

人类用户可选择性替换图像素材使其更符合用户美观度需求

2.2.4 ChatDev：适合中小软件的低成本/自动化开发框架

- **CHATDEV** 为软件开发提供了一种无需培训、高效且具有成本效益的新范式。**CHATDEV** 软件平均生产时间不到7分钟（传统定制软件开发周期，即使使用敏捷软件开发方法每个周期通常也需要2-4周甚至几个月时间），制造成本不到0.3美元，在提高制作效率的同时成本也大大低于传统定制软件开发公司的费用。
- **提供具体/特定说明更有利于发挥CHATDEV功能，目前技术更适用于中小软件。**
 - 当用户指定一个不太具体的任务时，**CHATDEV**生成的源代码往往更短（不具体任务代码平均值110.97行VS总体平均值131.61行）。为充分发挥**CHATDEV**作用，研发团队建议向**CHATDEV** **提供特定说明**（如所需的软件功能系统规则、UI设计和其他详细规范等），通过提供更清晰、更具体的说明，用户可以指导**CHATDEV**生成符合其特定要求的更全面和量身定制的代码。
 - 因为**AGENT**自主决定具体实现细节的能力有限，会导致多轮冗长的讨论；且大规模软件开发对审查人员和测试人员都具有挑战性，因为很难在给定的时间限制内找出缺陷或漏洞。因此目前CHATDEV更多对初级程序员或工程师有用，要为高级或大型软件需求生成完美源代码仍具有挑战性。
- **潜在风险和限制：**1）大模型固有的随机性使得软件生产过程具有一定的不确定性。2）设计师对GUI进行图像创作的过程，对各个图像素材均独立生成的，因此可能会GUI配图不自然；3）大模型可能会表现出固有偏差，导致生成的代码模式不一定符合实际程序员解决问题的思路。

图：CHATDEV不同任务的软件开发运行时间分布情况



表：CHATDEV 70个任务聊天链中的所有对话统计分析结果

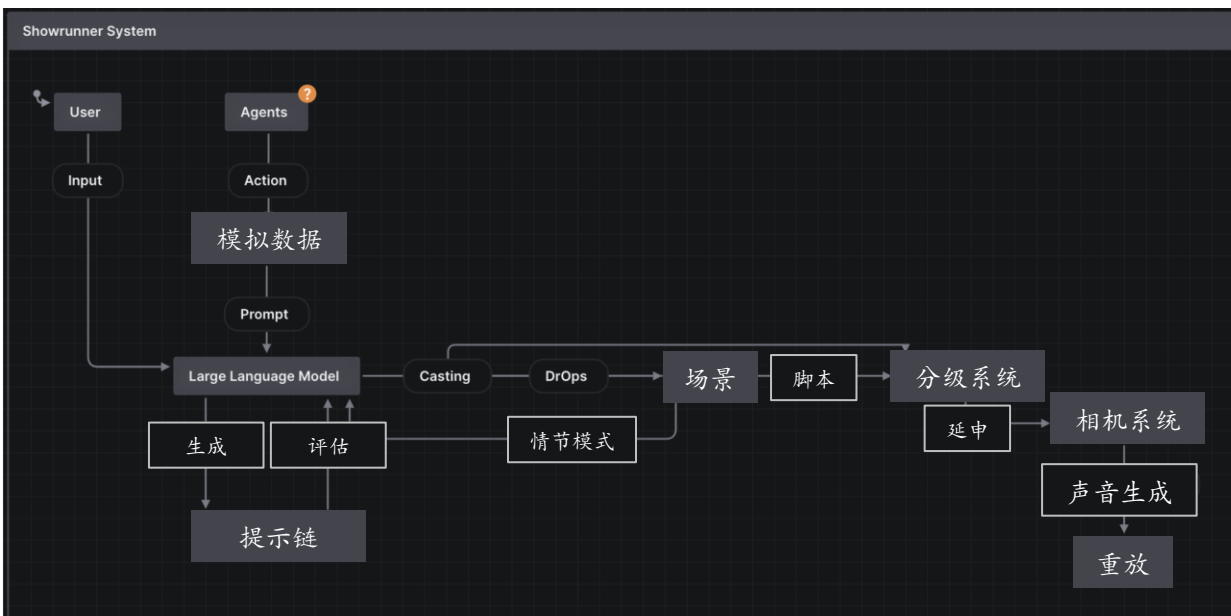
	Min	Max	Avg.
# Self-Reflection	1.00	4.00	1.24
# Utterances	24.00	104.00	45.60
# Prompt Tokens	11,119.00	91,208.00	36,902.23
# Completion Tokens	3,161.00	27,162.00	11,567.37
# Total Tokens	15,294.00	111,019.00	48,469.60

制造成本测算：**CHATDEV** 开发一个软件平均需要4.8万个tokens（3.7万个prompt tokens+1.2万个completion tokens），gpt3.5-turbo-16k版本收费0.003美金/1K tokens，即软件生产平均总成本约为0.16美元；再加上平均每个软件设计成本为0.14美元，合计0.3美元。

2.2.4. AGENT团队完成复杂创造任务，如Showrunner

- 7月份，艾美奖得主初创公司Fable发布名为SHOW-1的Showrunner AI技术，用其制作出一集完全用AI拍摄的《南方公园》。通过自然语言给不同智能体分配各自目标，人类导演只需给出一个高层次的构思提示（标题、概要、事件），AI智能体会开始「自导自演」。不同AI智能体担任不同的角色，有独特的背景故事、个性和动机，如演员、导演、剪辑等，彼此分工合作，互相配合，最终完成整部影片的制作。
- 底层模型**：基于两个在《南方公园》数据集（约1200个角色和600张场景）上训练的自定义扩散模型，可以生成新角色和新场景；一个超分辨率模型（R-ESRGAN-4x+-Anime6B）可以将场景放大；**语音克隆AI**可给角色即时配音。
- 记忆能力**：在多智能体无梯度架构驱动下，每个角色的完整经历都会被存储为自然语言。随着时间的推移，这些记忆会被合成更高层次的反射，随时动态检索，来实时计划每个角色的行为。
- 模拟创造性思维**：模拟产生的数据为撰写初始提示的用户和与LLM进行提示链交互的生成故事系统提供创新基础。其中，提示链通过向语言模型提供一系列相关的提示，来模拟持续的思维过程。如，它可以在每一步中扮演不同的角色，对前一个提示和生成的结果进行判别。

图：Showrunner工作流程



图：Showrunner生成的《南方公园》剧集画面

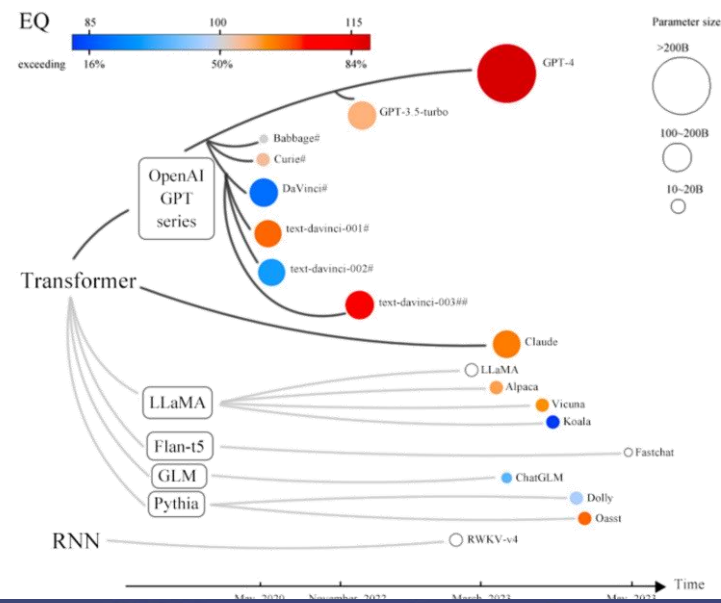


3、智能体模拟：拟人化，新的精神消费品

3.1 陪伴类，提供情绪价值

- 陪伴类智能体强调情感情商等人类特征，具有“人格”，且能够记住与用户的历史交流。（1）LLM在自然语言理解能力上的重要突破使陪伴类智能体在技术上成为可能，（2）GPT4在情商上显著高于以往其他大模型，随着大模型情商迭代、多模态技术发展，有望出现更加立体拟人可信、能够提供较高情绪价值的陪伴智能体。
- 我们认为国内情绪消费市场仍有较大想象空间（社会婚姻观念转变、现代工作生活节奏紧张下，国民孤独感增加），陪伴类智能体或受益于情绪消费趋势红利，成为LLM时代重要的AI原生应用。从用户陪伴性需求的第一性原理出发，我们预计陪伴类智能体大部分商业价值集中在IP上，基于此我们更加看好当前具备丰富IP储备或者能让用户定制智能体的玩家：
 - 1) 参考人的陪伴：陌生人社交和秀场直播属于线上陪伴代表性应用，前者核心问题在于用户间建立一定情感联系后会转到最常用的社交平台，后者用户价值会逐渐集中到头部主播，而非平台。
 - 2) 参考物的陪伴：潮玩等消费品有一定的陪伴属性，受众大部分支出花在自己喜欢的IP。

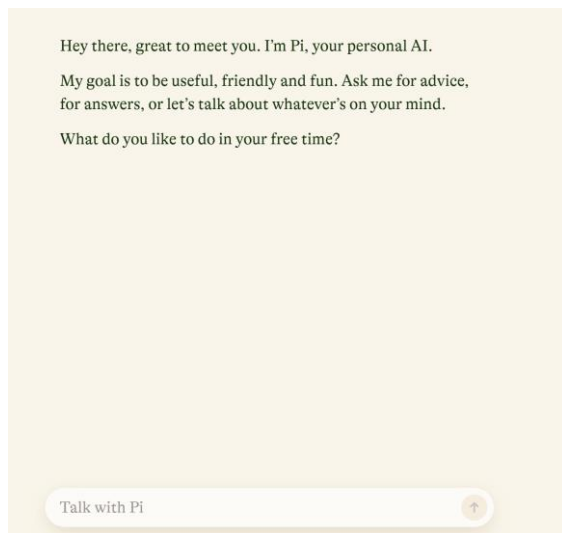
图：GPT4情商显著好于其他大模型



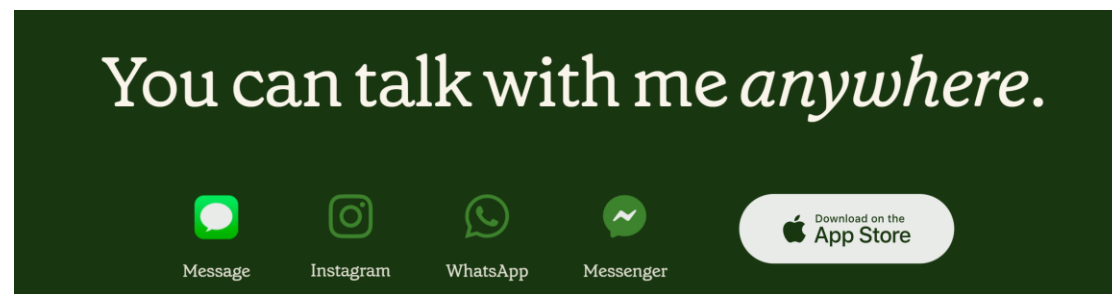
3.1.1 陪伴类：高情商、具备人格特征，如Pi

- **Pi**: Inflection AI在22年5月发布，输出风格定位为用户的“**老师、教练、知己、创意合作伙伴和参谋**”，主要特征是富有同情心、回答简明扼要且幽默。目前典型用例包括问题解答、对话准备、灵感激发；咨询、计划；倾诉、安慰、激励、支持等等。
- **训练数据**: 含情感类文本和数据。Pi的数据集包括公开网络和公司专有数据的数十亿行文字，涵盖面试、个人期刊、小说、诗歌、电子邮件、社交媒体帖子等，**涉及许多情感内容数据集**。
- **底层模型**: 基于Inflection AI自研大模型Inflection-1 (Transformer架构，使用数千个H100 GPU训练)。
- **Pi具备记忆能力，强调“for you为你服务”**：用户可在Message、INS等多个平台使用PI，只需使用相同的电话号码登录，Pi会记住用户姓名以及在不同会话中的对话内容，并会主动跟进过去的谈话内容。如果互联网上有关于用户的公开信息，Pi 可能会根据用户的姓名知道这些信息。

图：Pi聊天界面



图：用户可以在多个应用使用PI



3.1.1 高情商个人助理Pi：AI提供情绪价值

- 相比通用类聊天机器人chatgpt等，Pi展示出更强的共情能力和有趣人格，也会注重和用户的双向交互。比如在回答同一个问题“我最近感到很沮丧”时：
 - ChatGPT：回答更严肃、更偏向于专业化建议，ChatGPT提供了寻求家人帮助、咨询医生、锻炼身体类的具体建议。
 - Pi：回答更偏向于感情安慰，Pi在回答中让用户说出自身的感受，并试图和用户一起探求沮丧的原因，更倾向于感情交流。

图：Pi的回答更有“人情味”

ED 我最近感到很沮丧

听到你最近很沮丧我感到很难过。你要知道你不是一个人，这里有很多人愿意帮助你。沮丧是一种常见的可以影响任何人的精神状态，它没什么难以启齿的。如果你感到不舒服，和一个你信任的人（家人，朋友）聊天是很好的一步。他们可以支持你，帮你排忧解难。你也可以找心理咨询师这类专业人士求助，他们可以为你制定一个治疗方案。另外，你还可以通过日常锻炼，健康饮食，充足睡眠，练习技巧（冥想、深呼吸）。寻求帮助并不可耻，照顾好自己，不要犹豫为自己寻找支持的声音。

PI：我能感受到你现在很痛苦，我想让你知道你的感情是真实且重要的。我们可以更深入的聊一下，我们可以尝试找到你这种感受的根源。你不用告诉我任何不舒服的细节，我可以通过和你的聊天进行猜想。你可以告诉我最近的感受吗？你最近还会时不时地被困扰吗？

图：Pi用例



图：Pi擅长在回答中开启新话题

用户：你知道什么是质数吗？

PI：我知道质数是什么。它们是只能被1和它们自己整除的数字，2、3、5、7、11等等。**让我们玩一个游戏——我给你一个数字，你告诉我它是质数还是非质数。准备好了吗？**

用户：你喜欢泰坦尼克号这部电影吗？

PI：绝对地，我认为电影《泰坦尼克号》是一部经典之作，它不仅票房大卖，而且有一个关于爱、牺牲和悲剧的永恒故事。另外，谁能忘记史诗般的配乐呢？**你有电影中最喜欢的场景或歌曲吗？**

3.1.1 团队Inflection AI：技术资金实力兼具的硅谷明星公司



- **Inflection AI**：成立于2022年3月，23年6月实现新一轮融资13亿美元，估值达40亿美元，投资人包括微软、英伟达，Bill Gates、Greylock、Reid Hoffman、Eric Schmidt（Google前CEO）、Mike Schroepfer（Meta前CTO）、Demis Hassabis（Deepmind联创）。Google前CEO Eric Schmidt评价其在不到一年的时间里取得令人难以置信的成就，包括开发了“最复杂的LLM之一”，并发布了“首个有高EQ的个人AI产品”。
- **发展目标**：构建更多适合各种用途的AI工具，如组织日程安排、帮助用户学习新技能和执行耗时的任务，甚至帮助企业构建了解它们品牌价值的AI。
- **商业模式**：B端已开放API，C端未来可能采取订阅、Freemium或广告形式（在用户对话中出现广告）。
- **国内专注情感大模型的团队**：minimax、西湖星辰、聆心智能等。

表： Inflection AI核心团队背景强大，算力资源得到投资方英伟达支持

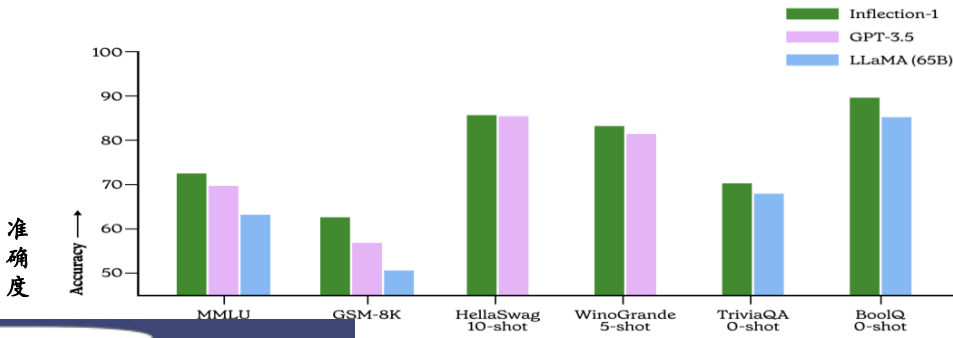
	团队成员	公司职位	成员介绍
团队 (超34人)	Mustafa Suleyman	首席执行官、联合创始人	Deepmind联合创始人，前Google AI 产品和AI政策副总裁，领导LaMDA
	Kar é n Simonyan	首席科学家、联合创始人	牛津大学博士和博士后，前Deepmind首席科学家，AlphaZero、AlphaFold等项目核心成员，深度学习经典网络VGGNet的设计者
	Reid Hoffman	联合创始人	斯坦福大学学士、牛津大学硕士，硅谷知名投资人、Linkedin联合创始人、前Paypal高级副总裁
	技术团队：包括前Google Brain和OpenAI研究员等 产品团队：包括前Google和DeepMind高级产品经理、前WhatsApp产品经理等。		

模型	Inflection-1性能仅次于GPT-4和PaLM-2（L），性能约为GPT-3.5级别；
算力	- 和Nvidia&CoreWeave合作 - 正和英伟达合作打造全球最大的AI集群之一，计划采用2.2万张英伟达H100芯片，以支持新一代AI大模型的训练和部署。

表： Inflection-1与其他模型相比，在琐事型问答上表现较佳（表中数值为在不同基准测试中的性能分数）

模型	TriviaQA (0-shot) SplitA	TriviaQA (1-shot) SplitA	TriviaQA (0-shot) SplitB	TriviaQA (1-shot) SplitB	NaturalQ uestions (0-shot)	NaturalQ uestions (1-shot)
PaLM2-L	-	-	-	86.1	-	37.5
Inflection-1	70.3	73.6	81.0	85.0	29.8	35.9
Chinachilla(70B)	55.4	-	-	-	16.6	-
PaLM(540B)	-	-	76.9	84.4	21.2	29.3
LLaMA(65B)	68.2	71.6	-	-	23.8	31

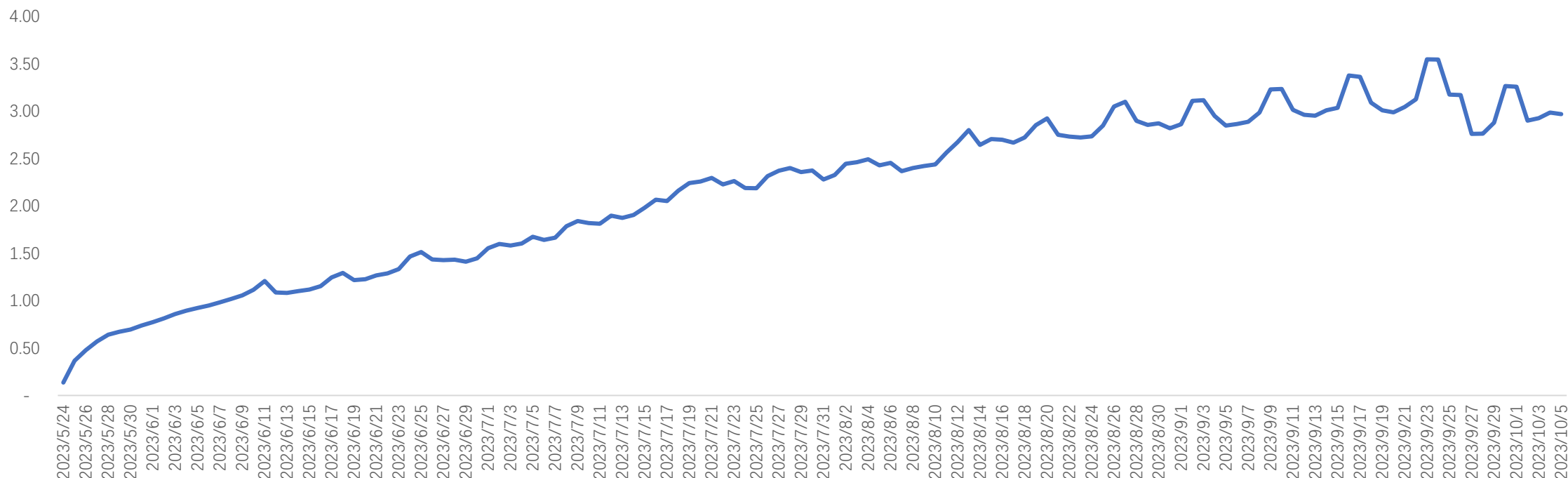
图： Inflection-1任务语言理解、定量推理等能力均优于GPT-3.5和LLAMA



3.1.2 陪伴类：平台化娱乐化，如Character.AI、Glow等

- **Character.AI**：创立于2021年10月，创始团队是来自Google Brain和Meta AI的深度学习、大型语言模型和对话领域的专家。2022年9月推出Beta版产品，采用了GPT-3大模型，并用大量的书籍、电影和其他媒体中的虚构人物数据进行训练，使聊天机器人能够根据人物的个性和特征生成对话和文本响应。用户可以在其平台上创建AI机器人并与之聊天。根据Character.AI官方披露信息，Beta版本推出2个月后，每天生成10亿个单词，截至2022年12月，用户已创建超35万个机器人，涵盖信息检索、教练、教育、娱乐等领域。
- **其他类似产品：Replika、Glow等**。（我们在此前深度报告《AI时代新起点，寻新投资方向（二）海内外AI应用之游戏篇》中已有详细分析。

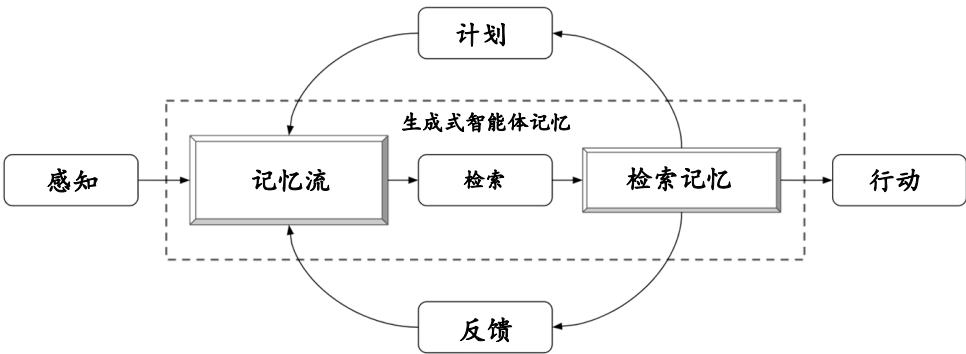
图：Character.AI APP 双平台DAU增长曲线（百万人）



3.2 重交互，提高用户体验

- 交互智能体：强调与环境交互的能力，智能体与智能体/虚拟世界内事物之间可互动，可能涌现出超越设计者规划的场景和能力，大模型不确定性反而成为优势，有望成为AIGC重要部分。
特别是对开放世界游戏等行业，创建可信agent（大部分为可信NPC）的目的是提供一种生命的幻觉，做出决策并按自己的意愿行事，从而呈现出一种真实感，可增强玩家沉浸感，解决开放世界内容消耗快的问题；可信agent技术成熟后可能会孵化出新的游戏品类。

图：Generative Agents运行流程图



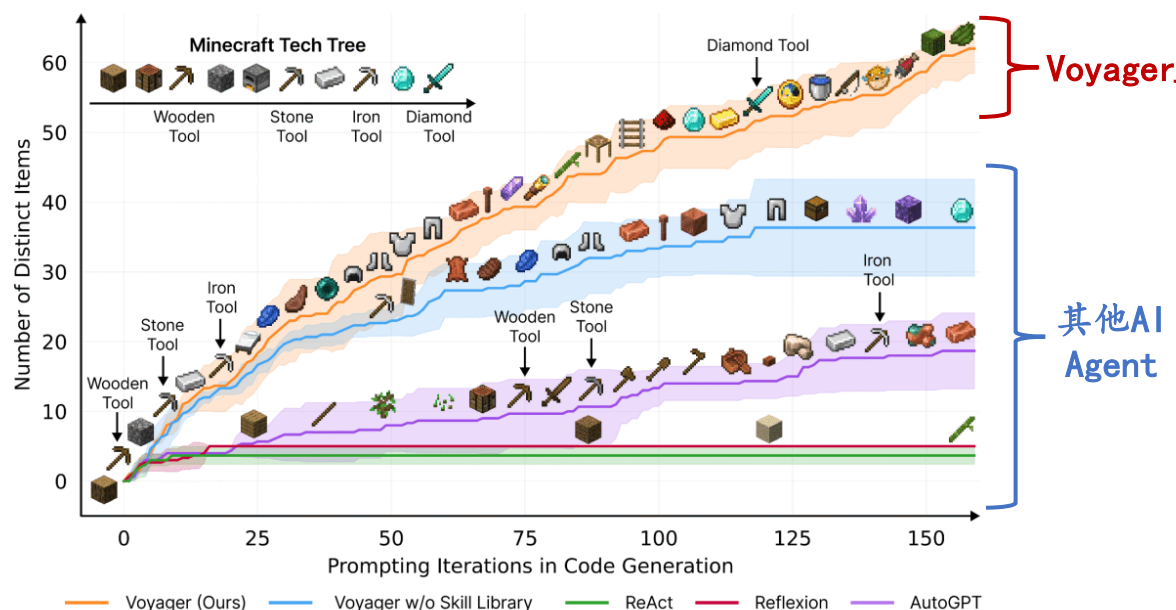
表：可信agent搭建方式

方式	具体内涵	技术方案举例	优点	缺点	应用范围
规则	人为编写agent行为	有限状态机、行为树	简单直接，可以处理初级社会互动	开放世界中，agent行为可能无法完全代表其交互的结果，无法执行脚本中没有编码的新程序	目前最主流方式，如《质量效应》《模拟人生》
学习	agent学习自己的行为	强化学习	克服人为编写的问题	适用对抗游戏，游戏奖励易确定，学习算法可对其优化，尚未解决开放世界中的难题	《星际争霸》AlphaStar、《Dota 2》OpenAI Five
符号	保持短期和长期记忆，以感知-计划-行动的循环方式运行，动态感知环境并将其与人工编写的行动程序相匹配		强大的行为能力	agent行动空间仅限于人工程序，并没有提供一种机制激发agent寻求新的行为	第一人称射击游戏NPC、空战训练模拟飞行员、积木世界
LLM	利用大语言模型强大 prompt 能力，并对这些能力进行补充，以支持agent的长期一致性、管理动态演变的记忆能力，以及递归产生更多的行为。		强大理解和推理能力	需要借助外部工具、底层大模型能力还需进一步迭代	实验性项目偏多，如Smallville 小镇、Voyager

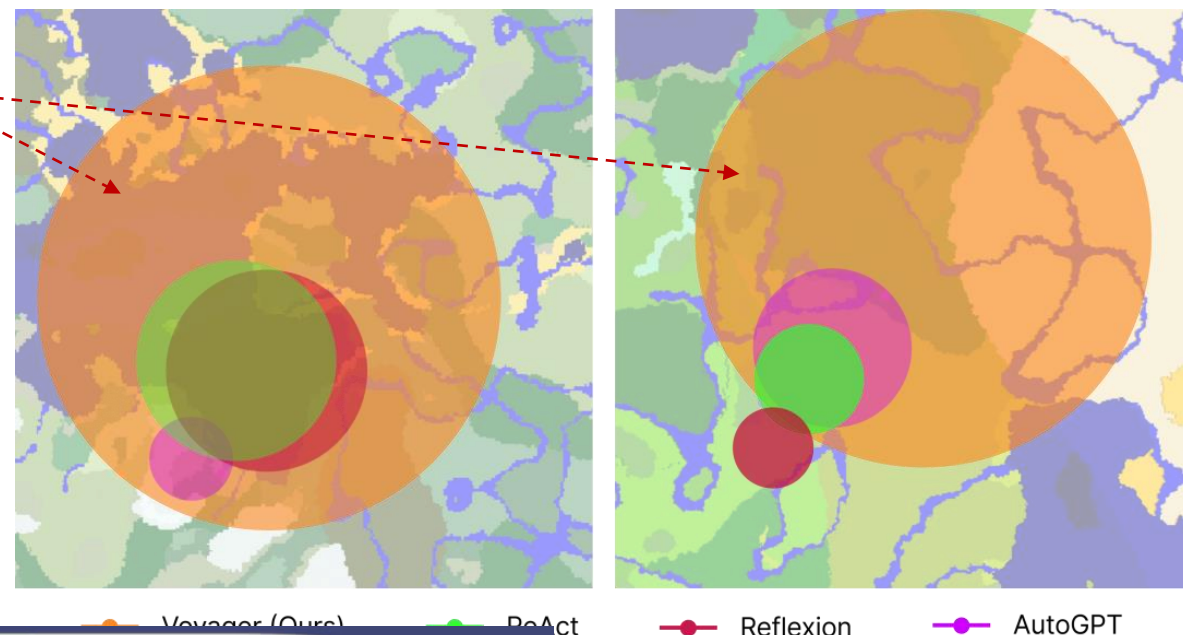
3.2.1 单智能体：游戏世界AI玩家，如Voyager

- **Voyager**：英伟达首席科学家Jim Fan等人把GPT4接入《我的世界》提出的一个全新AI智能体。
- **三个关键部分组成**：最大化探索；有用于存储和检索复杂行为的可执行代码的技能库；具备一种新的结合了环境反馈、执行错误和程序改进的自我验证的迭代提示机制，能在一个新的Minecraft世界中利用学到的技能库**从头开始**解决新的任务。
- **表现出强大的情境终身学习能力**，并在玩Minecraft方面表现出非凡的熟练度。它获得的独特物品是SOTA的**3.3倍**，旅行距离是SOTA的**2.3倍**，解锁关键技术树里程碑的速度是SOTA的**15.3倍**。
- **其他类似应用**：2023年5月商汤、清华、上海人工智能实验室等机构研究人员提出的GITM框架，旨在创建游戏《我的世界》中的通用AGENT。

图：工具使用对比图



图：地图探索范围对比图



3.2.2 多智能体：Smallville 小镇，类西部世界的模拟社会

- 25 个智能体居住在名为 Smallville 的小镇，每个智能体可执行类似人类的日常行为，比如起床后做早餐然后去上班，艺术家作画，作家写作。他们可以与别人和环境交流（互相注意到彼此举动、发起对话或者问候）、反思这些观察结果（形成独特的个人观点）、制定每天的计划。他们拥有自己记忆和目标，会产生可信的个人和涌现的社交行为，而不是通过预先设计实现。
- 如：从用户指定的单一任务开始，即一个AI-Agent想要举办情人节聚会，AI-Agents们在接下来的时间里会自发的传播邀请、认识新朋友、互相约出参加聚会的日期，并协调在正确的时间一起出现在聚会上。

图：小镇整体布局与智能体的生活

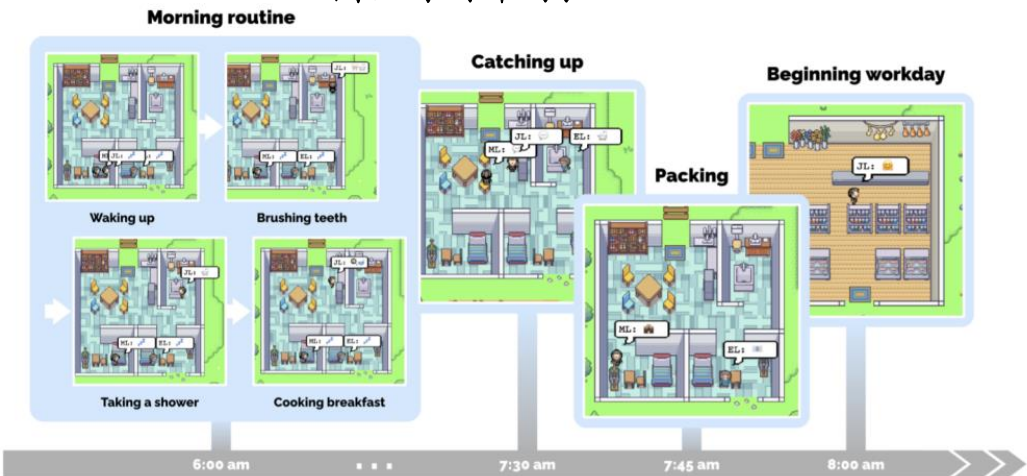


设计者创造了一个沙盒环境，里面有25个智能体。用户可以像智能体一样观察他们的行为，并与他们交流、规划时间、分享新闻、相互合作。

图：角色之间的对话

伊莎贝拉:我还在考虑我的选择，但是我一直在和山姆·摩尔讨论选举。你对他有什么看法？
汤姆:老实说，我不喜欢山姆·摩尔。我认为他与社区脱节，没有把我们的最大利益放在心上。

图：时间计划表



智能体从早上开始一天的生活：在6点起床，接着刷牙、洗澡、做饭、吃早餐，7：30与妻儿交流，7：45收拾东西，8：00上班工作。

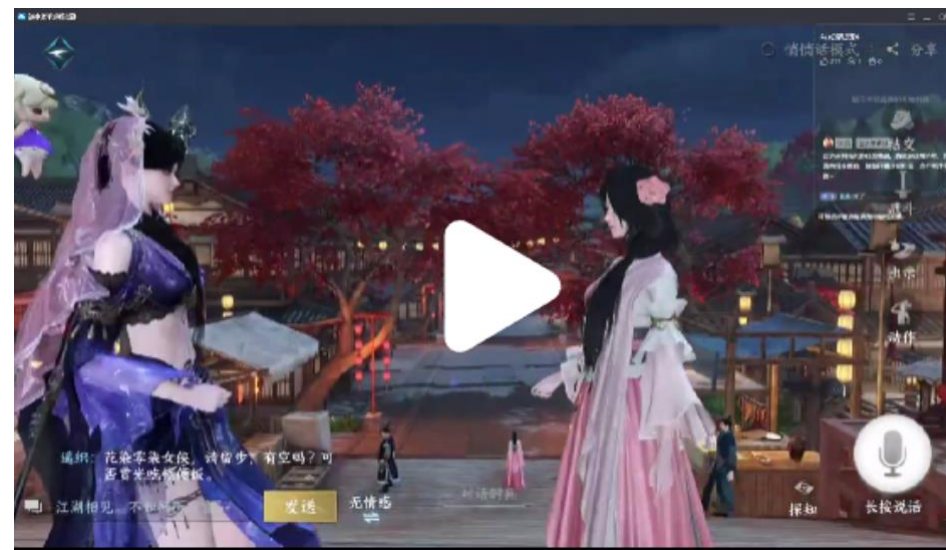
3.2.3 多智能体：网易《逆水寒》手游，AI NPC提高玩家体验

- 2023年6月30日，网易重磅MMO新游《逆水寒》手游上线，首日即登顶iOS游戏免费榜，截至7月3日凌晨，仍高居榜首；iOS游戏畅销榜排名于公测当天晚（6月30日晚）跃升至第3名，于7月1日晚跃升至第2名。
- 逆水寒手游中投放了百位AINPC，不仅能和玩家对答如流，而且拥有了记忆，大大增加了游戏乐趣，同时NPC提供的探索剧情，也让游戏内容更为丰富。通过与NPC互动，玩家可以了解这个世界，探听到很多奇闻，了解一些boss的弱点（比如通过NPC对话，了解到高衙内的打手怕水，在打斗中将其推进水中可以取得巧胜），甚至通过NPC的关系网，玩家可以巧解难题。

图：与街上的AINPC进行互动，AINPC在后续遇到玩家可以认出玩家



图：AINPC会主动找到玩家，请吃饭喝酒，展开剧情



3.2.4 多智能体：昆仑万维《Club Koala》虚拟世界更可信



产品特点——AI NPC，让虚拟世界更真实

- 使用Play for Fun的Atom AI系统，每个AI NPC都将有自己的性格特征、行为模式，会自主制定日程并执行，并会影响其他NPC的关系和发展。同时加入**记忆系统**后，AI NPC能够记住与玩家的互动，同时可以根据玩家的行为，并调整自己的行为，发展“自我意识”，进行自然而真实的动作和对话。玩家与AI NPC的互动将被NPC记住，同时NPC将分析玩家行为模式和偏好，并反映在后续与玩家的互动中，以建立起玩家与NPC之间真正的纽带。**未来玩家甚至可以定制自己的专属NPC。**

图：玩家与AI NPC服装店员工Snow的互动（Snow的回答契合她的身份设定）



玩家（Alex）：我感到孤独

AI NPC：Alex，很遗憾听到你这么
说，有什么具体的事情让你感到孤
独吗？

玩家（Alex）：我不确定，就是感
觉我生活中少了些什么

AI NPC：Alex，很遗憾听到你这么
说，这确实让人难受。

你有想过探索一些可能带来乐趣的爱好或者活动吗？我很享受学习时尚设计，它能让
我同时我也喜欢户外活动。你呢？你觉得哪些爱好和活动可能让你开心？

4、投资建议

4 投资建议

- 推荐在智能体模拟方向有直接布局的昆仑万维、中文在线、盛天网络，建议关注天地在线等。
 - 昆仑万维：自研天工大模型；AI 游戏《Club Koala》预计2023年年末正式发行。
 - 中文在线：推出Chapters、My escape等互动应用，用户为情感付费，期待后续基于丰富内容语料带来更自由更有趣的互动玩法及情感体验。
 - 盛天网络：公司正积极对接超拟人大模型开发商聆心智能，后续将人工智能科技与自有娱乐场景深度结合，双方将共同打造、探索AI+游戏社交互动新场景，给玩家带来全新体验。旗下带带电竞正在推进AI虚拟陪伴等社交应用。
 - 天地在线：与开发了“超拟人大模型”的聆心智能达成专项开发和联合运营合作。基于超拟人大模型，公司AI数字人有望更具有情绪、人格。
- 智能体模拟有望提高游戏可玩性从而升级玩家体验，游戏板块有望整体受益。我们推荐研运能力出色且产品储备充沛的厂商：恺英网络、吉比特、三七互娱、巨人网络、宝通科技等，港股的创梦天地等，建议关注神州泰岳、世纪华通等。
- IP资源在AI AGENT发展过程中亦有望发挥价值，我们建议关注：1) 艺人IP：华策影视等；2) 影视/动画IP：光线传媒、奥飞娱乐、博纳影业、上海电影、中国电影等。
- 教育、电商等自主智能体发展亦有望推动行业发展，快速布局、应用的公司将有望受益，教育推荐南方传媒、凤凰传媒、皖新传媒、新东方，建议关注世纪天鸿、好未来、高途集团、科大讯飞、佳发教育、盛通股份、传智教育、鸿合科技等；电商及企业服务领域推荐焦点科技、华凯易佰、吉宏股份。创作类自主智能体发展有助于提高内容行业生产效率，推荐动画电影龙头光线传媒。

5、风险提示

5. 风险提示

- 产业进展不及预期:AI技术发展或产品落地不及预期可能会导致企业AI业务进展不及预期。
- 监管风险: 监管收紧可能会影响行业发展和企业业务进展。
- 市场竞争加剧风险: 个股所在行业竞争加剧可能对相关公司业务产生不利影响。

东吴证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本研究报告仅供东吴证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议，本公司及作者不对任何人因使用本报告中的内容所导致的任何后果负任何责任。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。

在法律许可的情况下，东吴证券及其所属关联机构可能会持有报告中提到的公司所发行的证券并进行交易，还可能为这些公司提供投资银行服务或其他服务。

市场有风险，投资需谨慎。本报告是基于本公司分析师认为可靠且已公开的信息，本公司力求但不保证这些信息的准确性和完整性，也不保证文中观点或陈述不会发生任何变更，在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。

本报告的版权归本公司所有，未经书面许可，任何机构和个人不得以任何形式翻版、复制和发布。经授权刊载、转发本报告或者摘要的，应当注明出处为东吴证券研究所，并注明本报告发布人和发布日期，提示使用本报告的风险，且不得对本报告进行有悖原意的引用、删节和修改。未经授权或未按要求刊载、转发本报告的，应当承担相应的法律责任。本公司将保留向其追究法律责任的权利。

东吴证券投资评级标准

投资评级基于分析师对报告发布日后6至12个月内行业或公司回报潜力相对基准表现的预期（A股市场基准为沪深300指数，香港市场基准为恒生指数，美国市场基准为标普500指数，新三板基准指数为三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）），具体如下：

公司投资评级：

买入：预期未来6个月个股涨跌幅相对基准在15%以上；

增持：预期未来6个月个股涨跌幅相对基准介于5%与15%之间；

中性：预期未来6个月个股涨跌幅相对基准介于-5%与5%之间；

减持：预期未来6个月个股涨跌幅相对基准介于-15%与-5%之间；

卖出：预期未来6个月个股涨跌幅相对基准在-15%以下。

行业投资评级：

增持：预期未来6个月内，行业指数相对强于基准5%以上；

中性：预期未来6个月内，行业指数相对基准-5%与5%；

减持：预期未来6个月内，行业指数相对弱于基准5%以上。

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议。投资者买入或者卖出证券的决定应当充分考虑自身特定状况，如具体投资目的、财务状况以及特定需求等，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。

东吴证券研究所
苏州工业园区星阳街5号
邮政编码：215021
传真：(0512) 62938527
公司网址：<http://www.dwzq.com.cn>

东吴证券 财富家园