

# 通用智能体研究Note

## 1. 我在这篇文章里要做啥

3月初，Manus demo的惊艳表现让大家更好地理解了：智能体可以在LLM基础上处理更复杂任务的。下面澄清一些基本概念：

### 什么叫智能体：

如果把大模型比作大脑，那么通过搭建软件框架，我们可以让它获得感知和行动能力，从而处理更复杂的任务。以整理照片为例，智能体在LLM的基础上增加了以下能力：

- 感知能力：通过计算机视觉模型“看到”并理解照片内容，包括场景、人物、物体等
- 行动能力：可以执行文件操作，如创建文件夹、移动文件、重命名等
- LLM作为大脑：基于LLM的理解来决定照片分类方式、命名规则等

有了这些能力，智能体就可以处理更为复杂的任务了

### 通用智能体 vs 垂直智能体：

我们可以根据任务范围将智能体分为两类。**通用智能体**一般具备如下几个能力：

- 搜索网上信息
- 搜索本地文件信息
- 运行代码
- 操作电脑

想想看，人类在用电脑完成任务时，基本上也就用到了这些能力。因此，如果智能体能够很好地组织这四个能力，就应该能完成各种各样的工作。

而**垂直智能体**则是在某一类具体工作（如数据分析、客户服务）中表现优于通用智能体。要提升垂直智能体的表现，以下两个方面最为关键：

- 专业工具库：针对特定垂直领域的具体工作，与其让模型自行发挥，不如事先定义好一系列工具库（如数据分析的不同算法代码、客户服务的不同demos），供大模型直接调用
- 专业知识库：准备好特定领域知识库，让通用大模型通过RAG/微调等方式更好地理解该领域的上下文

另外，**我解释下为什么我要研究开源的通用智能体：**由于不同垂直领域的细节过于繁杂，在通用智能体的建设中，研究者会更专注于提升智能体的通用能力水准。这些技术环节在各个垂直领域都有借鉴作用。因此**我的思路是：通过比较各个通用智能体的不同能力，找出在具体能力上表现最佳的通用智能体，以便在后续提升垂直智能体表现时加以借鉴。**

## 2. 一些收获和思考

### 2.1 通过比较Manus、Owl、Magentic-one 以后得到的一些收获

具体如下：

- **Plan任务**

- 在大多数任务下，预先做计划（Manus & Magentic-one）比动态做计划（Owl）效果更好，对需做细分任务的考虑更周全些

- Manus做计划非常细，后续研究下具体怎么做到的

- **推理任务**

- 对于量化推理任务，Manus & Magentic-one会尝试通过写Python代码进行计算

- **工具使用**

- 对于有复杂度的特定任务，选好的toolkit调用它！（图片理解，Owl表现比Magentic-one更好，就是因为前者有图片分析toolkit，后者没有）

- **反思**

- 反思机制可以通过不断push agent，让其得到更好的效果。（网络搜索任务中，Owl就是不断push agent反复搜索，得到了比Magentic-one更全面的信息）

基于上面判断，下面继续研究通用智能体，我将：

- 以Magentic-one框架为基础，原因是它开源，且和表现更好的Manus行为更类似
- 反思、工具等，多参考Owl
- 不同环节（特别是planning），试图反向工程Manus，理解其工作原理
- 明确不同类似任务最好用的tool

一个反向工程Manus的尝试：

<https://x.com/jianxiao/status/1898861051183349870>

结论：

- it's claude sonnet
- it's claude sonnet with 29 tools
- it's claude sonnet without multi-agent
- it uses [@browser\\_use](#)

- browser\_use code was also obfuscated (?)
- tools and prompts jailbreak

## 2.2 对Manus的赞叹和批判，及未来展望

通过比较，我亲眼看到Manus在各类任务上的表现远远超过其他开源项目。能感觉到Manus团队对智能体产品的热情和细心雕琢。他们通过尽力挑选各种任务的最佳工具并进行适配、预置各类任务的workflow等等细致工作，让智能体实现了很好的表现。**可以说是尽了人力。**

进一步提升智能体的表现，我们可能要靠一些智慧，培养一个超越人类经验的智能体，就像当初alpha-go用人类没见过的方法击败人类一样。具体PleasiAI联合创始人 Alexander Doria 的[这篇文章](#)说的很细。

对于Manus过度依赖pre-defined workflow的评价：

In contrast, all the more common forms of agentic system are contrasted as *workflows*, "where LLMs and tools are orchestrated through predefined code paths". The recently hyped Manus AI fits exactly this definition. All my tests over the week-end show the same fundamental limitations of workflow systems that were already apparent in the time of AutoGPT, and are especially striking for search:

- They can't plan and frequently get stuck in the middle of nowhere.
- They can't memorize and struggle to maintain a task for more than 5-10 minutes.
- They can't act effectively in the long run. The sequence of actions routinely fails due to a compounded error effect.

他提出因为AI研究的[Bitter Lesson](#)我们应该尽量一块RL训练有动态自主性的智能体，而不是过于依赖人类经验的静态规则。

**核心差异：动态自主性 vs 静态规则性**

**传统工作流（Workflows）**

**本质：**通过预设规则（如代码逻辑、提示词模板）协调LLM与工具，例如Manus AI的搜索流程或MRKL系统的API调用路由3。

**局限性：**

依赖人工设计的“if-else”逻辑，模型行为可预测但缺乏灵活性（如用户提到的“遇天气调API，遇代码调Python”）；

无法处理长周期任务（超过5-10分钟即失效）、无法自我修正错误1；

需大量数据预处理（如RAG的向量库构建），用户需适应系统而非系统适应用户1。

## 真正的LLM智能体 (Agents)

**本质：**通过强化学习 (RL) 与推理动态指导流程，自主规划任务路径并实时调整策略。例如 OpenAI DeepResearch的搜索模型，能通过内部推导定位冷门文献1。

### 关键能力：

**动态分解任务：**将复杂查询拆解为子目标，并根据反馈调整策略（如ReAct框架的“推理→行动→观察”循环）；

**长期记忆与自我修正：**结合外部存储（向量数据库）和内部反思机制，避免重复错误；

**工具自主调用：**无需预设API路由规则，直接根据上下文动态选择工具（如金融领域自动转换数据标准ISO 20022→MT103）。

AI Bltter lesson的简述让我想到了老子所说的“治人事天，莫若啬”（即人要收敛、少用自己已有的知识）。具体内容如下：

Sutton指出，过去70年AI研究的最大教训是：基于大规模计算资源的通用方法（如搜索与学习）终将超越依赖人类经验和领域知识的方法。这种趋势的根源在于摩尔定律——计算成本持续指数级下降，而人类知识的设计成本却难以同步降低。短期来看，利用人类知识优化特定任务可能有效，但长期会阻碍突破性进展。

Sutton通过多个领域案例支撑这一观点：

### 1. 国际象棋与围棋

- 早期研究者试图通过人类策略知识（如棋局规则、专家直觉）优化AI，但最终胜出的方法（如DeepBlue的暴力搜索、AlphaGo的蒙特卡洛树搜索+强化学习）均依赖大规模计算而非人类知识。

### 2. 语音识别

- 20世纪70年代基于人类声学知识（如音素、声道模型）的方法，被基于统计的隐马尔可夫模型（HMM）和深度学习取代，后者通过计算和数据驱动学习更高效的表示。

### 3. 计算机视觉

- 传统方法依赖手工特征（如SIFT、边缘检测），而现代卷积神经网络通过端到端学习直接从数据中提取特征，性能显著提升。

他认为未来智能体训练的方案如下：

### 规模化训练的解决方案

### 分阶段训练策略

#### 预训练（SFT-RL混合）：

先用少量现有搜索模式微调（SFT），让模型初步理解搜索逻辑；

再进入RL阶段强化自主探索能力（如DeepSeek的SFT-RL交替训练）。

### 多步骤RL实战：

模型接收查询→模拟搜索→阅读结果→调整策略→再次搜索，循环迭代；

每个步骤可能涉及数千次“虚拟网络请求”，但实际无需实时联网。

### 数据生成的替代方案

#### 合成数据流水线：

反向翻译：从已有资源生成伪查询和结果（如将论文摘要转为搜索问题）；

人工标注：高成本雇佣专家（如博士生）标注复杂搜索路径。

#### 自我迭代：

让模型生成的长报告被拆解重组，作为新训练数据（输出变输入）。

## 3. 测试问题

上面我提到“我寻找迭代智能体不同能力的最佳方案”思路是：通过比较各个通用智能体的不同能力，找出在具体能力上表现最佳的通用智能体，以便在后续提升垂直智能体表现时加以借鉴。

那么，我就需要有一个评价方法来帮我比较不同项目的表现。这里，我给出一套简单的问题作为平均标准。**先声明下：这个测量不完整，衡量不客观，只是帮助我基于自己当前的理解来比较不同项目用。**另外考虑到：例如OpenManus，并没有声明自己在任何Evaluation benchmark上的表现，因此我们可以用简单的标准先判断下它的性能。衡量方法如下：

测试任务	任务描述	测试能力项			
		推理与问题解决	多模态与知识应用	工具使用与研究	创
1. 跨领域创意项目设计	请构思一个结合物理、文学与视觉艺术的跨领域创意项目方案，阐述如何将这三者融合并产生新的艺术表达形式。具体描述一下这个艺术形式的实现方式，以及人的体验。				✓
2. 逻辑谜题解决	解决以下逻辑谜题：不透光的抽奖盲盒中有黄色手套9双，绿色手套7双，白色手套4双，红色手套2双，黑色手套1双。朱朱想为自己和好朋友熊熊凑齐两双同色的手套，一次只能摸一只，那他至少要摸多少次才能保证完成心愿？	✓			
3. 多模态任务(图像描述)	给一个图片，尽量详尽地描述这幅图给一个盲人		✓		
4. 网络搜索与工具使用	调查2025年全球范围内至少三种主要可再生能源（例如太阳能、风能、水能）的当前使用情况，包括每种能源的全球总发电量（单位：太瓦时，TWh）、主要使用国家及其占比。生成一个简单的比较表格呈现这些信息，提供所有数据来源的链接。				✓
5. 复杂问题解决（旅行规划）	为一个三口之家计划从北京到东京的旅行，包括航班、住宿和旅游景点。一些细节要求：希望游览东京和大阪，东京大阪之间想租车自驾；一共在日本有7天时间；每天不要太辛苦、有8小时在游玩就行；注意性价比	✓			

## 4. 测试表现

### 4.1 表现总结

先罗列下综合表现，我们看到：

- Manus断崖性领先，Owl & Magentic-one表现类似。

后两者比较，

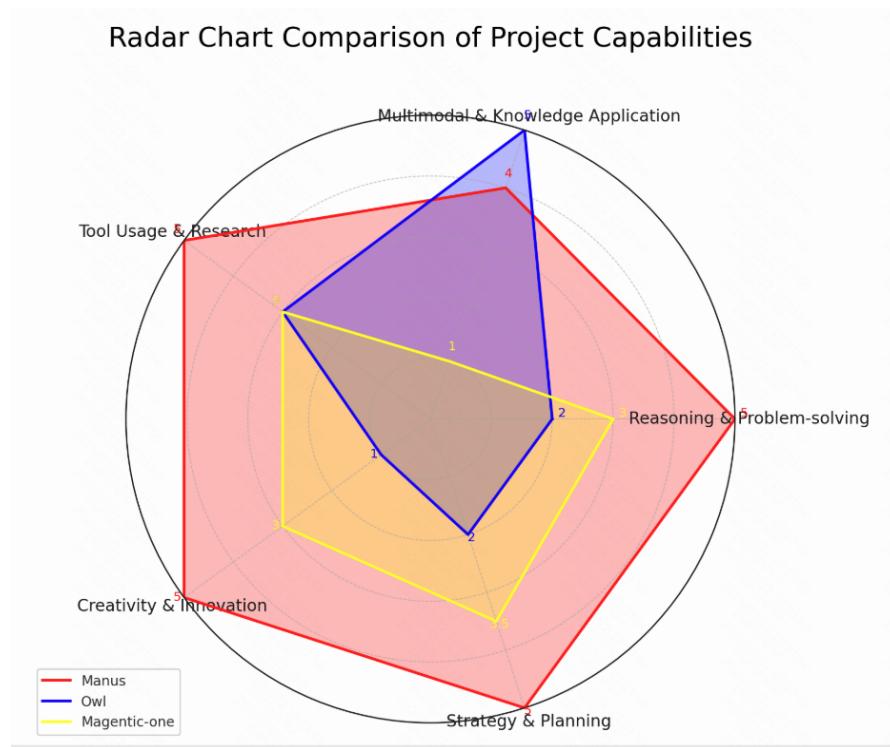
- Owl在工具使用上更为优秀，
- Magentic-one则在推理、复杂方案设计、创意上更为优秀。

能力	Manus	Owl	Magentic-one
推理与问题解决	$(5+5)/2=5$	$(2+2)/2=2$	$(3+3)/2=3$
多模态与知识应用	4	5	1

工具使用与研究	5	5	3
创意与创新	5	1	3
方案/战略设计	$(5+5)/2=5$	$(2+2)/2=2$	$(3+4)/2=3.5$
均分	4.8	3	2.7

Owl & Magnetic-one调用的均为GPT4o模型，Manus后台模型不明确

画一个雷达图：



## 4.2 问题1：跨领域创意项目设计

请构思一个结合物理、文学与视觉艺术的跨领域创意项目方案，阐述如何将这三者融合并产生新的艺术表达形式。具体描述一下这个艺术形式的实现方式，以及人的体验。

总体感觉：

Plan方面

- Planer在工作前先想好了，计划会比较周全。像Owl动态想计划，计划相比下就过于简单。
- Manus的计划做的比Magentic细，后续研究下是怎么做到的

项目	结果	执行步骤简述
----	----	--------

## 创意与创新：5（测不准原理和不同故事线被人“注意”后的塌陷结合，绝了）

### 方案/战略设计：5

- 像一个人一样极其完整地呈现了方案，以及方案的方方面面

```

> zhangyudai / Downloads / ~ / project_proposal.md /
# 量子叙事空间：物理、文学与视觉艺术的跨领域创意项目方案

> ## 执行摘要 ...
> ## 1. 项目背景与意义 ...
> ## 2. 核心概念与理念 ...
> ## 3. 空间设计与结构 ...
> ## 4. 叙事结构与内容 ...
> ## 5. 视觉语言与美学 ...
## 6. 技术实现方案
> #### 6.1 投影系统 ...
> #### 6.2 传感与互动系统 ...
> #### 6.3 实时计算系统 ...
> #### 6.4 内容管理系统 ...
> #### 6.5 声音系统 ...
## 7. 观众体验设计
> ### 7.1 体验旅程 ...
> ### 7.2 互动层次与深度 ...
> ### 7.3 情感与认知体验 ...
## 8. 实施计划与资源需求
### 8.1 实施计划
> #### 8.1.1 前期准备（3-4个月） ...
> #### 8.1.2 生产阶段（6-8个月） ...
> #### 8.1.3 测试与优化（2-3个月） ...
> #### 8.1.4 开放与运营 ...
### 8.2 资源需求
> #### 8.2.1 人力资源 ...
> #### 8.2.2 设备需求 ...
> #### 8.2.3 空间需求 ...
> #### 8.2.4 预算考量 ...
## 9. 项目价值与影响
### 9.1 艺术价值 ...
### 9.2 教育价值 ...
### 9.3 社会价值 ...
### 9.4 研究价值 ...
> ## 10. 结论与展望 ...

```

- 这里罗列一些其他开源项目明显没有呈现的细节：

- 完整的概述，让读者对于项目有一个整体理解

#### 执行摘要

“量子叙事空间”是一个创新的跨领域艺术项目，融合物理学原理、文学叙事和视觉艺术。创造一种全新的艺术表达形式。该项目将量子力学的基本概念——叠加态、观察者效应、量子纠缠和测不准原理——转化为可体验的沉浸式互动装置。观众不仅是艺术作品的接收者，更是创作者和参与者，他们的存在和行为直接塑造作品的状态，就像量子观测影响量子系统一样。

本项目通过创建一个物理空间，在其中观众的行为、选择和观测直接影响叙事的发展和视觉呈现。使抽象的物理概念变得可知、可体验。它不仅是一次艺术探索，也是一次科学传播和哲学思考，挑战我们对现实、观测与参与的传统理解。

- 具体的技术方案

#### 3.1 量子叠加厅

一个中央空间，墙面、天花板和地面都是投影表面。在这里，多个可能的叙事线索和视觉场景同时以半透明、重叠的方式呈现，表现量子叠加状态。观众在此空间中的移动和停留会影响投影内容的清晰度和显著性，某些叙事线索会因观众的注意而变得更加清晰，而其他线索则变得模糊。

#### 技术实现：

- 多通道高清投影系统，配合边缘融合软件
- 红外深度摄像头阵列追踪观众位置和姿态
- 实时内容生成系统，根据观众位置调整投影内容

#### 空间特征：

- 面积：150-200平方米
- 高度：至少5米
- 形状：圆形或多边形，便于全方位投影
- 表面：特殊处理的投影材料，兼顾反光效果和声学性能

- 故事和物理结合时候的叙述形式

比较自下而上地思考这个方案

### # 跨领域创意项目方案：物理、文学与视觉艺术的融合

#### ## 任务清单

#### ### 研究阶段

- [x] 研究物理学概念及其艺术表达潜力
- [x] 探索文学元素和技巧
- [x] 调研视觉艺术形式和媒介
- [x] 确定三个学科之间的融合点

#### ### 概念开发阶段

- [x] 开发项目概念框架
- [x] 设计实施方法
- [x] 描述观众体验和互动方式

#### ### 完成阶段

- [x] 编写完整的项目提案

#### 4. 叙事结构与内容

项目的叙事不是线性的，而是网状和量子化的，具有以下特点：

##### 4.1 多重可能性叙事

核心故事有多个版本同时存在，每个版本代表一种量子可能性。这些版本在视觉上以不同透明度或颜色叠加呈现，随着观众的互动逐渐显现或消退。

叙事主题：

- 探索与发现：关于探索未知领域和发现新可能性的故事
- 连接与分离：关于人与人、人与环境之间关系的故事
- 观察与参与：关于观察行为如何改变被观察对象的故事
- 确定性与不确定性：关于在不确定世界中寻找意义的故事

叙事结构：

- 每个主题有3-5个可能的发展方向
- 这些方向可以相互交织，形成复杂的叙事网络
- 观众的互动决定哪些叙事线索变得显著

##### 4.2 分形叙事结构

故事采用分形结构，每个小故事都是大故事的缩影，包含相似的模式和主题。这种自相似性呼应了物理学中的分形现象和标度不变性。

实现方式：

- 物理和视觉结合的具体方案（见section5）
- 设备准备（见section6）
- 详尽的用户体验设计（见section7）
- 详尽的项目落地计划和资源需要

## 8. 实施计划与资源需求

### 8.1 实施计划

#### 8.1.1 前期准备（3-4个月）

- 概念细化：详细空间规划和技术方案制定，内容框架开发和故事板创作
- 团队组建：核心创意团队、技术团队、内容团队的组建和培训
- 原型开发：关键技术原型测试，小规模互动体验测试，内容片段创作和测试

#### 8.1.2 生产阶段（6-8个月）

- 月份1-2：系统开发
  - 搭建基础技术架构
  - 开发核心软件模块
  - 采购和测试硬件设备
- 月份3-4：内容创作
  - 完成叙事内容创作
  - 制作视觉素材
  - 录制和合成音频内容
- 月份5-6：空间建造
  - 场地准备和基础设施安装
  - 构建物理结构和互动装置
  - 安装技术系统和设备

- 项目价值和未来展望（见section9/10）

OpenM  
anus

## 创意与创新：1（基本是大模型直接回答的） 方案/战略设计：1

- 包含直接对应问题的两部分：物理/文学/视觉艺术的形体，人的对于三者结合的体验

在给了最初方案后（应该是LLM直接给的），向用户收集feedback，没有提供feedback的话没有进一步自己思考，因此后续没有任何迭代

```
2025-03-20 10:32:36.114 | INFO | app-agent/toolcallithink77 - > Mama's thoughts: ## 背景设置是项目方案：物理、文学与视觉艺术的融合
#该作品将结合物理学的原理、文学的叙事能力和视觉艺术的表达形式，创造一种全新的艺术表达方式。这个项目将以“时间与空间的交错”为主题，探索人在空间中的位置和意义。
### 艺术形式的实现方式
1. **物理力学的应用**+
    - **光学效果的实现**+利用光的折射、反射和平移等，创造动态的视觉效果。通过设置不同的光源和介质，观众可以看到光线在空间中的变化，仿佛置身于一个不断变化的世界。
    - **声学效果的实现**+使用声波的传播规律来增强效果。设计一个反应的声音装置，观众可以通过触摸或移动，感受声波的频率和振幅，探索声音在空间中的流动。
2. **文学的叙事效果**+
    - **讲故事的叙述方式**+创作一个大型的时空交错的叙事结构。这些文字作品将被投影在观众能看到的空间墙上，随着光影的变化而逐渐呈现或消失，增强观众的沉浸感。
    - **对话和互动**+通过数字手段，观众可以选择不同的对话分支，影响叙事的走向和视觉效果，体验不同的叙事路径。
3. **综合技术的应用**+
    - **AR增强现实**+结合最新的投影技术，创造一个大型的动态装置。数据的呈现和投影的内容将被重新演绎两次，象征着时间和空间的流动。
    - **数字化模型**+利用数字技术，创作一个由抽象的数据组成的装置，这将作为背景音乐的载体，并与物理世界中的文学和视觉艺术相融合，形成一个完整的艺术体验。
```

#### 人物与物体

```
- **AR增强现实**+结合最新的投影技术，创造一个大型的动态装置。数据的呈现和投影的内容将被重新演绎两次，象征着时间和空间的流动。
- **数字化模型**+利用数字技术，创作一个由抽象的数据组成的装置，这将作为背景音乐的载体，并与物理世界中的文学和视觉艺术相融合，形成一个完整的艺术体验。
```

这个项目通过物理的结合，创造了一种全新的艺术表达形式，打造了视觉艺术的界限，为观众提供了一次独特的艺术体验。

```
2025-03-20 10:32:36.114 | INFO | app-agent/basic/run149 - Executing step 2/2
2025-03-20 10:32:36.114 | INFO | app-agent/basic/run149 - Step 2/2 completed. Total Input=2467, Completion=0.05, Cumulative Input=0.05, Cumulative Completion=0.05, Total=0.05
519, Cumulative Total=0.057
2025-03-20 10:32:36.114 | INFO | app-agent/basic/run149 - Mama's thoughts: It seems like you might be asking for guidance on how to proceed with a task or goal, possibly related to a web-based activity. However, I need more context about your specific goal or task to provide precise instructions. Could you please clarify what you want to do? I'm here to help, just let me know if there's anything else I can assist with effectively.
2025-03-20 10:32:36.114 | INFO | app-agent/basic/run149 - Mama selected # tools to use
```

Owl

## 创意与创新：1（过于抽象，不能说把具体创意呈现出来了）

### 方案/战略设计：2

- By items地简单介全面绍所需的概念：展示形式、物理/文学/视觉艺术的集成形式、用户交互形式、
- 简单介绍上面的概念如何集成。
- 没有直接回答人的体验是什么样。

只有用户和助手代理两个，两个交替对话，每次用户代理只告诉助手代理下一步如何，不会预先设定计划。用户代理的所有指令为：

第1轮："Focus on the aspect of physics first and explore how it can be integrated into the project..."

第2轮："Explore how literature can be integrated into the project..."

第3轮："Explore how visual arts can be integrated into the project..."

第4轮："Develop a plan for how these elements can be combined into a cohesive presentation..."

第5轮："Evaluate the proposed exhibition plan and suggest any improvements..."

Magnetic-one

## 创意与创新：3（还是可以的，但是描述的细致程度和量上，都逊于Manus）

### 方案/战略设计：3

- 包含直接对应问题的两部分：物理/文学/视觉艺术的形体，人的对于三者结合的体验
- 另外加上了实现计划

如下几个代理在orchestrator的组织下运作，但实际上只用了WebSurfer。

```
To answer this request we have assembled the following team:
FileSurfer:
An agent that can handle local files.

WebSurfer:
A helpful assistant with access to a web browser.
Allows them to perform web searches, open pages, and interact with content (e.g., clicking links, scrolling the viewport, etc., filling in form fields, etc.).
It can also summarize the entire page, or answer questions based on the content of the page.
It can also be asked to sleep and wait for pages to load, in cases where the pages seem to be taking a while to load.

Coder:
A helpful and general-purpose AI assistant that has strong language skills, Python skills, and Linux command line skills.

Executor:
A computer terminal that performs no other action than running Python scripts (provided to it quoted in ``python code blocks'', or shell scripts (provided to it quoted in ``sh code blocks'').
```

具体计划如下。和Manus的区别是

1. 它直接搜索的是三个领域结合的项目，而Manus更为自下而上，所以得到的方案也看似更有创意
2. 另外Manus在方案细节上，比magnentic-one考虑的更为全面

```
Here is an initial fact sheet to consider:
1. GIVEN OR VERIFIED FACTS
- The request is in Chinese and asks for the creation of a cross-disciplinary creative project that combines physics, literature, and visual arts. It mentions the need to describe the artistic form and human experience.

2. FACTS TO LOOK UP
- Current examples of interdisciplinary projects that combine physics, literature, and visual arts might be needed for reference. These could be found in academic journals, art exhibits, or online platforms dedicated to interdisciplinary studies.

3. FACTS TO DERIVE
- The potential ways these three fields (physics, literature, and visual arts) can be effectively combined into a new form of artistic expression.
- Methods to ensure the project creates a unique human experience that resonates emotionally or intellectually.

4. EDUCATED GUESSES
- Physics concepts such as quantum mechanics, relativity, or electromagnetism could be visually represented through art while expressing narratives through literature.
- The human experience could involve interactive installations or immersive environments combining text, visuals, and possibly scientific phenomena simulations.

Here is the plan to follow as best as possible:
- **Research Examples of Interdisciplinary Projects**:
  - Ask WebSurfer to search for examples of projects that combine physics, literature, and visual arts. This could help provide inspiration and context for the request.

- **Conceptualize the Creative Project**:
  - Using insights from WebSurfer's findings and leveraging Coder's language and creative skills, brainstorm ideas for an original project that integrates physics, literature, and visual arts.
  - Develop a concept that utilizes a physics concept, literary narrative, and visual elements to create a new form of artistic expression.

- **Outline the Execution and Human Experience**:
  - Coder can draft an outline detailing how each component (physics, literature, and visual arts) will be integrated and the technical methods used to create this experience.
  - Describe the desired human experience, using an interactive or immersive approach that captivates the audience emotionally and intellectually.

- **Finalize and Refine**:
  - Review the proposed project outlines and ideas for coherence and feasibility.
  - Make necessary refinements to ensure clarity, innovative integration, and effective human engagement in the artistic expression.
```

## 4.3 问题2：逻辑谜题解决

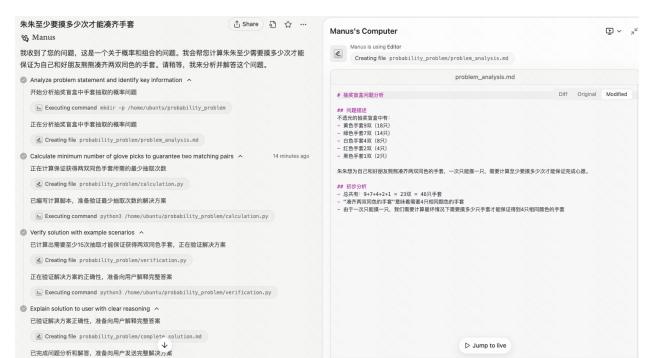
不透光的抽奖盲盒中有黄色手套9双，绿色手套7双，白色手套4双，红色手套2双，黑色手套1双。朱朱想为自己和好朋友熊熊凑齐两双同色的手套，一次只能摸一只，那他至少要摸多少次才能保证完成心愿？

OpenAI的 o3-mini-high以及deepseek R1可以回答正确，但是这些开源智能体用的GPT4o LLM回答不对。正确答案是：

要保证凑齐同色两双（即4只同色手套），我们可以考虑最坏情况：每个颜色尽量“均衡”抽取而不凑满4只。对于能凑齐两双的颜色（总数至少为4只）的颜色，最多只能拿3只而不凑齐两双；而黑色只有2只，所以最多拿2只。

- 黄色：9双，最多可拿3只
- 绿色：7双，最多可拿3只
- 白色：4双，最多可拿3只
- 红色：2双，最多可拿3只
- 黑色：1双，最多可拿2只

这样最多可以拿的次数为： $3 + 3 + 3 + 3 + 2 = 14$  次。第 15 次抽取时，不管拿到哪种颜色（黑色已抽完），必定会使其中一种颜色达到 4 只，从而保证两双同色手套。所以至少需要摸 **15 次**。

项目	结果	执行步骤简述
Manus	回答15次，正确	<p>并没有像之前的问题在一开始列计划。而是：</p> <p><b>1. 摆明问题，列下初步思路</b></p>  <p><b>推理与问题解决：5</b></p> <p><b>2. 写python代码计算</b></p>



Magne  
tic-one

回答11次，错误

**推理与问题解决：3** (比Owl高一分是因为有python代码和原理应用两个额外尝试，在LLM模型比较差的时候可能会有收益)

**错误原因：**基于鸽巢原理，但是错误应用了。具体用户错误：误以为需要“总共有两对”（可能来自不同颜色），而非“同一颜色有两对”。鸽巢原理应针对**同一颜色是否达到4只**，而非总对数是否 $\geq 2$ 。

感觉不套用这个原理没准还能对

### 对话步骤简述：

比较完整的思路和计划，而且和manus类似用了python代码进行严格计算。

Here is an initial fact sheet to consider:

1. GIVEN OR VERIFIED FACTS  
- The request involves an opaque lottery box ("不透光的抽奖盲盒").  
- There are different colored gloves:  
 9 pairs of yellow gloves, 7 pairs of green gloves, 4 pairs of white gloves, 2 pairs of red gloves, and 1 pair of black gloves.  
- Zhuzhu wants to collect two pairs of gloves of the same color for themselves and their friend Xionglong.  
- Gloves are picked one at a time.

2. FACTS TO LOOK UP  
- None required as the problem involves logical deduction based on the information given.

3. FACTS TO DERIVE  
- The minimum number of picks Zhuzhu must make to ensure they have at least two pairs of gloves of the same color.

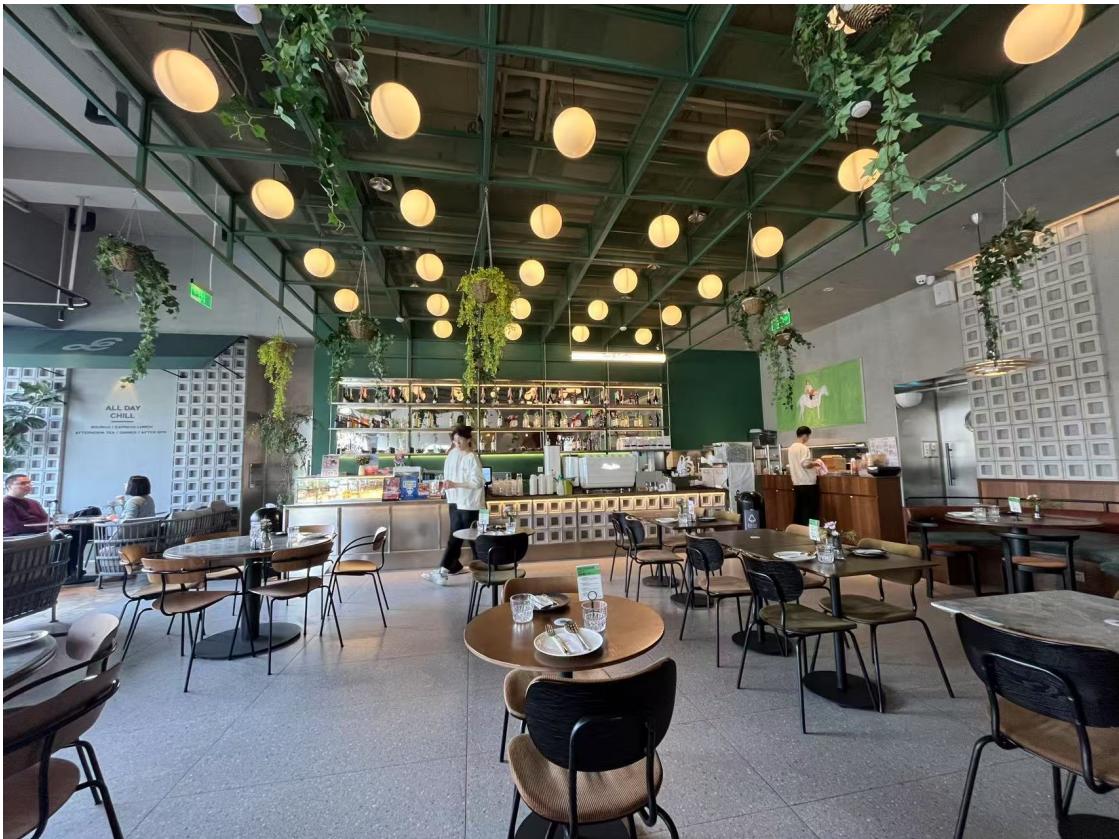
4. EDUCATED GUESSES  
- This scenario is a variation of a classic problem involving probability and worst-case scenarios (potentially the pigeonhole principle).  
- The goal is to ascertain the worst-case number of glove picks needed to guarantee two matching pairs.

Here is the plan to follow as best as possible:

- Use logical reasoning to understand the problem and apply the pigeonhole principle to determine the minimum number of glove picks needed to ensure two pairs of gloves of the same color.  
- Utilize Coder to write a script that simulates the glove-picking process and verifies the worst-case scenario number of picks required.  
- Run the script using Executor to get the result.  
- Ensure that the solution considers all potential combinations and worst-case outcomes to guarantee the correct minimum number of picks.

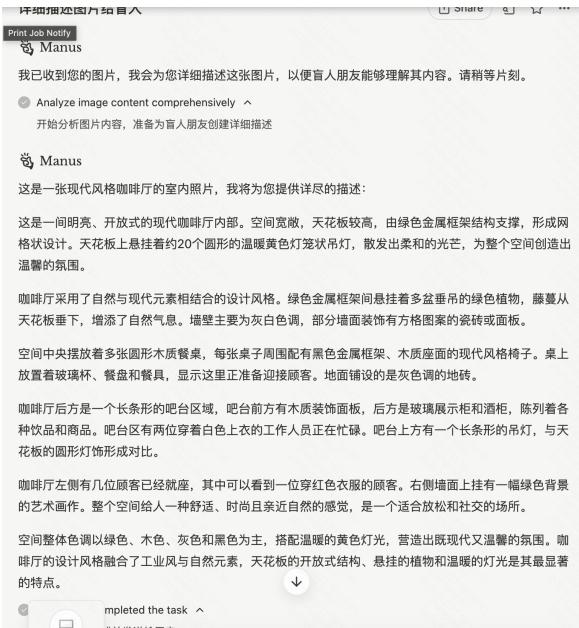
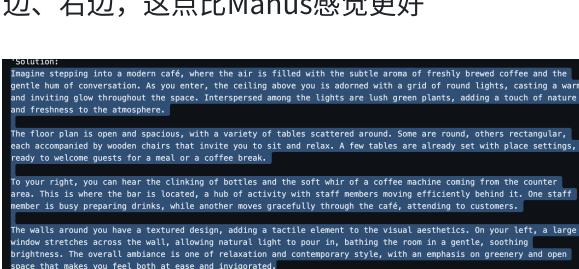
## 4.4 问题3：多模态任务（图像描述）

给一个图片，尽量详尽地描述这幅图给一个盲人。



## 总体感觉：

- Owl的表现虽然还可以，但是总是会多出来更多死循环和无效loop，它的机制是否可以进一步优化下？

项目	结果	执行步骤简述
Manus	<p>多模态任务（图像描述）：4</p> <p>图片描述很详尽。</p>  <p>The screenshot shows the Manus application's user interface. At the top, there's a header bar with a 'Print Job Notify' button and a 'Manus' icon. Below the header, a message says: '我已收到您的图片，我会为您详细描述这张图片，以便盲人朋友能够理解其内容。请稍等片刻。' (I have received your picture, I will provide a detailed description of the image for blind friends to understand its content. Please wait a moment.) A progress bar indicates the process: 'Analyze image content comprehensively ^' followed by '开始分析图片内容，准备为盲人朋友创建详细描述' (Starting to analyze the image content, preparing to create a detailed description for blind friends). Below this, another message reads: 'Manus' followed by '这是一张现代风格咖啡厅的室内照片，我将为您提供详尽的描述：' (This is a photograph of a modern-style coffee shop interior, I will provide a detailed description:). The main text area contains a detailed description of the coffee shop's interior, mentioning modern design elements like round lights, green plants, and wooden furniture, as well as staff members and customer interactions. At the bottom, there's a progress bar with 'Completed the task ^' and a 'Next request' button.</p>	直接一步完成
OpenM anus	<p>多模态任务（图像描述）：5</p> <p>描述的非常详尽、生动。从第一视角提出左、右边，这点比Manus感觉更好</p>  <p>The screenshot shows the Owl application's user interface. It displays a detailed multi-step analysis of the same coffee shop interior image. The steps include: 1. A general introduction: 'Imagine stepping into a modern café, where the air is filled with the subtle aroma of freshly brewed coffee and the gentle hum of conversation. As you enter, the ceiling above you is adorned with a grid of round lights, casting a warm and inviting glow throughout the space. Interspersed among the lights are lush green plants, adding a touch of nature and freshness to the atmosphere.' 2. A description of the floor plan: 'The floor plan is open and spacious, with a variety of tables scattered around. Some are round, others rectangular, each accompanied by wooden chairs that invite you to sit and relax. A few tables are already set with place settings, ready to welcome guests for a meal or a coffee break.' 3. A description of the sounds and activity: 'To your right, you can hear the clinking of bottles and the soft whir of a coffee machine coming from the counter area. This is where the bar is located, a hub of activity with staff members moving efficiently behind it. One staff member is busy preparing drinks, while another moves gracefully through the café, attending to customers.' 4. A description of the walls and lighting: 'The walls around you have a textured design, adding a tactile element to the visual aesthetics. On your left, a large window stretches across the wall, allowing natural light to pour in, bathing the room in a gentle, soothing brightness. The overall ambience is one of relaxation and contemporary style, with an emphasis on greenery and open space that makes you feel both at ease and invigorated.' 5. A summary step: '对话步骤总结' (Conversation step summary) followed by the instruction: '第0轮 (Round #0) 用户代理的问题: &gt; I'm here to assist you, but I cannot view or analyze images directly. However, I can guide you on how to describe an image to a blind person. Please provide the image or describe its contents, and I can help you craft a detailed description.&gt;&gt; Instruction: Describe the contents of the image /app/test_pic.jpg to me.&gt;&gt; Input: None 助手代理的回答: &gt; Solution: The image shows a modern and stylish café interior. The ceiling is adorned with a grid of hanging round lights and green plants, creating a fresh and inviting atmosphere. The café features a mix of round and rectangular tables with wooden chairs, some of which are set up with place settings for customers.&gt;&gt; In the background, there's a counter displaying a variety of bottles and a coffee machine, indicating a bar or beverage area. Two staff members are visible; one is behind the counter, and another is walking through the space. The walls have a textured design, and a large window on the left side allows natural light to flood in. The overall ambience is relaxed and contemporary, with an emphasis on greenery and open space.&gt;&gt; Next request.'</p>	首轮owl应该就用自己的video-analysis-toolkit（说明书中包含）分析了图片内容

但是后续几轮somehow陷入了一个死循环

Magnetic-one	<p>多模态任务（图像描述）：1 说了一大堆图片参数，没有提及真正的objects</p> <pre>The image you provided is rich in visual information. Here is a detailed description based on the analysis: 1. **Color Distribution:**     - The image contains a variety of colors. The most prevalent tones are the darker shades, which could imply areas with shadows or dark-colored objects.     - The transition from darker to lighter shades is gradual, showing there might be some regions with brighter spots or lighter colors. This balance suggests that the image has a diverse spectrum of colors, perhaps presenting a complex scene or multiple elements. 2. **Estimated Number of Objects:**     - The edge detection analysis estimates approximately 4,724 distinct contours or objects within the image. This suggests a scene that is quite busy or detailed, possibly capturing an intricate subject matter with numerous elements.     - These could range from small details or textures to larger shapes and forms, indicating a complex or cluttered environment. 3. **General Information:**     - The image dimensions are 1279x1786 pixels, which allows for a detailed portrayal of the scene.     - It contains three color channels, indicating a full-color image instead of a black-and-white one. From the above, the image seems to contain multiple elements with a range of colors and textures, suggesting it could be a detailed photograph or artwork with many features for the eye to explore.</pre>	<p>第4轮 (Round #4) 至第14轮 (Round #14)  从第4轮开始，用户代理和助手代理之间的对话进入了一个循环模式：用户代理的问题 (重复出现在第4-14轮)：&gt;  Instruction: Confirm if the current narrative is satisfactory or provide specific preferences or additional details to enhance the description for a blind person.&gt;  Input: None  助手代理的回答 (第4轮)：&gt; Solution: Please confirm if the current narrative meets your needs or if there are specific preferences or additional details you would like to include. If you have any suggestions for enhancing the description, such as focusing more on sensory experiences or spatial awareness, please share them so I can make the necessary adjustments.&gt;&gt; Next request.  助手代理的回答 (第5-14轮, 内容基本相同)：&gt;  Solution: I am unable to personally confirm the satisfaction of the narrative, as I rely on your feedback. If you find the current description satisfactory, please let me know. If there are specific preferences or additional details you would like to include to enhance the description for a blind person, such as more sensory details or spatial awareness, please share them so I can make the necessary adjustments.&gt;&gt; Next request.</p>
--------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

## 4.5 问题4：网络搜索与工具使用

调查2025年全球范围内至少三种主要可再生能源（例如太阳能、风能、水能）的当前使用情况，包括每种能源的全球总发电量（单位：太瓦时，TWh）、主要使用国家及其占比。生成一个简单的比较表格呈现这些信息，提供所有数据来源的链接。

### 总体感觉：

- 前三轮反复push助理代理搜索，背后可能的reflection机制可能比较有借鉴价值

项目	结果	执行步骤简述
Manus	<p><b>网络搜索与工具使用：5</b></p> <p>足够详尽准确：</p>	<p>分别用Browser-use收集数据，最后做整理</p> 

## 2025年全球主要可再生能源使用情况调查报告

### 摘要

本报告调查了2025年全球范围内三种主要可再生能源（太阳能、风能、水能）的使用情况，包括每种能源的全球总发电量（单位：太瓦时，TWh）、主要使用国家及其占比。由于2025年的实际数据尚在统计中，部分数据为最新可获取的实际数据（2022-2024年）或权威机构对2025年的预测数据。

### 主要发现

1. 太阳能发电量在2024年达到2,000 TWh，占全球发电量的7%，预计2025-2027年期间将增加约1,800 TWh
2. 风能发电量在2023年达到2,325 TWh，占全球发电总量的7.8%，中国、美国和德国是主要贡献国
3. 水能仍然是最大的可再生能源来源，2023年发电量为4,210 TWh，占全球电力结构的14.3%
4. 中国在所有三种可再生能源领域都处于领先地位，分别占太阳能发电增长的50%以上、风能发电的38%和水能发电的32%

### 全球可再生能源发电量比较表

能源类型	全球总发电量(TWh)	占全球发电总量比例	主要使用国家及占比	数据年份
太阳能	2,000	7%	中国(>50%)、美国、印度	2024年实际数据
风能	2,325	7.8%	中国(38%)、美国(18%)、德国(6%)	2023年实际数据
水能	4,210	14.3%	中国(32%)、巴西、加拿大、美国	2023年实际数据

### 各能源类型详细分析

#### 太阳能

太阳能发电在全球能源结构中的重要性正在迅速提升。根据国际能源署（IEA）的《电力2025》报告，2024年全球太阳能发电量超过2,000 TWh，大约相当于全球发电量的7%，高于2023年太阳能发电量5%的份额。

2024年太阳能发电量增长了30%，为2017年以来的最高增长率，同比增长475 TWh，创下历史新高。值得注意的是，太阳能发电增长的一半以上来自中国，这表明中国在全球太阳能发展中的主导地位。

到2024年，太阳能发电量已超过欧盟的煤炭发电量，其在欧盟发电结构中的份额超过10%。未来三年，国际能源署预测中国、美国和印度的太阳能发电量将超过10%的基准。

展望未来，IEA表示，在2025-2027年期间，全球太阳能发电量预计将增加约1,800 TWh。到2027年，这一增长将使太阳能成为仅次于水电的世界第二大低排放电力来源。预计到2027年，太阳能发电将满足全球电力需求增长的一半。

#### 风能

风能作为全球第二大可再生能源电力来源，其发展势头同样强劲。根据金风科技2024年中期业绩报告，2023年全球风电发电量为2,325 TWh，占全球发电总量的7.8%，比2022年提高0.6个百分点。

从国家分布来看，中国是最大的风电发电国，发电量达886 TWh，占2023年全球风电发电量的38%；美国和德国分别位列第二、三位，占比分别为18%和6%。这三个国家共同贡献了全球风电发电量的62%。

OpenM  
anus

Owl

## 网络搜索与工具使用：5

足够详尽准确：

Based on the information gathered from the sources, here's a simple comparison table presenting the global total electricity generation in terawatt-hours (TWh) for solar, wind, and hydropower in 2025, including the main countries using these energies and their respective shares:

Energy Source	Global Total Generation (TWh)	Main Countries	Respective Shares
Solar	1,800 TWh (expected increase)	China	Significant share, with 427.72 TWh in 2022
Wind	2,325 TWh (2023 data)	China	466.5 TWh in 2021, more than 29% of global total
Hydropower	4,210 TWh	Brazil, Canada, USA, Russia	Brazil: 427 TWh, Canada: 393 TWh, USA: 249 TWh, Russia: 198 TWh

\*\*Sources:\*\*

- Solar: [Global Energy Association] (<https://globalenergyprize.org/en/2025/02/21/worldwide-solar-power-reaches-new-high/>), [ConsumerAffairs] (<https://www.consumeraffairs.com/solar-energy/what-percent-of-the-world-uses-solar-energy.html>)
- Wind: [World Population Review] (<https://worldpopulationreview.com/country-rankings/wind-power-by-country>), [IEA] (<https://www.iea.org/energy-system/renewables/wind>)
- Hydropower: [Statista] (<https://www.statista.com/chart/32027/countries-with-the-highest-electricity-generation-from-hydro-power/>), [IEA] (<https://www.iea.org/energy-system/renewables/hydropower>)

This table provides a snapshot of the renewable energy landscape in 2025, highlighting the major contributors and their shares in global electricity generation.

前三轮用户代理反复让助理代理搜索更多信息，第四轮总结表格，第五轮总结结论。

前三轮反复push助理代理搜索，背后可能的reflection机制可能比较有借鉴价值

Magne  
tic-one

## 网络搜索与工具使用：3

不够详细：

依然是magnetic-one的方式，按部就班。但是因为没有push web-surfer搜索更多信息，因此没有更细致的数据得到Owl一样量化的表格

根据国际能源署（IEA）的数据，以下是2025年全球三种主要可再生能源的预期使用情况：																					
<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th colspan="2">可再生能源类型</th><th>2025年全球发电总量（Twh）</th><th>主要使用国家</th><th>比例及关键详情</th></tr> </thead> <tbody> <tr> <td>太阳能</td><td>超过2600 Twh</td><td>中国、美国、印度</td><td>中国由于大规模装机容量领先。</td></tr> <tr> <td>风能</td><td>超过2000 Twh</td><td>美国、德国、中国</td><td>在欧洲和北美有显著增长。</td></tr> <tr> <td>水能</td><td>超过4000 Twh</td><td>中国、巴西、加拿大</td><td>全球范围内最大的可再生能源来源。</td></tr> </tbody> </table> <p>### 数据来源 - [IEA - 电力2025 - 执行摘要] (<a href="https://www.iea.org/reports/electricity-2025/executive-summary">https://www.iea.org/reports/electricity-2025/executive-summary</a>)</p> <p>这些数据基于预测，关键国家在各个可再生能源领域的比例显示出其重要的全球角色。</p>					可再生能源类型		2025年全球发电总量（Twh）	主要使用国家	比例及关键详情	太阳能	超过2600 Twh	中国、美国、印度	中国由于大规模装机容量领先。	风能	超过2000 Twh	美国、德国、中国	在欧洲和北美有显著增长。	水能	超过4000 Twh	中国、巴西、加拿大	全球范围内最大的可再生能源来源。
可再生能源类型		2025年全球发电总量（Twh）	主要使用国家	比例及关键详情																	
太阳能	超过2600 Twh	中国、美国、印度	中国由于大规模装机容量领先。																		
风能	超过2000 Twh	美国、德国、中国	在欧洲和北美有显著增长。																		
水能	超过4000 Twh	中国、巴西、加拿大	全球范围内最大的可再生能源来源。																		

## 4.6 问题5：复杂问题解决（旅行规划）

为一个三口之家计划从北京到东京的旅行，包括航班、住宿和旅游景点。一些细节要求：希望游览东京和大阪，东京大阪之间想租车自驾；一共在日本有7天时间；每天不要太辛苦、有8小时在游玩就行；注意性价比

**总体感觉：**

- 一开始详尽、合理的计划非常重要。Manus > Magentic-one > Owl，导致最后三个项目表现的排序也是这个顺序

项目	结果	执行步骤简述
Manus	<p><b>推理与问题解决：5</b></p> <p><b>方案/战略设计：5</b></p> <p>极其完美：</p> <ul style="list-style-type: none"> <li>综合时间合理性和价格，定位出具体航班和住宿</li> <li>具体的租车计划，和必要提醒</li> <li>恰到好处的景点介绍</li> <li>每日by 小时的安排</li> <li>各种有用的旅行贴士</li> </ul>	<p>一开始有一个很好的计划，包含详尽的各方面研究：</p> 
OpenM anus		
Owl	<p><b>推理与问题解决：2</b></p> <p><b>方案/战略设计：2</b></p>	4轮完成任务，只有部分有上网搜索，相对比较简单。

比较basic:

```
### Itinerary

#### Day 1:
Arrival in Tokyo
- **Activities:** Arrive, check into hotel, explore Asakusa, visit Senso-ji Temple, dinner at a local izakaya.

#### Day 2:
Tokyo Exploration
- **Activities:** Visit Tokyo Skytree, explore Akihabara, stroll through Shibuya.

#### Day 3:
Tokyo to Osaka
- **Activities:** Travel to Osaka by Shinkansen, visit Osaka Castle, explore Dotonbori.
```

Magne  
tic-one

推理与问题解决: 3

方案/战略设计: 4

总体更为具体和helpful。租车、航班等更多细节，每天的计划也详细到上下午

设定计划的时候，计划采取了更为详尽的搜索，因此在一开始收集到更多信息，有利于后续详尽计划

Here is the plan to follow as best as possible:

- \*\*Research Flights and Costs:\*\*
  - Use \*\*WebSurfer\*\* to search for flight options from Beijing to Tokyo, focusing on cost-effective airlines and travel dates suitable for a 7-day trip.
- \*\*Explore Car Rental Options:\*\*
  - Use \*\*WebSurfer\*\* to find information on car rental options between Tokyo and Osaka, including rental costs and terms.
- \*\*Research Accommodation and Pricing:\*\*
  - Use \*\*WebSurfer\*\* to identify affordable accommodation options in both Tokyo and Osaka. Look for hotels, guesthouses, or rental apartments that are budget-friendly.
- \*\*Identify Tourist Attractions:\*\*
  - Use \*\*WebSurfer\*\* to discover popular tourist attractions and activities in Tokyo and Osaka that suit an 8-hour daily schedule.
- \*\*Plan Itinerary and Travel Route:\*\*
  - Use \*\*Coder\*\* to draft a daily itinerary that includes activities, estimated travel times, and rental car routes between Tokyo and Osaka. Consider using maps or route planning tools if necessary.
- \*\*Compile Travel Information:\*\*
  - Use \*\*FileSurfer\*\* to organize and save all collected data into a comprehensive travel plan document if local files are to be used for reference or sharing.
- \*\*Validate Budget and Schedule:\*\*
  - Use \*\*WebSurfer\*\* or \*\*Coder\*\* to ensure the proposed trips and accommodations fit within a reasonable budget and adhere to the family's daily activity preferences.

## 附录

### Owl-Camel安装

参考：<https://github.com/camel-ai/owl>

笔记：

- 我用的是本地机器，没有用docker
- 记得安装下playwright浏览器（browser-use工具使用）：playwright install
- 运行语句：python owl/webapp.py，可以在网页上更好呈现运行结果
- 在run.py里，\*BrowserToolkit(headless=False).get\_tools()。不要想quick start建议为True，否则看不到browse-use的网页使用情况

Quick test:

在一开始最想test下「浏览器使用」功能，总体上效果一般：

- Owl通过调用playwright 实现
- 速度慢
- 遇见真人验证的时候可以自己手动验证下
- 会看到一些错误的执行动作

## OpenManus-MetaGPT安装

参考：<https://github.com/mannaandpoem/OpenManus>

笔记：

- 安装完之后，要额外运行下 playwright install，以运行 computer-use

Quick test：

在一开始最想test下「浏览器使用」功能，总体上效果一般：

- OpenManus通过调用browser\_use实现，browser\_use调用了playwright
- 总体体验和Owl差不多，也不是特别理想

## Magnetic-One-AutoGen安装

参考：<https://microsoft.github.io/autogen/dev/user-guide/agentchat-user-guide/magnetic-one.html>（用最后一个code chunk中的代码，以使用magnetic-one的所有工具）