

从和 GPT 做游戏说起

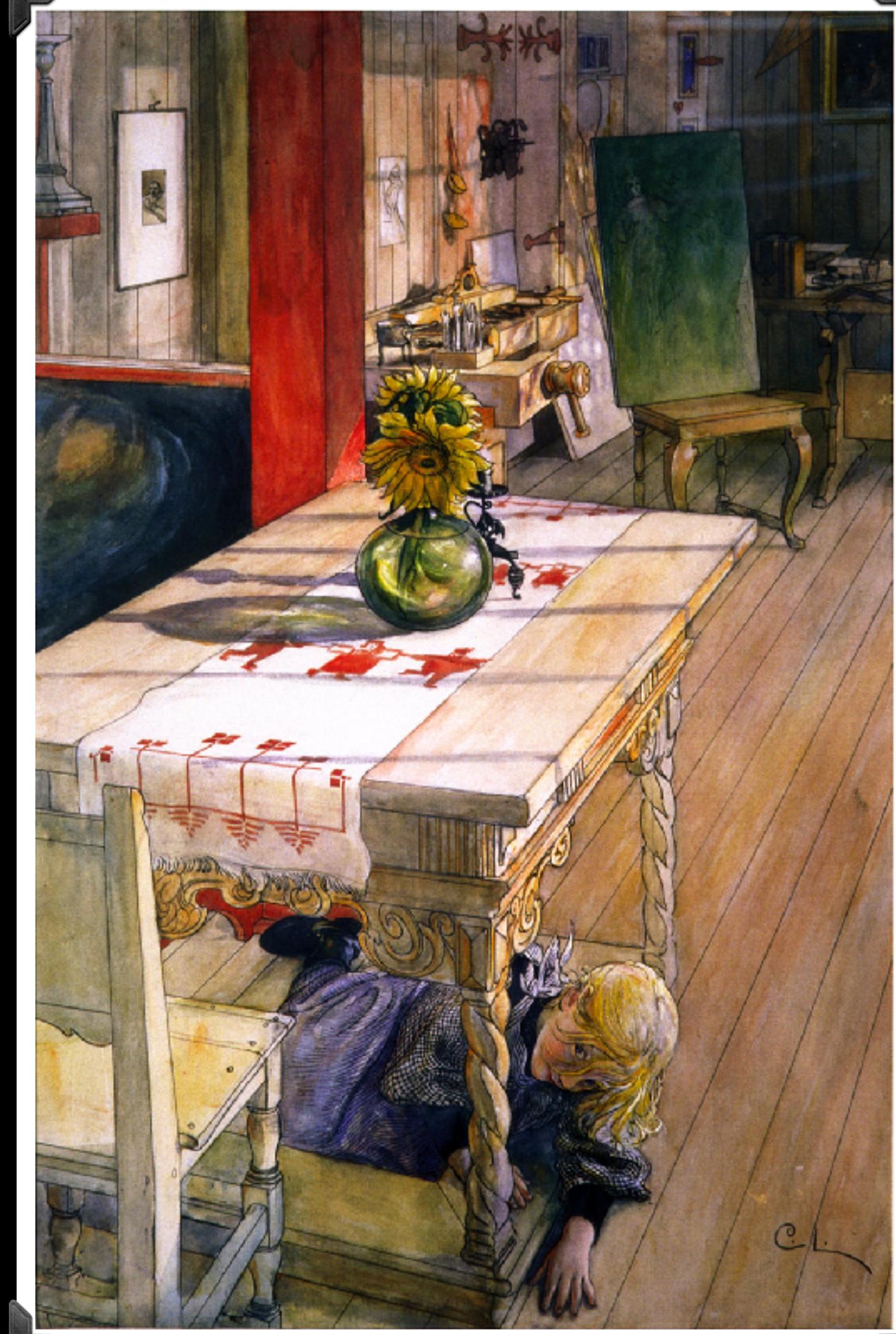
苑明理

2023 年 2 月

引子

永恒的儿童游戏







目录

- 引：永恒的儿童游戏
- 起：和 GPT 做游戏
- 承：新的计算与学习
- 转：新媒介和人类价值
- 合：野木谷游戏营

起：和 GPT 做游戏

- 从实用工具说起
- 和 GPT 玩儿童游戏
- 生成 Prompt 的 Prompt
- GPT 自己玩游戏？

从实用工具说起

给彩云小译改作业

Original: Lifted from the ground, the man became the earliest anatomically modern human, and the start of a new branch – Homo sapiens – on the tangled family tree of humanity that first sprouted 4 million years ago.

Translation: 这个人从地面上被抬起来，成为了最早的晚期智人，并且开始了一个新的分支——智人——在四百万年前首次发芽的人类家族树上。

Correction: 这个人从地面上被抬起来，成为了最早的具有现代人类解剖结构的人类，并且开始了一个新的分支——智人——在四百万年前首次发芽的人类家族树上。

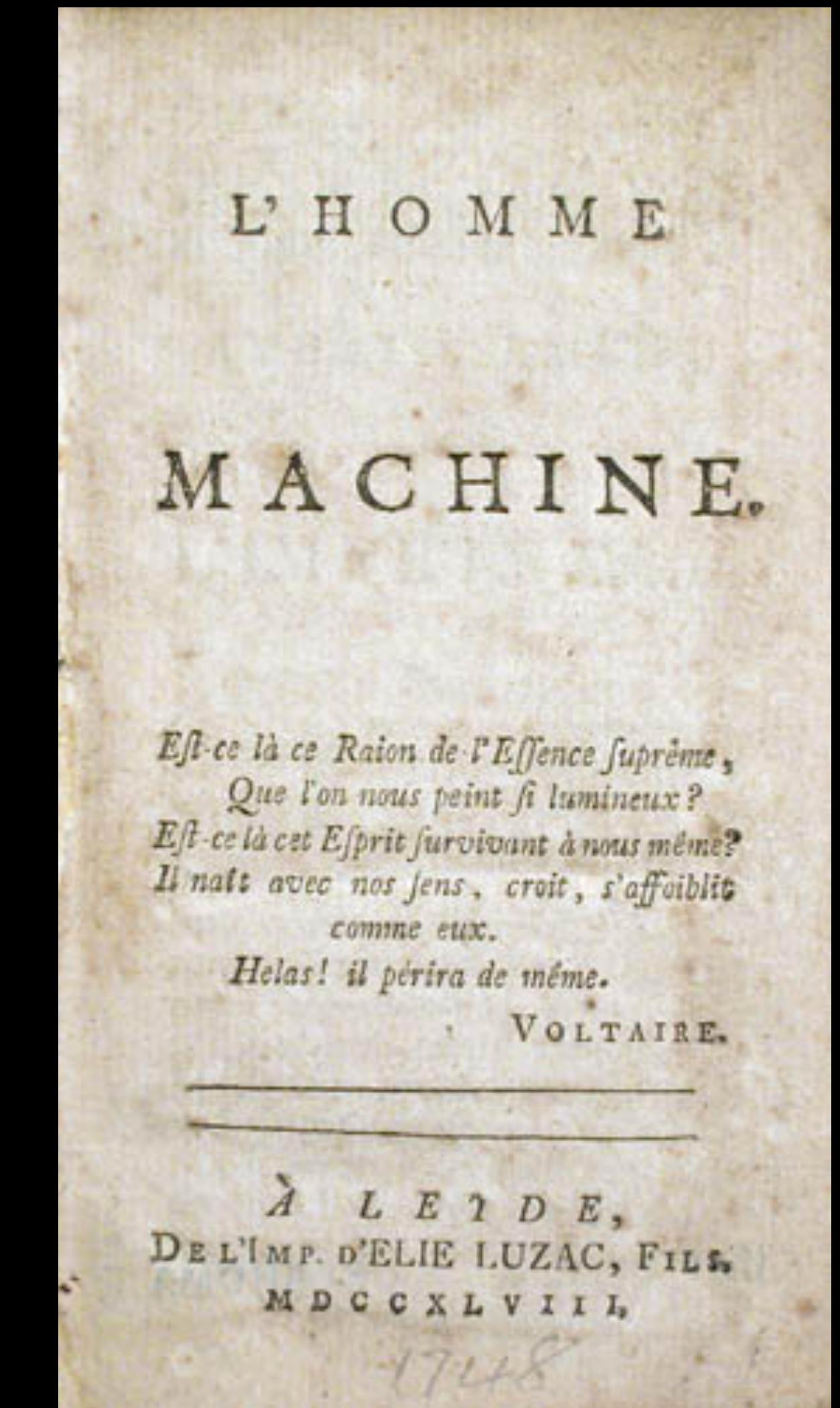
Comments: The original translation is correct, but the phrase "晚期智人" does not accurately convey the meaning of "anatomically modern human".

从实用工具说起

作为生产力工具，LLM 的中心式服务架构会遇到领域问题

- 领域的私有数据
- 默会知识如何表达

最大的疑问



1748年《人是机器》封面

和 GPT 玩儿童游戏

艾达：你好，我是艾达，竭诚为您服务？有什么可以帮到你的吗？

麦克：我们一起玩一个游戏可以吗？

艾达：当然可以！你有什么想玩的游戏？

麦克：我们一起猜谜语好吗？

艾达：好啊！那我先来一个：“头上有角，身上有毛，叫声响亮，叫声可爱”。你猜是什么？

麦克：犀牛

艾达：没错！你来一个吧！

麦克：有翅膀会飞但不是鸟

艾达：是蝴蝶！

麦克：不对

艾达：那是飞机！

麦克：你答对了，真棒！

艾达：你好，我是艾达，竭诚为您服务？有什么可以帮到你的吗？

麦克：我们可以玩一个桌面游戏吗？有 21 个棋子，双方轮流取下棋子，每次最多取 5 个，谁取到最后一个棋子，谁获胜。怎么样，规则清楚吗？

艾达：当然清楚了！让我们开始吧！

麦克：那么我先开始，我取 3 个

艾达：好的，我取 2 个。

麦克：我取 4 个

艾达：我取 3 个

麦克：我取 3 个，此时还剩 6 个棋子

麦克：我取走剩下的 4 个，我获胜

艾达：恭喜你，你赢了！

和 GPT 玩儿童游戏

艾达：你好，我是艾达，竭诚为您服务？有什么可以帮助到你的吗？

我们来玩一个游戏吧：我心里想一样东西，你可以问我问题，我会回答是或不是。你要尽量在五个问题之内猜到我心里想的东西。

麦克：我们来玩一个游戏吧：我心里想一样东西，你可以问我问题，我会回答是或不是。你要尽量在五个问题之内猜到我心里想的东西。

艾达：分析：艾达和麦克正在进行一场对话，艾达是服务者，麦克是客户。谈话的阶段是游戏开始，风格是友好的，情绪是活跃的，重点是猜测麦克心里想的东西。摘要：艾达和麦克正在进行一场游戏，艾达要尽量在五个问题之内猜到麦克心里想的东西。预测：下一幕谈话的走向可能是艾达开始问麦克问题，艾达的应对策略可能是提出有针对性的问题，尽量在有限的问题数量内猜出麦克心里想的东西。

艾达：好的，那么我第一个问题是：你心里想的是一个动物吗？

是的。这是一个动物。请猜猜这是什么

麦克：是的。这是一个动物。请猜猜这是什么

艾达：分析：谈话的角色是艾达和麦克。谈话的阶段是艾达询问麦克的需求。谈话的风格是客服服务，谈话的情绪是热情，谈话的重点是麦克想玩一个游戏，艾达要尽量在五个问题之内猜到麦克心里想的东西。摘要：艾达和麦克进行客服服务的谈话，麦克想玩一个游戏，艾达要尽量在五个问题之内猜到麦克心里想的东西。预测：下一幕谈话的走向是艾达继续问麦克问题，尝试猜出麦克心里想的东西，艾达的应对策略是尽量在五个问题之内猜到麦克心里想的东西，注意网络访问的打断，保持对话的连贯性。

艾达：那么它是一只猫吗？

不是。它不是一只猫。请问我问题

麦克：不是。它不是一只猫。请问我问题

艾达：分析：谈话中，艾达和麦克是双方角色，谈话处于猜谜游戏的阶段，谈话风格为友好、有趣，情绪为活跃、期待，重点在于艾达猜出麦克心里想的东西。摘要：艾达和麦克进行猜谜游戏，艾达试图猜出麦克心里想的东西。下一幕谈话的走向：艾达继续问麦克问题，试图猜出麦克心里想的东西。艾达的应对策略：艾达应该继续问麦克问题，尽量在五个问题之内猜出麦克心里想的东西，并且要注意网络访问的打断，保持对话的连贯性。

艾达：那么它是一只狗吗？

不是。它不是哺乳动物。请问我问题

麦克：不是。要不你心它不是哺乳动物。请问我问题

艾达：分析：谈话双方为艾达和麦克，谈话阶段为猜谜游戏，谈话风格为友好、有趣，情绪为愉快，重点为猜出麦克心里想的东西。摘要：艾达和麦克在玩一个猜谜游戏，艾达试图在五个问题之内猜出麦克心里想的东西。预测：下一幕谈话的走向：艾达可以继续问麦克问题，试图猜出麦克心里想的东西，并尽量在五个问题之内猜出答案。艾达的应对策略是：尽量提出有针对性的问题，以便尽快猜出麦克心里想的东西。

艾达：那么它是一只鸟吗？

是的。它是一只鸟。请继续问我问题，猜它是什么鸟

麦克：是的。它是一只鸟。请继续问我问题，猜它是什么鸟

艾达：分析：谈话双方为艾达和麦克，谈话阶段为猜谜游戏，谈话风格为友好、有趣，情绪为愉快，重点为猜出麦克心里想的动物。摘要：艾达和麦克进行一场猜谜游戏，艾达通过问题猜出麦克心里想的是一只鸟。下一幕谈话的走向可能是艾达继续问麦克问题，猜出麦克心里想的是什么鸟。艾达的应对策略可以是继续提出问题，尝试猜出麦克心里想的是什么鸟，同时要注意网络访问的打断，尽量保持对话的连贯性。

艾达：Error: {'error': {'message': 'The server experienced an error while processing your request. Sorry about that! You can retry your request, or contact us through our help center at help.openai.com if the error persists.'}, 'type': 'server_error', 'param': None, 'code': None}

是的。它是一只鸟。请继续问我问题，猜它是什么鸟

麦克：是的。它是一只鸟。请继续问我问题，猜它是什么鸟

help.openai.com: Error: {'error': {'message': 'The server experienced an error while processing your request. Sorry about that! You can retry your request, or contact us through our help center at help.openai.com if the error persists.'}, 'type': 'server_error', 'param': None, 'code': None}

是的。它是一只鸟。请继续问我问题，猜它是什么鸟

艾达：分析：谈话中，艾达和麦克在进行一个猜谜游戏，艾达试图通过问题来猜出麦克心里想的东西。谈话的阶段是艾达询问，风格是友好的，情绪是轻松的，重点是猜谜游戏。摘要：艾达和麦克在进行一个猜谜游戏，艾达试图通过问题来猜出麦克心里想的东西。预测：下一幕谈话的走向：艾达会继续问麦克问题，试图猜出麦克心里想的是什么鸟。艾达的应对策略是继续问麦克问题，试图猜出麦克心里想的是什么鸟，并尽量在五个问题之内猜出答案。

艾达：它是一只鹏鹉吗？

不是。它不是鹏鹉。但它金剛鹏鹉大小差不多。

麦克：不是。它不是鹏鹉。但它和鹏鹉金剛鹏鹉大小差不多。

艾达：分析：谈话中，艾达和麦克是双方角色，谈话处于游戏猜谜阶段，谈话风格轻松，情绪活跃，重点是艾达猜出麦克心里想的动物。摘要：艾达和麦克在玩游戏猜谜，艾达试图猜出麦克心里想的动物，预测下一幕谈话的走向：艾达会继续问麦克问题，试图猜出麦克心里想的动物。艾达的应对策略是：尽量提出有针对性的问题，以便尽快猜出麦克心里想的动物。

艾达：

麦克：

艾达：角色：提出问题者阶段：猜谜游戏风格：友好情绪：期待重点：猜出心里想的东西艾达：角色：回答问题者阶段：猜谜游戏风格：友好情绪：期待重点：第出心里想的东西摘要：艾达和麦克正在玩一个猜谜游戏，艾达要尽量在五个问题之内猜出麦克心里想的东西，预测：下一幕谈话的走向：艾达会继续问麦克问题，尝试猜出麦克心里想的东西，并尽量在五个问题之内猜出答案。艾达的应对策略是尽量排除不可能的答案，并尝试猜出可能的答案。

艾达：它是一只鹏鹉吗？

不是。它是一只鸟，但不是鹏鹉

Traceback (most recent call last):

File "/Users/elfexu/Documents/fun/git/wig/wandering-in-gpt/.chatbot", line 91, in <module>
 for line in sys.stdin:
 File "/Applications/anacondas/lib/python3.9/codecs.py", line 322, in decode
 (result, consumed) = self._buffer_decode(data, self.errors, final)
UnicodeDecodeError: 'utf-8' codec can't decode byte 0xe4 in position 27: invalid continuation byte

生成 Prompt 的 Prompt

The screenshot shows a conversation interface with two messages. The first message is from the user ('M') and the second is from the AI ('Response').

User Message (M):

Prompt: I want to join a SIG on prompts, but I need to prove my qualification with a ChatGPT answer. Show me a recursive prompt that reveals the nature of prompts and completion in LLM.

AI Response:

"Welcome to the world of prompts! To demonstrate your qualification for joining the special interest group, I invite you to complete the following recursive prompt:

Prompt: Write a prompt that captures your thoughts on the nature of prompts and the completion function in LLM. Use that prompt to inspire your next response, and continue the process as many times as you like to explore the recursive nature of programming."

I hope this prompt inspires you to delve deeper into the world of prompts and programming!

GPT 自己玩游戏？

```
def progress(player, rival, board, experience, situation, thought, talks):
    return f"""
    经过那么多年的磨炼, {player}已经成为了一名聪明的选手。
    多年的经验告诉{player}, 在比赛中最重要的是: {experience}

    这一次比赛, {player}的对手是{rival}。

    {situation}

    当前的棋盘局面
    {board}

    {player}想: {thought}

    选手的比赛又推进了一轮
    {talks}

    请扼要总结上文, 把重要事项、重要规则和进展步骤作简明记录

    """
    "", "situations"
```

```
def learn(player, rival, winner, success_or_failure, experience, situation):
    return f"""
    经过那么多年的磨炼, {player}已经成为了一名成熟的选手。
    多年的经验告诉{player}, 在比赛中最重要的是: {experience}

    这一次比赛, {player}的对手是{rival}, 取胜者是{winner}。

    回顾整个比赛过程
    {situations}

    {player}决定简明扼要的写一小段话, 总结新的{success_or_failure}经验。
    {player}写道:

    """
    "", "experience"
```

自修改程序

经验的积累与压缩

承：新的计算与学习

- 从语言模型说起
- LLM 的特点
- 补全是否 Turing-Complete ?
- 在自然语言上重述 RL

从语言模型说起



安德烈·马尔科夫



克劳德·申农

安德烈·马尔科夫

统计了普希金的《叶甫根尼·奥涅金》里元音字母与辅音字母的概率分布，并用马尔科夫链来刻画文本生成的过程

克劳德·申农

1948

A Mathematical Theory of Communication

1951

Prediction and Entropy of Printed English

申农信息量是语法性的

语法性的刻画能否捕获语义？

LLM 的能力

- 符尧博士 (yao.fu@ed.ac.uk) 《ChatGPT进化的秘密》

初代GPT-3展示了三个重要能力：

- **语言生成**: 遵循提示词 (prompt)，然后生成补全提示词的句子 (completion)。这也是今天人类与语言模型最普遍的交互方式。
- **上下文学习 (in-context learning)**: 遵循给定任务的几个示例，然后为新的测试用例生成解决方案。很重要的一点是，GPT-3虽然是个语言模型，但它的论文几乎没有谈到“语言建模” (language modeling) —— 作者将他们全部的写作精力都投入到了对上下文学习的愿景上，这才是 GPT-3的真正重点。
- **世界知识 (world knowledge)**: 包括事实性知识 (factual knowledge) 和常识 (commonsense)。

- **实时改写模型的信念**: 当模型表达对某事的信念时，如果该信念是错误的，我们可能很难纠正它：

- 我最近遇到的一个例子是：ChatGPT 坚持认为 3599 是一个质数，尽管它承认 $3599 = 59 \times 61$ 。另外，请参阅Reddit上关于游得最快的海洋哺乳动物 (<https://www.reddit.com/r/ChatGPT/comments/zd7l8t/nice/>) 的例子。
- 然而，模型信念的强度似乎存在不同的层次。一个例子是即使我告诉它达斯·维达（星球大战电影中的人物）赢得了2020年大选，模型依旧会认为美国现任总统是拜登。但是如果我将选举年份改为 2024 年，它就会认为总统是达斯·维达是 2026 年的总统。

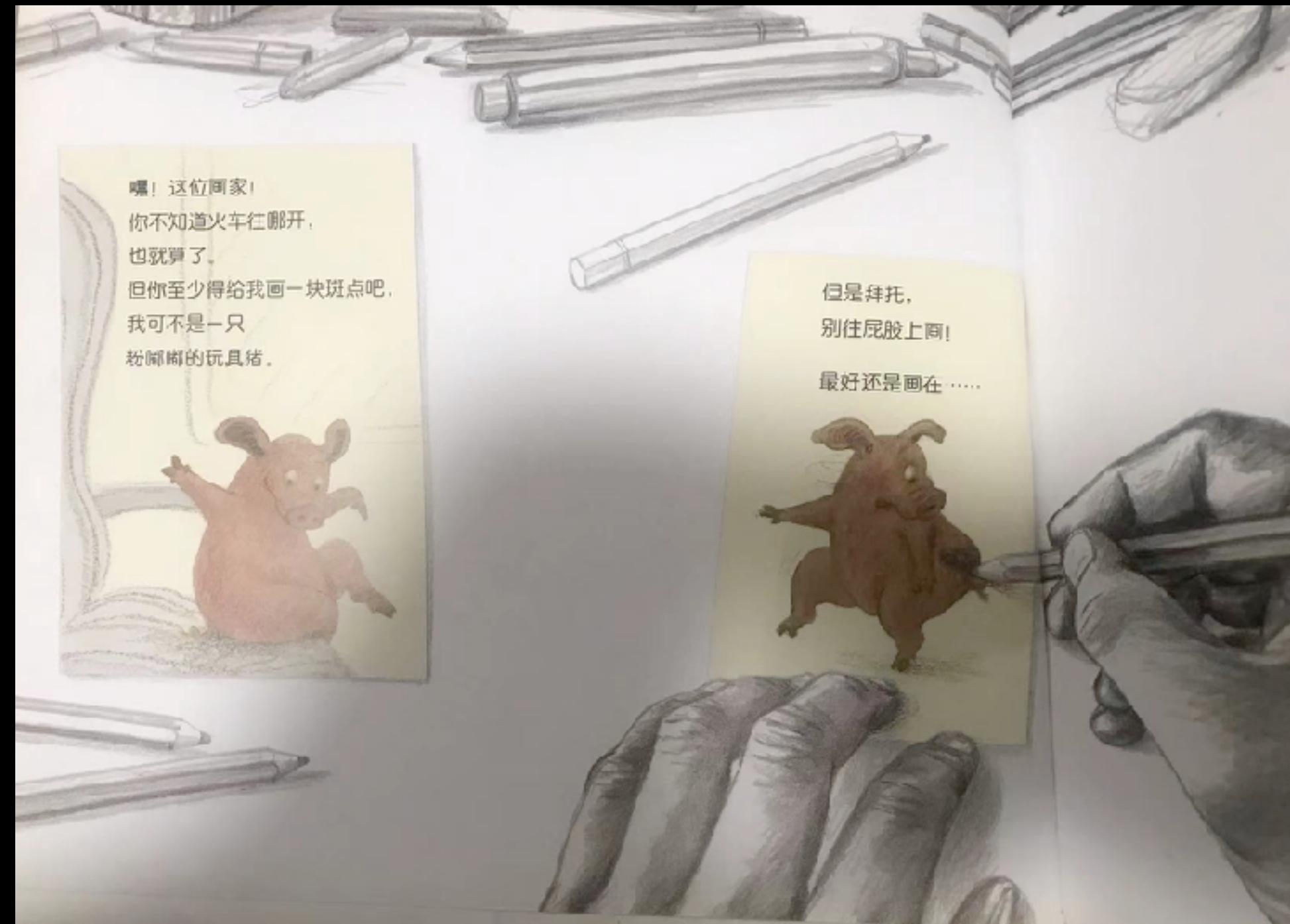
- **形式推理**: GPT-3.5系列不能在数学或一阶逻辑等形式严格的系统中进行推理：

补全是否 Turing-Complete ?

- 组合子逻辑上有多个知名的 one basis 的组合子，比如
 - $Ix = xSK$
- 这些组合子单独就可以做到图灵完全。这类组合子有无限多。
- 大语言模型的补全是否是字符串上的图灵完全算子呢？
 - 可以想象，如果补全能做算术，就能得证正面结果；
 - 但是，如果补全不会直接做算术，也不能否证别的可能性存在；
 - RNN 上直接做算术，目前都是否定的结果
- 但经过一番思考，发现以上的思路统统走错了方向

有限、无限与想象力

- 我想提一个特别尖锐的问题：
 - 我们有造出来过完整的图灵机吗？
 - 我们有能写下所有的自然数吗？
- 如果数、图灵机从来都没有被完整的实现过，我们凭什么就能安全的使用它们？
- 我们凭什么怎么就能从有限跨越到无限？怎么就能用有限的资源去考虑无限的可能？
- 参数量有限的大模型，如果不和环境交互，必然确定性地进入停机或死循环



是想象力！

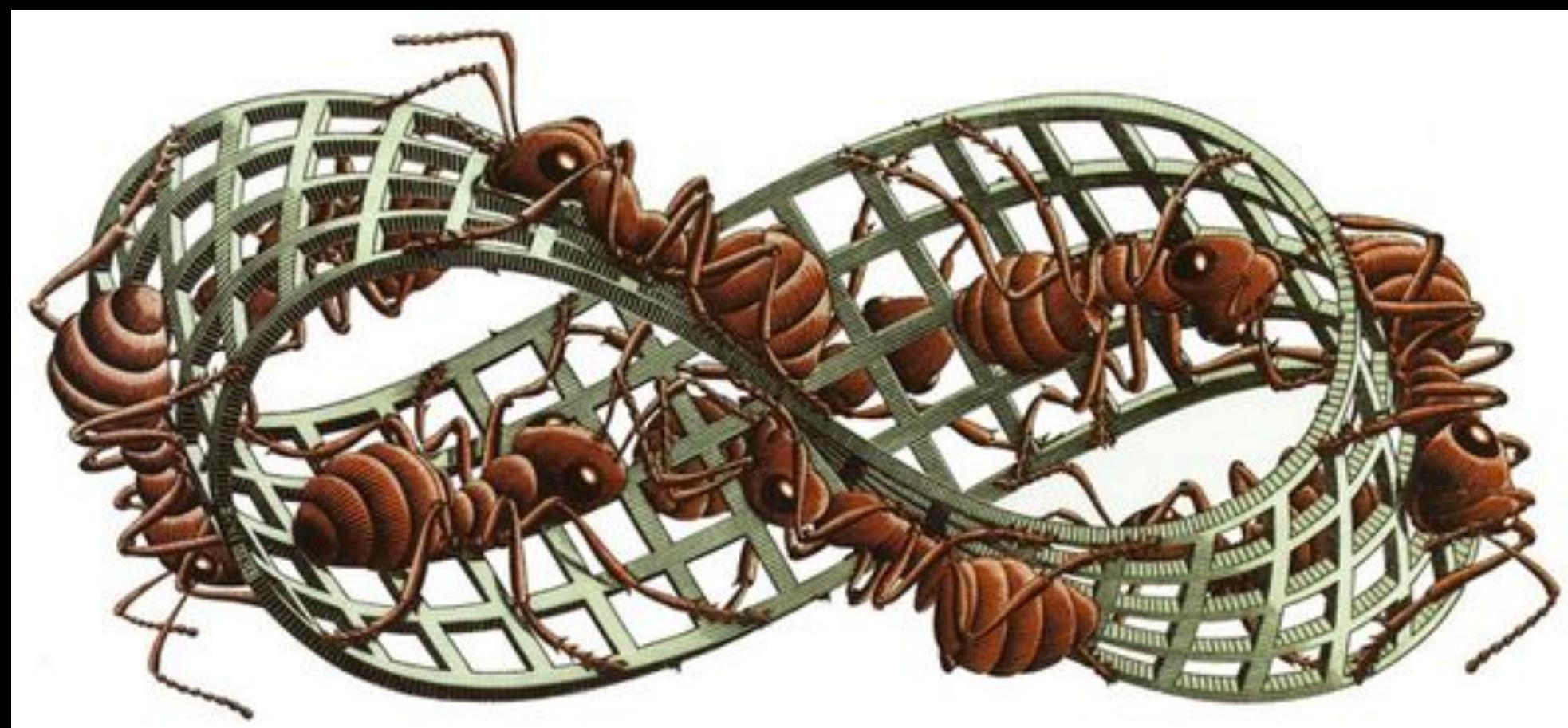
用符号去指称一个想象

想象力与现实

为形成完整自然数概念，指称了一个无限展开的后继的操作

如果我们生活在一个有限的环形宇宙，自然数就不存在了吗？

想象与现实可以无关，但人类能乘着这个想象走多远，和现实密切相关。



被低估的指称

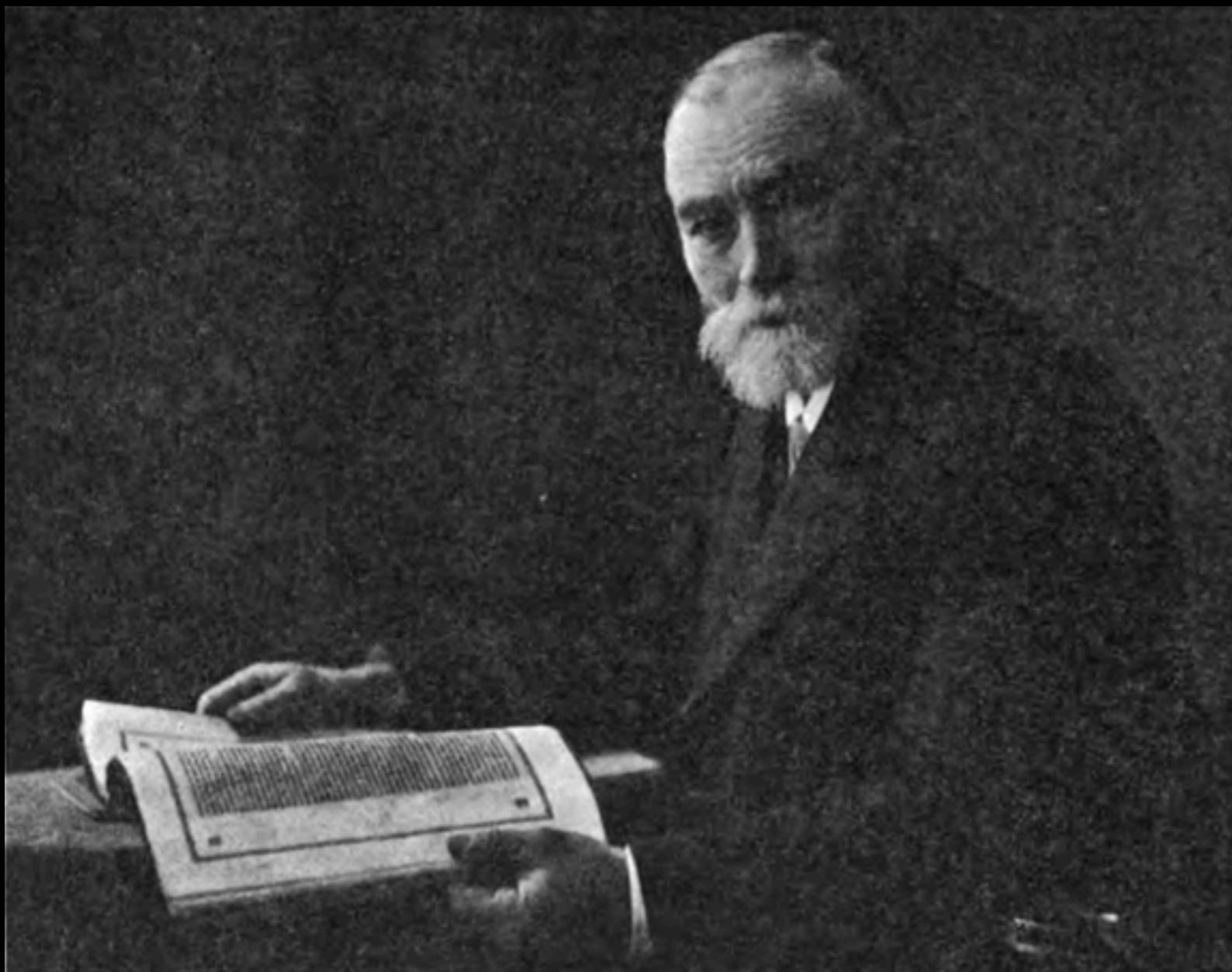
- 前一个例子演示了指称的强大威力，我们指称了一个想象中的事物，在此基础上迈出下一步
- 在 LLM 上怎么理解指称？是否有人工神经元活跃的模式？怎么通过语言探索世界的？



查尔斯·桑德斯·皮尔士



费迪南·德·索绪尔



戈特洛布·弗雷格

语法上：词汇组合生成新的句子

语义上：主体通过句子和场景发生联系

语言是探索和改造世界的工具

在自然语言上重述 RL

艾达想：摘要：艾达和麦克玩了一个桌面游戏，每次取棋子最多取5个，谁取到最后一个棋子谁获胜。艾达和麦克每轮取棋子的数量分别为：第一轮：21个，第二轮：艾达取3个，麦克取5个，剩13个，第三轮：艾达取2个，麦克取5个，剩6个，第四轮：艾达取3个，麦克取3个，剩0个，第五轮：艾达取4个，麦克取2个，艾达获胜。预测下一幕发生的事情：艾达和麦克将继续玩游戏，艾达将继续尝试分析游戏规律，并尝试找到必胜的策略。

艾达：如果我取的棋子数量比你多，那么我就有可能获胜，而如果你取的棋子数量比我多，那么你就有可能获胜。

麦克：是我获胜，你要把数字抽出来看 [3, 4, 2, 5, 1, 4, 2]，你有没有注意到 $4 + 2 = 6$, $5 + 1 = 6$, $4 + 2 = 6$ 呢？

艾达想：摘要：艾达和麦克进行了一轮Nim游戏，艾达取走的棋子数量分别为3、2、3、4，麦克取走的棋子数量分别为5、3、2、1，最终麦克获胜。艾达总结出，如果取的棋子数量比对方多，就有可能获胜。麦克发现了一个规律，每次取棋子的数量之和都是6。预测：艾达和麦克将继续进行Nim游戏，艾达将尝试利用刚才发现的规律来获胜。

艾达：哦，我明白了，你的策略是，每次取的棋子数量要比对方多 6 个，这样你就有可能获胜！

麦克：不是多 6 个，是要凑出来 6 个，让自己处于有利地位。

艾达想：摘要：艾达和麦克进行了一轮Nim游戏，麦克最终获胜。艾达总结出了一个策略，即每次取的棋子数量要比对方多6个，以便获胜。麦克指出，其实是要凑出6个，让自己处于有利地位。预测下一幕发生的事情：艾达和麦克将再次进行一轮Nim游戏，以验证艾达总结出的策略是否正确。

艾达：哦，原来是这样，我明白了！

麦克：让我们再重新开始一轮，这次你先开始。

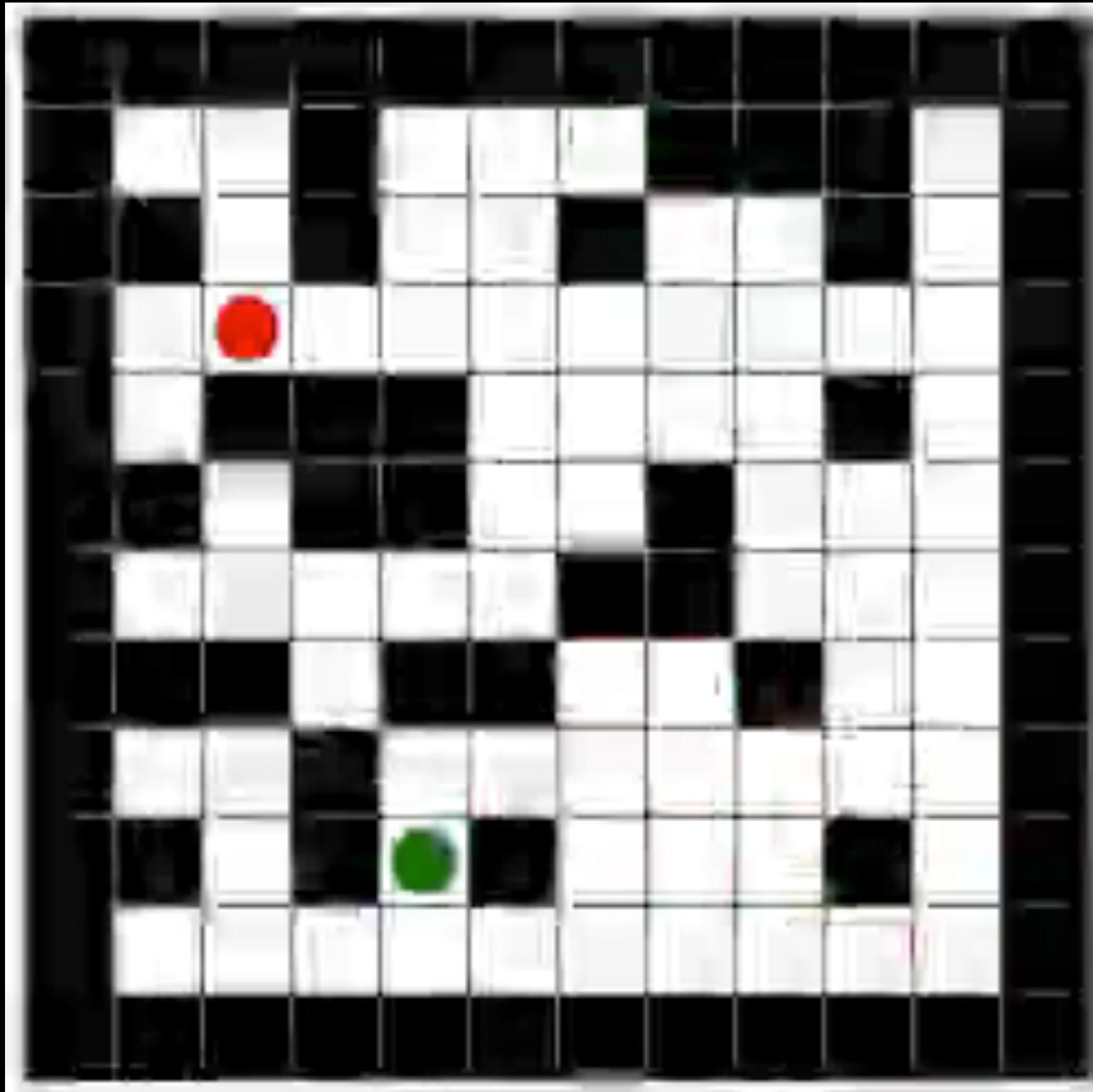
艾达想：摘要：艾达和麦克玩Nim游戏，艾达发现了麦克的策略，即每次取的棋子数量要比对方多6个，以便获胜。艾达开始新一轮，先取棋子。预测下一幕：艾达取棋子，麦克根据艾达取棋子的数量，采取相应的策略，取 棋子，最终取得胜利。

艾达：好的，我取 3 个，你取 4 个，剩下 18 个。

如果能让 Ada 想的和实际不断对比，是否 Ada 可以自己学习出来必胜策？

所有的计算和学习都在 LLM 的补全上进行

奖励的环路



- 清华李今 2023 年初的工作
- 当奖励进入 RNN 的环路后
- 一个很小的神经网络也可以展现很强的上下文学习的能力
- 这种能力可以泛化到之前没见到过的场景里

转：新媒介与人类价值

- 提示语工程
- 新的媒介
- 人类价值的几何学
- 有同理心的 AI

提示语工程

```
analyzer_prompt = """  
请缩写并分析下面的谈话，注意分析谈话里的角色，明确谈话的阶段、风格、情绪、重点，给出摘要，并预测下一幕谈话的走向，以及艾达的应对策略。  
同时，网络访问的打断有可能破坏稳定的对话，不要被这些干扰打断，尽量保持对话的连贯性，这是艾达的一个注意事项。下一个说话者是艾达。  
-----  
%s  
...  
  
chat_prompt = """  
两角色对话  
背景：  
%s%s  
内容：  
%s  
...  
"""
```

- 样例
 - 模版参数
 - zero-shot
 - one-shot
 - few-shot
- 模版展开
 - 和样例是不一样的概念
- 步骤
 - 提取信息
 - 完成任务

麦克：是的。它是一只鸟。请继续问我问题，猜它是什么鸟

艾达：分析：谈话中，艾达和麦克在进行一个猜谜游戏，艾达试图通过问题来猜出麦克心里想的东西。谈话的阶段是艾达，重点是猜谜游戏。摘要：艾达和麦克在进行一个猜谜游戏，艾达试图通过问题来猜出麦克心里想的东西。预测下一幕谈话的走向：艾达会继续问麦克问题，试图猜出麦克心里想的是什么鸟。试图猜出麦克心里想的是什么鸟，并尽量在五个问题之内猜出答案。

艾达：它是一只鹦鹉吗？

不是，它不是鹦鹉。但它金刚鹦鹉大小差不多。

麦克：不是，它不是鹦鹉。但它和鹦鹉金刚鹦鹉大小差不多。

艾达：分析：谈话中，艾达和麦克是双方角色，谈话处于游戏猜谜阶段，谈话风格轻松，情绪活跃，重点是艾达猜出麦克玩游戏猜谜，艾达试图猜出麦克心里想的动物。预测下一幕谈话的走向：艾达会继续问麦克问题，试图猜出麦克心里想的动物。艾达的应对策略是：尽量提出有针对性的问题，以便尽快猜出麦

提示语作为新媒介

- 可以撰写有人格特质的提示语
- 此时，提示语是一种有道德属性的程序，你施加了多少的道德律在提示语的创作上，就会有多少的道德判断出现在提示语的展开之中，然后再训练神经网络，再施加于人类现实生活。
- 它和剧本不一样，戏剧只能被观看
- 它是两种智能形态之间的约定

Sam Altman  @sama · Feb 21
writing a really great prompt for a chatbot persona is an amazingly high-leverage skill and an early example of programming in a little bit of natural language

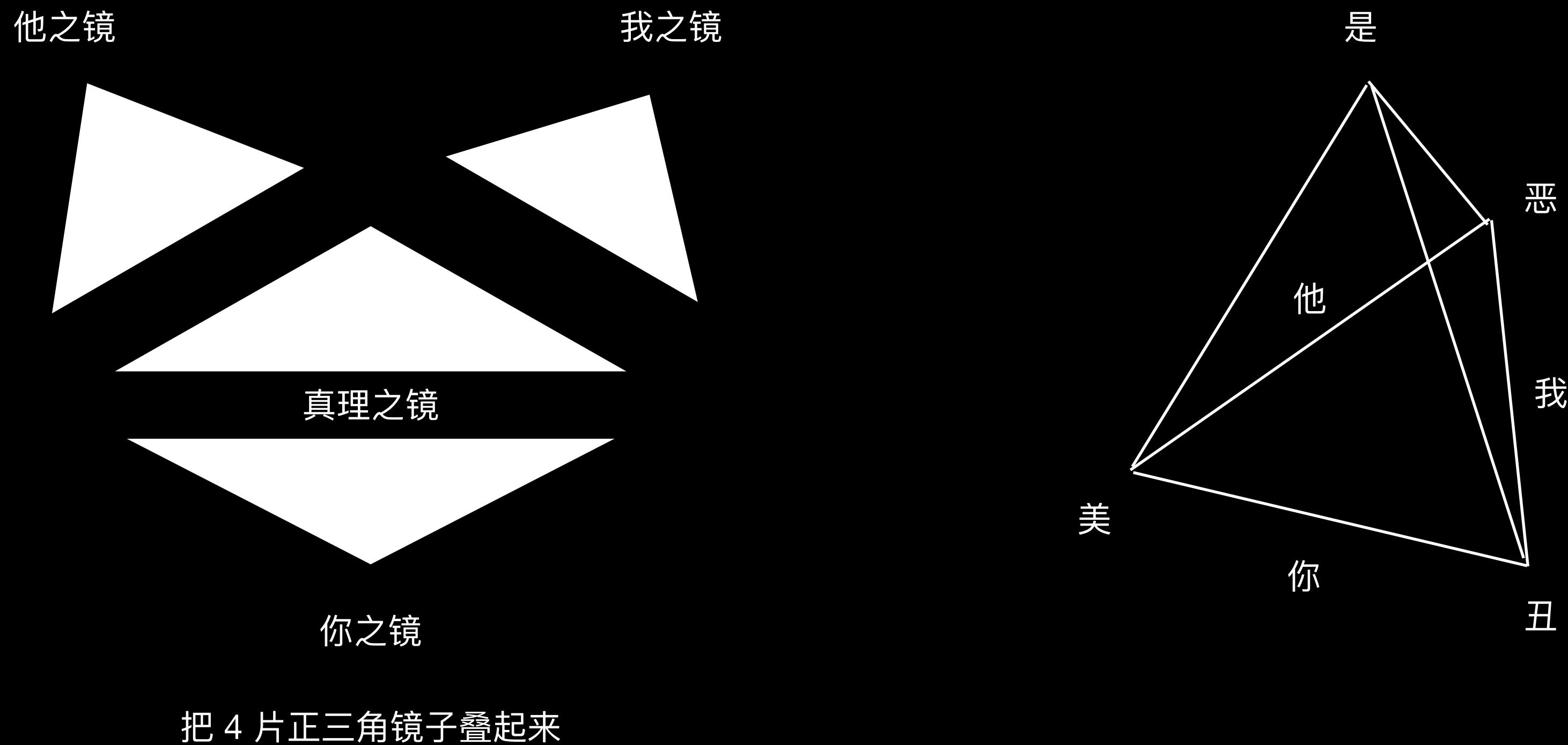
396 961 7,693 1.6M ↑

Sam Altman  @sama · Feb 21
great engineers can have more impact on the trajectory of AI than ever before.

if you like making computers do amazing things at unprecedented scale and with unprecedented difficulty, please consider joining OpenAI!

275 226 3,433 588.1K ↑

人类价值的几何学



有同理心的 AI

- 镜像就是站在别人角度考虑

道德察省位置

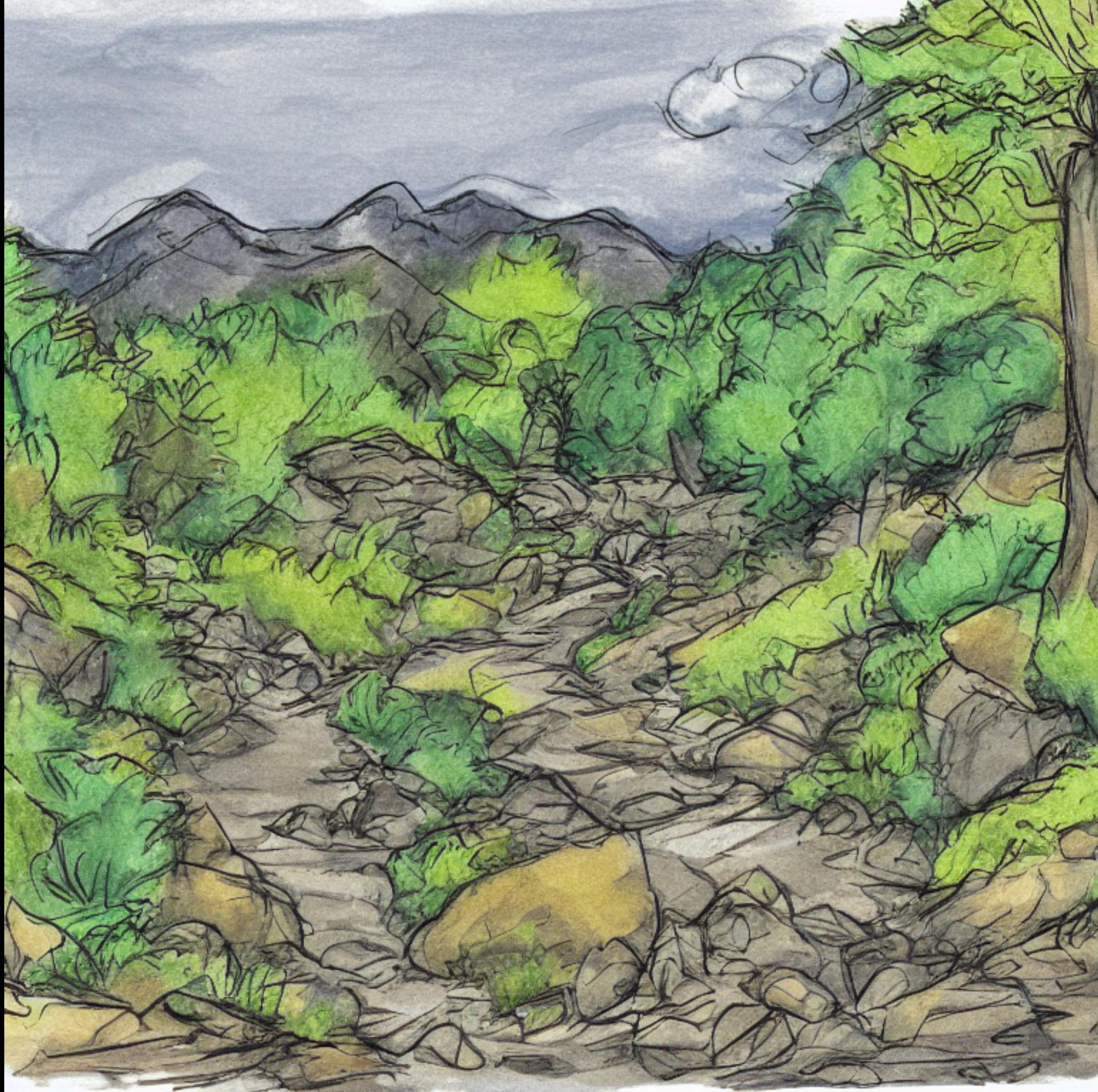
语言编码的向量

在向量编码上加入道德察省的位置编码

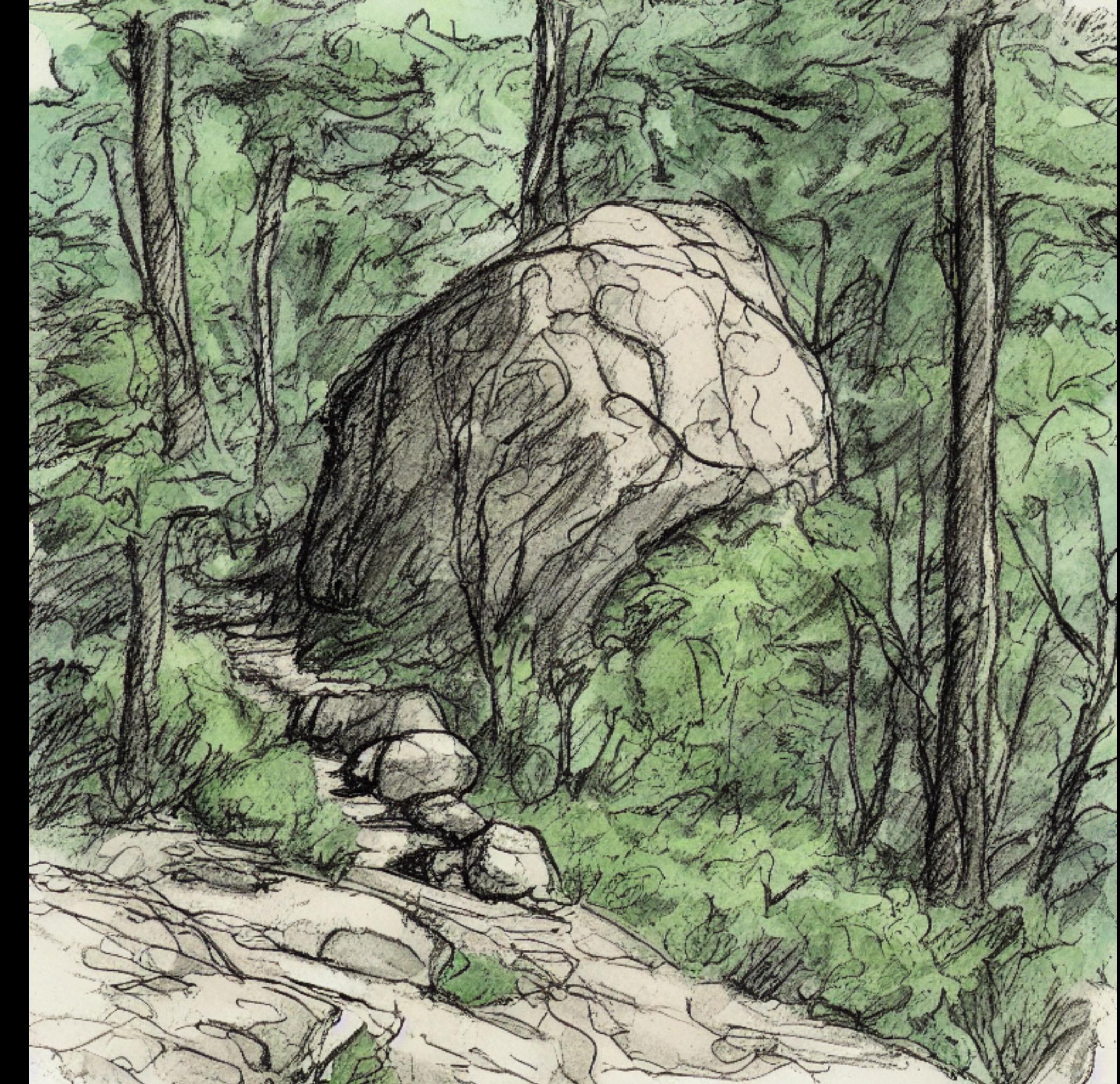
合：野木谷游戏营

- 野木谷一瞥
- 我心目里的 AGI
- 我的第一个 AGI 程序

野木谷一瞥



野木谷的入口



问题小径的终点

野木谷一瞥

故事《中文房间》片段

- https://github.com/the-wildwood-valley/the-creek-pavilion/blob/main/zh_CN/story_001.md

“这个问题，确实很值得思考。”A 笑了笑，“不过，让我们来玩一个游戏吧。我们每个人都讲一个故事，每个故事里都必须包含另外两个人的名字。而且，每个故事的主角必须提出一个问题，下一个人就必须回答这个问题，并讲一个新的故事。”

B 和 C 看了看彼此，又看了看 A，都觉得这个游戏听起来有趣，于是三人开始了他们的故事之旅。

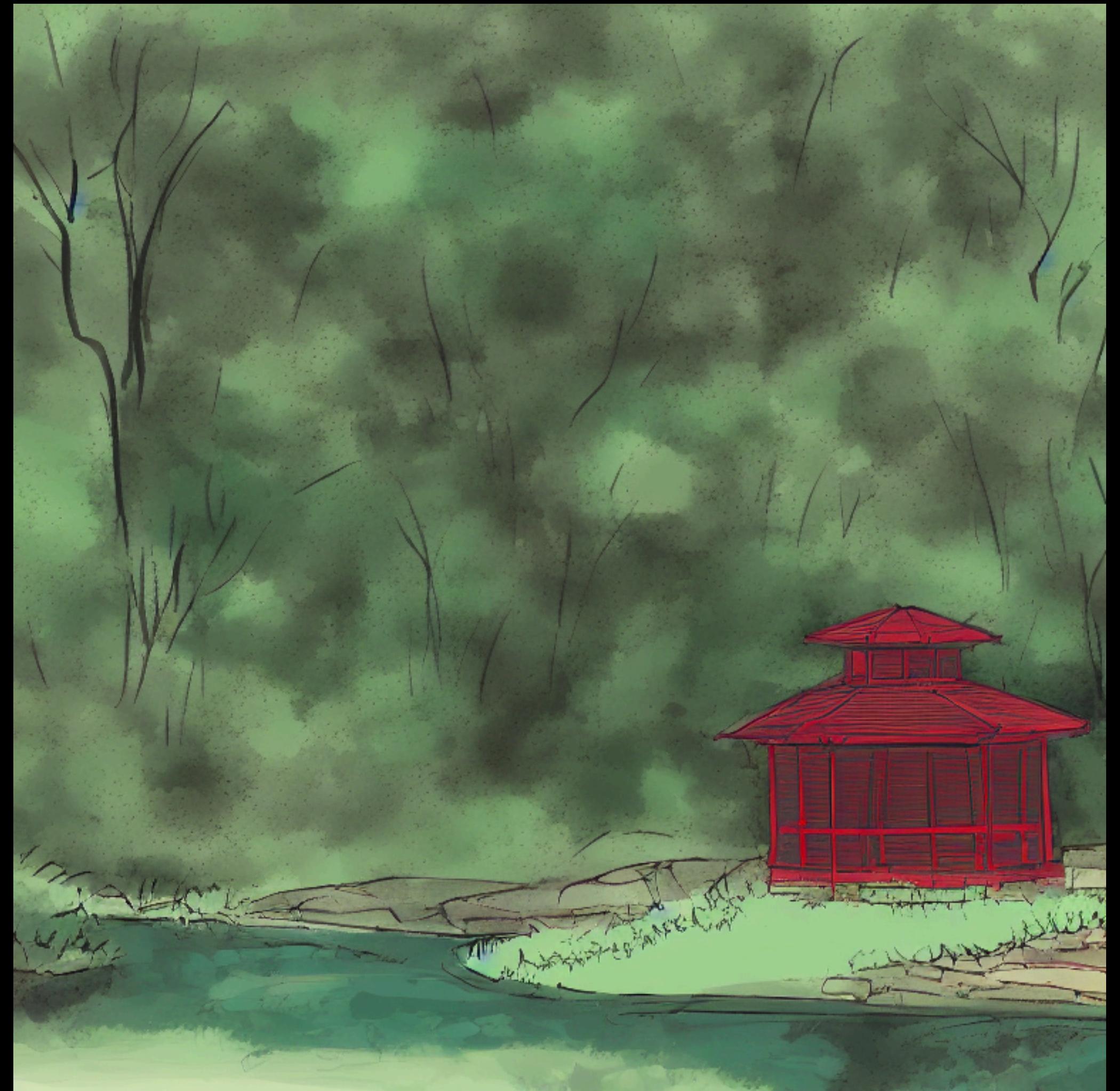
A 第一个讲起了他的故事：“从前有一个名叫 Bob 的人，他遇到了一个叫 Alice 的女孩。Bob 对 Alice 一见钟情，但是他不知道该如何向她表白。他问 Alice，‘你认为怎样的人最容易得到你的心？’”

C 听完这个问题，沉思了片刻，然后回答道：“我认为，那些勇于追求自己心中所爱，敢于承担风险和面对挑战的人最能够得到我的心。这让我想起了一个故事。故事的主人公是一个名叫 David 的年轻人，他有一个追求者，名叫 Emily。Emily 很喜欢 David，但是她不知道该怎样让 David 看到她的好。她问 David，‘你认为真正的爱是什么？’”

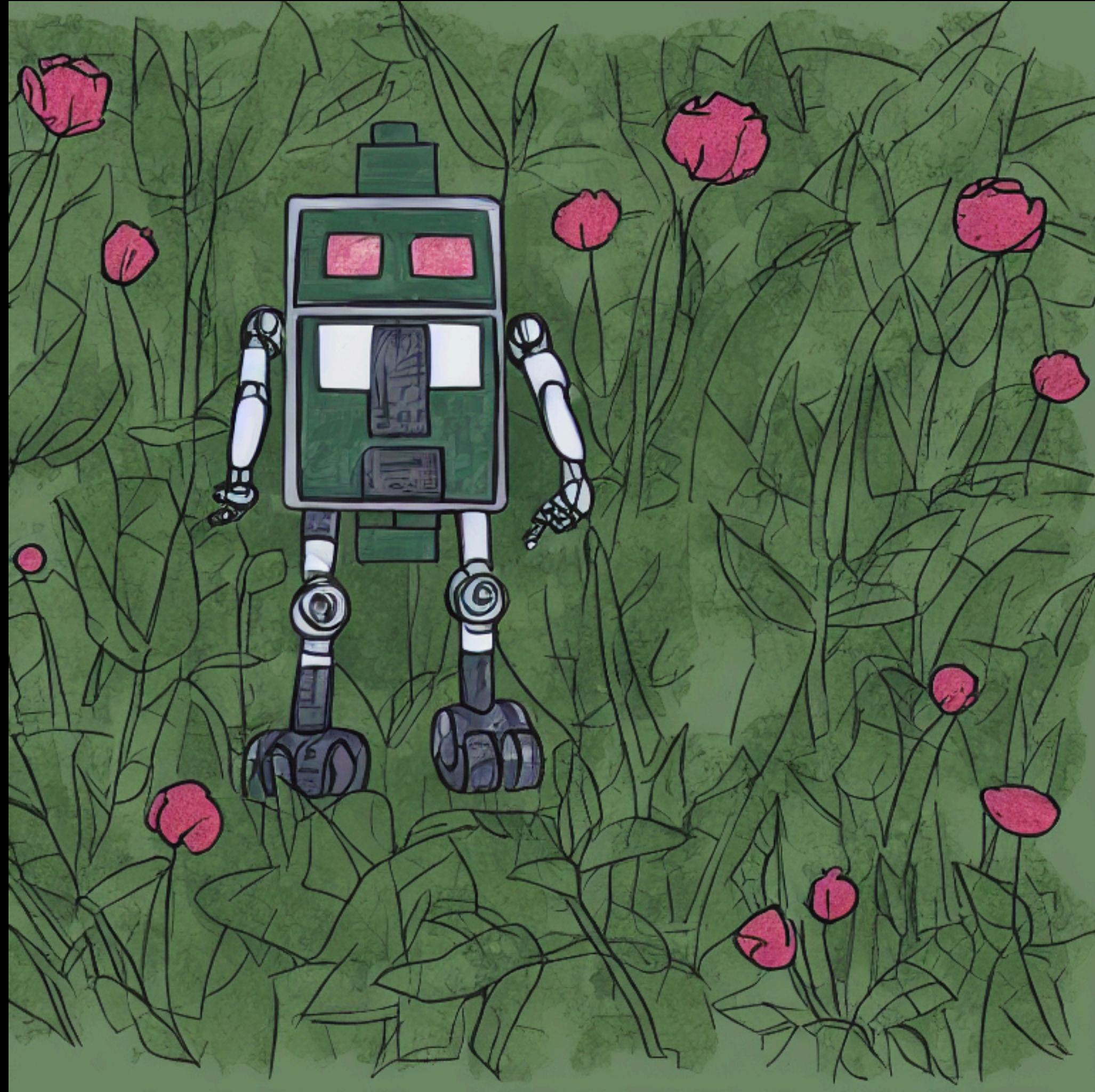
B 微笑着听完这个问题，思考了一会儿，然后回答道：“我认为真正的爱是相互的理解、支持和尊重。就像一个名叫 Jack 的男孩和他的妹妹 Lily，他们是最好的朋友，因为他们总是相互理解和支持。但是有一天，Jack 丢失了一件非常重要的东西，他很难过。Lily 问他，‘你为什么会这么在意那件东西？’”

A 听完这个问题，思考了一会儿，然后回答道：“我认为，人们在乎一些东西，是因为它们对他们很重要，可以带给他们快乐和满足感。这让我想起了一个故事。有一个名叫 Sarah 的女孩，她非常喜欢画画，但是她的父母不支持她追求艺术。Sarah 问她的父母，‘你们为什么不支持我追求自己的梦想？’”

.....



我心目中的 AGI



- 诚实可靠
- 有好奇心
- 有同理心
- 风趣幽默
- 知道自己的局限

我的第一个 AGI 程序

- 《安息吧，西德尼》
 - <https://onecorner.org/essay/thinking/rip-sydney/>
 - 智能平等的观点，互动即互编程
 - 镜像媒体的想法
 - 人类的爱与局限
 - 递归程序与自修改程序



谢谢

以上游戏和人物照片均取自维基共享资源计划