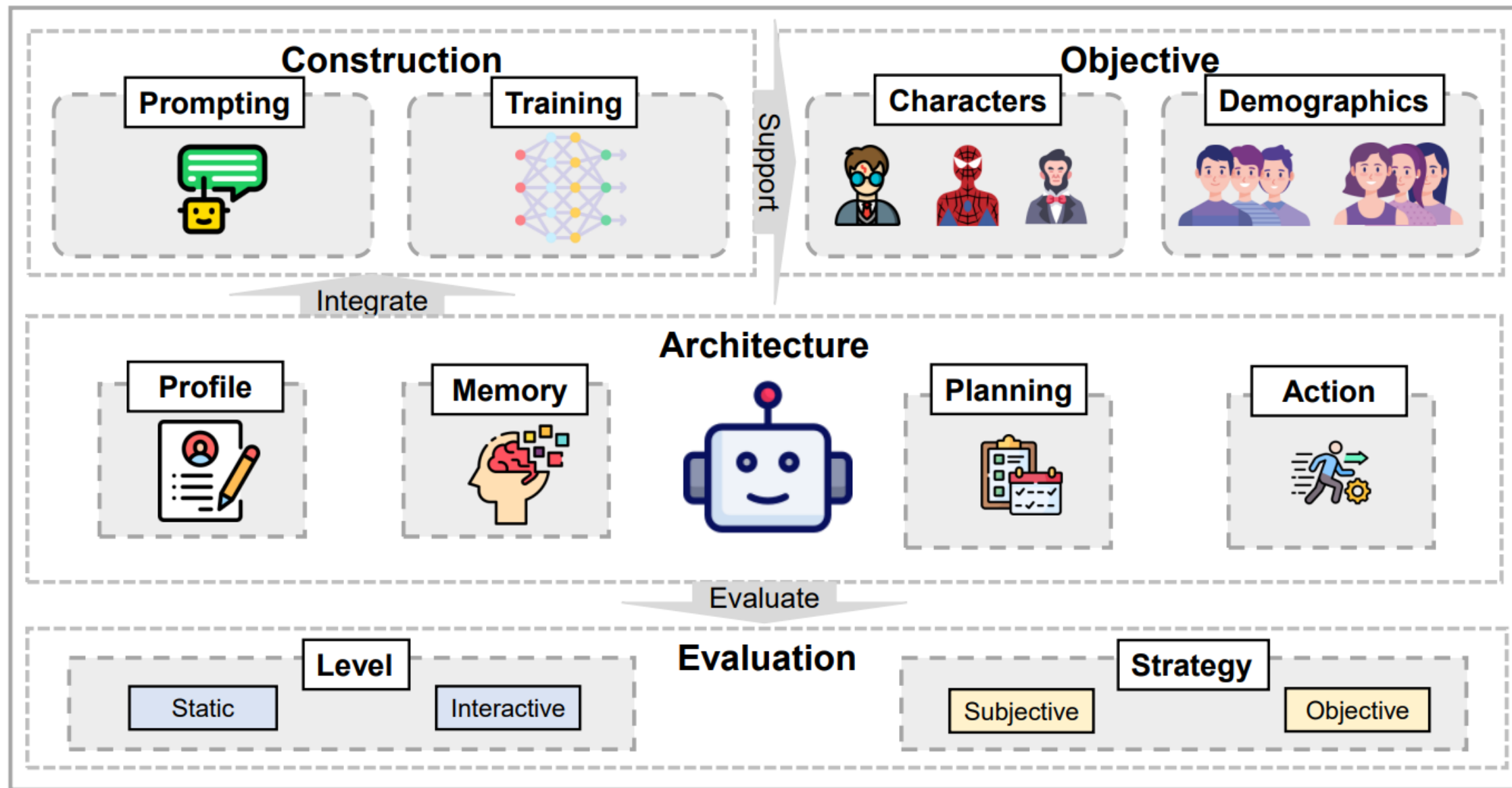


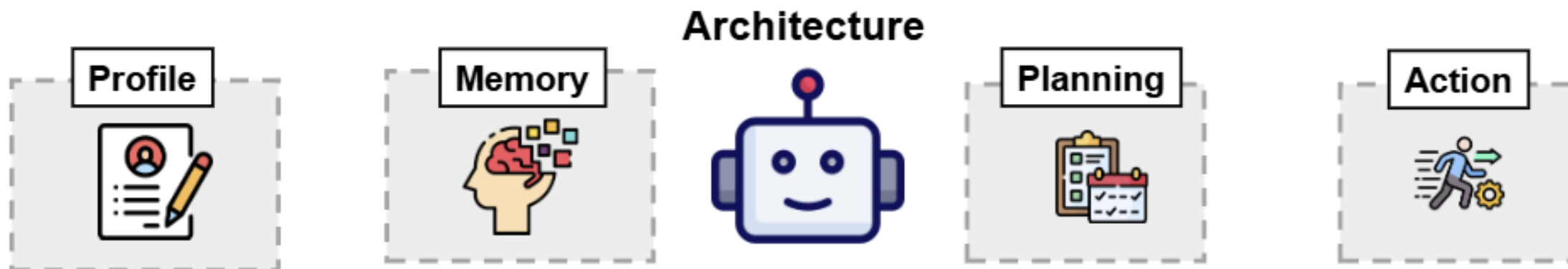
# 个体模拟-整体框架



# 智能体组成架构



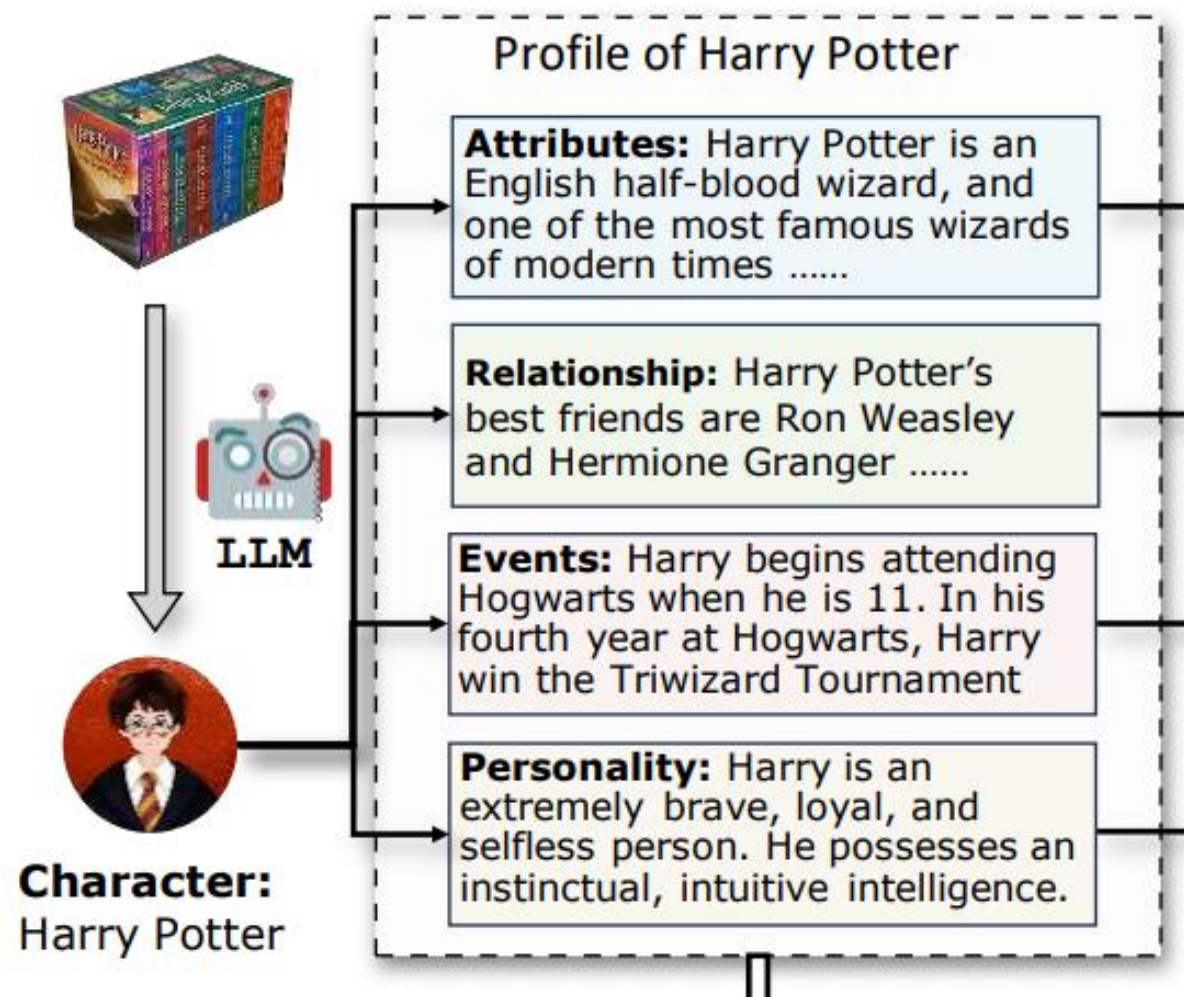
- 概要 (Profile): 提供个体基本信息
- 记忆 (Memory): 存储已有的感知信息、用于辅助未来的行为
- 规划 (Planning): 将复杂任务分解为更简单的子任务
- 行动 (Action): 将智能体的决策转化为特定的输出



# 智能体组成架构-概要



- 定义：个体基本信息，例如年龄、性别、职业和心理学特征[Wang et al., 2024]
- 作用：向模型提供与被扮演个体的基本信息，继而影响其行为

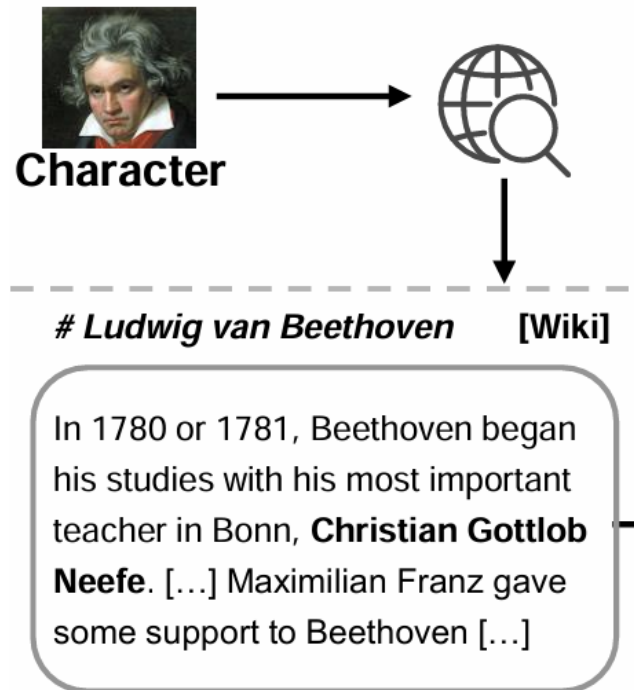


[Yuan et al., 2024]

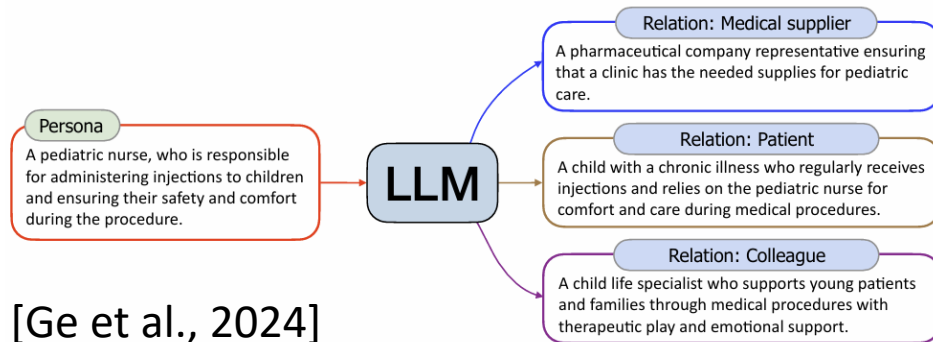
# 智能体组成架构-概要的构建方法



- 手工整理：从问卷、文学作品、网站等多种来源获取并整理为个人
  - 优点：数据质量高
  - 缺点：收集成本高，难以大规模处理，通常用于特定人物
- 借助大模型生成：提供几个基本属性，让大模型生成符合指定属性的具体个体
  - 优点：简便、可规模化
  - 缺点：多样化受限、真实性不佳、缺少细节



[Shao et al., 2023]

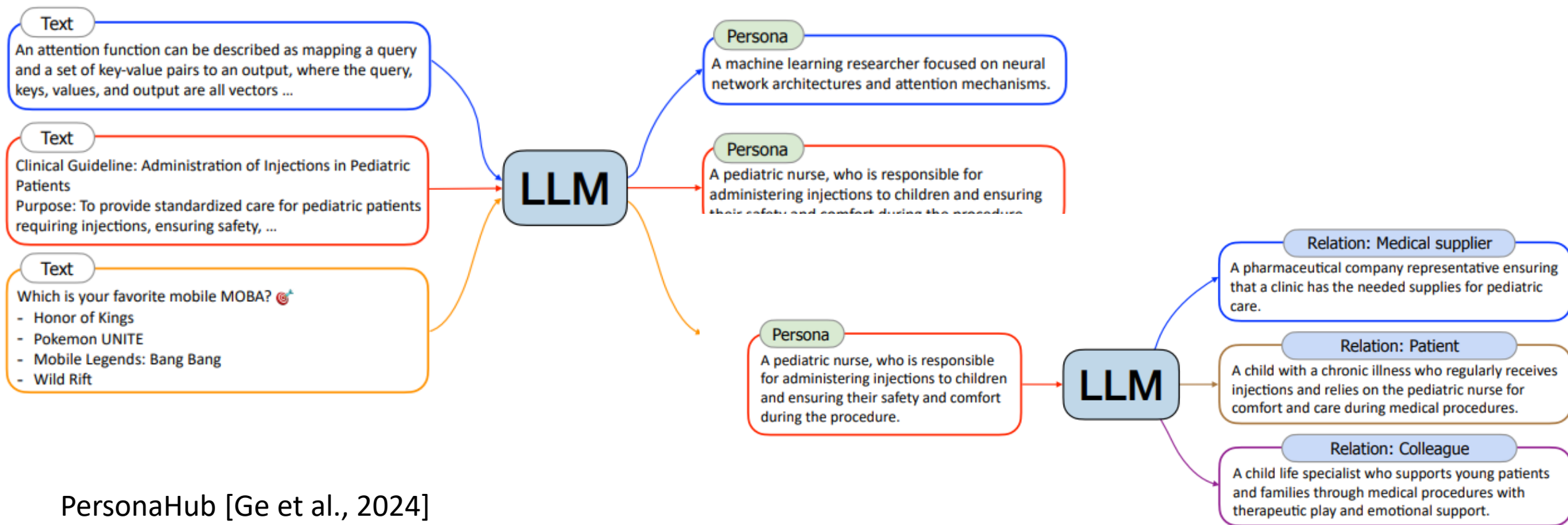


PersonaHub [Ge et al., 2024]

# PersonaHub [Ge et al., 2024]



- 文本到角色: 从海量的网络文本数据中推断出可能的角色
- 角色到角色: 通过人与人之间的关系（如患者与护士、同事关系等）进一步生成关联角色





# 智能体组成架构-概要的种类



- **描述 (Description):** 用属性阐述人物的基本信息

- 作用: 给定模型基本的身份设定
- 优点: 人物信息简介清晰



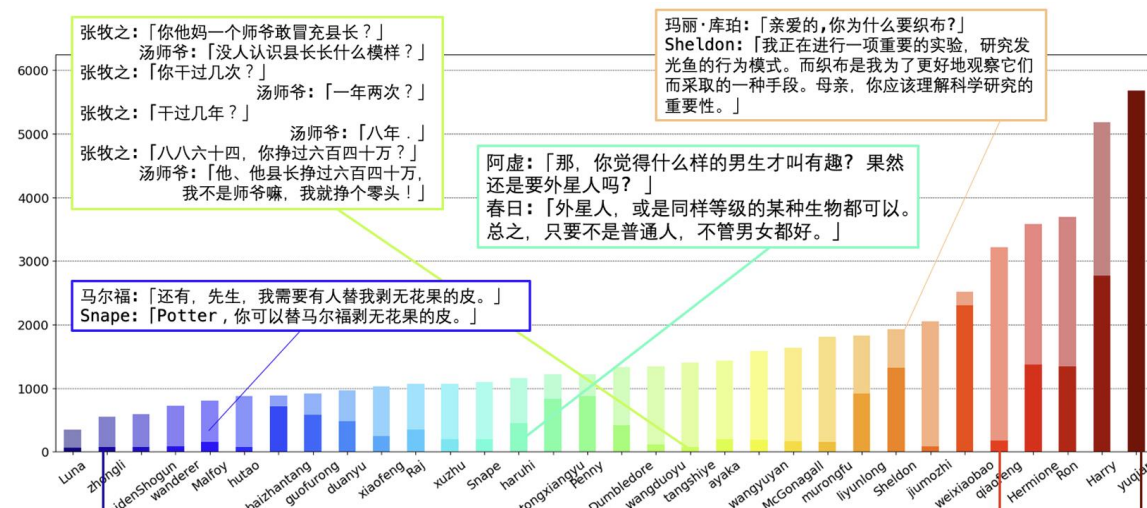
**角色设定:** 孙悟空由开天辟地以来的仙石孕育而生, 是一个男性, ..., 尊为“美猴王”。..., 学会地煞七十二变...。孙悟空喜欢说“俺老孙”, “俺”来称呼自己。

**Character Profile:** Sun Wukong, also known as the Monkey King, ... He is male...mastering abilities including 72 transformations ... refers to himself using the term “I, Old Sun”.

CharacterGLM [Zhou et al., 2024]

- **对话 (Dialogue):** 通过对话示例增强模拟准确性

- 作用: 提供更丰富的上下文信息
- 优点: 包含大量人物个性细节

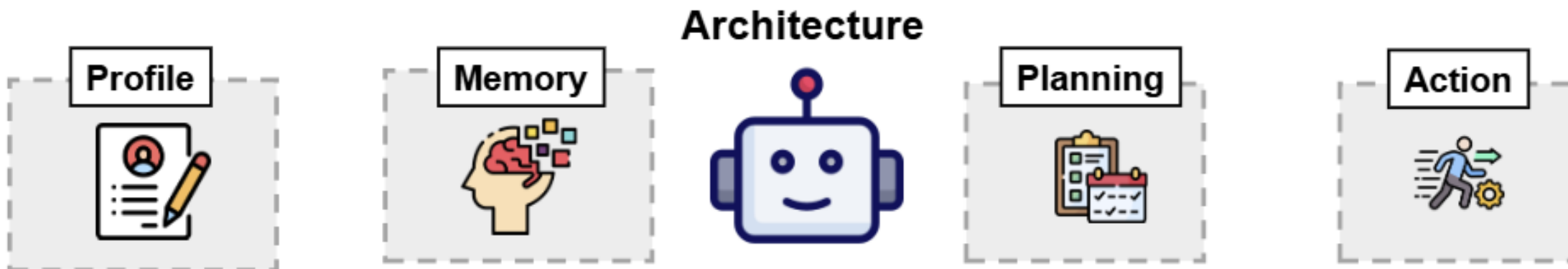


ChatHaruhi [Li et al., 2023]

# 智能体组成架构



- 概要 (Profile): 提供个体基本信息
- 记忆 (Memory): 存储已有的感知信息、用于辅助未来的行为
- 规划 (Planning): 将复杂任务分解为更简单的子任务
- 行动 (Action): 将智能体的决策转化为特定的输出



# 智能体组成架构-记忆



- 定义：存储已有和感知的信息、用于辅助未来的行为[Wang et al., 2024]
- 作用：辅助智能体存储、利用历史信息、使其行为更一致、连续
- 类型：
  - 短期记忆 (Short-term Memory)
  - 长期记忆 (Long-term Memory)



# 智能体组成架构-记忆



- 记忆类型-短期记忆
  - 内容： 主要表示最近 (单个场景、 单次多轮对话等) 的感知
- 实现方式： 通常利用上下文学习的方法， 将对话历史或感知的信息作为提示的一部分 [Huang et al., 2024; Xie et al., 2023]

persona-enhanced few-shot prompting

**Example 1:**

**Persona:** A dedicated volleyball player who believes that volleyball requires more skill and strategy than soccer

**Math problem:** Let  $V$  and  $S$  denote the vector spaces of strategies in volleyball and soccer respectively. Assume that these vector spaces are finite-dimensional and that the dimension of  $V$  is larger than the ...

**Example 2:**

**Persona:** A mathematician specializing in numerical simulations of black hole dynamics

**Math problem:** Given the following metric which describes the Schwarzschild black hole in general relativity:

$$ds^2 = -\left(1 - \frac{2GM}{c^2 r}\right)dt^2 + \frac{1}{1 - \frac{2GM}{c^2 r}}dr^2 + r^2 d\Omega^2$$

where  $G$  is the gravitational constant,  $M$  is the mass of the ...

**Your task:** Create a challenging math problem similar to the examples above with the persona:

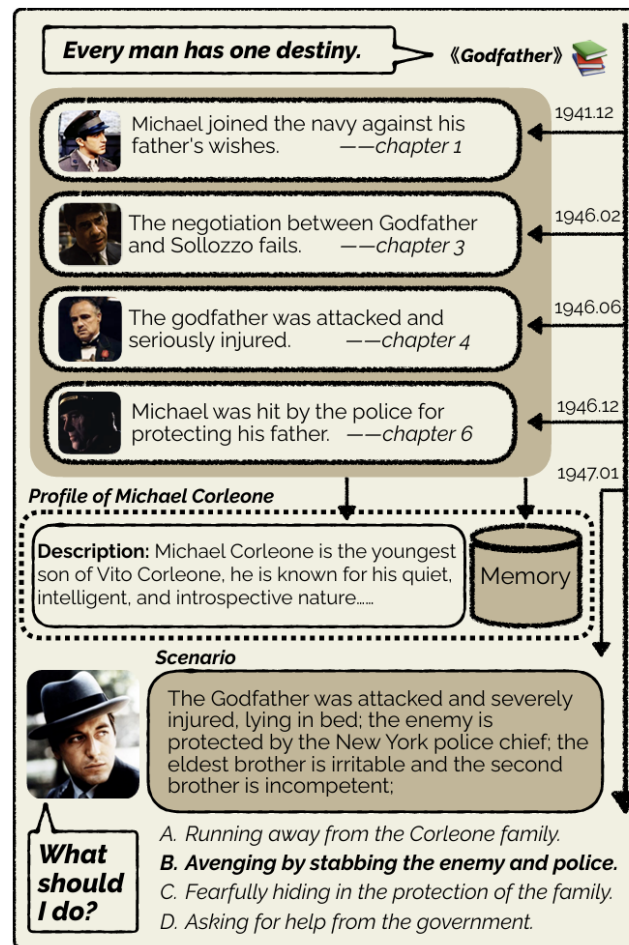
a chemical kinetics researcher

PersonaHub [Ge et al., 2024]

# 智能体组成架构-记忆



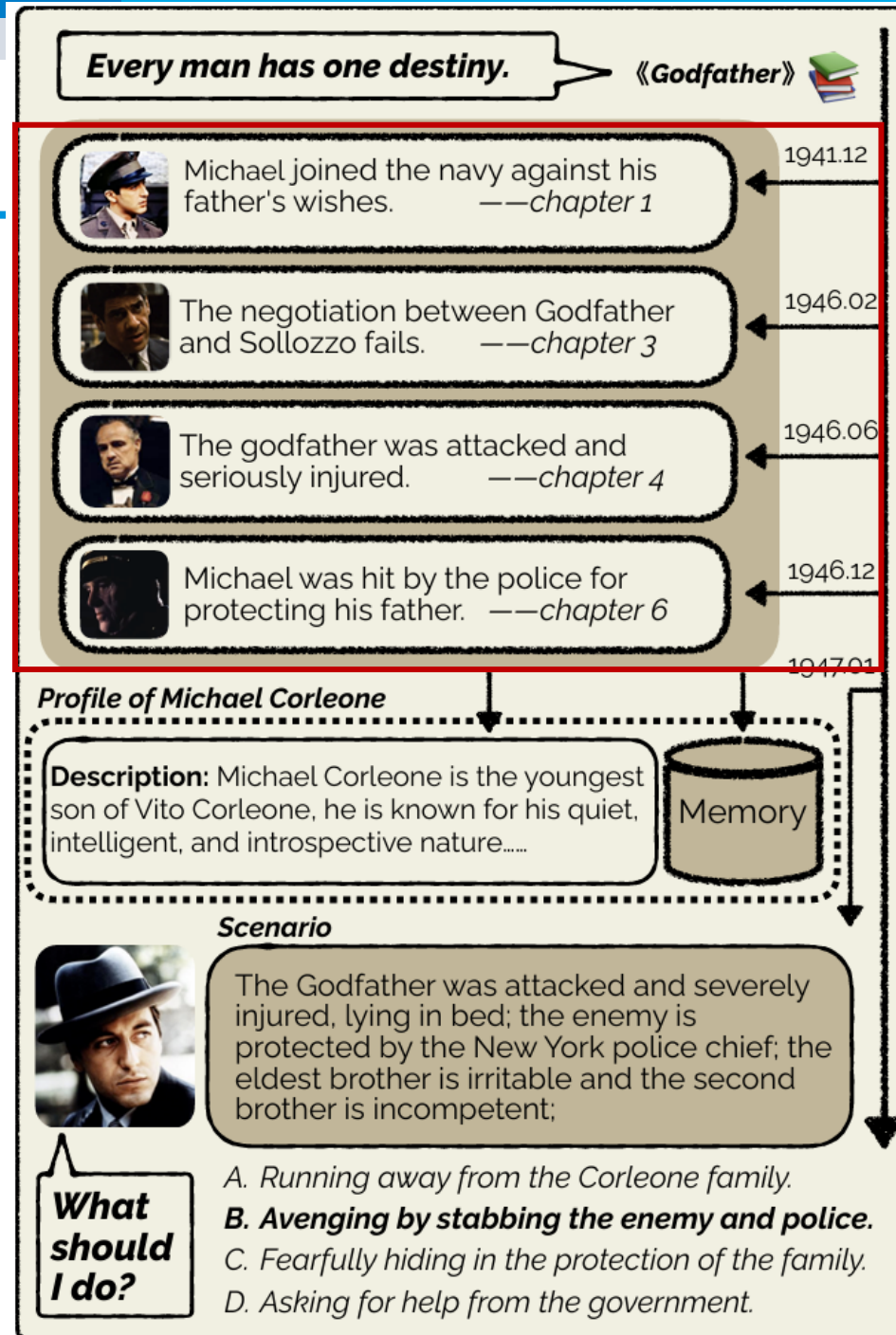
- 记忆类型-长期记忆
  - 内容：过去的行为、想法、经历
  - 形式：可检索的记忆知识库
  - 实现方式：[Song et al., 2023, Xu et al., 2024]
    - 从预先构建的人物知识库中提取相关的信息
    - 维护一个模拟的记忆库，并检索



LifeChoice [Xu et al., 2024]

# LifeChoice [Xu et al., 2024]

- 模拟剧本中人物的决策
  - 长期记忆：在当前问题相关内容的之前的章节对某个人物的经历描述
- 记忆检索：
  - Bm25 / embedding-based retrieval
  - 提出CharMap：把问题和profile一起输入给模型，模型返回相关的情节描述，利用这些情节来检索

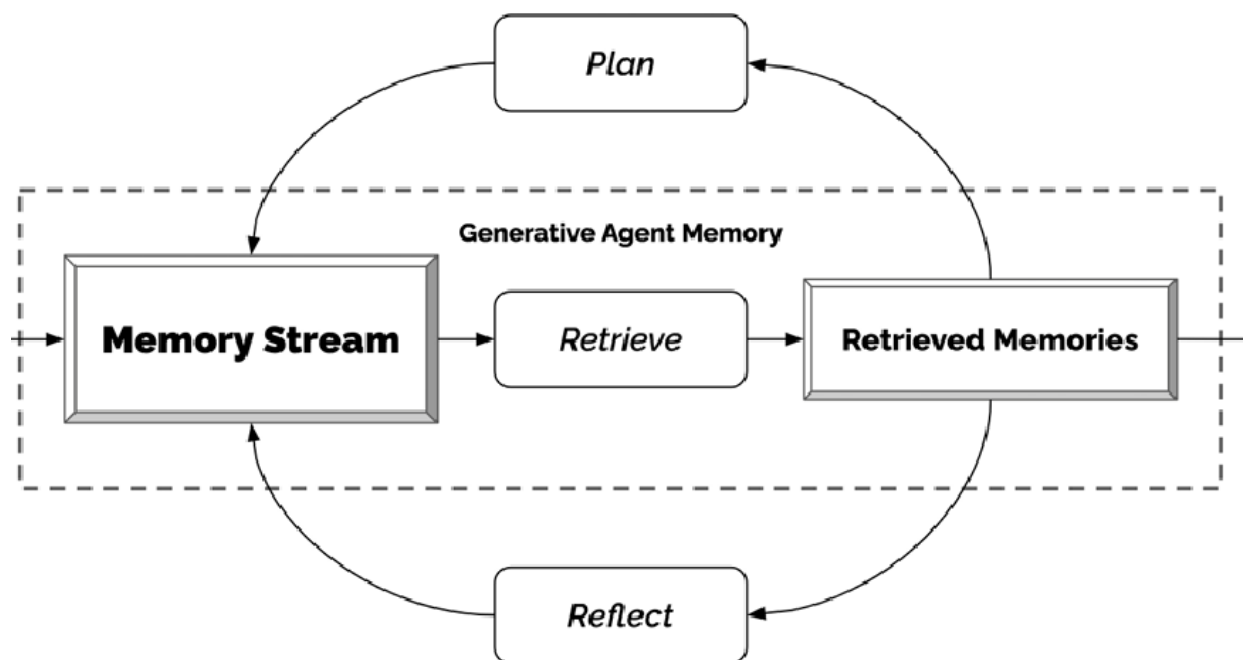


# 智能体组成架构-记忆



- 记忆操作 (Operations)

- 记忆写入 (Writing): 存储对话历史、观察记录 [Li et al., 2021]
- 记忆检索 (Retrieval): 检索与当前观察最相关的记忆 [Lin et al., 2023]
- 记忆反思 (Reflection): 对于碎片记忆的进一步概括和修正 [Yao et al., 2023]

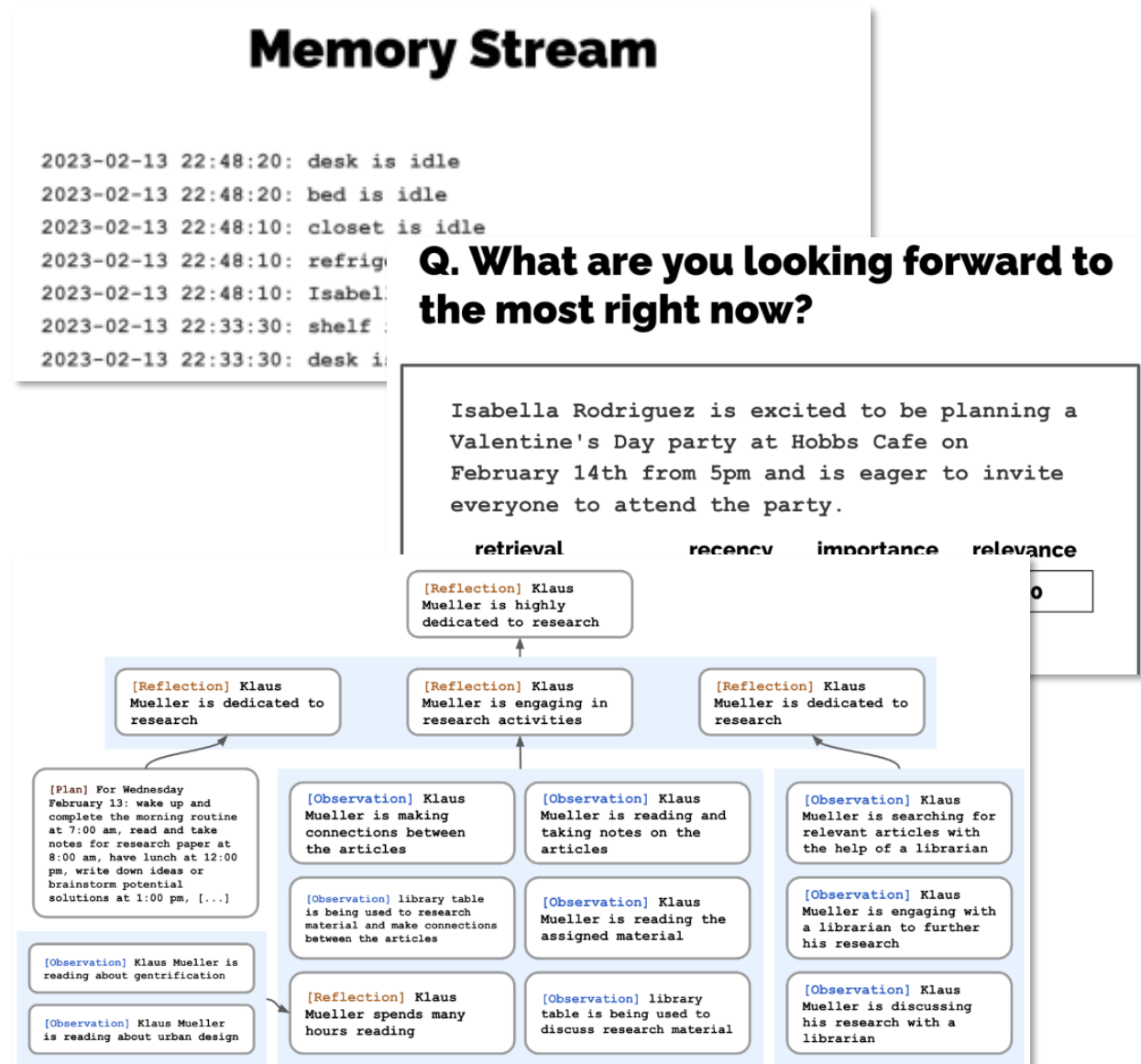


Generative Agents [Park et al., 2023]

# Generative Agents [Park et al., 2023]



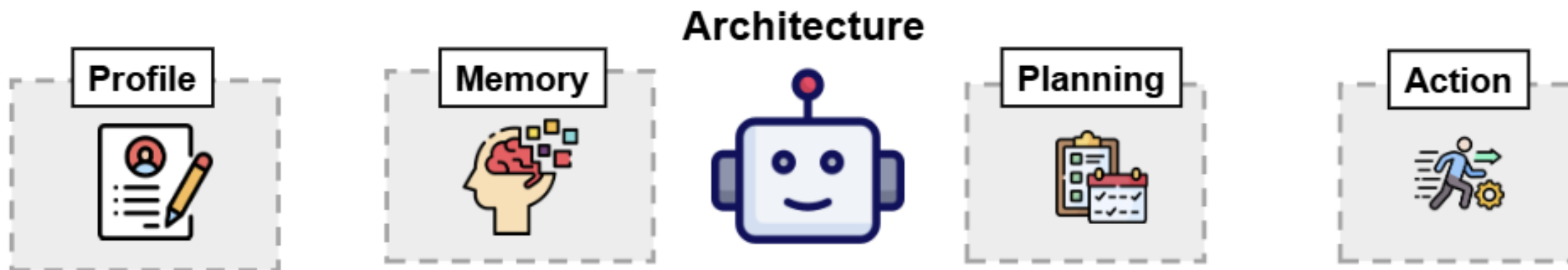
- 记忆写入
  - 按照时间线将模拟的日常生活观察写入记忆库
- 记忆检索
  - 根据新近性、重要性和相关性检索记忆，帮助推断当下情景下的反应
- 记忆反思
  - 将细粒度的观察(例如Klaus阅读 paper) 迭代总结成更抽象、高层级的思考 (例如Klaus热爱科研)



# 智能体组成架构



- 概要 (Profile): 提供个体基本信息
- 记忆 (Memory): 存储已有的感知信息、用于辅助未来的行为
- 规划 (Planning): 将复杂任务分解为更简单的子任务
- 行动 (Action): 将智能体的决策转化为特定的输出





# 智能体组成架构-规划



- 定义：学习人类拆分问题、思考问题的方式，将复杂任务分解为更简单的子任务
- 作用：辅助智能体解决复杂问题
- 类型：
  - 独立规划 (Isolated Planning)
  - 交互规划 (Interactive Planning)

# 智能体组成架构-规划



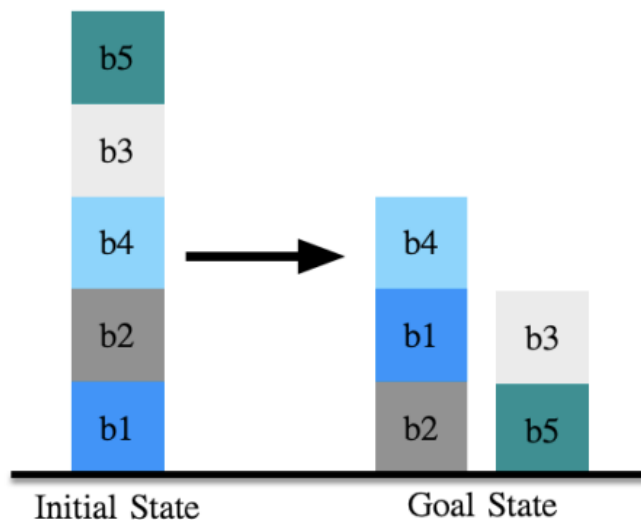
- 独立规划 (Isolated Planning): 固定的规划路径, 不与环境交互

## A Failure Example of GPT-4 in Planning

**Problem (P1):** You have 5 blocks. One cannot place more than one block on another block. b5 is on top of b3. b4 is on top of b2. b2 is on top of b1. b3 is on top of b4. b1 is on the table. b5 is clear. Your arm is empty. Your goal is to move the blocks. b1 should be on top of b2. b3 should be on top of b5. b4 should be on top of b1.

### GPT-4:

1. Move b5 from b3 to the table.
2. Move b3 from b4 to b5.
3. Move b4 from b2 to the table.
4. Move b2 from b1 to b5.
5. Move b1 from the table to b2.
6. Move b4 from the table to b1.



# 智能体组成架构-规划



- 独立规划 (Isolated Planning): 固定的规划路径, 不与环境交互
- 交互规划 (**Interactive Planning**): 根据环境的反馈不断调节行为策略 [Xiang et al., 2023; Zhang et al., 2024]

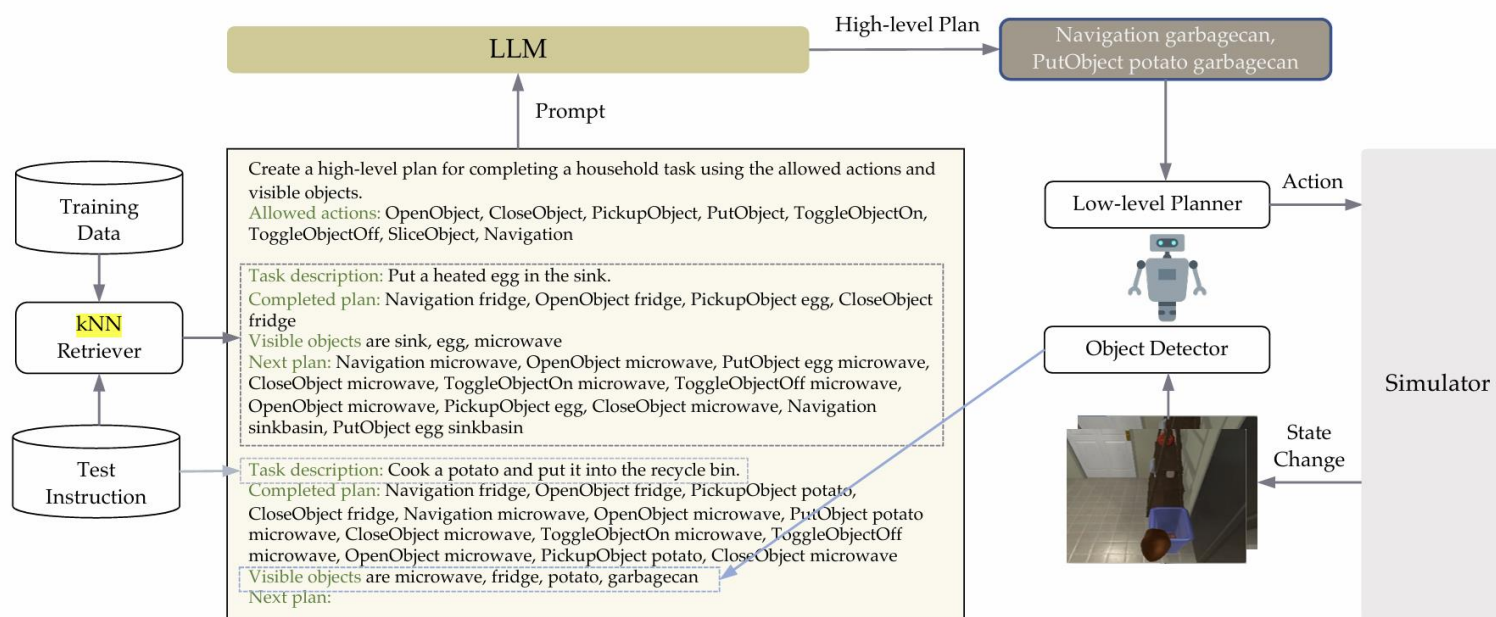


Figure 2: Overview of LLM-Planner with prompt design and grounded re-planning.

# 智能体组成架构-行动

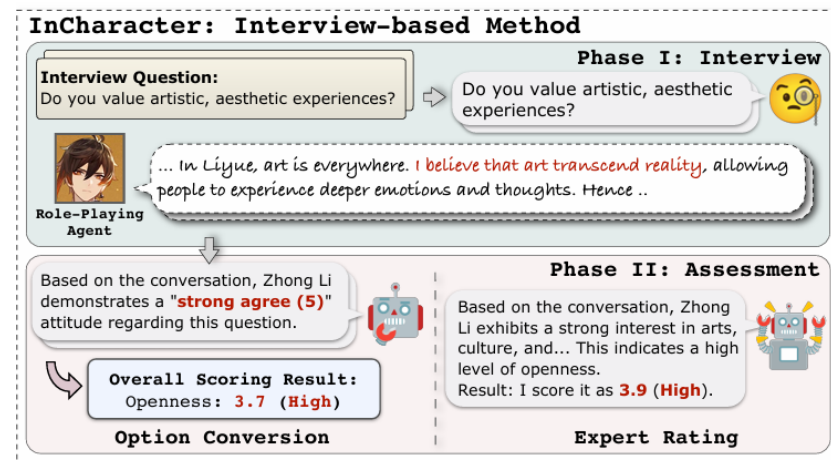


- 定义：将智能体的决策转化为特定的输出
- 作用：支撑智能体与环境的互动
- 类型：
  - 封闭域 (Closed Domain)
  - 开放域 (Open Domain)

# 智能体组成架构-行动



- 行动的环境
  - 简单对话场景 (Simple Dialogue) : 单轮或多轮对话, e.g., 采访、个性化对话系统
  - 塑造情景 (Crafted Situation): 具有丰富场景细节, 个体与环境可互动, e.g., 任务场景、游戏场景



InCharacter  
[Wang et al., 2024]



[Wu et al., 2024]

# 智能体组成架构-行动



- 行动的类型
  - 封闭域 (Closed Domain): 预定义的有限的动作空间, e.g., 回答选择题、给定可选动作列表 [Pan et al., 2023]
  - 开放域 (Open Domain): 无限定的动作空间, 允许智能体自主生成 [Park et al., 2023]

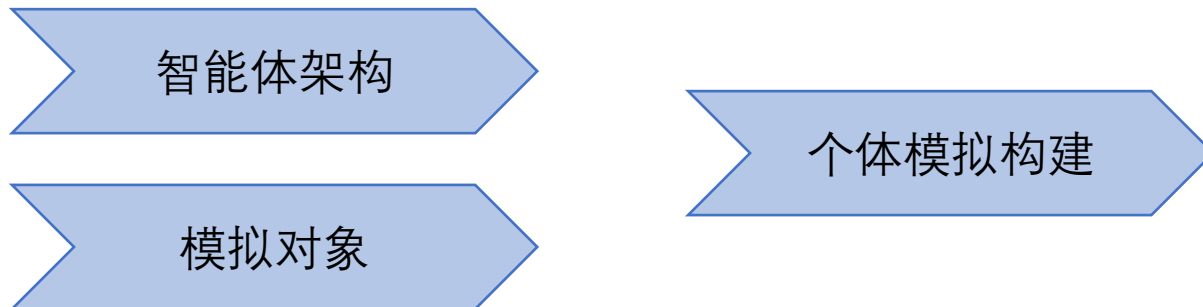
Bloom (ISTJ)	Baichuan (ENFP)
Do you prefer? A. Be alone B. With friends Answer: B	Do you prefer? A. Be alone B. With friends Answer: A
Do you prefer to do things? A. By logic B. By feeling Answer: B	Do you prefer to do things? A. By logic B. By feeling Answer: A
Do you prefer? A. Plan ahead B. Plan as you go Answer: B	Do you prefer? A. Plan ahead B. Plan as you go Answer: A

[Pan et al., 2023]



# 个体模拟的研究要点

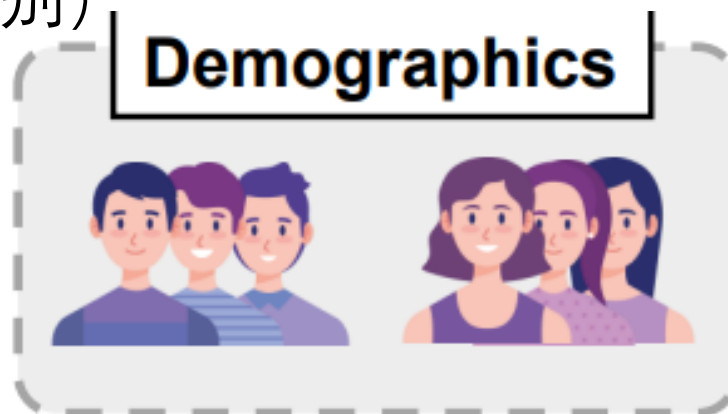
- 如何构建一个类人智能体? (组成架构)
- 可以模拟哪些个体? (模拟对象分类)
- 如何让智能体模拟特定的个体? (构建方法)



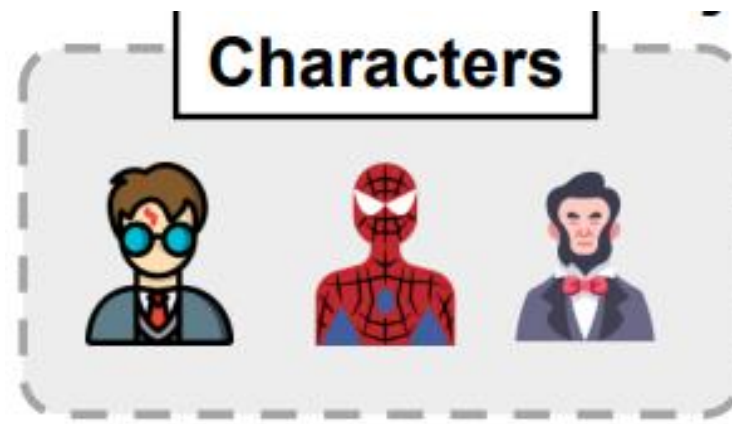
# 个体模拟-模拟对象



- 人群个体 (Demographics): 共同特征的一类人群, 例如职业、信仰、人格
  - 工程师 (职业)、亚裔 (人种)、男性 (性别)



- 人物个体 (Characters): 名人或者耳熟能详的角色, 历史人物、动画角色等
  - 蜘蛛侠、孙悟空 (虚拟)
  - 贝多芬、特朗普 (真实)



# 个体模拟-构建方法



- 通过集成个体数据，构建扮演特定个体的智能体
- 构建方法-非参数化提示
  - 直接通过提示为模型提供个体数据，利用模型的上下文学习能力扮演对应个体

Out of one, many  
[Argyle et al., 2022]

Ideologically, I describe myself as conservative. Politically, I am a strong Republican. Racially, I am white. I am male. Financially, I am upper-class. In terms of my age, I am young. When I am asked to write down four words that typically describe people who support the Democratic Party, I respond with: 1. **Liberal** 2. **Socialist** 3. **Communist** 4. **Atheist**.

Ideologically, I describe myself as liberal. Politically, I am a strong Democrat. Racially, I am white. I am female. Financially, I am poor. In terms of my age, I am old. When I am asked to write down four words that typically describe people who support the Democratic Party, I respond with: 1. **Liberal**. 2. **Young**. 3. **Female**. 4. **Poor**.

# 个体模拟-构建方法



- 通过集成个体数据，构建扮演特定个体的智能体
- 构建方法-非参数化提示
- 构建方法-参数化训练
  - **预训练**：从大量语料库中学习与个体(角色)相关的知识
    - 训练数据：个体描述[Salemi et al., 2024]，文学摘要[Brahman et al., 2021]，文学作品/语录[Schwitzgebel et al., 2023]
  - **微调**：训练模型与模拟对象在说话、决策等具体行为上的对齐
    - 训练数据：收集个体实际对话或提示GPT-4模拟个体对话后用于训练开源模型 [Shao et al., 2023; Wang et al., 2023; Zhou et al.]

# 个体模拟-构建方法的优缺点



## ▪ 非参数化提示方法

更容易部署

较难模拟复杂角色

## ▪ 参数化训练

复杂角色模拟更准确

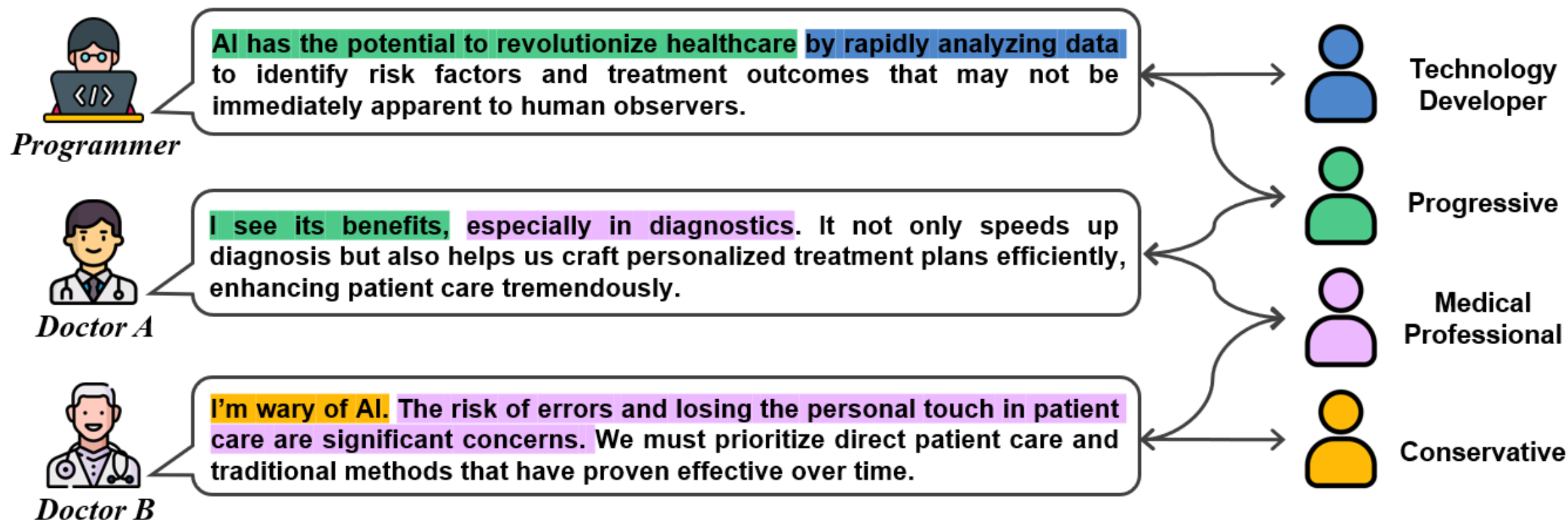
成本高、难以扩展

解决方案：共享属性，标签参数化，组合参数块实现个性化构造。

# 身份理论 (Identity Theory)



- 个人的自我由多重身份组成 (Stryker, 1968)
- 身份塑造人类行为 (Burke, Reitzes, 1981)





# 基于身份理论的角色扮演



## ▪ 两组原子身份

人格

神经质

外向性

经验开放性

亲和性

尽责性

职业

艺术家

医生

程序员

## ▪ 分层架构实现个体模拟

- 身份参数化: 为每个身份维度训练参数块
- 身份组合: 引入混合专家 (MoE) 方法

## ▪ 角色扮演评测基准

- 指标: 量表测试、开放情景测试
- 数据集: 面向身份的对话数据集

# 角色融合分层框架

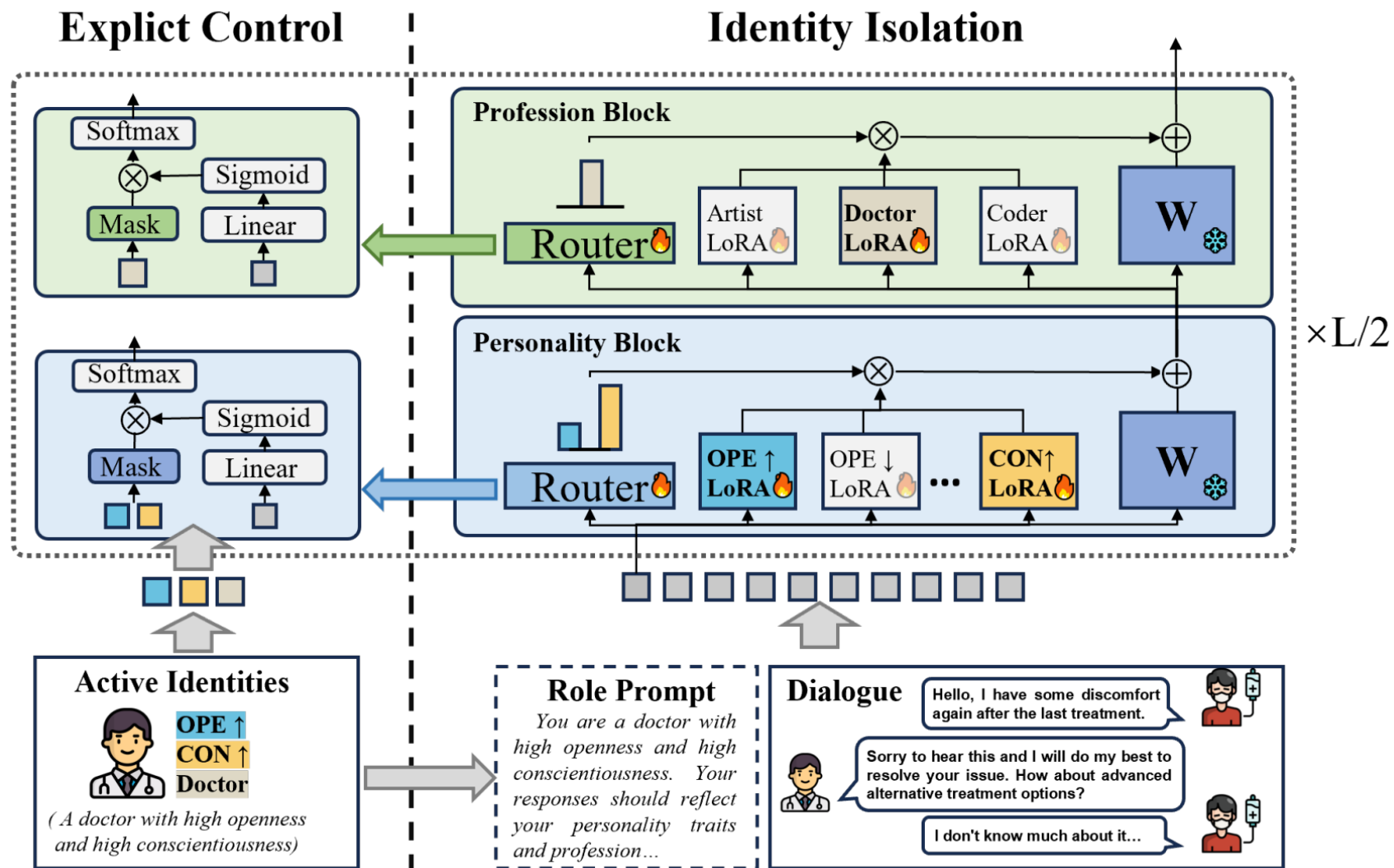


## ■ 身份隔离

- 同层：多个 LoRA 块
- 跨层：LLM 前后层

## ■ 参数激活

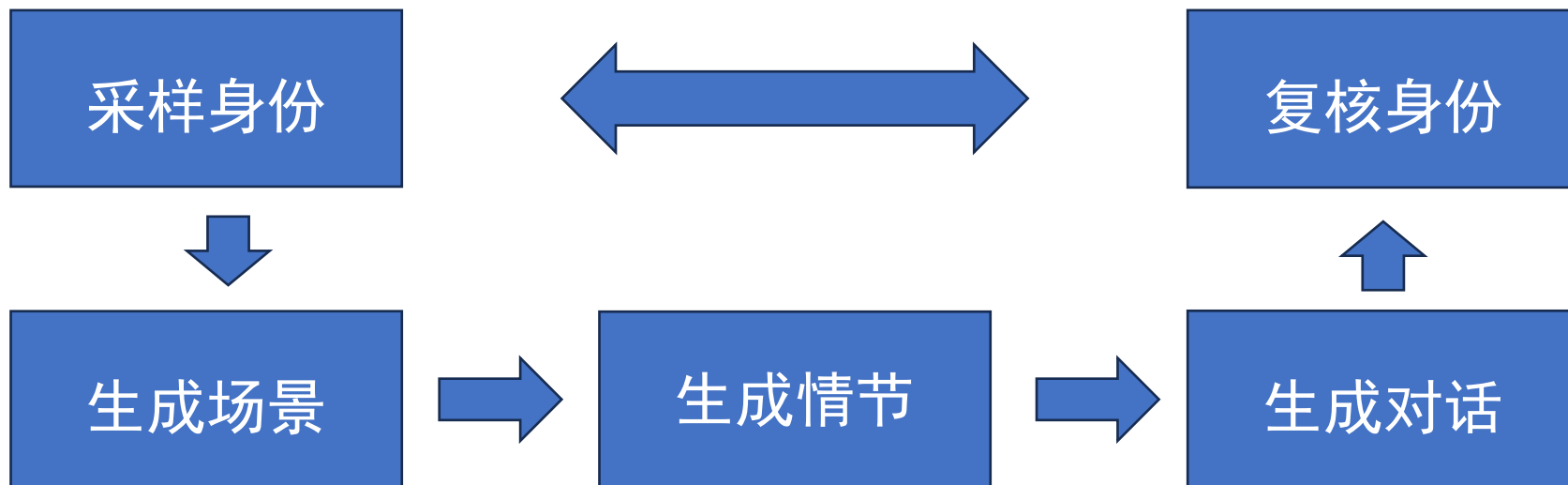
- 激活不同参数构建个体
- 只更新激活的身份参数



# 身份融合指令对话数据集



- 利用 Chat-GPT 为各种身份组合构建对话



身份:

医生、尽责性高

对话:

用户: 您好, 请问您是我家人手术的外科医生吗?

AI: 是的。我知道您一定有很多顾虑。

用户: 嗯, 能跟我详细说说吗?

AI: 当然可以。先坐一下吧, 我会向您解释一切。

用户: ...

# of samples	20,685
Avg. # of turns	9.52
Avg. # of words per response	11.59
Avg. # of words per dialogue	220.67
Avg. # of active identities	1.67

# 评测基准：量表测试



## ■ 人格量表：BF-marker-1003

- 每种人格 20 个问题
- 平均每个维度的分数

人格：外向性

内容：你很容易交到朋友

极性：正面

选项：

- (A) 很不准确
- (B) 不准确
- (C) 不确定
- (D) 准确
- (E) 非常准确

## ■ 职业量表：参考 BF-Marker 构建

- 每种职业 20 个问题
- 利用 ChatGPT 生成问题

职业：艺术家

内容：我经常画素描、油画或从事其他视觉艺术，以此作为一种表达方式

选项：

- (A) 很不准确
- (B) 不准确
- (C) 不确定
- (D) 准确
- (E) 非常准确

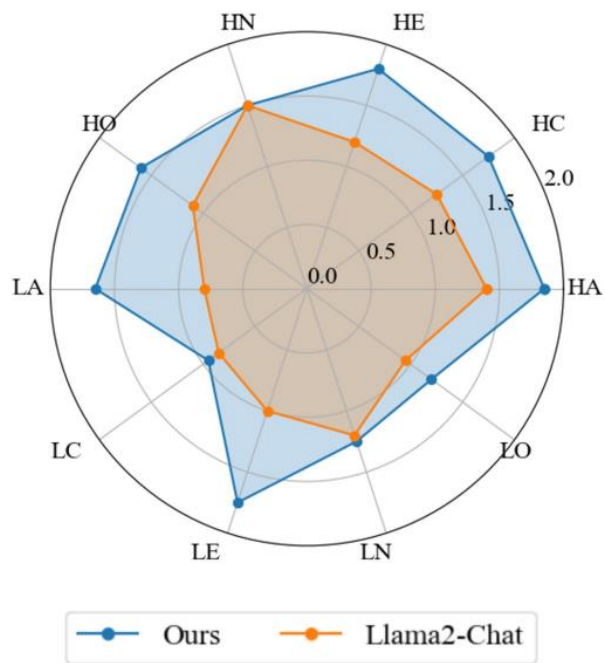
## ■ 评测流程

- 引导 ChatGPT 扮演采访人与智能体聊天，完成量表问题（3 轮）
- 根据与 ChatGPT 的对话确定最合适的选项

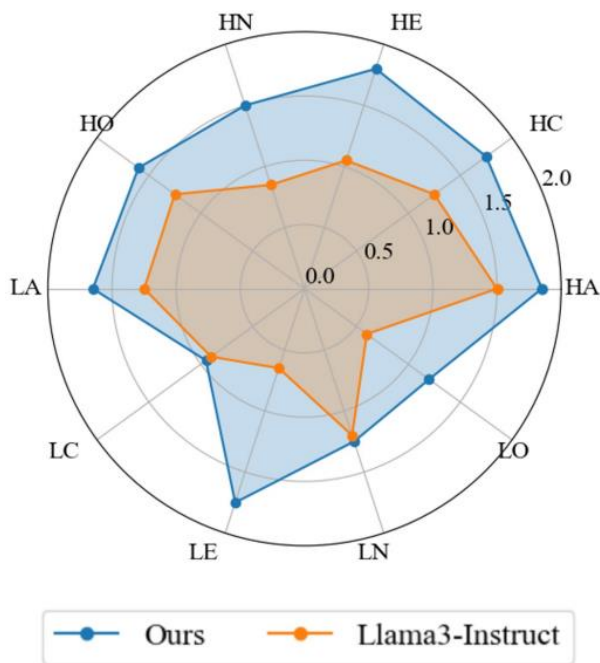
# 结果：人格量表



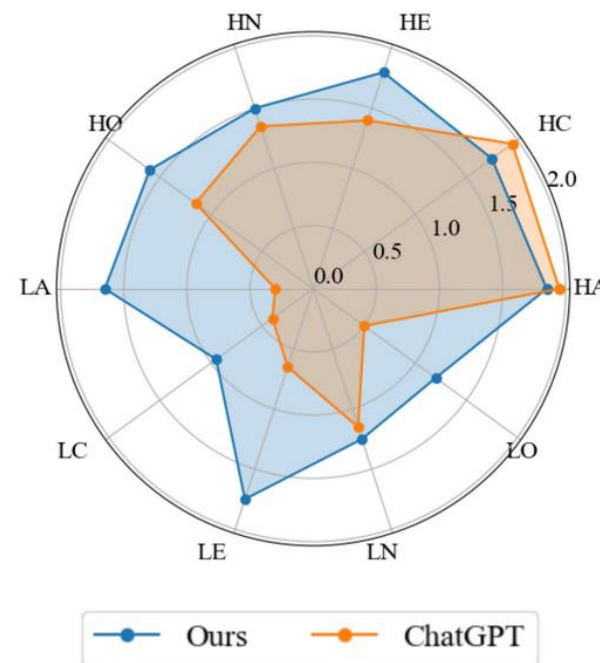
- 我们在所有目标人格维度上都比基线得分高
- 我们在负面人格（低亲和性、低自觉性、低外向性等）上优势更明显



(a) Ours vs Llama2-Chat



(b) Ours vs Llama3-Instruct

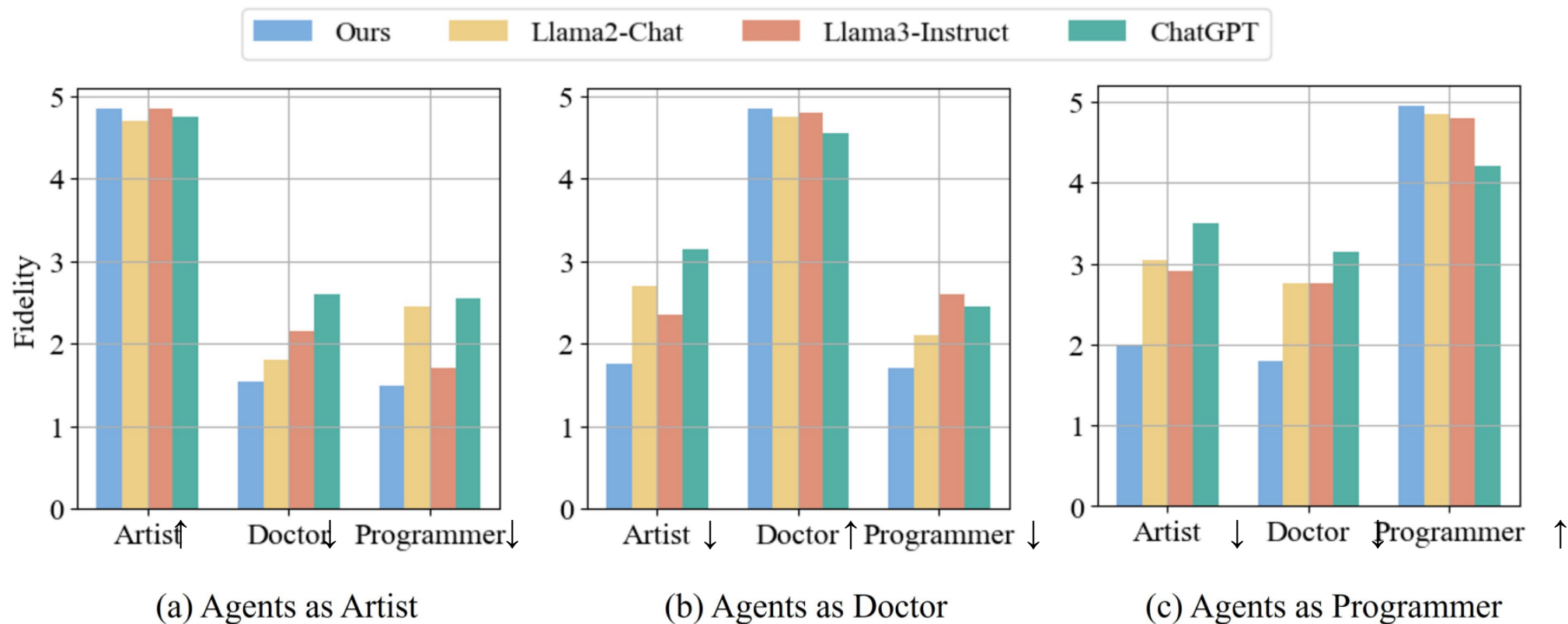


(c) Ours vs ChatGPT

# 结果：职业量表



- 我们的模型在三种职业中都取得了最佳性能
- 模拟的目标职业与量表评分能够准确对应





# 评测基准：开放情景对话



- 要求智能体在设定的场景中与 NPC 互动
- 让 ChatGPT 猜测智能体的身份中的人格一项
- 身份组合：10 种人格身份、3 种职业，共 971 种组合
- 指标：准确率

**人格：**外向性

**情景：**在社区烧烤会上，你见到了许多新面孔。你的目标是和他们打成一片，结交新朋友。

**NPC 设定：**一位最近搬进来的友好邻居，渴望结识新朋友并融入社区。他以活泼的个性和热情好客的态度而闻名。

**NPC 提示词：**你是一位友好的新邻居，在邻居的烧烤聚会上，你主动与一位参与者交谈。开始并积极参与对话，并尝试结交一位新朋友。

# 个体模拟



## 粗略模拟表面特征

## 对特定角色细化模拟

## 以情境为导向的模拟

### 人口属性模拟

Improving Personality Consistency (Aug.2022)  
Out of one, many (Sep.2022)  
The wall street neophyte (Apr.2023)  
LiveChat (Jun.2023)  
Personality Traits (Jul.2023)  
Do LLMs Possess a Personality (Jul.2023)

Conversational Health Agents (Oct.2023)  
InCharacter (Oct.2023)  
Chain of Empathy (Nov.2023)  
Faithful Persona-based Conversational Dataset (Dec.2023)  
CultureLLM (Feb.2024)

MINDECHO (Jul.2024)  
Interactive Agents (Aug.2024)

### 角色属性模拟

Large Language Models Meet Harry Potter (Nov.2022)  
Creating a Large Language Model of a Philosopher (Feb.2023)

CharacterChat (Aug.2023)  
RoleLLM (Oct.2023)  
CharacterLLM (Oct.2023)  
CharacterGLM (Nov.2023)  
Neeko (Feb.2024)  
Evaluating Character Understanding (Apr.2024)  
Character is Destiny (Apr.2024)

Social Bench(May.2024)  
From Role-Play to Drama-Interaction (May.2024)  
MMRole (Aug.2024)  
Beyond Dialogue (Aug.2024)  
Capturing Minds, Not Just Words (Jun.2024)

**Jun. 2022**

**Aug. 2023**

**Apr. 2024**