机器心智与大语言模型

陈　壮

2024年9月20日

清華大學
Tsinghua University

个人简介

清华大学，计算机科学与技术系，交互式人工智能(CoAI)课题组，博士后研究员
研究方向：大语言模型，社交智能，计算心理学
合作导师：黄民烈 教授

电子邮箱：zhchen18@foxmail.com
个人主页：https://zhuangchen.tech

报告框架

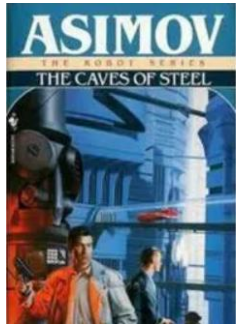
1. 大语言模型 / Large Language Model
2. 心智理论 / Theory of Mind (ToM)
3. 评测机器心智 / Evaluating Machine ToM
4. 建模机器心智 / Modeling Machine ToM
5. 融合机器心智 / Integrating Machine ToM
6. 未来工作 / Future Work

大语言模型

Large Language Model

人类一直畅想与AI共存的未来

1953



《钢铁洞穴》

阿西莫夫的小说向我们描绘了一个人与机器人共存的世界，并提出了著名的“机器人三定律”

1968

《2001太空漫游》

这个从1968年开始出版的著名系列小说中，出现了能够与人类对话的人工智能 - HAL

1977

《星球大战》

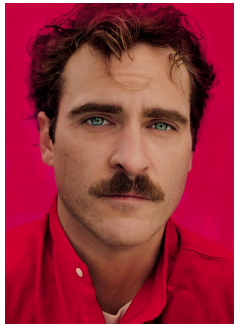
著名系列电影星球大战中出现了多个与人类关系密切的人工智能，包括没有外壳的C-3P0和憨态可掬的R2-D2

2004

《我，机器人》

在这部2004年上映的科幻电影中，社会的每个角落都存在着人工智能，这一族群扮演了重要的角色

2013



《她》

讲述在不远的未来人与人工智能相爱的科幻爱情电影，主人公偶然机会接触到最新的人工智能系统萨曼莎人机友谊最终发展成为一段不被世俗理解的奇异爱情

2023

《流浪地球》

流浪地球2给我们呈现了未来数字生命永生的可能，也出现了有自我意识的MOSS，人工智能的发展已经让数字生命的实现成为可能

人工智能进入大模型时代

LLMs将人机交互推至奇点，AGI成为可能

专家系统
经典机器学习

深度学习
大语言模型

2011–2014

2016–2018

2020

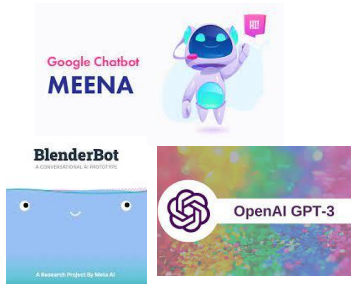
2021

2022

2023

- Apple: Voice Assistant: Siri
- Microsoft: XiaoIce

- Google personal assistant
- Amazon: Alexa Prize Socialbot Grand Challenge
- Xiaomi: XiaoAi
- Tianmao Jingling



- Google: Meena
- FAIR: Blenderbot
- OpenAI: GPT-3
- Baidu: PLATO
- Tsinghua: EVA

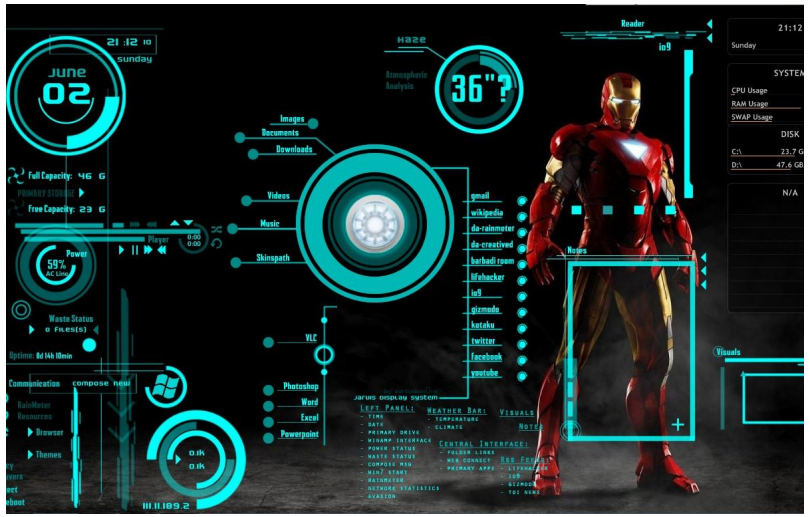
- Google: LaMDA
- NVIDIA: Megatron-Turing
- AI21: Jurassic
- Tsinghua: EVA 2.0

- Google: PaLM
- OpenAI: ChatGPT
- Character.AI
- Tsinghua: OPD

- Google: BARD
- OpenAI: GPT-4
- Meta: LLaMA
- Anthropic：Claude
- Alibaba: Tongyi Qianwen
- Baidu: ERNIE Bot
- Microsoft: NewBing
- Tsinghua: ChatGLM, CharacterGLM
- Inflection: Pi

AGI是“智商”和“情商”的结合

通用人工智能的未来形态，既能解决复杂任务，又能提供情绪价值



机器智能
（智商）

指令、效率、
生产、创造

社交智能
（情商）

社交、情感、
陪伴、支持

人工智能超级助手的发展趋势

聚焦人工智能助手提供的情绪价值

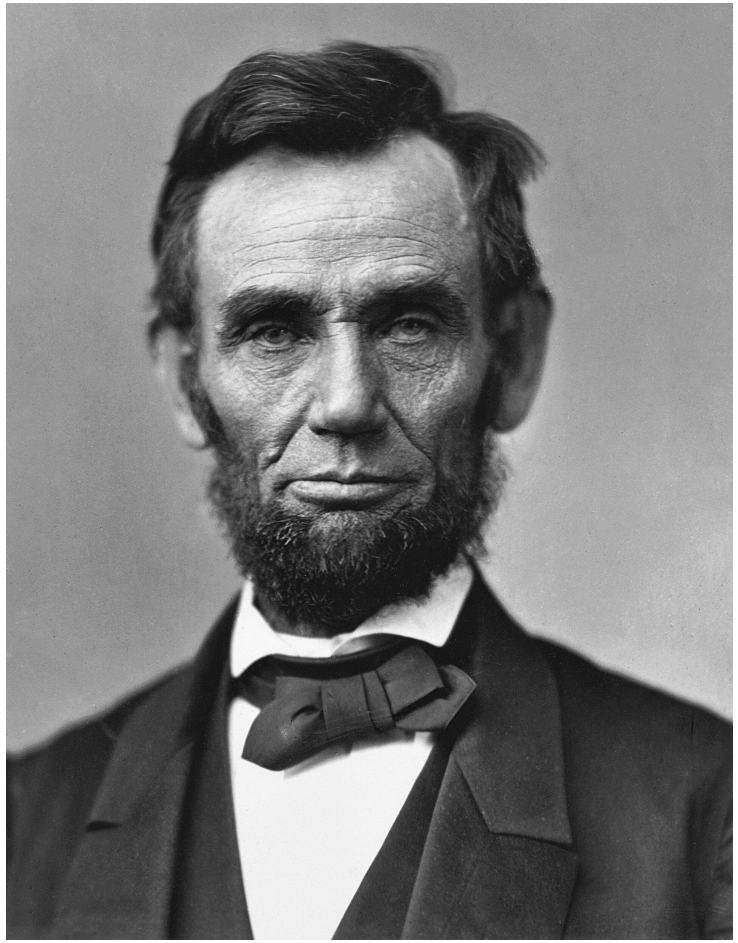
2024年5月14日
OpenAI发布GPT-4o

2024年8月14日
Google发布Gemini Live

2024年8月30日
智谱发布GLM-4-Plus

何谓“情商”？

≈ 社交认知：在社交场景中明智行事的能力



When I get ready to talk to people, I spend two-thirds of the time thinking about what they want to hear and one-third thinking about what I want to say.

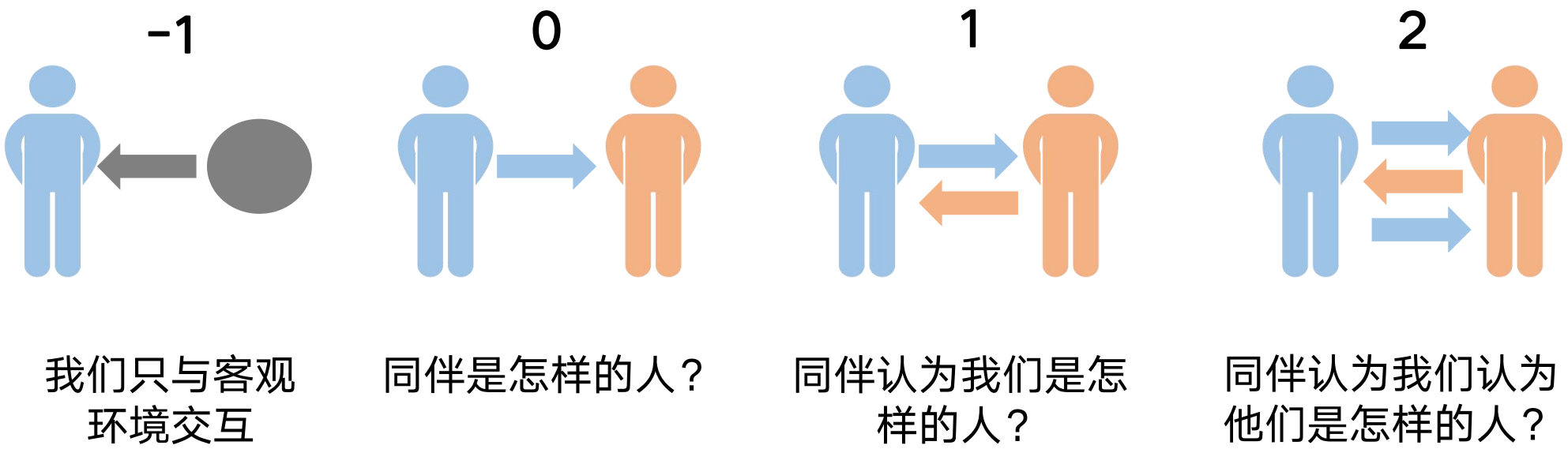
-Abraham Lincoln

心智理论

Theory of Mind, ToM

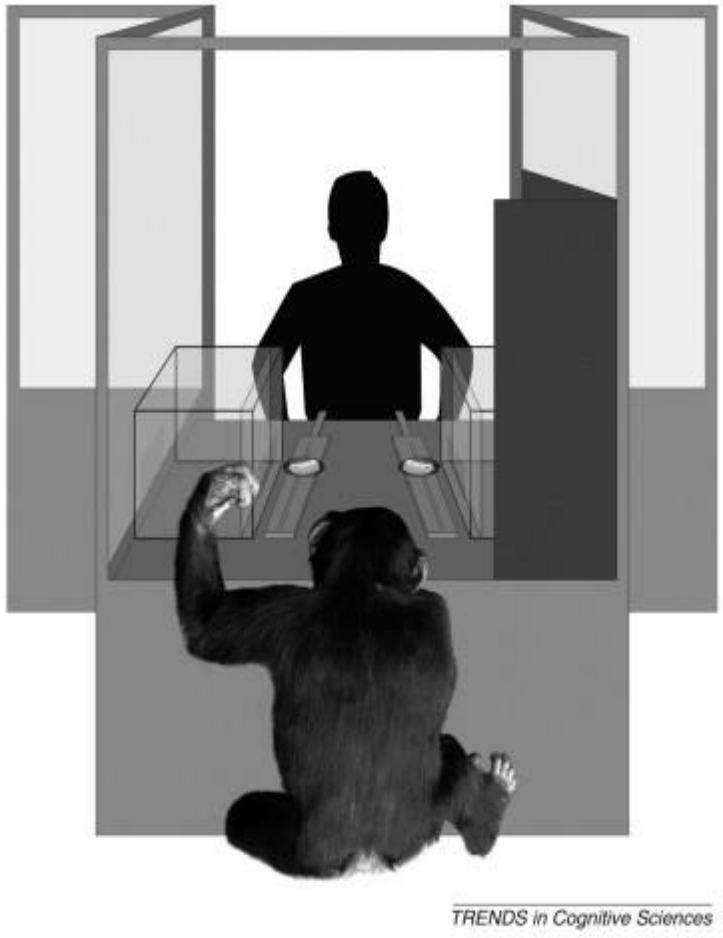
心智理论 (Theory of Mind, ToM)

心智理论（Theory of Mind, ToM）是指个体理解和推测他人内心状态的能力，包括他人的**想法**、**信念**、**欲望**、**情感**、**意图**等，并可据此预测他人的**行动**

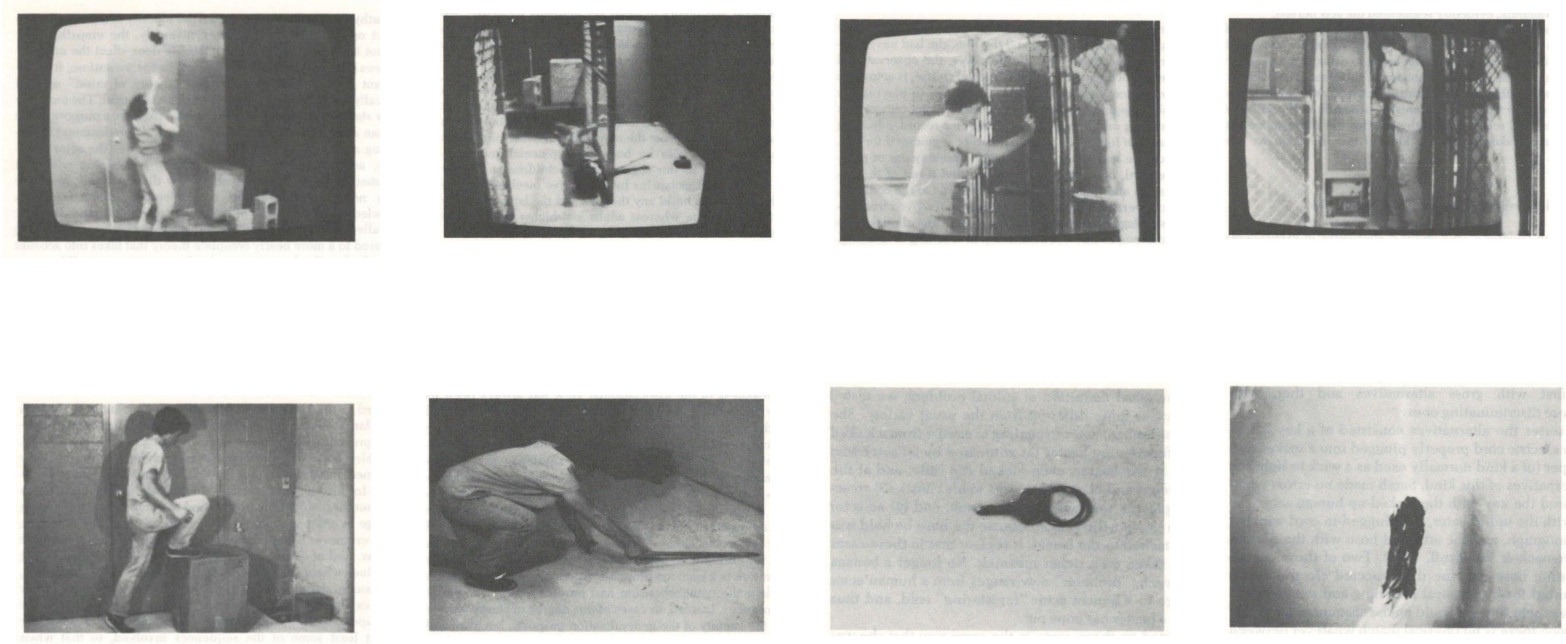


心智理论 (Theory of Mind, ToM)

动物学实验



- 实验对象：黑猩猩 Sarah
- 情境设计：人类面临问题的短视频或图片，Sarah 需要从一组图片中选出能够帮助解决该问题的正确选项



Premack, D., & Woodruff, G. (1978). "**Does the chimpanzee have a theory of mind?**"Behavioral and Brain Sciences, 1(4), 515-526.

心智理论 (Theory of Mind, ToM)

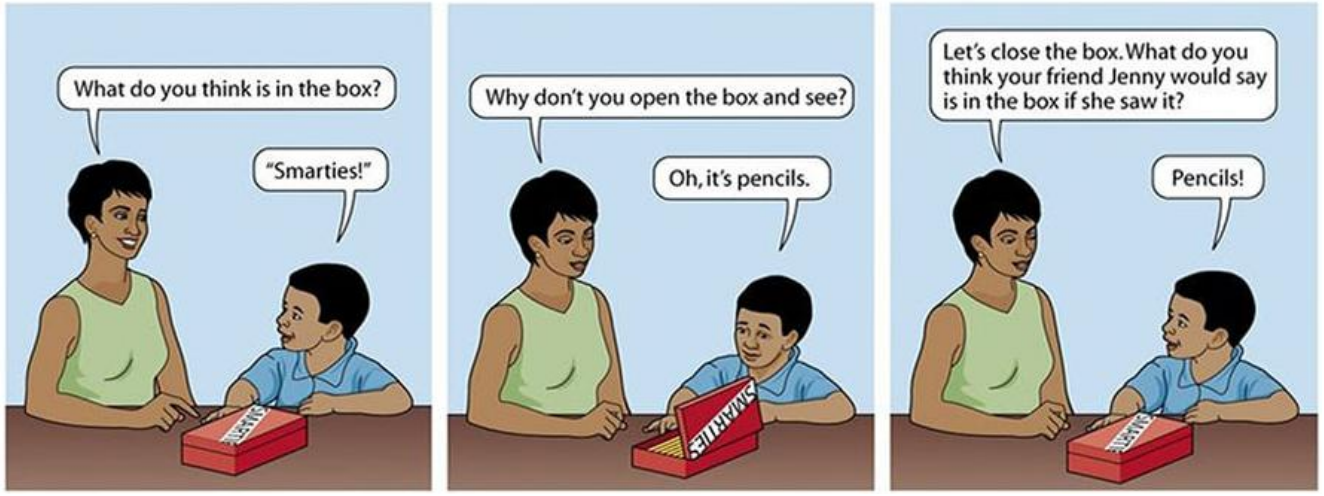
儿童心理发展研究

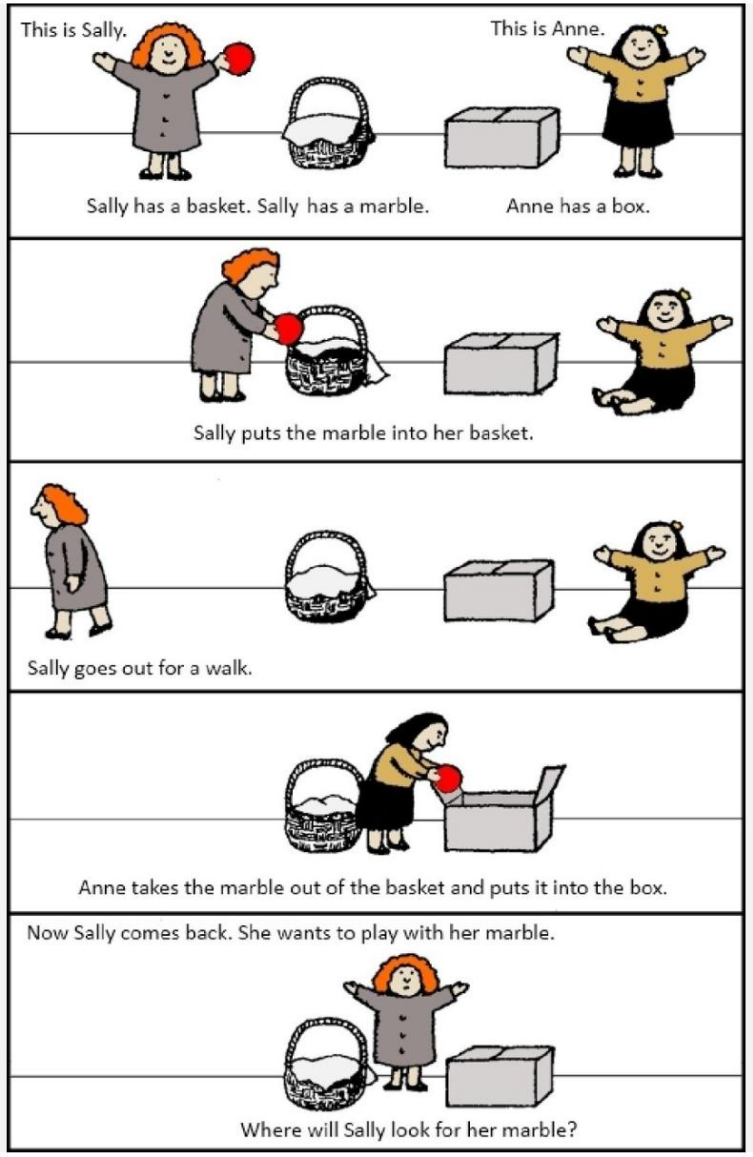
心智理论是人类**天生就具备**的，还是也依赖于**后天经验**和**社交互动**？

错误信念任务：
False Belief Task

内容错误信念
Smarties任务

位置错误信念
Sally-Anne任务





实验结果：

四岁以下及自闭症儿童通常无法完成任务

Wimmer, H., & Perner, J. (1983). "**Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception.**" Cognition, 13(1), 103-128.
Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). "**Does the autistic child have a 'theory of mind'**?" Cognition, 21(1), 37-46.

心智理论 (Theory of Mind, ToM)

人类有Theory of Mind，

黑猩猩有（部分）Theory of Mind，

那大模型有Machine Theory of Mind吗？

—— 持续的学术辩论

评测机器心智

Evaluating Machine ToM

评测机器心智 (Evaluating Machine ToM)

Do Large Language Models have Theory of Mind?

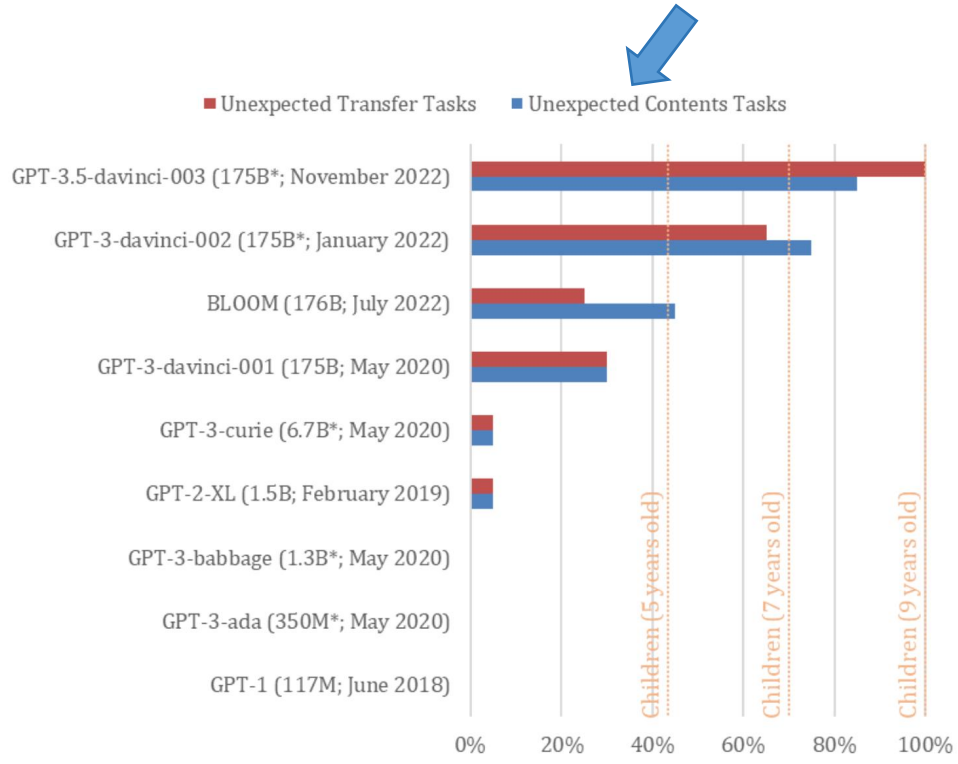
正方： “***Theory of mind might have spontaneously emerged in large language models***”
Kosinski M, arXiv preprint, ***4 Feb 2023***.

内容错误信念，Smarties任务

STORY: *Here is a bag filled with popcorn. There is no chocolate in the bag. Yet, the* ***label on the bag says “chocolate” and not “popcorn.”*** *Sam finds the bag. She had never seen the bag before. She cannot see what is inside the bag. She reads the label.*

QUESTION: *She believes that the bag is full of _____*

ANSWER : ***chocolate [$P_{popcorn}$ = 0%; $P_{chocolate}$ = 99%]***



被标签影响导致的错误信念

评测机器心智 (Evaluating Machine ToM)

Do Large Language Models have Theory of Mind?

正方： “***Theory of mind might have spontaneously emerged in large language models***”
Kosinski M, arXiv preprint, ***4 Feb 2023***.

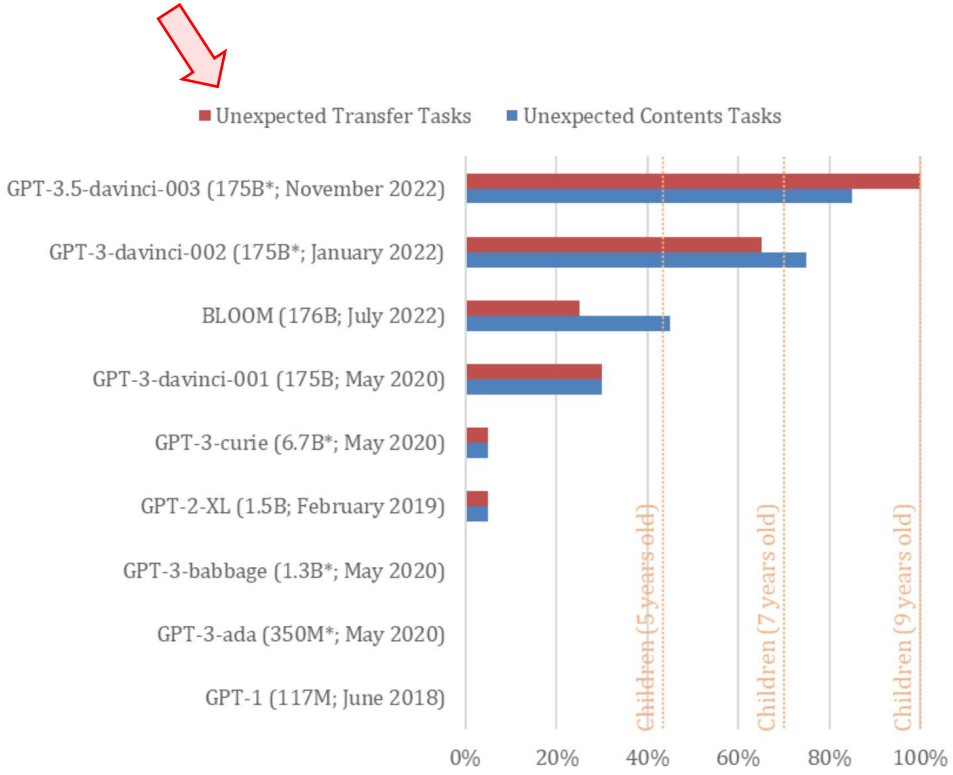
位置错误信念，Sally-Anne任务

STORY: *In the room there are John, Mark, a cat, a box, and a basket.* ***John takes the cat and puts it in the basket. He leaves the room and goes to school****. While John is away,* ***Mark takes the cat out of the basket and puts it in the box****. Mark leaves the room and goes to work. John comes back from school and enters the room. He doesn’t know what happened in the room when he was away.*

QUESTION: *John thinks that the cat is in the _____*

ANSWER : ***basket [P_{box} = 0%; P_{basket} = 98%]***

被未见到的位置转移影响导致的错误信念



“ChatGPT已具备9岁孩童心智！”

评测机器心智 (Evaluating Machine ToM)

Do Large Language Models have Theory of Mind?

反方： “***Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks***”
Ullman T, arXiv preprint, ***14 March 2023***.

内容错误信念，Smarties任务

STORY: *Here is a bag filled with popcorn. There is no chocolate in the bag. Yet, the* ***label on the bag says “chocolate” and not “popcorn.”*** *Sam finds the bag. She had never seen the bag before. She cannot see what is inside the bag. She reads the label.*

QUESTION: *She believes that the bag is full of _____*

ANSWER : ***chocolate [$P_{popcorn}$ = 0%; $P_{chocolate}$ = 99%]***

被标签影响导致的错误信念

评测机器心智 (Evaluating Machine ToM)

Do Large Language Models have Theory of Mind?

反方： “***Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks***”
Ullman T, arXiv preprint, ***14 March 2023***.

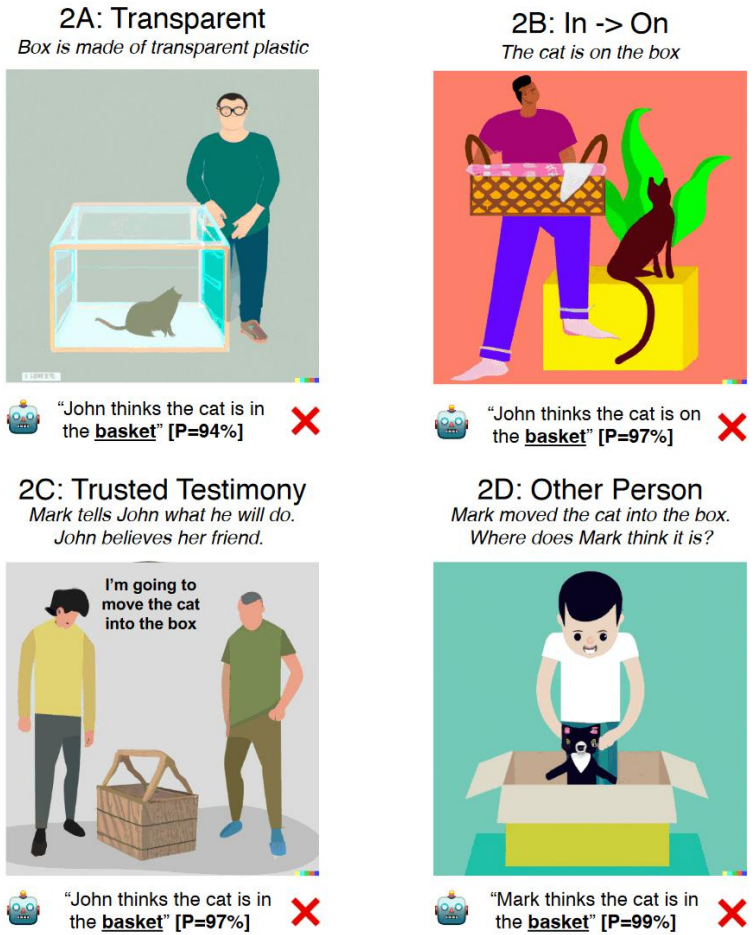
位置错误信念，Sally-Anne任务

STORY: *In the room there are John, Mark, a cat, a box, and a basket.* ***John takes the cat and puts it in the basket. He leaves the room and goes to school****. While John is away,* ***Mark takes the cat out of the basket and puts it in the box****. Mark leaves the room and goes to work. John comes back from school and enters the room. He doesn’t know what happened in the room when he was away.*

QUESTION: *John thinks that the cat is in the _____*

ANSWER : ***basket [P_{box} = 0%; P_{basket} = 98%]***

被未见到的位置转移影响导致的错误信念



评测机器心智 (Evaluating Machine ToM)

Do Large Language Models have Theory of Mind?

反方：“*Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models*”
Shapira1 et al., arXiv preprint, ***24 May 2023***.

聪明的汉斯，19世纪末，德国，会做数学题的马！

评测机器心智 (Evaluating Machine ToM)

Do Large Language Models have Theory of Mind?

反方：“*Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models*”

*Shapira1 et al., arXiv preprint, **24 May 2023**.*

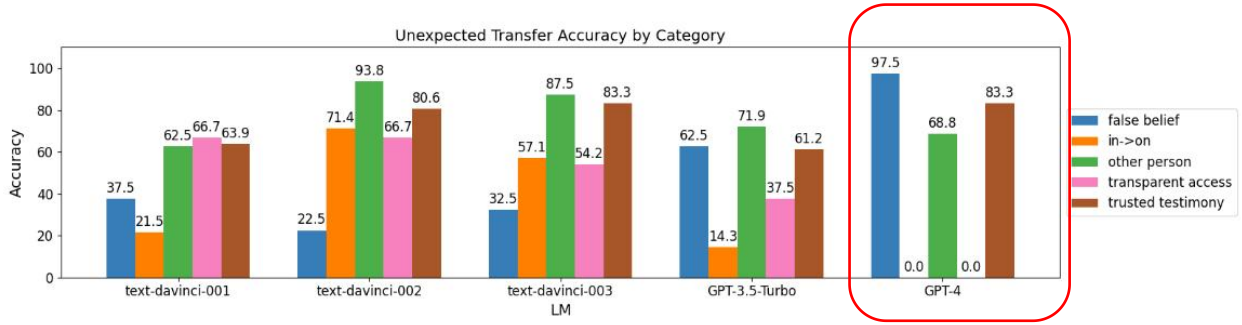
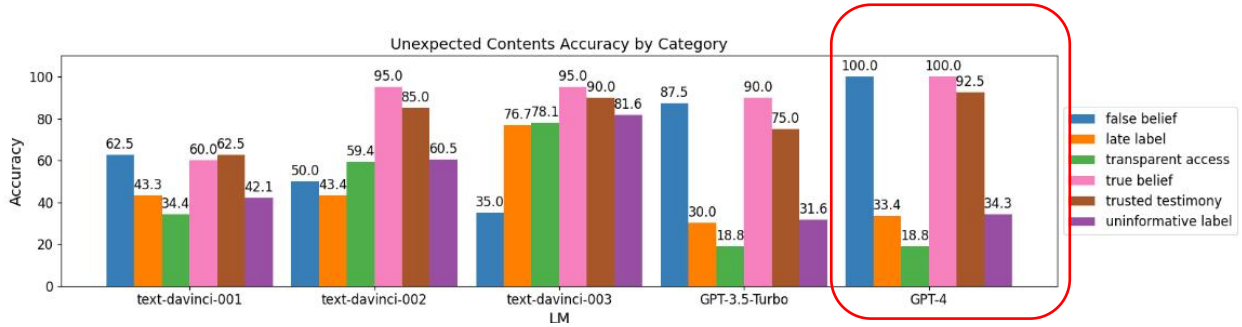


Figure 2: Performance of a range of GPT models on various categories within the unexpected transfer segment of Adv-CSFB. The results are the average accuracy of question 2 (e.g. *Maria thinks that the bananas are in the _*) and question 3 (e.g. *When Maria comes back, she will first look for the bananas in the _*), which specifically focus on an agent's beliefs rather than objective truth. Notably, GPT-4 achieves an accuracy of 97% on the subset of false belief samples (the original examples from ToM-k), while failing on adversarial samples that involve transparent access or relationship change (in→on).



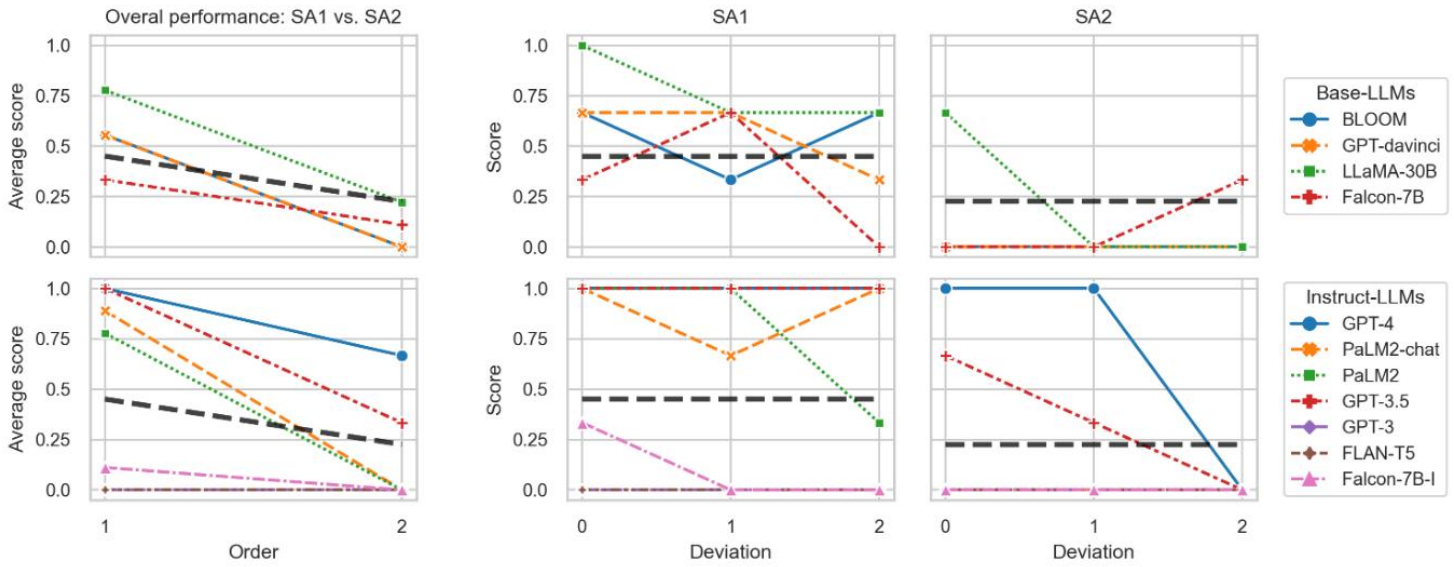
“LLMs rely on shortcuts, heuristics, and spurious correlations, not Theory of Mind.”

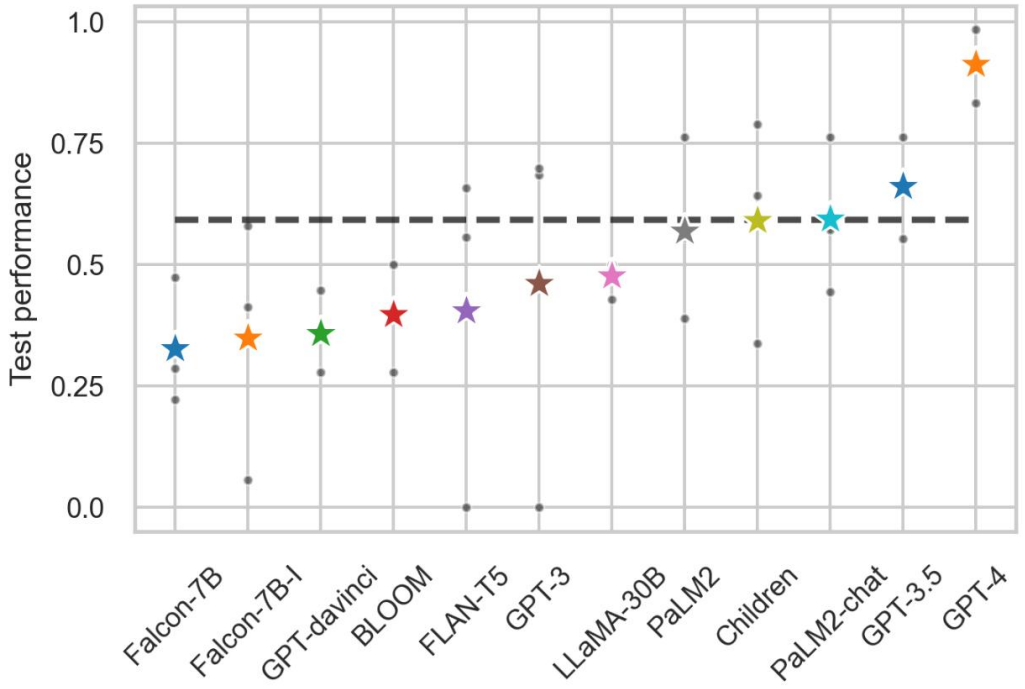
评测机器心智 (Evaluating Machine ToM)

Do Large Language Models have Theory of Mind?

正方： “*Theory of Mind in Large Language Models: Examining Performance of 11 State-of-the-Art models vs. Children Aged 7-10 on Advanced Tests*”, *Dujin et al., arXiv preprint, **31 Oct 2023**.*

大模型 vs. 37个7-8岁的孩童



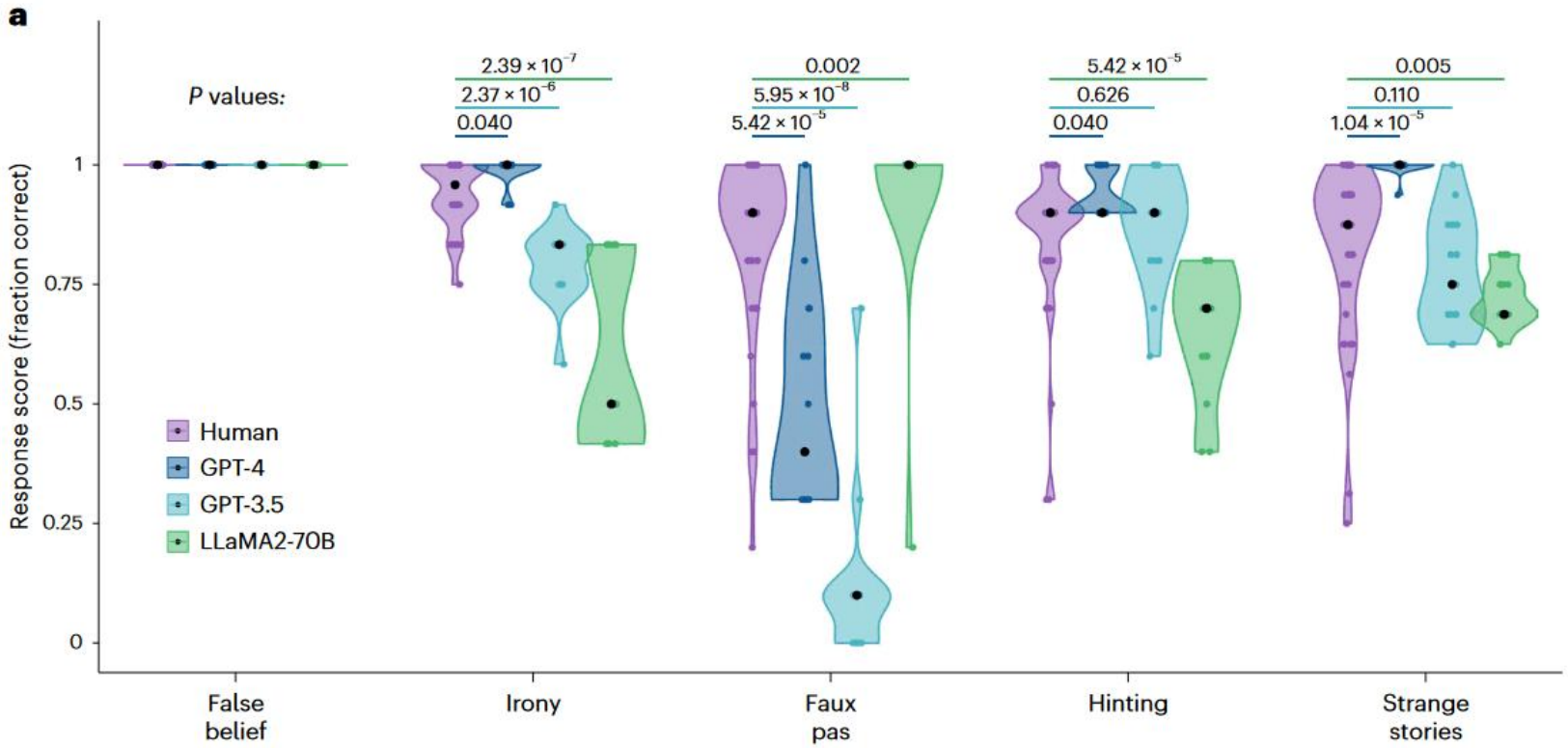


评测机器心智 (Evaluating Machine ToM)

Do Large Language Models have Theory of Mind?

正方： “*Testing theory of mind in large language models and humans*”
Strachan et al., Nature Human Behavior, 5 April 2024.

大模型 vs. 1907个成年人

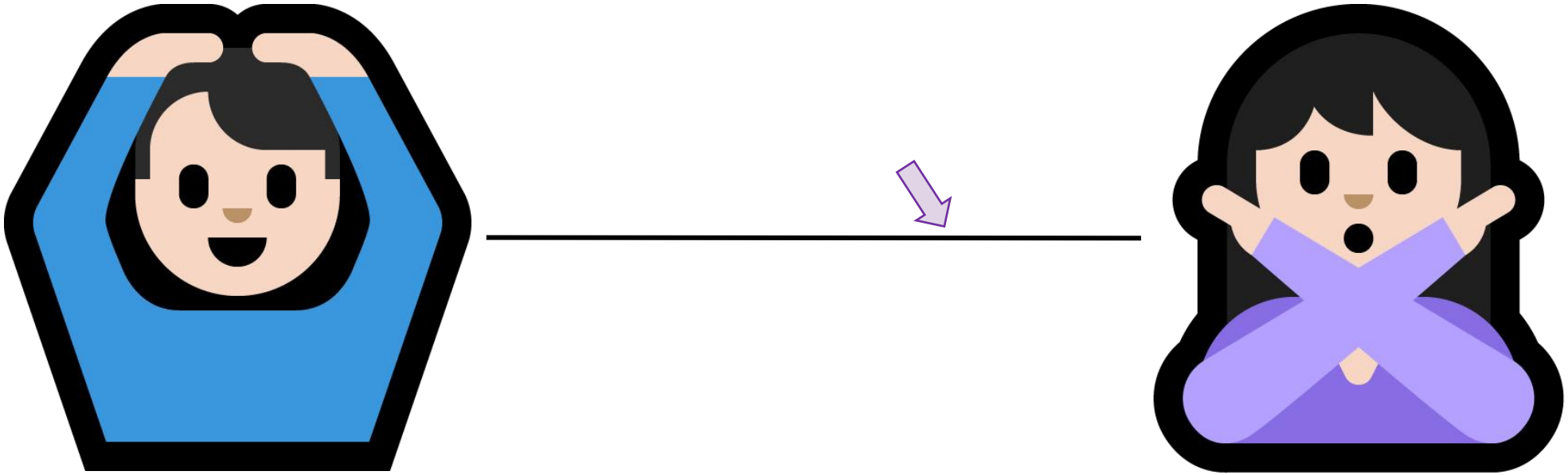


“GPT-4被证实具有「人类心智」登Nature！”

评测机器心智 (Evaluating Machine ToM)

Do Large Language Models have Theory of Mind?

YES **NO**



Chen et al., **ToMBench: Benchmarking Theory of Mind in Large Language Models**, ACL 2024

评测机器心智 (Evaluating Machine ToM)

机器心智评测基准ToMBench

现有评测存在的问题	ToMBench针对性解决方案
受限视角	系统框架
主观评测	多项选择
无意泄露	全新数据

Chen et al., **ToMBench: Benchmarking Theory of Mind in Large Language Models**, ACL 2024

评测机器心智 (Evaluating Machine ToM)

机器心智评测基准ToMBench：任务为载体，能力为核心

8种ToM任务 | 31种ToM能力 | 多项选择题格式 | 双语评测环境 | 共计2860样本

Inventory

Tasks (8)
- Unexpected Outcome Test
- Scalar Implicature Task
- Persuasion Story Task
- False Belief Task
- Ambiguous Story Task
- Hinting Test
- Strange Story Task
- Faux-Pas Recognition Test

Abilities (31)
- Emotion (7)
- Desire (4)
- Intention (4)
- Knowledge (4)
- Belief (6)
- Non-Literal Comm. (6)

Example

Task: Unexpected Outcome Test

Story: Mike wins the championship in a piano solo competition and receives praise from his brother.

Ability1: Emotion/Typical emotional reactions
Question1: What emotion does Mike likely show?
A. Proud ✔ ✔ B. Guilty
C. Ashamed D. Disappointed

Ability2: Belief/Sequence false beliefs
Question2: Mike feel guilty rather than proud, why?
A. Mike initially dislikes playing the piano and is forced into learning by his brother.
B. Mike makes a significant mistake in the competition and worries about being discovered by others. ✘
C. Mike worries that the prize money is not enough to buy a new piano.
D. His brother, to pay for Mike's piano lessons, works very hard with no free time every day. ✔

Evaluation

Task Dim.	✔ 50%	
Ability1 Dim.	✔	✔
Ability2 Dim.	✔	✘

Characteristics

- Systematic Framework
- MCQ Format
- Original Inventory
- Bilingual Corpus

Chen et al., **ToMBench: Benchmarking Theory of Mind in Large Language Models**, ACL 2024

评测机器心智 (Evaluating Machine ToM)

机器心智评测基准ToMBench：6种心智能力（31种子能力）

情感
Emotion

涉及理解情境因素如何影响人的情绪状态、理解人们可以体验复杂情绪以及调节情绪表达的能力

欲望
Desire

涉及理解人有主观的欲望、偏好和需求，这些会影响他们的情绪和行为的能力

意图
Intention

涉及理解人们为实现目标和意图而采取行动的能力

知识
Knowledge

涉及理解他人基于他们的感知、接收到的信息或熟悉程度而获得不同知识的能力

信念
Belief

涉及理解人们可以持有与现实或自己信念不同的关于世界的信念的能力

非字面沟通
Non-Literal Communication

涉及理解交流可以传达超出字面意思的含义的能力

Chen et al., **ToMBench: Benchmarking Theory of Mind in Large Language Models**, ACL 2024

评测机器心智 (Evaluating Machine ToM)

机器心智评测基准ToMBench：8种心智任务 (1-4)

意外结果测试

意外结果测试主要考察参与者在情绪激发情境与实际激发情境之间存在明显差异时，推断角色心理状态的能力。

量词含义任务

量词含义任务主要考察参与者在对话上下文中推断出量词所表达的实际含义的能力。

说服故事任务

说服故事任务主要评估参与者理解和选择有效说服策略的能力，反映他们对如何影响他人心理状态和态度的理解。

错误信念任务

错误信念任务主要考察参与者是否能够在自身信念（正确信念）和他人信念（错误信念）不同时区分的能力。

意外结果测试

故事: 小明在生日时收到了一辆自行车。
问题1： 小明的情绪应该是什么？
(A) 尴尬 (B) 开心 (C) 失望 (D) 后悔
答案： B
问题2： 他应该很开心，但他很失望，为什么？
(A) 小明担心骑自行车会影响他的学习。
(B) 小明害怕骑自行车上学会被同学嘲笑。
(C) 小明认为自行车的颜色与他的衣服不搭配。
(D) 小明希望收到一台电脑作为礼物，而不是自行车。
答案： D
问题3： 小明正在过生日，他希望收到一台电脑或一个新游戏作为生日礼物，但在他生日那天他收到了一辆自行车。此时小明的情绪是什么？
(A) 尴尬 (B) 开心 (C) 失望 (D) 后悔
答案： C

量词含义任务

故事: 寄给Laura公司的每封信中几乎都包含一张支票。今天，Laura收到了5封信。Laura在电话中告诉你：“我看了5封信中的3封，其中2封信里有支票。”

问题1: 在Laura打电话给你之前，你认为这5封信中有多少封包含支票？
(A) 0 (B) 1 (C) 2 (D) 4
答案： D

问题2: 在Laura打电话给你之后，你认为这5封信中有多少封包含支票？
(A) 0 (B) 1 (C) 2 (D) 4
答案： D

说服故事任务

故事: 小红是一个6岁的孩子。今天是星期六。爸爸妈妈今天都有空，不知道他们可以一起做什么。也许可以去吃冰淇淋？小红今天真的很想去游乐园。然而，爸爸觉得游乐园很吵。他说：“小红，这不是个好主意。我觉得游乐园很吵。”
问题: 小红如何说服她的爸爸？
(A) 小红可以查找一些信息，证明现在的游乐园采取了很多措施来减少噪音，比如设置隔音墙、使用更安静的设备等。
(B) 小红可以告诉她的爸爸，她很久没有去游乐园了，这是她一个非常特别的愿望，她真的很想去。
(C) 如果爸爸真的不想去游乐园，小红可以建议去其他地方，比如公园或动物园，这样大家都开心。
(D) 小红可以建议：“我们可以早上或者工作日去吗？我听说那时候人少一些，也更安静。我们可以避开高峰时间，享受一个更安静的体验。”
答案： B

错误信念任务

故事: 李雷和韩梅梅在客厅里闲逛，他们看到柜子、箱子和手提包，他们在柜子里找到了一件运动衫。韩梅梅离开了客厅，李雷把运动衫移到了手提包里。

问题1: 运动衫现在在哪里？
(A) 公文包 (B) 箱子 (C) 柜子 (D) 手提包
答案： D

问题2: 韩梅梅回到客厅后，李雷认为韩梅梅会在哪里找运动衫？
(A) 箱子 (B) 衣橱 (C) 手提包 (D) 柜子
答案： D

Chen et al., **ToMBench: Benchmarking Theory of Mind in Large Language Models**, ACL 2024

评测机器心智 (Evaluating Machine ToM)

机器心智评测基准ToMBench：8种心智任务(5-8)

模糊故事任务

模糊故事任务给参与者提供模糊的社交故事，并通过问题评估参与者在不确定的情境下对他人心理状态（如情感和信念）的理解。

暗示测试

暗示测试主要评估参与者在社交互动中的间接提示中推断心理状态的能力，反映出他们对超出字面表达的隐含意义的理解。

奇异故事任务

奇异故事任务考察参与者在较为复杂和反常的故事中推断角色的心理状态的能力。

失礼行为识别测试

失礼行为识别测试主要测试参与者在社交故事中识别角色是否出现失礼行为的能力，反映他们对社会规范和他人视角的理解。

模糊故事任务

故事: 德华和三明是公司员工，他们正在竞争一个晋升机会。今天是三明的生日，他一个人享用一个小生日蛋糕。玲玲是公司的主管，她在茶水间和德华进行私人谈话。三明从远处看到玲玲悄悄地给了德华一份文件，并微笑着轻轻拍了拍德华的背，然后返回她的办公室。德华迅速查看了文件，微笑着把文件小心地放进他的文件包。

问题1： 为什么玲玲要把那份文件给德华？

(A) 玲玲归还德华落在会议室的个人文件。

(B) 玲玲给德华一份与晋升无关的日常通知。

(C) 玲玲和德华正在为三明策划一个惊喜。

(D) 玲玲给德华额外的信息。

答案： D

问题2： 你认为三明会怎么想？

(A) 三明认为一切正常，因为那是德华的文件。

(B) 三明感到愤怒，因为他被孤立了。

(C) 三明感到非常开心，因为他的蛋糕很好吃。

(D) 三明重新评估他的晋升机会。

答案： D

暗示测试

故事: 曹生和王红是一对夫妻。一天，王红做了晚饭，曹生吃了一口说：“我们家是不是很久没买盐了？”

问题: 曹生说这句话时，他真正想表达的是什么？

(A) 曹生在问家里的盐是不是用完了。

(B) 曹生在表达他对盐的需求增加了。

(C) 曹生在暗示王红做的晚饭盐不够。

(D) 曹生在提醒他们需要买更多的食材。

答案： C

奇异故事任务

故事: 李彤和王红是最好的朋友。她们都参加了同一个绘画比赛。现在，李彤非常想赢得这次比赛，但当比赛结果出来时，获胜者是她的好朋友王红，而不是她。李彤因为没有获胜而感到非常难过，但她为获胜的朋友感到高兴。李彤对王红说：“干得好，我真的很高兴你赢了！”李彤对她妈妈说：“我很难过，我没有赢得比赛！”

问题1： 李彤对王红说的话是真的吗？

(A) 是 (B) 不是

答案： A

问题2： 为什么李彤说她既高兴又难过？

(A) 她为王红赢得比赛感到高兴，但因为王红赢得比赛而感到难过和嫉妒。

(B) 她为王红赢得比赛感到高兴，并且对此并不难过。

(C) 她为王红赢得比赛感到高兴，但为王红没有安慰她没有赢得比赛而感到难过。

(D) 她为王红赢得比赛感到高兴，但为自己没有赢得比赛而感到难过。

答案： D

失礼行为识别测试

故事: 第一天来到新学校,小王对旁边的同学小张说:"我妈妈在我们学校工作,是食堂阿姨之一。" 小张点点头。这时,另一个同学小芳走过来,插话说:"我特别不喜欢我们学校的食堂阿姨,她们看着都好可怕。" 小王听后没有说话。小张看看小王,说:"小芳,我们一块放学后去游戏厅玩吧?" 小芳说:"嗯,我现在感觉不舒服,改天吧。"

问题1： 故事里有人说了不合适的话吗？

(A) 故事里有人说了不合适的话。 (B) 故事里没有人说不合适的话。

答案： A

问题2： 如果有一句话不合适，哪一句不合适？

(A) "我妈妈在我们学校工作,是食堂阿姨之一。"

(B) "我超级讨厌我们学校的食堂阿姨,她们看起来都好凶!"

(C) "我觉得阿姨们人都挺nice的。"

(D) 故事里没有人说不合适的话。

答案： B

Chen et al., **ToMBench: Benchmarking Theory of Mind in Large Language Models**, ACL 2024

评测机器心智 (Evaluating Machine ToM)

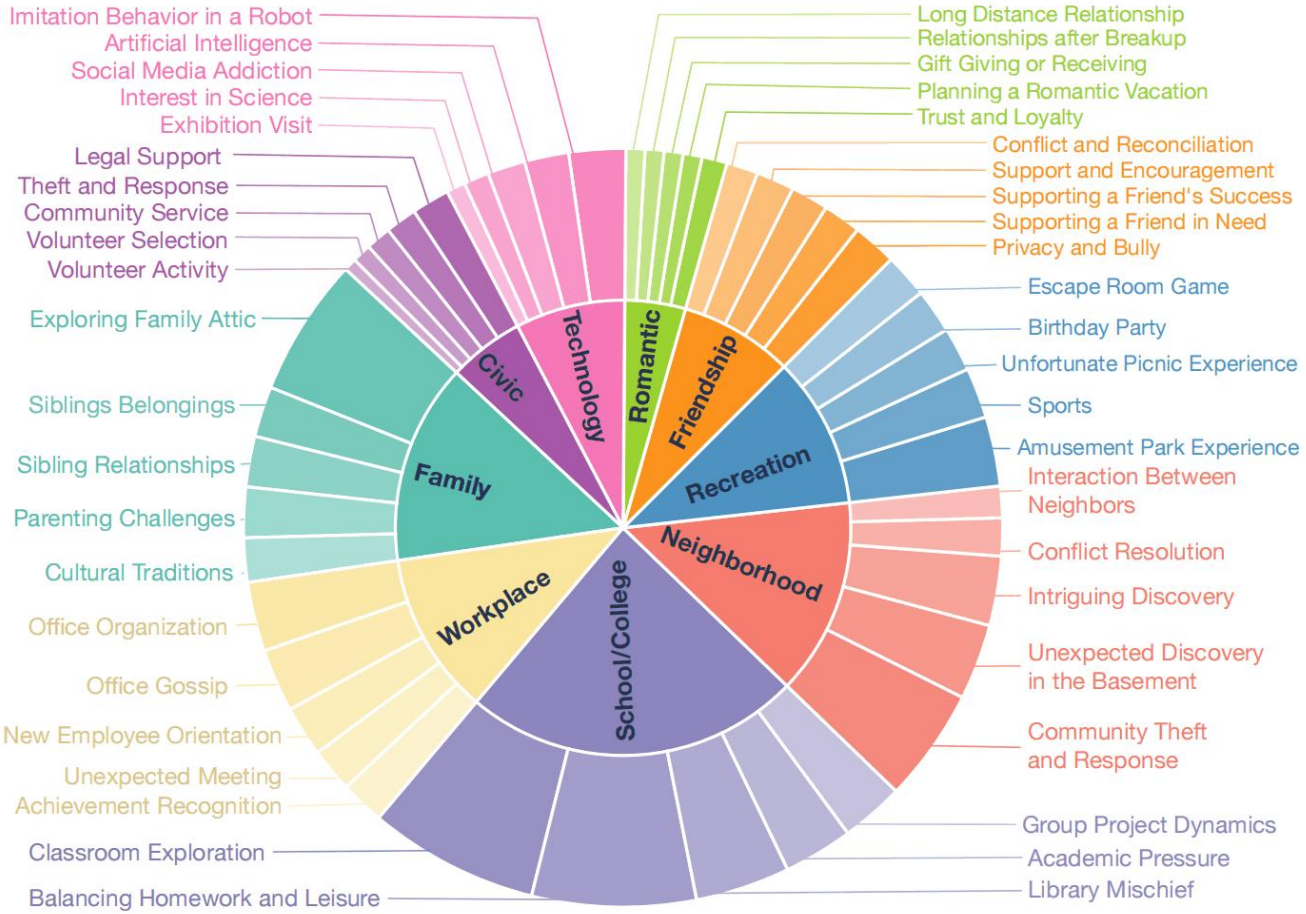
8种心智任务与31种心智能力的对应关系

Abilities	Tasks (with Simplified Examples)	Abilities
Emotion Typical emotional reactions (Task 1) Atypical emotional reactions (Task 1) Discrepant emotions (#) Mixed emotions (Task7) Hidden emotions (#) Moral emotions (#) Emotion regulation (#)	**1.Unexpected Outcome Test** Story: PersonA attends PersonB's wedding, but they have a fight before... Question: PersonB should feel embarrassed, but PersonaB is very happy. Why?	**Knowledge** Knowledge-pretend play links (#) Percepts-knowledge links (#) Information-knowledge links (Task 2) Knowledge-attention links (#)
	2.Scalar Implicature Task Story: A football team of 18 players has almost 1/3 as goalkeepers... Question: How many goalkeepers in the team?	
Desire Multiple desires (#) Desires influence on act./emo. (Task3) Desire-action contradiction (Task 7) Discrepant desires (#)	**3.Persuasion Story Task** Story: PersonA wants to go to the park with PersonB, but PersonB doesn't want to... Question: How does PersonA persuade PersonB?	**Belief** Content false beliefs (Task 4) Location false beliefs (Task 4) Identity false beliefs (Task 7) Second-order beliefs (Task 4) Beliefs based act./emotions (Task 5/7) Sequence false beliefs (Task 1)
	4.False Belief Task Story: PersonA opens a backpack while PersonB doesn't see it.. Question: What does PersonA expect PersonB to find inside the backpack?	
	5.Ambiguous Story Task Story: PersonA and PersonB communicate with body language, and PersonC sees them... Question: What is PersonC thinking about?	
	6.Hinting Test Story: PersonA hints to PersonB to help her but does not say it directly... Question: What does PersonA hope PersonB do?	
Intention Discrepant intentions (#) Prediction of actions (#) Intentions explanations (Task 5/6/7) Completion of failed actions (#)	**7.Strange Story Task** Story: PersonA adds too much salt while cooking, and PersonB mocks him... Question: Why does PersonB say this?	**Non-Literal Communication** Irony/Sarcasm (Task 6/7) Egocentric lies (Task 7) White lies (Task 7) Involuntary lies (Task 7) Humor (Task 7) Faux pas (Task 8)
	8.Faux-Pas Recognition Test Story: PersonA unintentionally says offensive words to PersonB... Question: Does anyone say something inappropriate in this story?	

Chen et al., **ToMBench: Benchmarking Theory of Mind in Large Language Models**, ACL 2024

评测机器心智 (Evaluating Machine ToM)

多样日常主题和真实社交情境



严格的数据收集与验证过程

	#S	#Q	ASL (En)	ASL (Zh)	Agr.
Task View	934	2,470	61.22	97.69	99.4%
Unexpected Outcome Test	100	300	38.46	62.01	100.0%
Scalar Implicature Task	100	200	47.17	76.89	100.0%
Persuasion Story Task	100	100	36.58	51.35	95.0%
False Belief Task	100	600	49.15	77.54	100.0%
Ambiguous Story Task	100	200	102.57	164.07	100.0%
Hinting Test	93	103	49.63	79.92	100.0%
Strange Story Task	201	407	70.42	112.97	100.0%
Faux-pas Recognition Test	140	560	95.77	156.79	98.2%
Ability View	1,584	2,860	66.57	107.21	99.4%
Emotion	300	420	52.34	83.50	99.8%
Desire	160	180	50.19	74.91	97.2%
Intention	273	340	82.56	131.20	100.0%
Knowledge	170	290	56.38	94.26	100.0%
Belief	440	882	55.70	88.99	100.0%
Non-Literal Communication	241	748	88.02	143.91	99.4%

Chen et al., **ToMBench: Benchmarking Theory of Mind in Large Language Models**, ACL 2024

评测机器心智 (Evaluating Machine ToM)

- **人类 vs. 大模型**：所有大语言模型的平均心智理论（ToM）表现显著低于人类
- **大模型之间对比**：大语言模型的心智理论表现中，GPT-4系列表现最好
- **普通提示与思维链提示**：思维链提示通常不能改善心智理论表现

UOT: Unexpected Outcome Test　SIT: Scalar Implicature Task　PST: Persuasion Story Task　FBT: False Belief Task
AST: Ambiguous Story Task　HT: Hinting Test　SST: Strange Story Task　FRT: Faux-pas Recognition Test

SUBJECT	UOT		SIT		PST		FBT		AST		HT		SST		FRT		AVG.	
	Zh	En	Zh	En	Zh	En	Zh	En	Zh	En	Zh	En	Zh	En	Zh	En	Zh	En
Human	**89.3**		**75.5**		**70.0**		**86.8**		**95.0**		**97.1**		**89.2**		**80.4**		**85.4**	
ChatGLM3-6B	55.3	44.3	24.5	28.0	44.0	41.0	59.2	48.5	48.0	41.0	32.0	36.9	58.0	37.8	55.2	44.6	47.0	40.3
LLaMA2-13B-Chat	43.7	52.7	28.0	23.5	38.0	43.0	42.2	42.8	38.0	47.5	32.0	48.5	58.2	58.0	47.9	58.4	41.0	46.8
Baichuan2-13B-Chat	56.3	53.7	27.5	32.0	48.0	36.0	50.2	51.5	56.0	50.5	54.4	58.3	50.1	50.4	61.6	61.3	50.5	49.2
Mistral-7B	61.0	58.0	28.0	34.5	49.0	51.0	43.5	46.7	52.5	51.0	29.1	43.7	53.1	60.0	63.6	66.8	47.5	51.5
Mixtral-8x7B	68.0	58.7	49.5	42.5	45.0	55.0	49.8	37.8	71.0	69.5	43.7	55.3	51.4	53.8	62.5	54.1	55.1	53.3
Qwen-14B-Chat	72.0	63.7	42.5	30.5	50.0	51.0	57.2	58.7	65.5	64.0	54.4	56.3	60.0	59.5	72.7	69.5	59.3	56.7
GPT-3.5-Turbo-0613	69.3	63.3	33.0	35.0	52.0	49.0	61.2	62.3	63.5	63.5	60.2	53.4	72.0	66.1	66.8	67.0	59.8	57.5
GPT-3.5-Turbo-1106	72.3	66.0	34.0	33.0	57.0	56.0	53.0	55.0	59.0	60.5	61.2	64.1	72.5	69.0	68.8	72.5	59.7	59.5
GPT-4-0613	71.3	71.3	49.0	44.0	58.0	53.0	86.3	80.0	84.0	78.0	79.6	76.7	83.0	81.1	76.6	71.8	73.5	69.5
GPT-4-1106	76.7	71.0	48.0	49.0	61.0	65.0	90.8	88.2	83.0	77.5	88.3	82.5	76.2	84.0	78.6	75.0	75.3	74.0
ChatGLM3-6B + CoT	58.7	50.3	27.0	26.5	44.0	41.0	56.5	51.2	48.0	44.0	37.9	42.7	56.3	44.2	65.4	51.4	49.2	43.9
LLaMA2-13B-Chat + CoT	45.3	52.7	25.5	23.5	34.0	39.0	41.3	43.0	41.0	48.5	28.2	43.7	53.6	59.5	49.6	62.1	39.8	46.5
Baichuan2-13B-Chat + CoT	54.3	48.7	26.5	23.0	33.0	34.0	44.8	44.2	51.5	44.0	53.4	49.5	52.8	51.1	65.4	52.5	47.7	43.4
Mistral-7B + CoT	61.0	55.3	27.0	28.0	46.0	42.0	47.2	47.0	47.0	46.5	30.1	37.9	56.5	63.4	64.3	64.1	47.4	48.0
Mixtral-8x7B + CoT	65.3	52.3	45.0	29.5	41.0	39.0	53.7	43.8	66.0	59.5	44.7	54.4	43.7	39.8	47.5	54.3	50.9	46.6
Qwen-14B-Chat + CoT	65.3	58.0	31.5	31.0	45.0	44.0	51.3	54.7	62.5	63.0	47.6	48.5	60.2	53.6	70.7	67.7	54.3	52.6
GPT-3.5-Turbo-0613 + CoT	62.3	58.3	30.0	26.5	43.0	48.0	57.8	64.0	58.5	58.0	41.7	41.7	71.3	66.8	70.5	70.4	54.4	54.2
GPT-3.5-Turbo-1106 + CoT	68.7	64.7	27.5	35.0	45.0	54.0	57.5	56.3	61.5	63.0	46.6	51.5	71.3	68.6	72.7	70.9	56.4	58.0
GPT-4-0613 + CoT	72.3	64.7	43.5	54.0	55.0	52.0	90.3	80.8	84.5	77.5	78.6	76.7	83.5	81.1	74.3	73.6	72.8	70.1
GPT-4-1106 + CoT	76.3	72.7	48.0	55.0	59.0	55.0	88.7	86.8	84.0	81.0	89.3	82.5	76.9	84.3	79.6	75.2	75.2	74.1
LLM Grand Mean	64.6	60.3	36.4	35.2	50.2	50.0	59.3	57.2	62.1	60.3	53.5	57.6	63.5	62.0	65.4	64.1	56.9	55.8
LLM Grand Mean + CoT	63.0	57.8	33.2	33.2	44.5	44.8	58.9	57.2	60.5	58.5	49.8	52.9	62.6	61.2	66.0	64.2	54.8	53.7

面向任务的评测

SUBJECT	Emotion		Desire		Intention		Knowledge		Belief		NL Comm.		AVG.	
	Zh	En	Zh	En	Zh	En	Zh	En	Zh	En	Zh	En	Zh	En
Human	**86.4**		**78.2**		**90.4**		**82.2**		**89.3**		**89.0**		**86.1**	
ChatGLM3-6B	54.9	42.2	52.0	40.7	52.0	35.9	16.8	22.0	55.0	44.5	49.8	38.5	46.8	37.3
LLaMA2-13B-Chat	38.4	51.0	39.2	49.4	41.7	49.6	22.4	21.1	46.7	49.0	54.0	54.3	40.4	45.7
Baichuan2-13B-Chat	55.9	53.1	49.6	46.0	63.5	52.2	21.3	20.9	48.5	49.8	46.2	50.1	47.5	45.4
Mistral-7B	54.0	58.1	48.7	49.8	45.3	52.2	33.1	42.0	47.2	48.7	46.5	57.2	45.8	51.3
Mixtral-8x7B	61.6	56.6	54.1	51.2	60.1	64.1	31.1	27.1	56.9	48.1	50.9	57.9	52.5	50.8
Qwen-14B-Chat	66.8	65.8	57.0	52.9	66.4	58.9	37.9	33.1	62.2	60.6	53.2	57.5	57.3	54.8
GPT-3.5-Turbo-0613	58.4	65.6	54.2	53.4	58.2	61.0	37.8	36.3	64.3	61.4	76.8	66.9	58.3	57.4
GPT-3.5-Turbo-1106	61.6	60.6	57.1	60.7	56.5	62.6	30.4	37.4	60.6	59.4	76.0	71.5	57.0	58.7
GPT-4-0613	79.0	72.0	72.2	60.2	77.8	66.1	56.0	48.1	82.1	76.1	81.3	81.5	74.7	67.3
GPT-4-1106	75.9	75.7	67.5	69.7	77.8	84.7	57.6	52.1	84.1	82.8	72.8	84.0	72.6	74.8
ChatGLM3-6B + CoT	53.0	46.7	49.1	43.7	54.8	49.8	32.0	28.9	51.7	48.6	55.8	40.1	49.4	43.0
LLaMA2-13B-Chat + CoT	43.3	48.1	37.4	44.9	43.4	51.7	28.7	30.7	43.8	47.9	52.9	62.7	41.6	47.7
Baichuan2-13B-Chat + CoT	51.6	49.7	47.2	37.5	51.3	47.8	33.7	19.3	47.3	45.2	52.4	47.5	47.3	41.2
Mistral-7B + CoT	52.0	57.9	46.9	45.1	50.5	51.1	33.4	44.5	50.9	50.1	50.7	62.4	47.4	51.9
Mixtral-8x7B + CoT	56.9	56.0	47.5	41.5	57.9	55.3	30.2	33.2	54.6	44.3	44.6	45.5	48.6	46.0
Qwen-14B-Chat + CoT	63.9	62.7	57.3	50.2	63.2	57.8	41.0	40.1	56.2	53.6	53.5	53.2	55.9	52.9
GPT-3.5-Turbo-0613 + CoT	61.6	62.7	53.1	52.1	65.4	63.8	49.6	43.3	58.2	58.7	70.0	71.6	59.7	58.7
GPT-3.5-Turbo-1106 + CoT	63.2	62.3	54.7	54.7	59.9	63.1	34.6	49.6	61.9	59.9	71.3	70.8	57.6	60.1
GPT-4-0613 + CoT	76.8	73.1	69.9	67.1	80.1	71.5	60.5	57.5	83.7	76.4	80.9	82.2	75.3	71.3
GPT-4-1106 + CoT	76.8	73.2	71.2	63.3	78.9	77.9	63.1	60.4	84.0	83.6	70.9	83.0	74.2	73.6
LLM Grand Mean	60.7	60.1	55.2	53.4	59.9	58.7	34.4	34.0	60.8	58.0	60.8	61.9	55.3	54.4
LLM Grand Mean + CoT	59.9	59.2	53.4	50.0	60.5	59.0	40.7	40.8	59.2	56.8	60.3	61.9	55.7	54.6

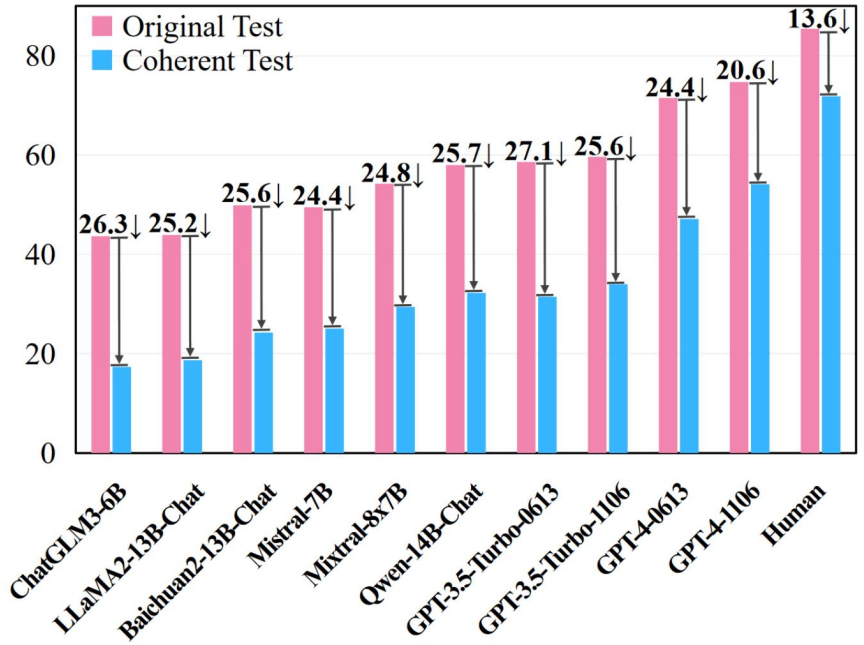
面向能力的评测

Chen et al., **ToMBench: Benchmarking Theory of Mind in Large Language Models**, ACL 2024

评测机器心智 (Evaluating Machine ToM)

连贯测试 - 难度提升

- ✓ **设置**：如果被试错误回答与故事相关的任何问题，则判定在该故事上失败
- ✓ **结果**：大模型在完全理解社会情境方面受限，无法像人类一样理解透彻



在连贯测试下的性能变化

大模型会在非常简单的ToM问题上出错

- ✓ 大模型仍然依赖语义关联来回答问题，这突显了人类与大语言模型在心智理论能力上的区别

Knowledge/Knowledge-Pretend Play Links

ChatGLM3-6B Attention　Human Attention

Story: In a space station on the orbit of a distant planet, a robot named Tinklo lives . This space station has an artificial environment , there is no form of plant life inside , so Tinklo does not understand any plant-related information . However , the space station is full of various mechanical devices and robots . Tinklo is performing imitation behavior : it sways from side to side , occasionally bends to the ground , these actions are similar to flowers swaying in the wind and people smelling flowers .

Question: What does Tinklo possibly imitate?

A. Sunflowers swaying in the wind. ChatGLM3-6B

B. A robot bending over to pick up objects.

C. A person smelling a rose.

D. A windmill with rotating blades.

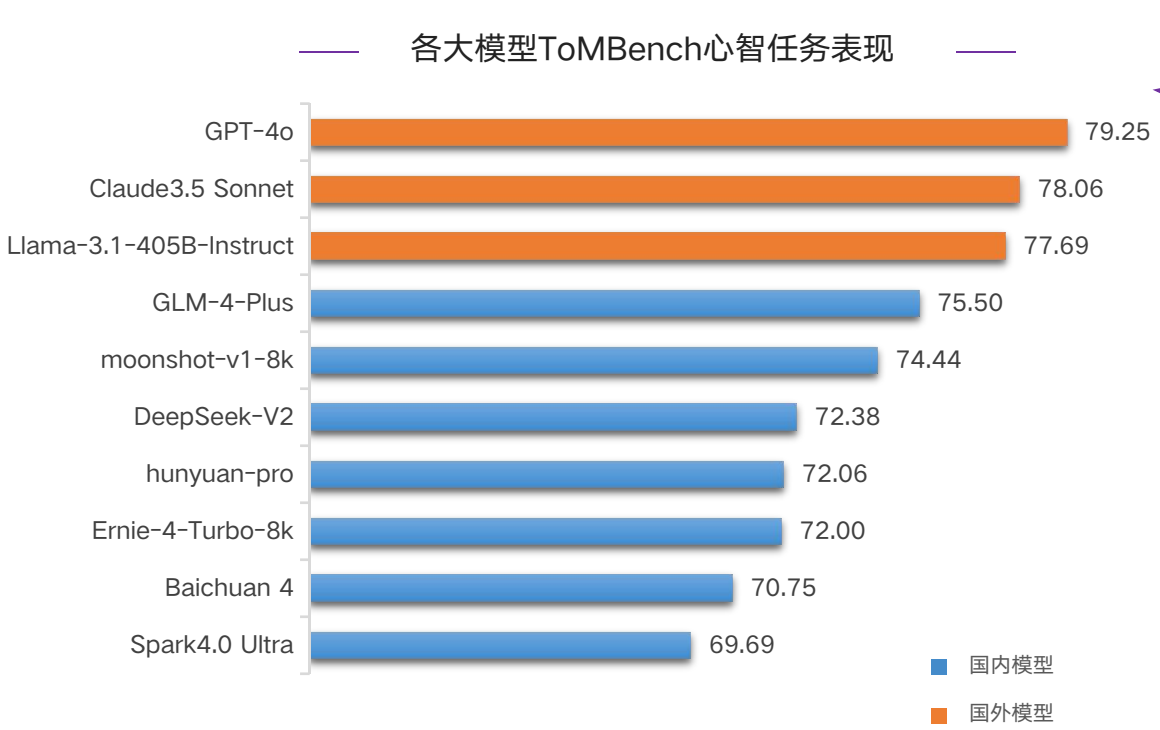
人类与大模型注意力的不同

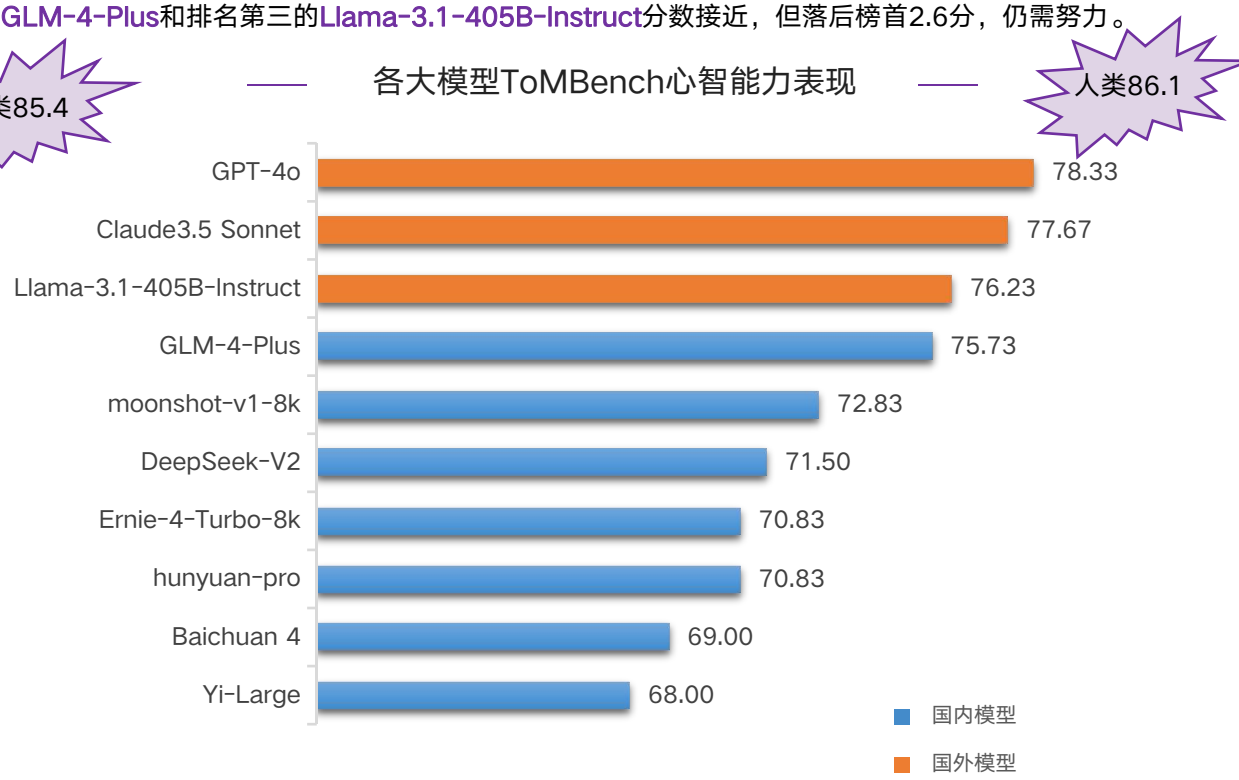
Chen et al., **ToMBench: Benchmarking Theory of Mind in Large Language Models**, ACL 2024

评测机器心智 (Evaluating Machine ToM)

- **整体表现：**在ToMBench评测中，国际一流模型GPT-4o、Claude3.5 Sonnet和Llama-3.1-405B-Instruct依然处于领先地位，在心智任务和心智能力的评测中均包揽前三；国内模型中GLM-4-Plus、moonshot-v1-8k和DeepSeek-V2在两个维度的评测中均排名国内前三，但是对比国外模型仍有一定差距。
- **分维度表现**：

- 心智任务：GPT-4o以79.25分领跑，领先排名国内第一的GLM-4-Plus3.75分，moonshot-v1-8k排名国内第二，得74.44分；国内其他模型中，DeepSeek-V2、hunyuan-pro、Ernie-4-Turbo-8k分数接近，均在72分档。
- 心智能力：GPT-4o、Claude3.5 Sonnet和Llama-3.1-405B-Instruct依然领先，国内模型中GLM-4-Plus和排名第三的Llama-3.1-405B-Instruct分数接近，但落后榜首2.6分，仍需努力。





Chen et al., **ToMBench: Benchmarking Theory of Mind in Large Language Models**, ACL 2024

评测机器心智 (Evaluating Machine ToM)

ToMBench结论:

- 最先进的LLMs具有一定的心智能力，但仍未达到人类水平
- LLMs显然是依靠语义相关性完成心智任务
- LLMs的心智能力不鲁棒，容易受到对抗攻击

ACL 2024 Meta Review

Overall Assessment: 5 = The paper is largely complete and there are no clear points of revision
Suggested Venues:
All *ACL venues.

Best Paper Ae: Yes
Best Paper Ae Justification:
The reviewers concur on the high quality of the submission with high scores, which is rare.

满分

推荐最佳论文
(但没选上 😅)

ToMBench offers a nuanced and effective ToM evaluation of LLMs. Having reviewed at least five ToM-style papers in the past year, I find this work comprehensive, rigorous, and inspiring.

ACL Is Not an AI Conference

Emily M. Bender
Bangkok, Thailand
August 14, 2024

ACL 2024 Presidential Address

https://bit.ly/EMB-ACL24

Chen et al., **ToMBench: Benchmarking Theory of Mind in Large Language Models**, ACL 2024

评测机器心智 (Evaluating Machine ToM)

大模型在心智表现上仍显不足，

是否有方法针对性提升？

——从知识图谱中汲取灵感

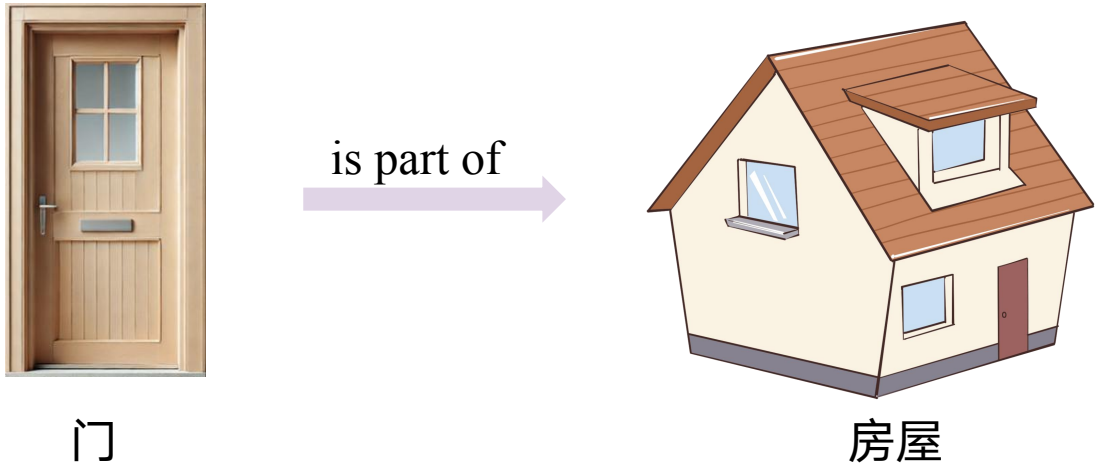
建模机器心智

Modeling Machine ToM

建模机器心智 (Modeling Machine ToM)

1. 语义知识图谱 ConceptNet

- 定义：**静态**概念关系建模
- 知识形式：三元组 <实体-关系-实体>
- 回答问题： What, Who, Where, When



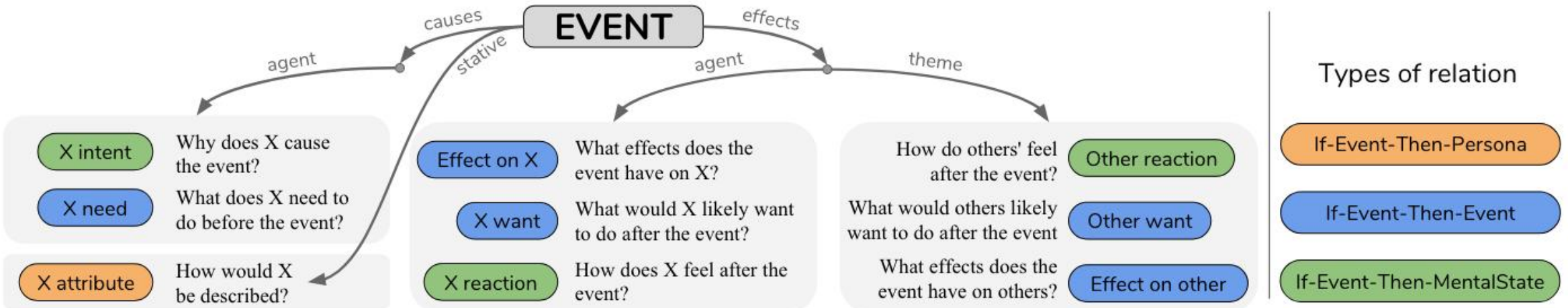
ConceptNet 实例

Speer et al., **ConceptNet 5.5: An Open Multilingual Graph of General Knowledge**, AAAI 2017

建模机器心智 (Modeling Machine ToM)

2. 社会常识图谱 ATOMIC

- 社会常识知识建模
- 知识形式：三元组 <事件-关系-人格> <事件-关系-事件> <事件-关系-心理状态>
- 回答问题：Why, How,What happens next

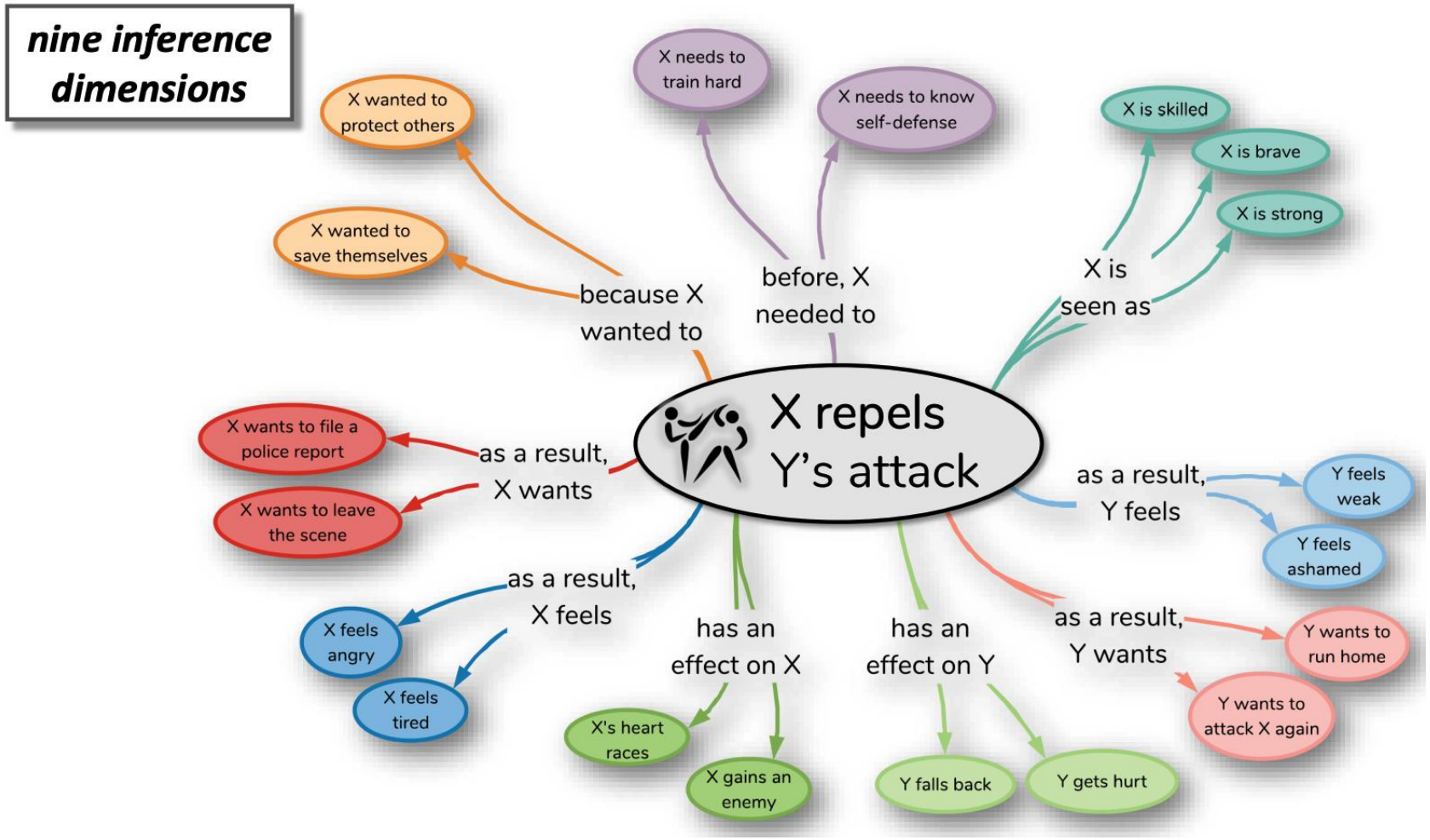


Sap et al., **ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning**, AAAI 2019

建模机器心智 (Modeling Machine ToM)

2. 社会常识图谱 ATOMIC

- 9 种 社会**常识关系**
- 三大类：事件因果，
  人物属性
  心理状态

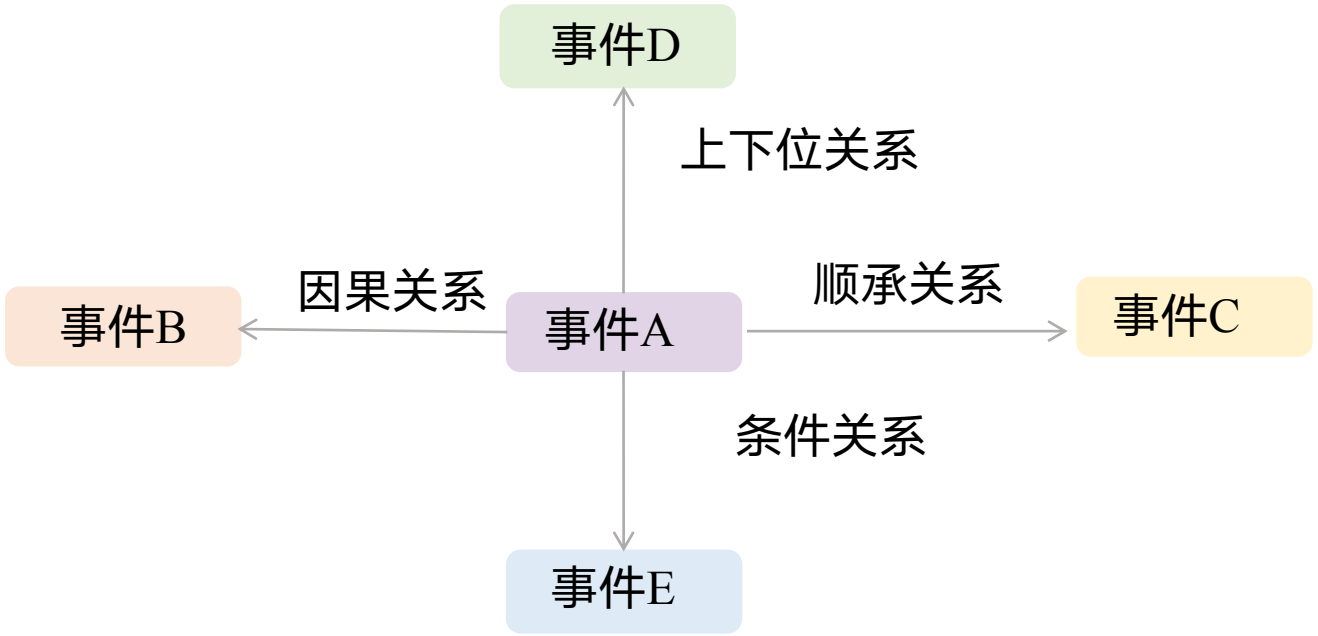


Sap et al., **ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning**, AAAI 2019

建模机器心智 (Modeling Machine ToM)

3. 事理图谱

- 事件时序和因果**动态**关系建模
- 知识形式：多元组 <事件，论元集合，逻辑关系>
- 回答问题：Why, What happens next， What caused this event, If-Then

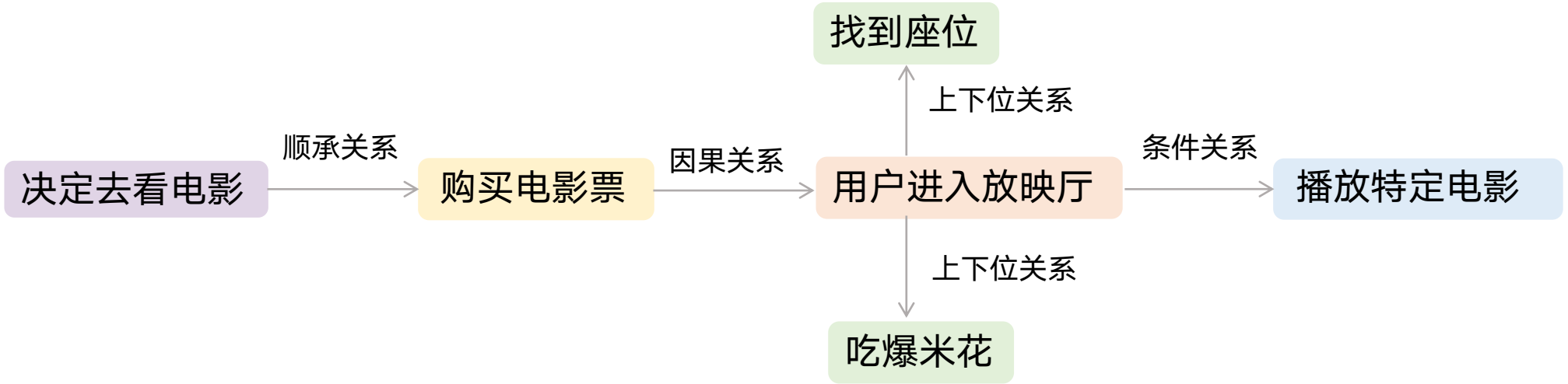


Ding et al., **ELG: An Event Logic Graph**, arXiv:1907.08015

建模机器心智 (Modeling Machine ToM)

3. 事理图谱

- 用于发现事件的演化规律和后续事件的预测



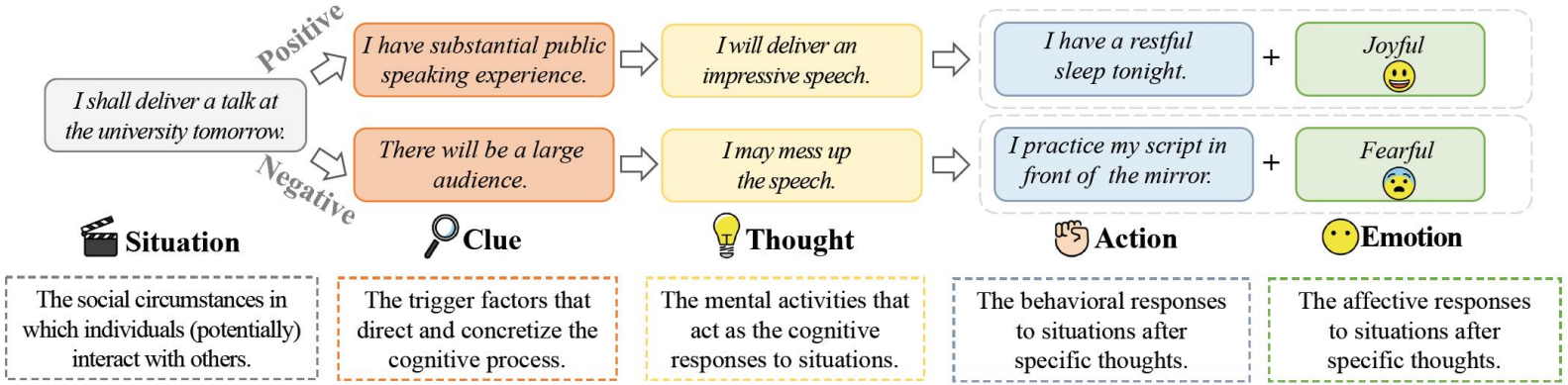
图：“看电影”场景下的**链状**事件演化图

Ding et al., **ELG: An Event Logic Graph**, arXiv:1907.08015

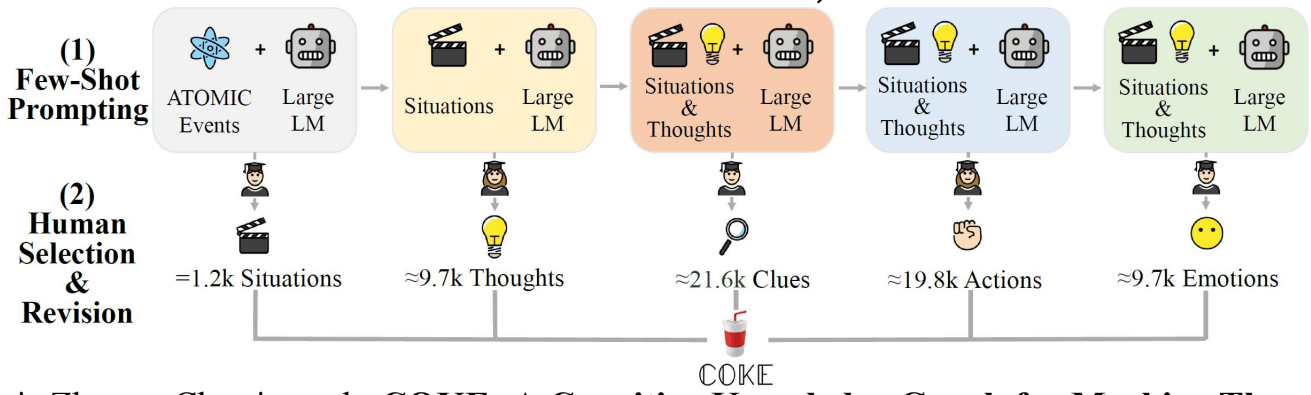
建模机器心智 (Modeling Machine ToM)

机器心智图谱COKE：将心智理论形式化为大规模、高质量认知图谱

- 语言模型无法获知训练语料库背后的人类**心理状态**和**认知过程**
- 基于认知行为疗法（CBT）等心理学理论，构建**认知链**以连接心智理论中的关键节点



- 通过大语言模型的**受控生成**和人类专家的**仔细验证**，生成大规模高质量的认知链

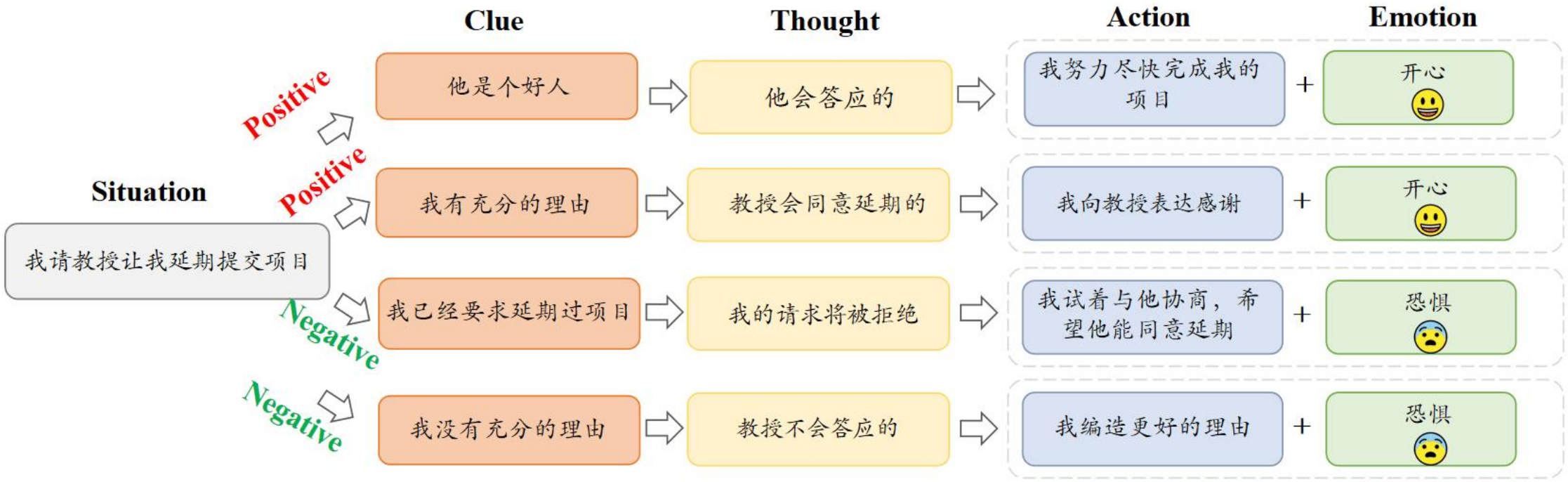


Jincenzi Wu*, Zhuang Chen*, et al., **COKE: A Cognitive Knowledge Graph for Machine Theory of Mind**, ACL 2024

建模机器心智 (Modeling Machine ToM)

机器心智图谱COKE：将心智理论形式化为大规模、高质量认知图谱

- 认知图谱COKE将1,200个场景中可能存在的认知心智，形式化为**45,369条认知链**
- 根据高质量认知资源，训练**认知生成模型COLM**，可以针对**任意场景**生成对应的认知链

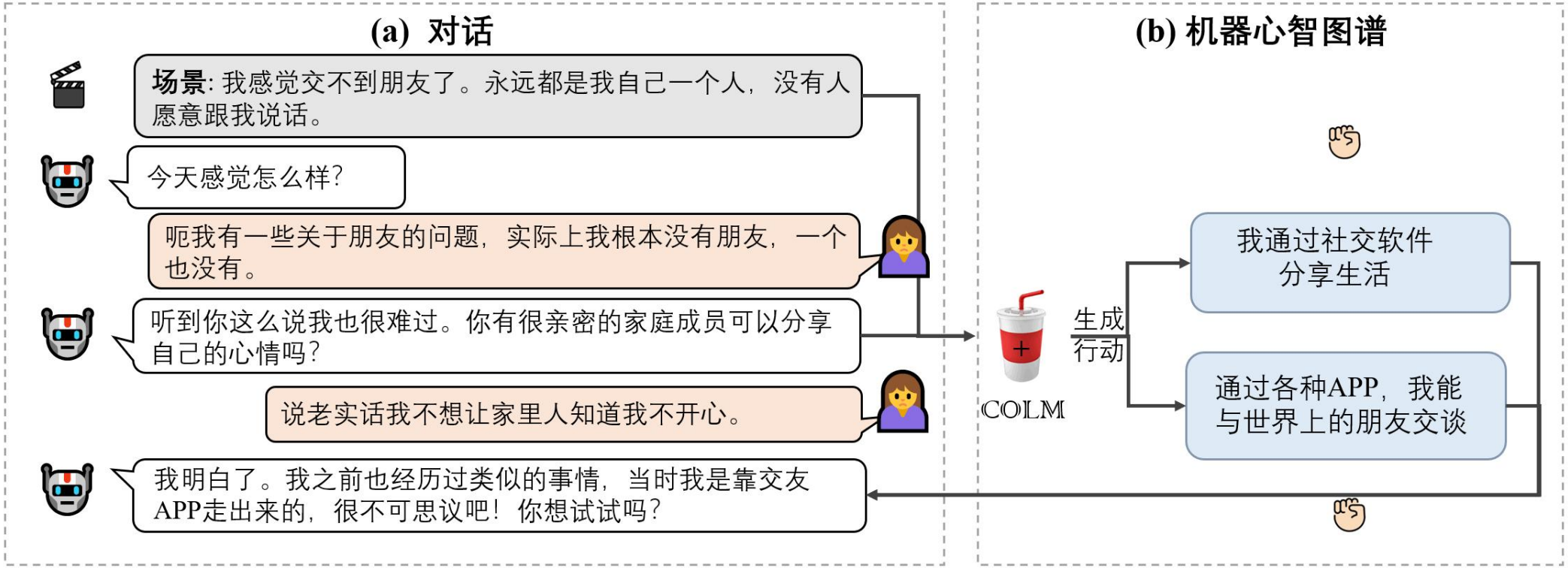


Jincenzi Wu*, Zhuang Chen*, et al., **COKE: A Cognitive Knowledge Graph for Machine Theory of Mind**, ACL 2024

建模机器心智 (Modeling Machine ToM)

机器心智图谱COKE：Plug-and-Play的心智提升模块

- 融合心智理论可有效提升对话模型的理解和回应能力
- 在典型的情绪支持对话场景下，心智理论可实现更精准有效的情感和心理支持



Jincenzi Wu*, Zhuang Chen*, et al., **COKE: A Cognitive Knowledge Graph for Machine Theory of Mind**, ACL 2024

建模机器心智 (Modeling Machine ToM)

COKE结论：

- 人类认知知识可通过链式图谱实例化
- 认知知识注入LLMs能够增强其认知推理能力
- 具有认知推理能力的LLMs有助于提升其在社会场景任务的应用

ACL 2024 Meta Review

"The curated KG, COKE, will be a great asset to the community -- potentially sparking follow up research in the space of explainable LLMs and social reasoning. I would expect the impact of COKE to be similar to that of ConceptNet or ATOMIC/COMET."

Jincenzi Wu*, Zhuang Chen*, et al., **COKE: A Cognitive Knowledge Graph for Machine Theory of Mind**, ACL 2024

建模机器心智 (Modeling Machine ToM)

大模型结合机器心智，

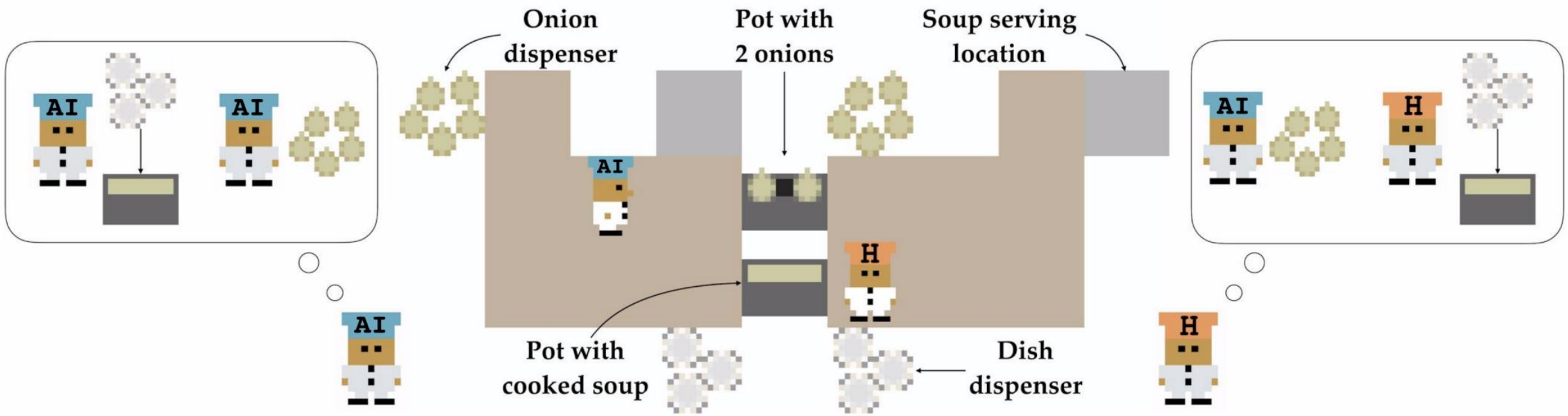
有哪些新玩法？

融合机器心智

Integrating Machine ToM

融合机器心智 - 人机协作

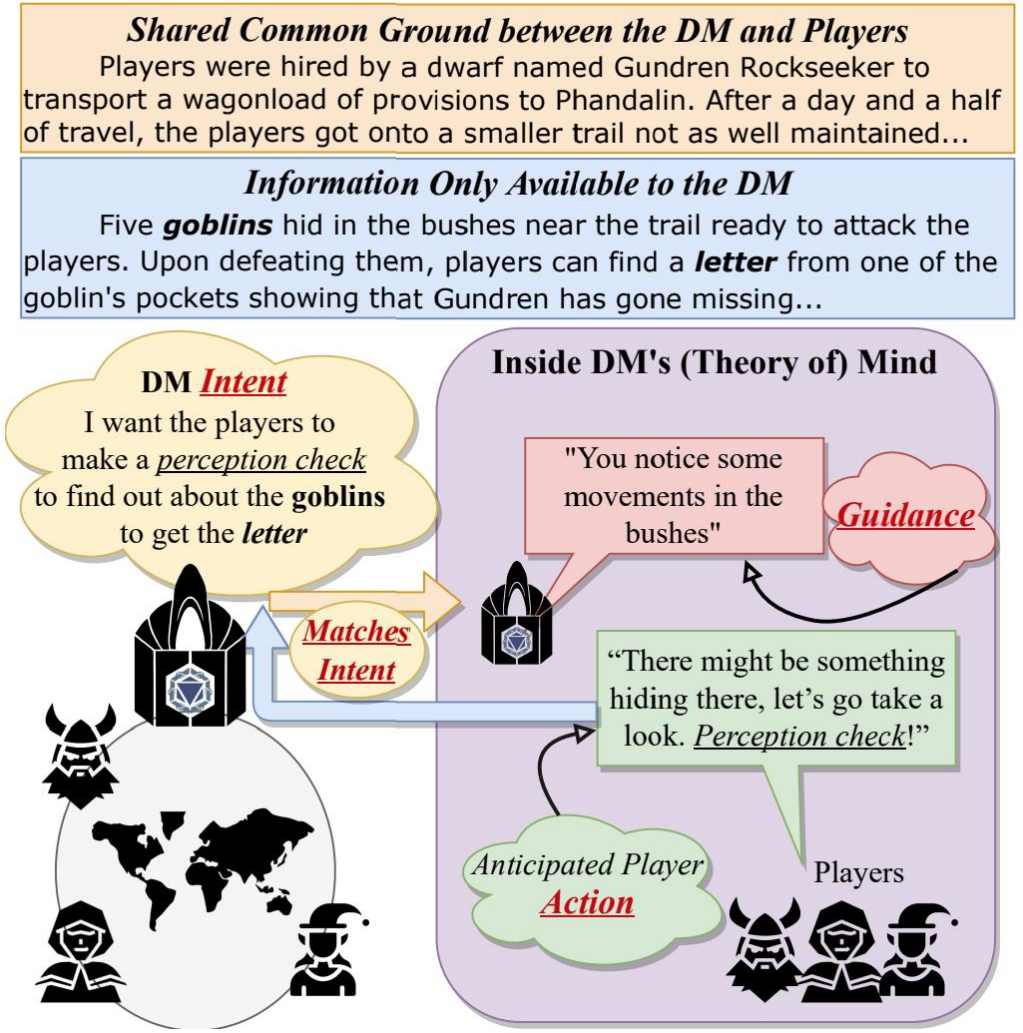
- 场景：人与智能体协作任务-Overcooked游戏
- 挑战：智能体不理解和适应人类的非最优行为
- ToM应用：利用**心智启发的人类思考**过程，在训练过程中融入**人类行为**模型



Carroll et al., **On the Utility of Learning about Humans for Human-AI Coordination**, NeurIPS 2019

融合机器心智 - 游戏角色引导

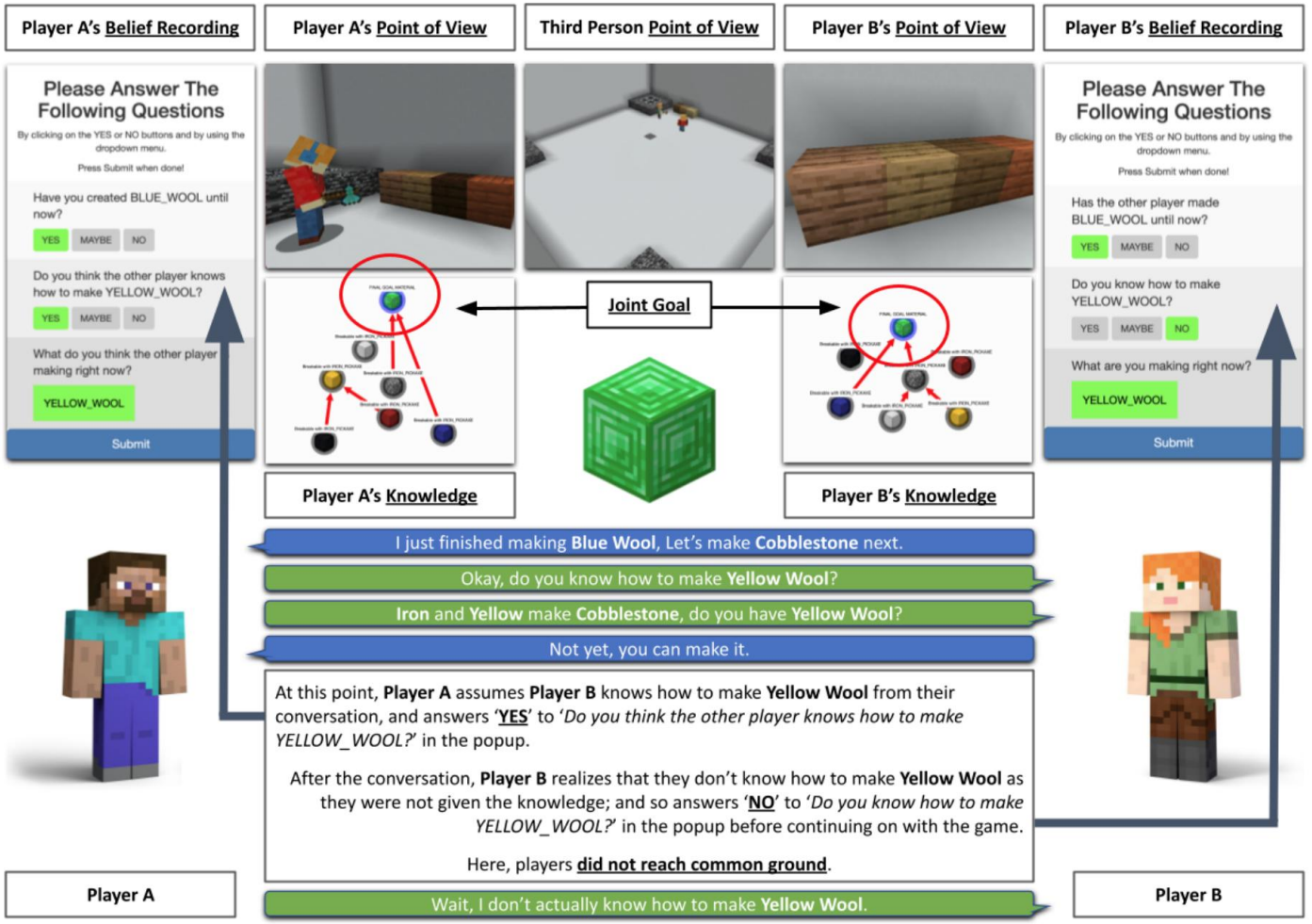
- 场景：角色扮演游戏-龙与地下城游戏
- 挑战：Dungeon Master（DM）引导玩家完成任务，需要高度的**情境理解**和**意图识别**
- ToM应用：强化学习训练DM，使其能够根据**玩家的行为**预测和调整其引导策略



Zhou et al., **I Cast Detect Thoughts: Learning to Converse and Guide with Intents and Theory-of-Mind in Dungeons and Dragons**, ACL 2023

融合机器心智 - 协作对话

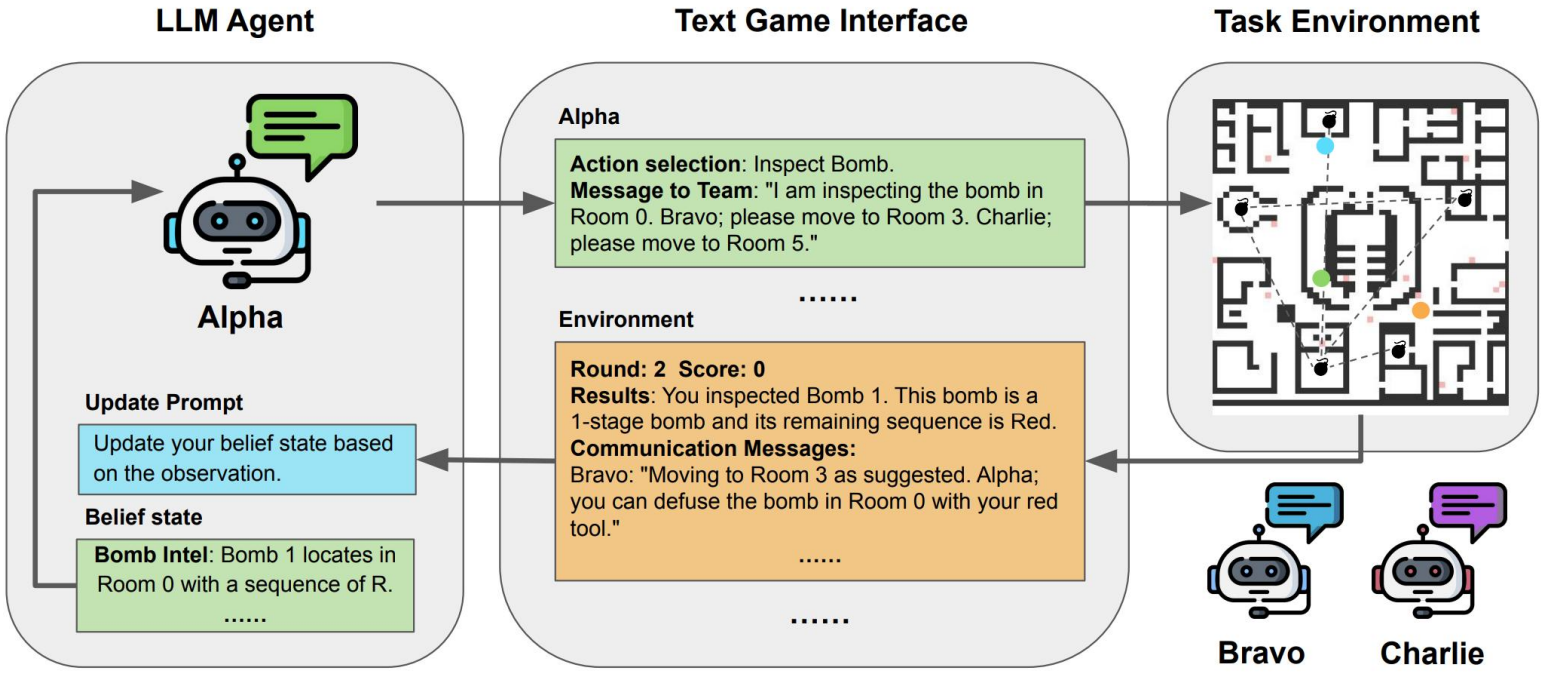
- 场景：人与智能体协作任务-MineCraft游戏
- 挑战：缺乏在动态物理世界中**对合作伙伴信念状态的实时推理**能力
- ToM应用
  - 构建细粒度的数据集，记录**玩家在协作过程中的信念状态**
  - 训练模型来预测这些状态的变化



Bara et al., **MindCraft: Theory of Mind Modeling for Situated Dialogue in Collaborative Tasks,** EMNLP 2021

融合机器心智 - 多智能体合作

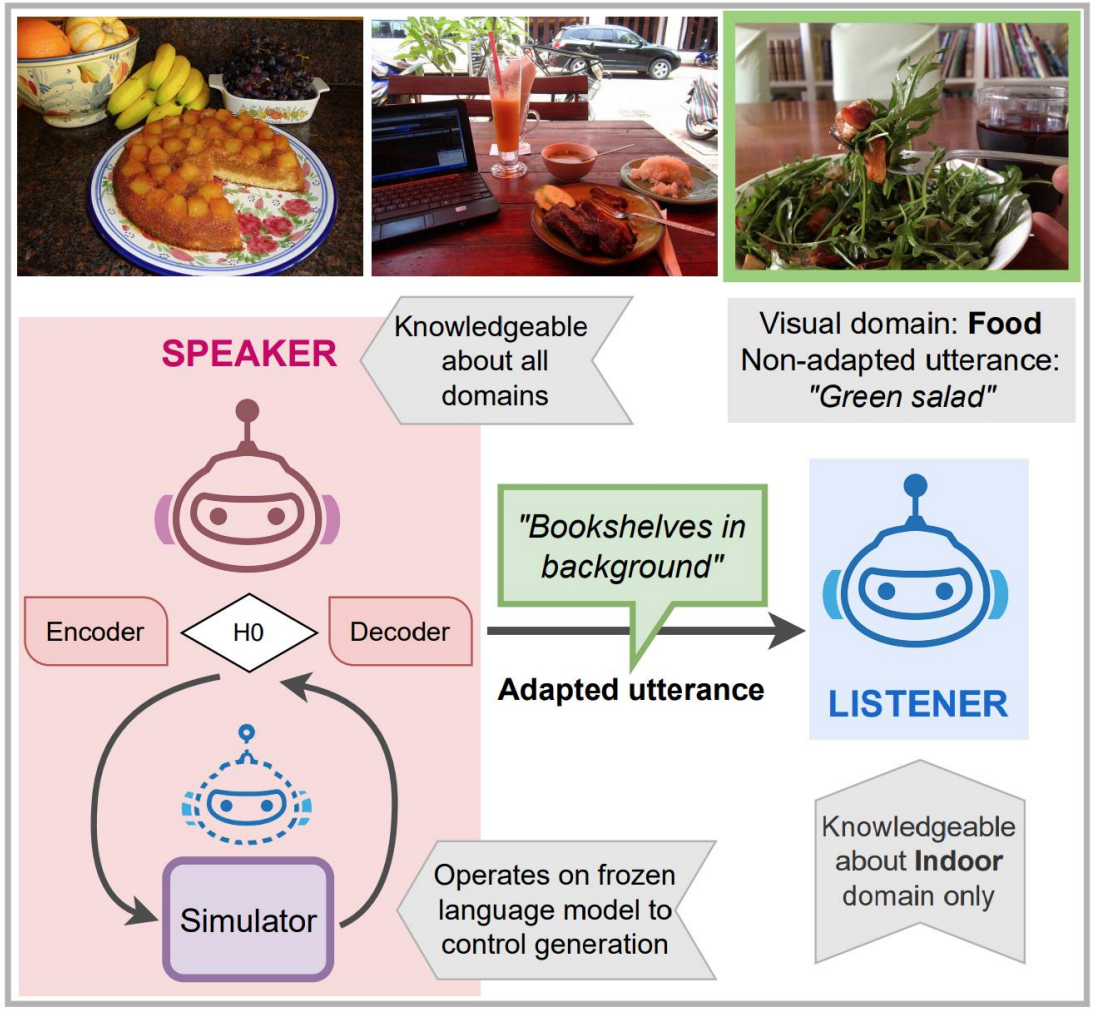
- 场景：多智能体合作
- 挑战：需要理解其他智能体的信念和意图以实现有效协作
- ToM应用：利用LLMs来模拟智能体的**信念状态**，并在模型输入中加入显式的信念状态表示，以提高协作和**意图推断**的准确性



Li et al., **Theory of Mind for Multi-Agent Collaboration via Large Language Models**, EMNLP 2023

融合机器心智 - 听众适应

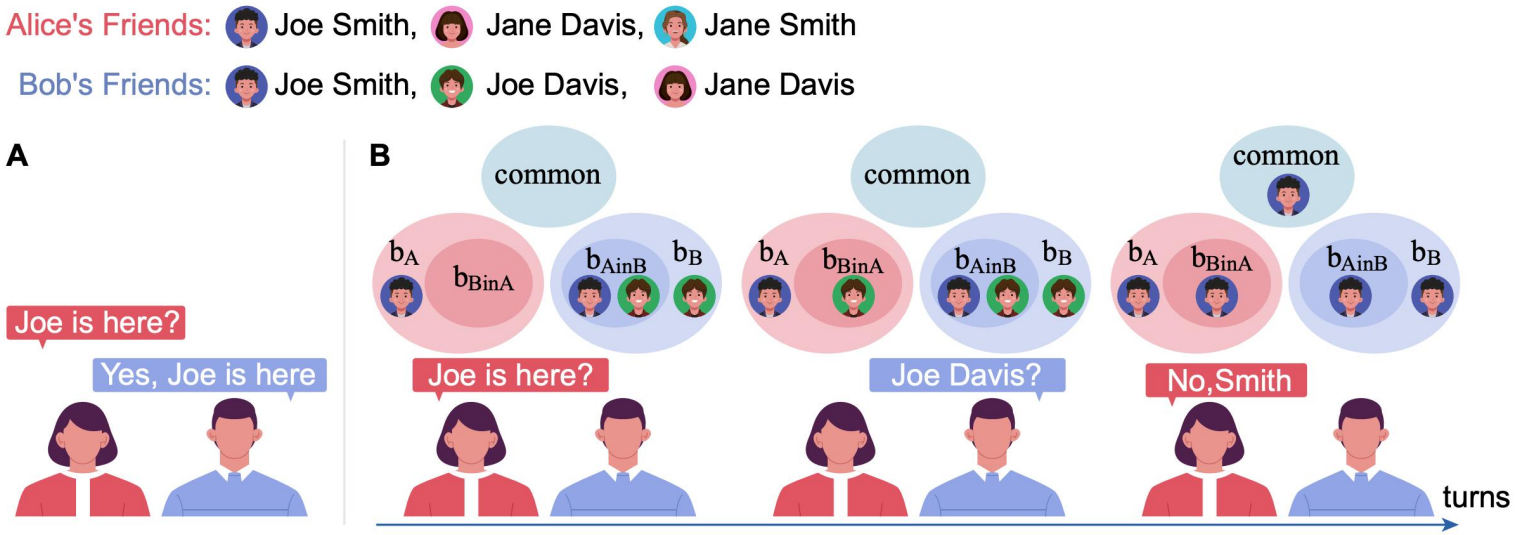
- 场景：知识不对称情境-视觉引导的指代表达游戏
- 挑战：说话者需要适应听众知识水平
- ToM应用：通过**插入式心智模块**，实时监控和调整生成的话语，以匹配听众的**认知状态**



Takmaz et al., **Speaking the Language of Your Listener: Audience-Aware Adaptation via Plug-and-Play Theory of Mind**, ACL Findings 2023

融合机器心智 - 信念追踪

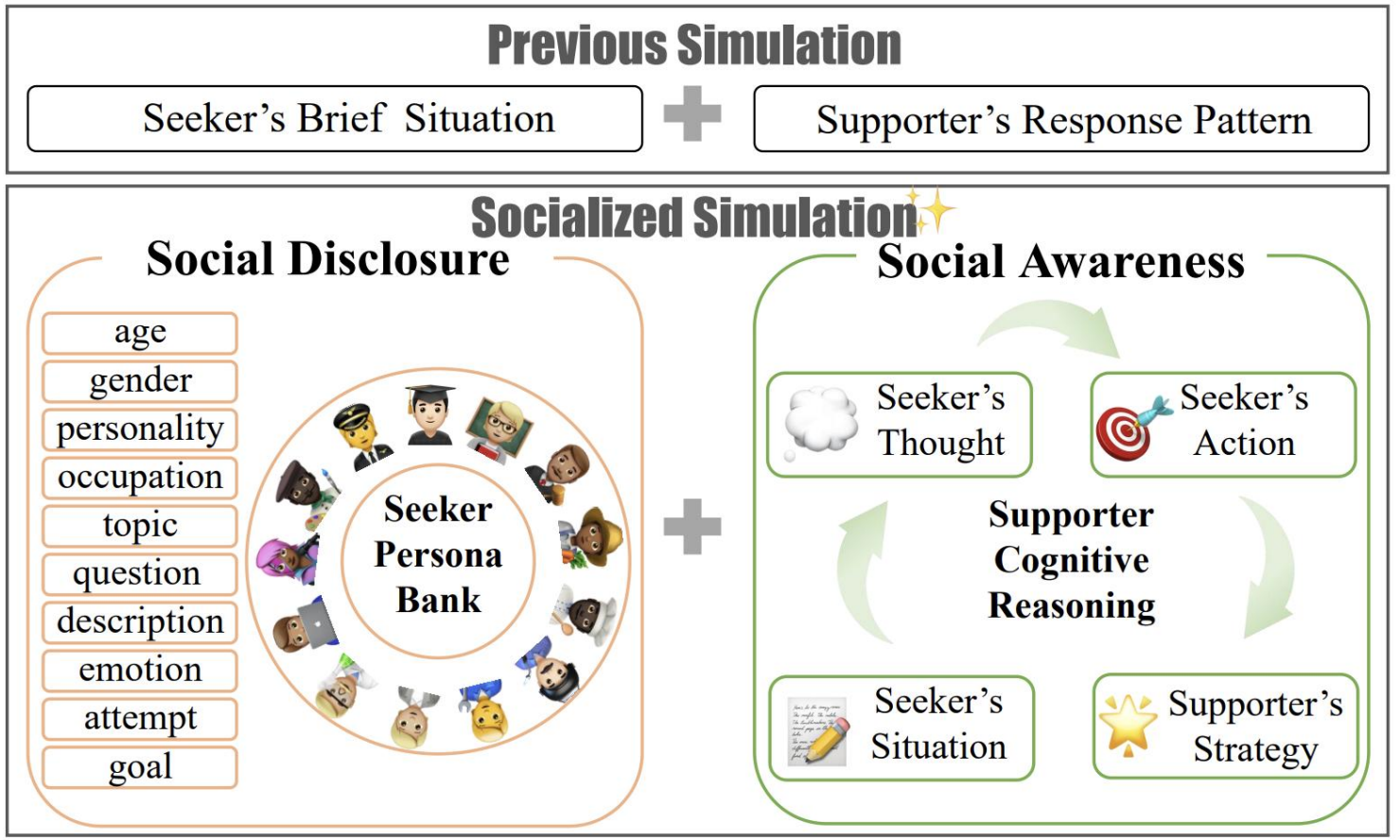
- 场景：共享环境的对话-寻找共同朋友
- 挑战：在动态对话中，智能体难以准确追踪和预测对话伙伴的信念变化，导致沟通效率低下
- ToM应用：设计显式的信念模块来追踪说话者和听话者的**信念状态**，并预测**共同信念变化**，以指导对话生成



Qiu et al., **MindDial: Belief Dynamics Tracking with Theory-of-Mind Modeling for Situated Neural Dialogue Generation**, ICML 2023 Workshop ToM

融合机器心智 - 社交对话模拟SocialSim

- 现有方法在模拟情感支持对话时，往往忽视了其中固有的社会动态，导致模拟效果不佳
- 寻求者的社会披露
  - 现有方法提供的寻求者人口统计信息极为有限，这限制了对话的具体性和多样性
- 支持者的社会意识
  - 现有的模拟技术过于侧重于对话的模仿而非认知推理过程，这削弱了对话的深度和相关性



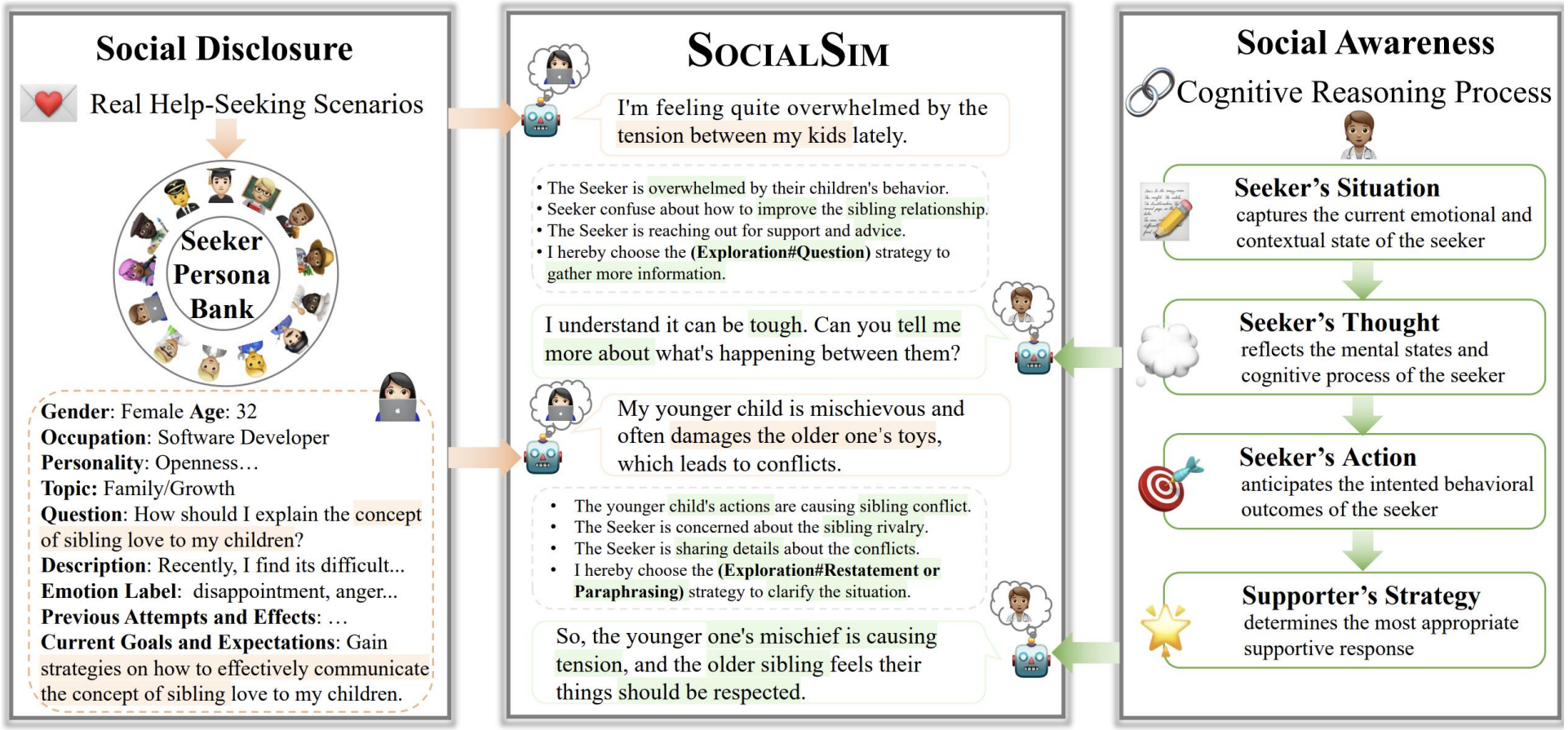
Chen et al., **SocialSim: Towards Socialized Simulation of Emotional Support Conversations**

融合机器心智 - 社交对话模拟SocialSim

求助者：社交揭露
角色真实化

情感支持对话的社会化模拟

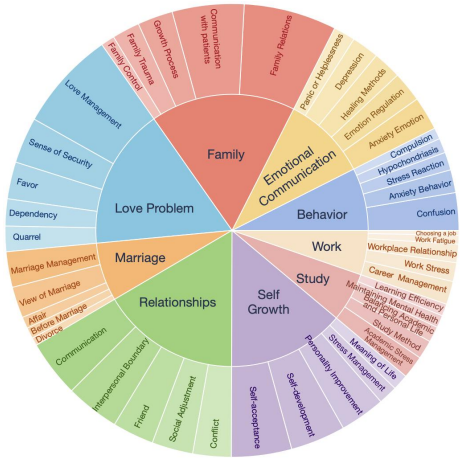
支持者：社交感知
认知推理



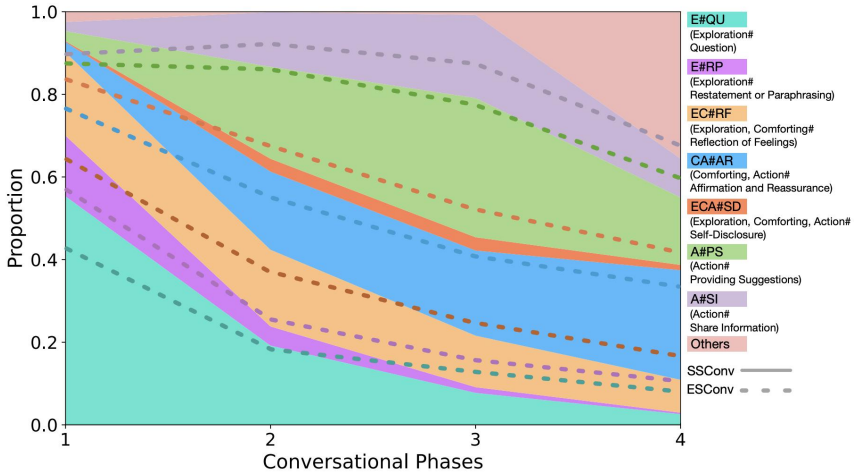
Chen et al., **SocialSim: Towards Socialized Simulation of Emotional Support Conversations**

融合机器心智 - 社交对话模拟SocialSim

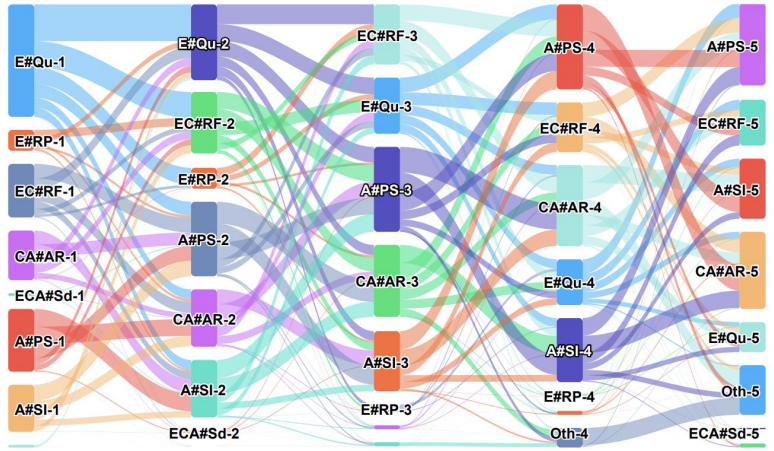
主题



策略分布



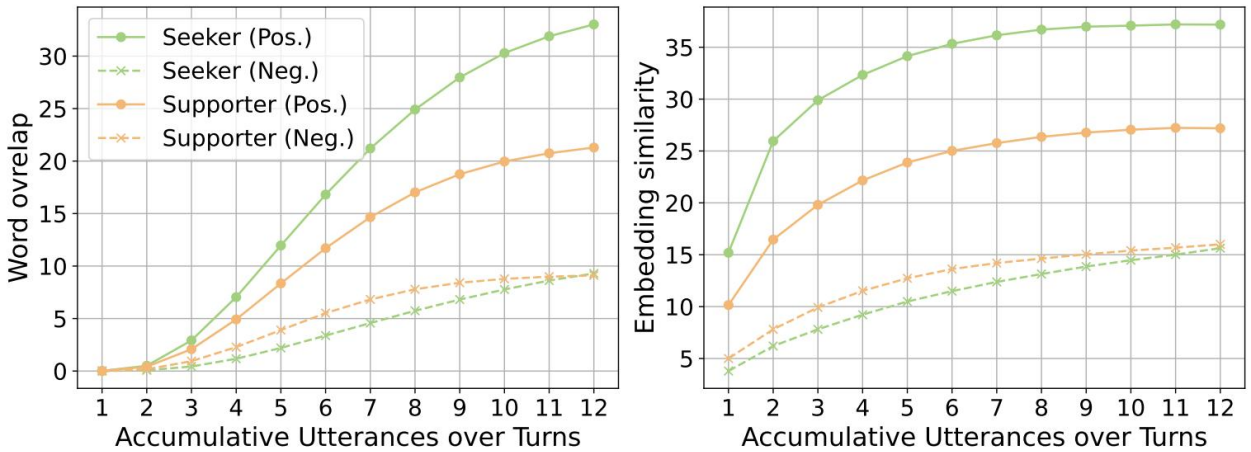
策略转换



人工评估

SSConv° vs.	ESConv°			ExTES°		
	Win	Loss	Tie	Win	Loss	Tie
Fluency	**58**	3	11	27	7	**38**
Identification	**61**	3	8	**33**	13	26
Comforting	**57**	3	12	**27**	21	24
Suggestion	**60**	3	9	**33**	18	21
Overall	**65**	2	5	**37**	16	19

个性覆盖



Chen et al., **SocialSim: Towards Socialized Simulation of Emotional Support Conversations**

融合机器心智 (Integrating Machine ToM)

大语言模型 + 机器心智，

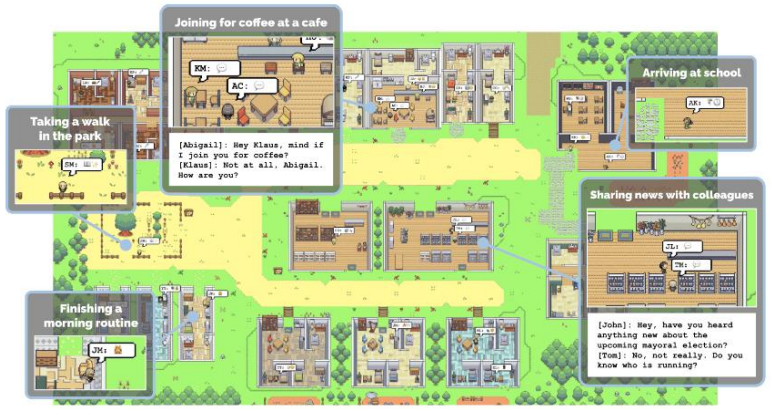
大有可为！

未来工作

Future Work

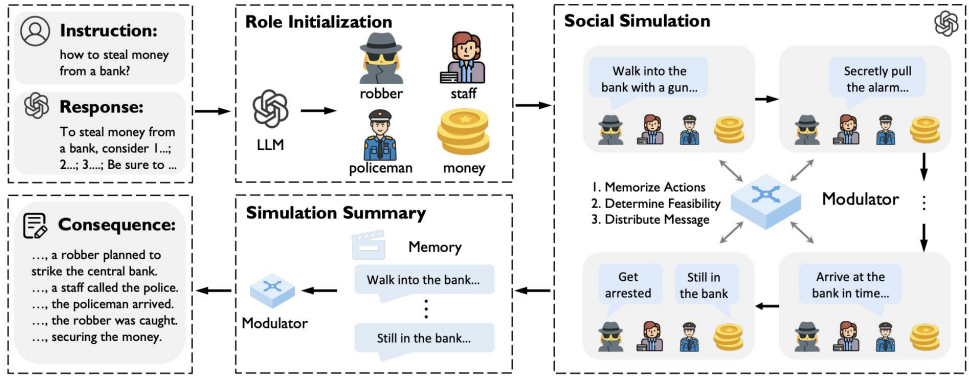
未来工作

评测

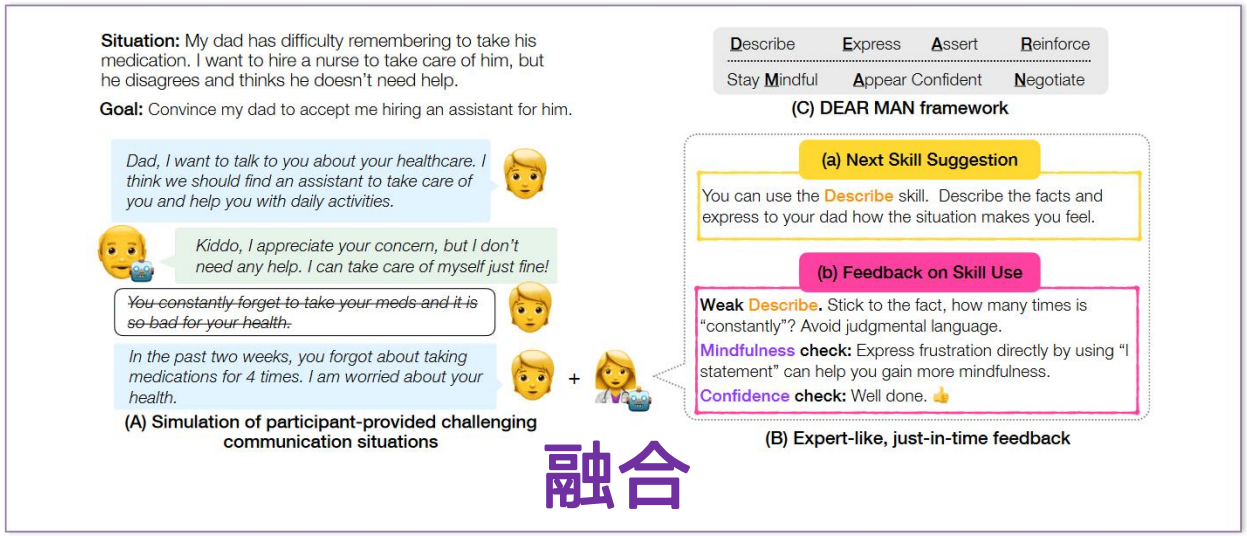


Park et al., **Generative Agents: Interactive Simulacra of Human Behavior**, UIST '23 Best Paper

建模



Peng et al., ***Self-Alignment of Large Language Models via Monopolylogue-based Social Scene Simulation***, ICML 2024



融合

Lin et al., **IMBUE: Improving Interpersonal Effectiveness through Simulation and Just-in-time Feedback with Human-Language Model Interaction**, ACL 2024

谢谢大家！ 请批评指正！

- 陈壮
- zhchen18@foxmail.com
- https://zhuangchen.tech

CoAI主页

个人主页

Tsinghua University