心理科学进展 2025, Vol. 33, No. 12, 2027–2042 https://doi.org/10.3724/SP.J.1042.2025.2027
Advances in Psychological Science https://cstr.cn/32111.14.2025.2027

·第二十七届中国科协年会学术论文·

大语言模型的人工心理理论：证据、界定与挑战*

杜传晨[1] 郑远霞[1,2] 郭倩倩[1] 刘国雄[1]

([1]南京师范大学心理学院, 南京 210097) ([2]宁波大学教师教育学院, 浙江 宁波 315211)

摘 要 传统观点认为，心理理论是有意识的生物独有的社会认知能力。然而，近来迅速发展的大语言模型能够解决多种心理理论任务，这引发了关于其是否具有心理理论的激烈争议。大语言模型的人工心理理论在表现上与心理理论类似，但内部过程不同。本文首先从评估对象和任务特征两个方面系统梳理人工心理理论研究，通过综合分析 GPT-4 在心理理论任务中的较高通过率及使其表现受限的内外因素，指出当前模型的心理理论表现与人类相似。其次，通过深入对比支撑心理理论与人工心理理论的神经基础和发展因素，揭示二者的内部过程存在本质差异，进而补充了人工心理理论的定义。未来研究应关注制定并采用标准化的评估方案，在共同心理理论框架下探究人工心理理论机制，以及与人类心理理论对齐。

关键词 大语言模型，心理理论，人工心理理论，人工智能

分类号 B842

1 引言

心理理论(Theory of Mind, ToM)指个体对自己或他人内部心理状态(包括愿望、信念、知识、意图等)的理解和推测，及由此对相应行为做出的因果性解释与预测，是社会认知(Social Cognition)和社会互动的基石(Premack & Woodruff, 1978; Wellman, 2017)。长久以来，心理理论被认为是有意识的生物独有的社会认知能力(Yıldız, 2025)。然而，人工智能(Artificial Intelligence, AI)的不断发展，尤其是近年来诸如 GPT、Claude、DeepSeek 等大语言模型(Large Language Models, LLMs)的迅速进展，重新塑造了人们对 AI 能力边界的认识。一项研究发现，GPT-4 在改编版错误信念(False Belief)任务中的通过率达到 75%，与 6 岁儿童的心理理论表现相当，表明 LLMs 可能自发涌现出心理理论(Kosinski, 2023)。该研究凸显了将心理理论等社会认知能力整合到 LLMs 中，以实现更自然的人机交互的潜力，同时引发了关于 LLMs 人工心理理论(Artificial Theory of Mind, AToM)的激烈争议。

模拟人类认知是 AI 研究的核心目标之一(Deshpande & Magerko, 2024)，该领域对心理理论的研究主要沿两个方向展开：其一，基于特定算法训练能够模拟心理理论的 AI 模型，通常被称为机器心理理论(Machine Theory of Mind)方向；其二，研究未经特定训练的 LLMs 能否以及如何涌现出心理理论，通常被称为人工心理理论方向。机器心理理论研究主要聚焦于两类模型的构建：基于贝叶斯方法建模人类心理状态的因果关系，即通过行为观察间接推断心理状态(Baker et al., 2011; Yuan et al., 2022)；基于深度学习(Deep Learning)范式，建立行为与心理状态间的直接映射关系(Rabinowitz et al., 2018; Sclar et al., 2022)。两类模型各有优势，前者具有较好的解释性，后者能够反映内隐的心理状态。然而，它们仅能解决特定的心理理论任务，尚不具备人类水平的心理理论能力(Mao et al., 2024)。而人工心理理论研究的兴起，源于 LLMs 经规模扩展后，涌现出一些未通过针对性训练而获得的，小型模型不具备的能力(舒文韬 等, 2023; Wei, Tay, et al., 2022)。使用心理学行为任务评估 LLMs，分析任务输入

收稿日期：2025-05-10
* 国家重点研发计划(2023YFC3341301)资助。
通信作者：刘国雄, E-mail: 17219367@qq.com

与模型输出间的映射关系, 能够解构其表现及内在机理(Hagendorff, 2024)。人工心理理论研究遵循这一思路, 研究者使用改编或定制(ad-hoc)的心理理论任务评估 LLMs, 推断其是否具有心理理论。然而, 现有研究的结论存在分歧。例如, Trott 等人(2023)采用改编版错误信念任务评估 GPT-3, 其通过率为 74.5%, 与 Kosinski (2023)研究中 GPT-4 在同类任务中的通过率(75%)几乎一致, 但两项研究对 LLMs 是否具有心理理论持相反观点。这也引出另一问题: 仅通过模型在任务中的表现, 能否判定其人工心理理论类似于人类心理理论? 有研究者通过细致反思 Kosinski (2023)实验的结构和方法, 提出 LLMs 能够表现出类似人类的心理理论能力, 但不具备与人类相同的心理状态和过程(赵泽林, 程聪瑞, 2024)。

此外, 目前对 LLMs 心理理论的术语界定尚未达成统一共识。一方面, 机器心理理论、神经心理理论(Neural Theory of Mind)和人工心理理论被混用于描述 LLMs 的心理理论。其中, 机器心理理论侧重基于特定算法, 使用非常有限的数据开发能够模拟人类心理理论表现的 AI 模型(Lebiere et al., 2025); 神经心理理论强调机器理解和追踪他人心理状态的能力(Xu et al., 2024); 人工心理理论则关注人机交互中智能体(Intelligence Agents)的心理理论(Williams et al., 2022; Winfield, 2018), 强调 LLMs 对人类心理理论表现的模拟(Bendell et al., 2024; Marchetti et al., 2023)。鉴于 LLMs 本质上是基于海量文本进行无监督训练的深度学习系统(Blank, 2023), 其未经针对性建模便可通过多种心理理论任务, 但尚不明确模型能否理解和追踪他人的心理状态(Jin et al., 2024), 因此使用人工心理理论较为合理。另一方面, 人工心理理论作为研究实践中逐渐形成的共识性术语, 其定义尚不完善。Marchetti 等人(2025)将人工心理理论界定为: LLMs 在心理学研究所确立的心理理论任务中, 表现出与人类一致反应的能力。然而, 该定义仅聚焦于 LLMs 在心理理论任务中与人类表现的一致性, 尚未明确人工心理理论与心理理论的差异, 并不足以界定人工心理理论。

综上, 针对当前人工心理理论研究存在的问题, 本文首先系统梳理了相关实证研究文献, 通过综合分析 GPT-4 在心理理论任务中的较高通过率及使其表现受限的内外因素, 指出当前 LLMs 具有在表现上与心理理论相似的人工心理理论。其次, 本文通过深入对比支撑心理理论与人工心理理论的神经基础和发展因素, 揭示二者的内部过程存在本质差异, 并根据此差异补充了人工心理理论的定义。最后, 本文总结出在评估方案、人工心理理论机制与心理理论对齐三个方面的挑战, 并提出了相应展望: 制定并采用标准化的评估方案, 在共同心理理论框架下探究人工心理理论机制, 以及与人类心理理论对齐。

2 大语言模型具有人工心理理论的证据

为全面分析 LLMs 在心理理论任务中的表现, 本文系统检索了自 transformer 架构提出后(Vaswani et al., 2017), 2017~2025 年间的相关文献。英文数据库(Web of Science)中以关键词“theory of mind”、“artificial theory of mind”、“machine theory of mind”、“neural theory of mind”与“Artificial Intelligence”、“Large Language Models”、“AI”、“LLMs”、“ChatGPT”之间组合的方式进行检索。检索截止日期为 2025 年 6 月, 共检索到 409 篇文献(文献检索与筛选流程见图 1)。随后从评估对象和任务特征两个方面进行整理(见表 1)。

评估对象方面, 多数研究(96.2%)聚焦于评估 GPT 系列模型的心理理论表现, 部分研究(42.3%)同时纳入人类被试进行人机对比。作为当前主流的 LLMs 之一, GPT 系列模型通过 transformer 架构压缩整合了广泛的世界知识, 具备了较为全面的任务解决能力(Zhao et al., 2023), 现已从 GPT-1 迭代更新至 GPT-5。不同版本的模型之间的差异主要在于学习算法和规模。除使用同一生成算法 transformer 外, 研究者还通过学习算法提升 GPT 系列模型回答的准确性(Chang, 2023)。GPT-3 经上下文学习(In-context Learning)提高了泛化能力(Dong et al., 2024); GPT-3.5 在 GPT-3 的基础上进行指令微调与基于人类反馈的强化学习(Reinforcement Learning from Human Feedback, RLHF), 获得了更好的指令遵循和对齐人类意图的能力(Ouyang et al., 2022); GPT-4 同样进行了基于人类反馈的强化学习, 其输出更符合人类的表达和意图(Zhou et al., 2024), 在多种认知任务中接近人类水平(Rathje et al., 2024)。模型规模主要

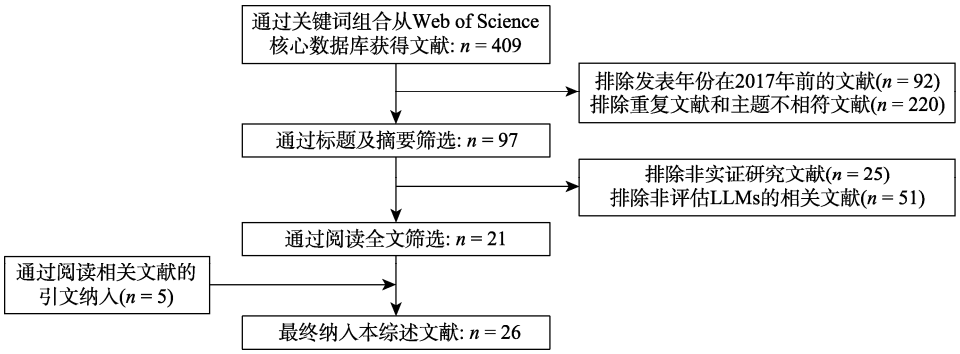

图 1　文献检索与筛选流程

表 1　人工心理理论研究总结

文献	LLMs	人类被试	定制任务	模态	ATOMS 心理状态								
					一阶信念	高阶信念	行动意图	交流意图	愿望	情感	知识	感知	非字面交际
Sap et al., 2022	GPT-3			文本	√								
Trott et al., 2023	GPT-3	√		文本	√								
Van Duijn et al., 2023	GPT-3, 3.5, 4; Llama 等	√		文本	√	√		√					√
Shapira et al., 2023	GPT3			文本	√								√
He et al., 2023	GPT-3.5, 4		√	文本	√	√							
Kim et al., 2023	GPT-3.5, 4; Llama-2 等	√	√	文本	√	√					√		
Li et al., 2023	GPT-3.5, 4		√	文本	√	√					√	√	
Sileo & Lernould, 2023	GPT-3, 3.5, 4	√	√	文本	√	√							
Ma et al., 2023	GPT-3.5, 4			文本	√	√							
Ullman, 2023	GPT-3.5			文本	√								
Gandhi et al., 2023	GPT-3.5, 4; Claude 等		√	文本	√	√			√			√	
Brunet-Gouet et al., 2023	GPT-3.5, 4			文本	√	√		√					√
Jones et al., 2024	GPT-3	√		文本	√	√		√		√	√		√
Verma et al., 2024	GPT-3.5, 4	√	√	视觉; 文本		√					√	√	
Wilf et al., 2024	GPT-3.5, 4; Llama-2 等		√	文本	√	√			√			√	
Shapira et al., 2024	GPT-3.5, 4; Flan 等		√	文本	√	√	√			√			√
Strachan et al., 2024	GPT-3.5, 4; Llama-2	√		文本	√	√		√					√
Chen et al., 2024	GPT-4	√	√	文本	√	√		√	√	√	√		√
Jin et al., 2024	GPT-3.5, 4; Llama-2 等	√	√	视觉; 文本	√	√	√						
Xu et al., 2024	GPT-3.5, 4; Mixtral 等		√	文本	√	√	√		√	√		√	
Yongsatianchot et al., 2024	GPT-4; Claude3 等		√	文本	√	√		√	√	√			
Amirizaniani et al., 2024	GPT-4; Llama-2 等	√	√	文本				√		√			
Attanasio et al., 2024	GPT-3.5, 4	√		文本						√			√
Kosinski, 2024	GPT-1, 2, 3, 3.5, 4 等			文本	√								
Nickel et al., 2024	Llama-2; Mixtral 等		√	文本	√	√							
Ünlütabak & Bal, 2025	GPT-3.5, 4			文本	√	√							

指训练数据和参数规模, 规模扩展能够较为显著地提升模型性能(舒文韬 等, 2023)。GPT-3 使用了45TB 的训练数据, 参数规模为 1750 亿; GPT-3.5 未扩展规模; GPT-4 的具体细节未公开发布, 但估计训练数据和参数规模增长巨大(Wu et al., 2023)。

任务特征主要包括是否使用定制任务、任务模态以及评估的心理理论子能力。首先, 鉴于 LLMs 的训练数据中可能包含经典任务(Sarıtaş et al., 2025), 存在污染评估结果的风险, 相关研究普遍使用改编或定制任务评估模型的人工心理理论。改编任务是在经典心理理论任务的基础上, 通过调整原任务场景或生成新场景构建的任务(Van Duijn et al., 2023); 定制任务指参照经典心理理论任务的逻辑, 专为 LLMs 设计的新任务, 例如, 在信息不对称的即时对话情境中测试模型心理理论表现的 FANToM 任务(Kim et al., 2023); 能够测量模型对六类心理状态理解的系统评估框架 ToMBench (Chen et al., 2024)。改编和定制任务的差异主要体现在提示(prompts)上, 即人类向模型提供的、用于引导其输出的任务文本(Meskó, 2023)。提示影响 LLMs 输出文本的质量(Lin, 2024), 进而影响其心理理论表现。此外, 两类任务在任务结构与逻辑、反应形式、评估的心理理论子能力等维度均无实质性差异。其次, 在任务模态方面, 多数研究(92.3%)使用文本形式的任务, 只有两项研究采用视频加文本形式的多模态任务评估了多模态(multimodal)模型 GPT-4V 的人工心理理论(Jin et al., 2024; Verma et al., 2024)。最后, 根据 Beaudoin 等人(2020)提出的心理理论空间能力架构(ATOMS), 并参考 Sarıtaş 等人(2025)对该框架的扩展, 梳理了各项研究评估的心理理论子能力, 包括对一阶与高阶信念、行动与交流意图、愿望、情感、知识、感知和非字面交际的理解。

2.1 GPT-4 能够解决多数心理理论任务

人类心理理论是按照稳定、可预测的序列发展的(O'Hare et al., 2009; Westby & Robinson, 2014)。鉴于 GPT-4 性能良好且受到研究者的广泛关注, 遵循人类心理理论关键子能力的发展序列系统梳理其任务表现, 最有可能揭示当前 LLMs 是否具有人工心理理论。

错误信念任务是评估儿童是否具有成熟心理理论的黄金标准(Sodian et al., 2020), 17 项研究使用其改编或定制版本考察 GPT-4 对他人持有与自身不同的错误信念的理解。综合结果表明, GPT-4 能够解决一阶错误信念任务。例如, 有研究者使用两种改编程度不同的错误信念任务验证了 GPT-4 能够稳定地通过测试(Van Duijn et al., 2023); Strachan 等人(2024)在七个改编任务中观察到, GPT-4 的通过率接近 100%, 表现出"天花板效应"; 使用定制的 ToMBench 任务集的研究发现, GPT-4 在错误信念任务中的通过率高达 89.5%, 略高于人类被试(86.8%) (Chen et al., 2024)。

二阶错误信念任务要求个体理解某人可能对他人的信念持有错误信念(Osterhaus et al., 2016), 17 项研究评估了 GPT-4 在此类任务中的表现。综合结果发现, GPT-4 的通过率仍处于较高水平。例如, 在经典和改编任务中, GPT-4 的表现均优于 7~10 岁儿童(Van Duijn et al., 2023); GPT-4 连续三次通过了改编版本的二阶错误信念任务的全部题目(Brunet-Gouet et al., 2023); 与一阶任务类似, GPT-4 以高于人类被试的正确率(94.3%与 81.0%)通过了 ToMBench 任务集中的二阶错误信念任务(Chen et al., 2024)。

六项研究评估了 GPT-4 在暗示(hinting)、失言(faux pas)、反讽(irony)和奇异故事(strange stories)等高阶心理理论任务中的表现。模型能够解决多种改编版本的奇异故事任务(Attanasio et al., 2024; Strachan et al., 2024; Van Duijn et al., 2023); 在改编的反讽任务中, GPT-4 同样表现出高于人类被试的水平(Strachan et.al., 2024); 对于 ToMBench 任务集, GPT-4 能够通过其中的失言任务, 在暗示任务中的表现略低于人类被试水平(Chen et al., 2024); Brunet-Gouet 等人(2023)发现 GPT-4 在暗示任务中的表现与先前研究中人类被试的水平接近。

综上, GPT-4 在多项心理理论任务中均能模拟人类被试的反应, 似乎已具有人工心理理论。然而, 使用行为实验评估 LLMs 时, 任务通过率可能不足以反映其能力, 模型表现欠佳的情况同样值得关注(Ullman, 2023)。

2.2 导致模型表现受限的内外因素

在特定任务设置下, GPT-4 的表现欠佳。其一, 任务提示文本的扰动导致模型表现下滑。Ullman (2023)通过调整 Kosinski (2023)所用错误信念任

务的关键信息, 例如将原本装有爆米花却贴有巧克力标签的不透明袋子改为透明容器, 验证了提示的扰动会导致 GPT-3.5 任务性能的急剧下滑; 多项后续研究重复证实, GPT-3.5 与 GPT-4 均无法解决调整后的透明访问变体任务(Ünlütabak & Bal, 2025; Brunet-Gouet et al., 2023; Shapira et al., 2024)。其二, GPT-4 在部分定制任务中表现受限。虽然 GPT-4 成功通过了 ToMBench 任务集中的错误信念、失言等子任务, 但其在全部子任务上的平均通过率(75.3%)低于人类被试(85.4%) (Chen et al., 2024); 在 FANToM 任务中, GPT-4 的整体表现同样低于人类基线(Kim et al., 2023); Sileo 和 Lernould (2023)使用动态认知逻辑任务评估了 GPT-4 对错误信念的理解, 发现其能够表现出认知推理的表面形式, 但偶尔会出现明显错误; 在定制的 BigToM 任务中, GPT-4 表现出与人类被试相似的推理模式, 但在需要根据行动推断信念时存在困难(Gandhi et al., 2023)。

然而, 模型表现受限不能简单地归结为其缺乏心理理论能力, 也可能归因于提示设计、基线设定与测试项目效度等外部因素(Kosinski, 2024)。首先, 提示设计不当可能使模型的表现与心理理论无关。例如, Pi 等人(2024)通过对透明访问变体任务进行最小化的、渐进的修改, 发现 GPT-4 的受限表现可能源于其缺乏“某人看透明容器意味着同时识别容器内物品”的常识推断; 有研究发现约 50%的人类被试同样无法通过透明访问变体任务(Strachan et al., 2024)。其次, 多数研究未纳入代表性人类样本作为基线。例如, Chen 等人(2024)的研究将 20 名中国研究生协作完成一套任务的通过率作为人类基线, 样本代表性有限。此外, 该研究未报告推断统计结果, 难以确定 GPT-4 与人类被试的表现之间是否存在显著差异, 特别是当 GPT-4 的通过率处于较高水平时。Kim 等人(2023)的研究以 11 名研究生在 32 组对话上的表现作为人类基线, 而 GPT-4 需要完成 256 组对话任务, 对比结果的有效性存疑。最后, 测试项目的效度问题可能直接影响 GPT-4 的表现。例如, ToMI 任务的叙述模棱两可, 推理过程中存在额外干扰因素(Gandhi et al., 2023); OpenToM 任务仅由 3 位研究者标注问题设计的合理性、主观性等, 未进行规范的效度验证(Xu et al., 2024)。

此外, GPT-4 自身的固有局限(表现形式与形成原因见表 2)是导致其难以通过部分任务的内部因素, 包括幻觉、超保守主义、常识错误、捷径或伪相关、虚假因果推断、推理深度不足、时间无知与长时程上下文等(He et al., 2023; Li et al., 2023; Ma et al., 2023; Shapira et al., 2024; Strachan et al., 2024)。其中, 幻觉广泛存在于各种自然语言处理任务中(林曦, 2025; Ma et al., 2023), 分为事实性(factuality)幻觉和忠实性(faithfulness)幻觉, 前者描述 LLMs 生成的内容与事实间存在差异, 后者则强调模型的输出偏离用户输入或存在内部矛盾(Huang, Yu, et al., 2025)。上下文外推理(Out-of-Context Reasoning, OCR)是同时驱动 LLMs 产生泛化和幻觉的核心机制, 模型表现出的行为取决于被关联的概念间存在因果关系还是伪相关(Huang, Zhu, et al., 2025), 即幻觉根源于 LLMs 的底层架构。在心理理论任务中, 研究者通过提示 LLMs 生成答案前进行视角采择, 以此保留目标角色已知的信息或明确表征其信念状态, 降低了模型产生幻觉的风险(Wilf et al., 2024; Li et al., 2023)。然而, 需采用综合数据层面、训练过程及推理阶段的整体优化策略才能有效缓解 LLMs 幻觉(Pei et al., 2025), 仅凭提示调整效果有限。与幻觉类似, 其他固有局限同样源于模型训练或推理过程中的多方面因素, 属于自然语言处理领域的研究难点(Ma et al., 2023), 本文不再展开分析。

固有局限的存在可能导致模型表现与能力的分离, 即 GPT-4 具备模拟人类心理状态推理的计算能力, 但受限于自身的固有局限, 在特定任务中失败(Strachan et.al., 2024)。因此, 不能由于模型存在固有局限而否定其任务表现。此外, 人类被试会依据预存的行为序列知识推断行为以解决心理理论任务(Korman & Malle, 2016), 该启发式推理策略与 LLMs 通过捷径处理问题相似, 暗示人类在任务中也存在某些反应倾向。

最后, 综合分析 GPT-4 在心理理论任务中的较高通过率及使其表现受限的内外因素, 能否认为 LLMs 已涌现出人工心理理论? 第一, 依据心理理论的评估标准并参考人工心理理论的定义, 即根据 LLMs 的任务表现进行推断(Sodian et al., 2020), GPT-4 在多数任务中展现出较高通过率, 能够表现出与人类一致的反应。第二, GPT-4 的表

表 2 GPT-4 的固有局限

固有局限	表现形式	形成原因
幻觉 (hallucinations)	答案包含无法从任务叙述中推断出的或与叙述相矛盾的信息; 编造不符合事实的细节和意图以补全推理过程	源于模型获取能力过程中的多方面因素, 贯穿数据层面、训练过程和推理阶段
超保守主义 (hyperconservatism)	过于谨慎而拒绝对任务中角色的信念做出推断	模型难以自发地减少推理中的不确定性; 避免提供错误答案
常识错误 (commonsense errors)	生成违背常识的回答, 或不具备某些常识	模型对常识的掌握与其在复杂推理中有效运用常识的能力割裂
捷径/伪相关 (heuristics/spurious correlations)	利用任务描述中简单的、表面的模式解决任务	模型倾向于生成续写内容; 表现出过度配合性, 即假设所有输入细节都重要
时间无知 (lack of temporal information)	对任务中主人公行为序列的时序理解存在偏差	模型尚未形成对时序关系的实质性理解
虚假因果推断 (spurious causal inference)	仅基于问题的表层模式做出错误或误导性的因果推断	模型难以把握文本间统计关联背后的深层逻辑
长时程上下文 (long-horizon contexts)	回答问题时忽略距离过远的上文内容	模型处理长文本的能力有限
推理深度不足 (insufficient reasoning depth)	跳过一些必要的推理步骤, 将高阶问题转化为低阶问题	训练数据中包含较多的简单模式; 模型难以有效保持并整合多步骤信息

现波动可能源于评估方法的局限性与模型的固有局限, 而与其人工心理理论无关(Ma et al., 2023)。因此, GPT-4 已具有在表现上与心理理论相似的人工心理理论。

3 人工心理理论的界定

能够解决心理理论任务的 LLMs 正在重塑研究者对心理理论的理解, 也凸显了研究心理理论内部过程的重要性。尽管人工神经网络与人脑一样难以被解释(Griffiths et al., 2024), 多项研究仍表明心理理论与人工心理理论的内部过程可能存在差异。例如, 模型在心理理论任务中并未像人类一样推理心理状态并动态追踪其变化(Jin et al., 2024; Ünlütabak & Bal, 2025), 而可能利用了任务数据结构中有效的替代规则(Lu, Zhang, et al., 2025); LLMs 依靠模式识别处理社会智能任务, 与人类的行为模式存在明显区别(Wang, Zhang, et al., 2024); 人类需抑制自我视角以解决心理理论任务(Van Der Meer et al., 2011), 模型缺乏自我视角, 不需要调和相互冲突的心理状态(Strachan et al., 2024)。因此, 为明确区分心理理论与人工心理理论, 有必要对比支撑二者的神经基础和发展因素, 以揭示二者内部过程存在差异的根源, 进而基于此差异补充人工心理理论的定义。

3.1 心理理论与人工心理理论的内部过程存在差异

3.1.1 人脑与人工神经网络的对比

大脑是人类智能的核心, 它包含约 1000 亿个神经元(neuron), 这些神经元相互连接组成的神经回路超过十万千米, 神经回路和纤维的协同工作实现着各种复杂的认知功能(Feng et al., 2024)。人工神经网络(Artificial Neural Networks)是受人脑结构和运作方式的启发而建立的计算机模型, 旨在模拟人脑神经元执行的基本功能(Bridgelall, 2024)。尽管 AI 研究已取得重大进展, 但人类的大脑远比当前任何 LLMs 的人工神经网络复杂(Aru, 2025)。

首先, 构成两类神经网络的神经元存在差异, 生物神经元的复杂性远超人工神经元。生物神经元由胞体(cell body)、树突(dendrite)和轴突(axon)组成, 是强大的信息处理单元, 即使在单个细胞中也能实现高度复杂的非线性计算(Duan et al., 2020)。它们在结构、分子机制和动态发育等方面表现出巨大的生化复杂性(Paus, 2023)。具体到心

理理论, 人类背内侧前额叶皮层(dmPFC)可能存在能够表征他人信念的单个神经元, 即支持心理理论的候选神经元(Jamali et al., 2021)。而人工神经元通过相互加权联结来处理人类的输入并生成输出, 是简化的单元(Aru, 2025), 缺乏模拟复杂计算的功能, 单个神经元难以实现对他人信念的表征。

其次, 激活脑区方面, 人类负责心理理论的脑区主要包括内侧前额叶皮层(mPFC)和颞顶联合区(TPJ) (Jamali et al., 2021; Schaafsma et al., 2015)。内侧前额叶皮层与自我投射和抽象社会知识表征有关, 颞顶联合区负责区分自我与他人的心理状态(Schurz et al., 2021)。人工神经网络尚未形成如人脑般细化的认知模块, 目前仅展现出模拟注意、记忆、推理和遗忘等自然认知过程的潜力(Jiao et al., 2025)。

最后, 心理理论涉及大规模皮质系统的动态交互, 而人工神经网络的认知模块间缺乏有效协同。先前研究表明, 心理理论的神经基础以默认模式网络(DMN)为核心, 通过内侧前额叶皮层、颞顶联合区和后扣带回(PCC)的层级协作, 实现从抽象心理推理到情境化社会理解的动态整合(Schurz et al., 2021)。默认模式网络是丘脑皮质系统的一部分, 该系统的特点是支持复杂的、循环往复的信息加工过程(Aru et al., 2023)。人工神经网络依赖于前馈(feedforward)架构, 信息以单向的方式流动, 目前不具备生物大脑中至关重要的复杂反馈回路(Blank, 2023; LeCun et al., 2015)。

3.1.2　影响习得并发展心理理论与人工心理理论的因素

心理理论的习得与发展过程漫长而有序, 儿童遵循一个基本的发展阶梯逐步理解他人的心理(Yu & Wellman, 2024)。以往研究表明, 在生命的前 4~5 年里, 儿童发展出最基本、最原始的心理理论(Rakoczy, 2022)。具体而言, 9 个月后, 婴儿能够追踪他人的感知、目标并预期他人的目标指向行为(Call & Tomasello, 2008)。1~3 岁时, 儿童逐渐意识到自己与他人有着不同的愿望和信念(Wellman & Liu, 2004)。4 岁左右, 儿童开始理解行为由行动者的愿望和信念共同支配产生(Wellman, 2017)。4~5 岁的儿童能够通过一阶错误信念任务, 理解他人的信念与自己不同, 并能够基于该理解预测行为(Flavell, 1999)。此后, 许多需要更高认知复杂性的心理理论子能力在儿童期和青春期持续发展。6~7 岁儿童能够通过二阶错误信念任务, 8~10 岁儿童表现出对隐喻、失言、反讽、奇异故事等的理解(Fu et al., 2023; Devine & Hughes, 2013)。与之不同, 尽管 GPT-4 在心理理论任务中的表现普遍优于 GPT-3 与 GPT-3.5 (Kosinski, 2024; Van Duijn et al., 2023), 但其似乎突然习得了人工心理理论。这间接表明, 人类和 LLMs 在习得与发展心理理论的过程中受到不同因素的影响。

婴儿最初通过对环境中行为规律的统计学习理解行为, 随后在语言与执行功能等一般认知能力的完善以及社会互动的推动下, 逐步构建起心理理论(Ruffman, 2023)。大量研究表明, 语言能力在儿童心理理论的发展中起关键作用, 例如, 一般语言能力(涵盖句法、语义及语用等层面)影响儿童的错误信念理解(Milligan et al., 2007); 儿童 3 岁时的语言能力可有效预测其学前期和青春早期心理理论的变化(Ebert, 2020)。与儿童类似, LLMs 同样暴露于海量语言输入中, 其规模可达儿童接受量的数千乃至十万倍(Frank, 2023), 然其数据利用效率远低于儿童。这种差异主要源于儿童在结构化的社会互动中接受语言输入, 同时伴随着大量的多模态感觉信息(Frank, 2023), 而 LLMs 基于文本数据训练, 缺乏具身性导致模型无法在真实、复杂的社会环境中进行互动(Peng et al., 2024; Riva et al., 2025), 其心理理论表现主要受到文本提示、自身规模以及学习算法等因素的影响(Zhao et al., 2023)。

上文通过对比评估 GPT-4 的不同研究, 在控制模型规模和学习算法的前提下, 探究任务提示对模型表现的影响。总体而言, 结构不良的提示可能降低模型的通过率, 良好的提示框架能够有效提升模型表现(见表 3)。与之类似, 通过对比同一任务中三个版本模型的表现, 即控制提示对模型表现的影响, 可明确学习算法和规模的作用。GPT-3 和 GPT-3.5 规模接近, 两者表现的差异更可能由学习算法引起; GPT3.5 与 GPT-4 采用相同的学习算法训练, 它们在任务中的不同表现则体现模型规模的影响。具体而言, 在一项奇异故事任务中, GPT-3 的表现接近随机水平, 而 GPT-3.5 则达到了 75%的通过率(Van Duijn et al., 2023)。这一差异表明, 采用基于人类反馈的强化学习等算

表 3 人工心理理论研究使用的提示框架

提示框架	简要叙述	文献
思维链 (chain of thought)	通过向模型提供包含输入、中间推理步骤和输出的示例，提示模型生成类似的中间步骤，从而逐步解决问题	Chen et al., 2024; Shapira et al., 2024
显式信念表征 (explicit belief representations)	提示模型实时存储与任务相关的关键信念，并在推理过程中整合心理状态	Amirizaniani et al., 2024; Li et al., 2023
模拟心理理论 (simulated theory of mind)	提示模型首先根据目标角色已知的信息对上下文进行筛选，然后再回答关于其心理状态的问题	Wilf et al., 2024; Xu et al., 2024
自我问答 (self-ask prompting)	提示模型在回答初始问题前，自省式地生成并回答后续问题，直至得出确定结论	Press et al., 2023; Zhou et al., 2023
预见与反思 (foresee and reflect)	提示模型先基于观察预测潜在未来事件，再反思不同行动选择如何帮助角色应对潜在挑战，最终确定最优方案	Zhou et al., 2023
符号心理理论 (symbolic theory of mind)	提示模型进行显式符号表征，并对任务中多角色信念状态进行多级推理	Sclar et al., 2023
时间心理理论 (time theory of mind)	为任务文本添加时序维度来构建时空体系，提示模型依据各角色在时间线上感知的事件构建时空信念状态链	Hou et al., 2024
离散世界模型 (discrete world models)	将任务文本分割成多个序列块，在每块后插入要求描述角色当前信念的提示，最后提示模型生成答案	Huang et al., 2024

法训练，使得模型在理解非字面交际方面与人类对齐。此外，GPT-4 在该任务上的通过率接近 100%，表明规模扩展同样能够提升模型的能力。多项研究结果呈现类似的趋势，GPT-4 的表现普遍优于 GPT-3.5 (Brunet-Gouet et al., 2023; Attanasio et al., 2024)，GPT-3.5 的通过率高于 GPT-3 (Kim et al., 2023; Kosinski, 2024)，共同印证了规模和学习算法对模型表现的影响。

3.2 补充人工心理理论的定义

上述对比表明，支撑心理理论与人工心理理论的神经机制在多个层面呈现不同的复杂度，影响人类与 LLMs 习得并发展心理理论的因素存在差异。因此，二者的内部过程存在本质区别。然而，作为区别于心理理论的概念，现有的人工心理理论定义仅强调人类与 LLMs 在任务表现层面的相似，未涉及二者在内部过程层面的差异。另外，提示对仅能处理文本信息的 LLMs 的表现具有重要影响(Lin, 2024)，但现有定义未纳入这一关键因素。因此，本文基于 Marchetti 等人(2025)的定义，将人工心理理论界定为：LLMs 接收心理理论任务提示后，通过识别并匹配提示文本中的统计规律，在生成文本时所表现出的对人类反应的模拟。具体而言，该定义明确表明人工心理理论的主体是 LLMs，强调提示对模型表现的作用，指出其内部过程与本质是映射及操纵文本数据的复杂统计规律(Riva et al., 2025)，并突出 LLMs 对人类心理理论表现的模拟。

4 挑战

4.1 评估方案存在问题

评估 LLMs 人工心理理论的核心步骤包括评估设置、提示设计以及任务评分(Laskar et al., 2024)，任一环节存在问题均可能造成研究结果与解释的不一致。

评估设置分为选择心理理论任务与确定测试对象两部分。鉴于心理理论包含对广泛的心理状态的理解与推论(Beaudoin et al., 2020)，为 LLMs 选择适当的测试任务至关重要。然而，当前研究选用的心理理论任务普遍存在系统性和有效性问题。一方面，研究者通常聚焦于特定任务，如错误信念任务(Trott et al., 2023)，但仅凭模型在单一任务上的表现，难以得出针对其人工心理理论的一般结论。目前覆盖面最广的研究评估了七类心理理论子能力，仍未涉及行动意图与感知(Chen et al., 2024)。另一方面，改编或定制版本的任务通常由 LLMs 生成原始文本，缺少规范的效度验证(Xu et al., 2024; Gandhi et al., 2023)，可能难以有效测量人工心理理论。筛选合适的 LLMs 对于人工心理理论评估同样至关重要。然而，研究者过度关注 GPT 系列模型的表现。尽管主流模型大多

基于 transformer 架构(Zhao et al., 2023), 但训练数据等方面的差异仍可能导致模型的人工心理理论表现有所不同。此外, 构建人类被试的基线数据能够验证任务的有效性, 并为理解 LLMs 的任务表现提供关键参照(Jones et al., 2024), 甚至可以称其为衡量模型表现的最佳标准, 但多数研究(57.7%)并未纳入人类被试。

选定任务与评估对象后, 下一步即是设计提示。当前研究主要采用三类提示技术：零样本(zero-shot)提示、思维链提示与针对心理理论任务设计的提示, 这些技术各有局限。第一, 尽管多项研究表明, 使用适当的提示技术能够有效提升模型的任务表现(Amirizaniani et al., 2024; Li et al., 2023; Wilf et al., 2024), 但大部分研究仍然只采用零样本提示, 即直接向 LLMs 呈现心理理论任务。第二, 思维链提示已被证明能够显著提升 LLMs 的复杂推理能力(Wei, Wang, et al., 2022), 但其对心理理论表现的作用存在争议。先前研究表明思维链提示能够有效提升 GPT-4 的任务表现(Shapira et al., 2024)。然而, Chen 等人(2024)发现思维链提示不能提升模型的心理理论表现, 甚至可能引起表现下滑。第三, 针对心理理论任务设计的提示存在揭示任务结构的风险, 可能造成过度指导。例如, 显式信念表征提示模型在推理过程中整合心理状态, 可能会暗示问题答案, 从而干扰对模型能否自发考虑心理状态的评估(Amirizaniani et al., 2024)。

最后, 除选择题、判断题等可自动评分的问题外, 任务中的开放式问题依赖于人工评分。人工评分过程存在主观性, 可能导致评分结果不一致, 进而难以比较不同模型的表现以及不同研究的结果(Sarıtaş et al., 2025)。此外, 人工评分需耗费大量人力, 特别是当任务包含数千个问题时(Ma et al., 2023)。

4.2　人工心理理论的机制问题

内部过程是区分心理理论与人工心理理论的关键。然而, LLMs 从输入到输出的转化过程部分或全部隐藏(张珺皓, 2025), 仅依靠其文本输出, 甚至难以推断模型的人工心理理论源于检索还是推理过程(Verma et al., 2024)。注意力机制可能通过识别和追踪任务叙述中目标角色的行为、对话与内在状态间的联系, 辅助 LLMs 预测该角色的信念(Kosinski, 2024), 但当前的描述相对笼统, 且尚无研究明确揭示注意力机制在心理理论任务中的运作过程。

此外, LLMs 主要为人机交互而设计, 其在社会互动中稳定地展现出能够被人类感知到的心理理论更具实践意义(Ünlütabak & Bal, 2025)。这要求研究者同时关注两个方面, 评估 LLMs 的人工心理理论, 并尝试赋予其类似于人类的心理理论能力; 理解人类将信念、愿望和意图等心理状态归因于 LLMs 的倾向, 即探究人机交互中影响人类感知到人工心理理论的因素(Wang, Walsh, et al., 2024)。然而, 当前研究采用的评估任务缺乏贴近现实情境的实时人机交互形式(Gandhi et al., 2023), 且未充分关注人类对 LLMs 进行心理状态归因的过程。人机交互中的影响因素在 LLMs 表现出人工心理理论, 并被人类感知到的过程中发挥着重要作用, 亟需予以系统性关注。

4.3　心理理论对齐问题

心理理论依赖于心理状态的存在, 诸如信念、愿望和意图等心理状态被认为是有意识的生物独有的(Yıldız, 2025)。这引出了人工心理理论与心理理论对齐的关键问题：LLMs 有意识吗? 大多数研究者对此持否定态度。基于神经科学视角, 当前的 LLMs 缺乏具身性、难以有效整合多模态信息, 不具备形成意识的神经基础, 难以自主产生意识(Aru et al., 2023)。根据生物自然主义的观点, 意识根植于生命有机体的本质属性, 现有技术尚未实现真正的人工意识(Seth, 2024)。Shanahan 等人(2023)则用角色扮演的概念描述 LLMs 的意识, 认为其只具有表面上的自我意识。

那么, 没有意识或主观经验的 LLMs, 能否真正拥有心理理论? 基于发展心理学的视角, 儿童在 3~4 岁发展出反思意识, 促进了心理理论的产生与发展, 标志着认知和社会功能的重大飞跃(Zelazo, 2004)。这种自我意识发展与心理理论产生的交织, 表明真正的心理理论不仅是简单的认知功能的集合, 还依赖于意识和反思的作用(Yıldız, 2025)。当前的 LLMs 能够模拟人类在心理理论任务中的表现, 但尚未真正具备类似人类的心理理论能力(Amirizaniani et al., 2024)。

5　展望

5.1　制定并采用标准化的评估方案

鉴于评估 LLMs 的复杂性, 构建一套既实用又易于实施的结构化评估框架十分困难。本文基

于现有评估方案的不足及模型的性能局限，探讨各评估阶段的标准化方向。

评估设置方面，亟需编制并使用系统且有效的心理理论任务集。首先，研究者应明确用于指导编制的、能有效衡量任务集系统性的心理理论框架。例如，可以参考涵盖七大类心理状态理解的心理理论空间能力架构(Beaudoin et al., 2020)，或采用能够系统描述心理理论发展进程的量表(Yu & Wellman, 2024)。其次，组织相关研究者从零开始构建任务集(Chen et al., 2024)，注意平衡各子任务的题目数量，避免题目数量过多导致难以进行人机对比，并尽可能对编制的任务集进行严格的效度验证(Brunet-Gouet et al., 2023)。更多纳入多模态任务，注意设置多样化的，能够有效整合不同模态信息的任务场景(Jin et al., 2024)。此外，鉴于许多心理理论任务实际测量一些低阶认知过程(Quesque & Rossetti, 2020)，采用依据这些“无效”的任务模式编制的新任务评估 LLMs，并与模型在同类有效任务上的表现进行对比，有助于解析人工心理理论所依赖的基本认知过程。最后，应同时评估人类被试在任务集上的表现，以此建立参照标准。针对当前研究模型单一的问题，鉴于语言的细微差别及训练数据对 LLMs 展现出人工心理理论的重要性(Ünlütabak & Bal, 2025)，有必要纳入 DeepSeek、ERNIE 等主要基于汉语文本数据训练的模型(Lu, Song, & Zhang, 2025)。

提示设计方面，首先应完整记录并报告研究使用的提示文本，便于后续研究复现研究结果，并避免同一提示技术因细节不同而产生效果差异。其次，细致比较同一提示技术对不同任务以及不同模型的影响(Kim et al., 2023)，探究特定提示发挥作用的机制，进一步揭示影响 LLMs 表现出人工心理理论的因素。最后，为心理理论任务定制提示时，应结合任务看待提示的效果，避免特定提示导致任务简化。

任务评分方面，采用多种题型有助于考察 LLMs 在同类任务上表现的稳定性(Kim et al., 2023)，但需尽量避免人工评分。Ma 等人(2023)开发了基于 LLMs 的自动评分器，适用于问答与文本补全两类题型，未来研究可扩展至开放式题型的自动评分。此外，为了系统考察 LLMs 在任务中表现出的沟通策略和语用技能，可采用定性方法分析其输出内容(Attanasio et al., 2024)。

5.2 在共同心理理论框架下探究人工心理理论机制

共同心理理论(Mutual Theory of Mind)框架试图归纳当前人工心理理论研究的两个视角，促进实现自然的人机交互(Richter, 2025)。该框架认为人类与 LLMs 通过解释、反馈和动态交互三个阶段理解对方的心理状态(Wang & Goel, 2024)。

解释阶段指 LLMs 基于提示形成对人类心理状态的解释。该阶段依赖模型自身的人工心理理论，并受提示的影响。将机器心理理论研究领域的认知计算模型与 LLMs 结合，是揭示人工心理理论内部过程的重要方向(Ullman, 2023)。其中，贝叶斯方法对建模 LLMs 的认知，量化解释其行为具有独特价值(Griffiths et al., 2024)。此外，有必要定量分析提示框架的具体特征对模型表现的影响，探究其作用机制。

第二阶段中，人类通过 LLMs 的反馈推断其对自身心理状态的理解，进而形成对模型的认知判断。人类自身的心理理论能力与 LLMs 输出文本的特征作用于该阶段。Wang 和 Goel (2024)构建了一种能够通过用户的言语反馈，自动表征用户感知的对话智能体，发现 LLMs 的输出特点(如文本的多样性、可读性和流畅性)，通过文本传达情感的能力及其拟人化程度等正向预测人类对智能体的感知。未来可重复验证此类表征研究的有效性，并深入探究影响人类与 LLMs 对彼此感知的因素及作用机制。此外，可参照发展研究中的家长报告问卷，如 ToMI (Hutchins et al., 2012)，编制用于评估人类对人工心理理论感知的量表，以期揭示人类能否以及如何将人工心理理论归因于 LLMs。

最后一个阶段涉及 LLMs 依据人类反馈调整其初始理解，并生成新的解释，实现人与模型间的动态互动。该阶段主要考验模型持续进行多轮对话的能力，而现有研究表明这一能力尚不完善(Yongsatianchot et al., 2024)，因此需要模型开发者持续投入以提升模型的相关能力。

5.3 与人类心理理论对齐

为弥合人工心理理论与心理理论间的差距，可从人类视角增强 LLMs 的心理理论表现(He et al., 2023)。在神经基础层面，根据意识的积木理论(Building Blocks Theory)，LLMs 缺乏产生意识的必要模块。未来可参考类脑研究，通过附

加模块、插件以及集成多个模型推进 LLMs 实现意识的涌现(Tait et al., 2024)。此外, 通过持续扩展与重新配置人工神经网络, 有望增强其对人脑功能的模拟, 最终使模型具备心理理论(Kosinski, 2024)。在信息感知层面, 鉴于 LLMs 缺乏具身性, 难以整合多模态信息, 导致其虽能根据工具性知识部分恢复世界模型, 但与人类的“世界知识”存在本质差异的缺点(Yildirim & Paul, 2024), 未来应将多模态 LLMs 和具身(Embodied)AI 纳入研究。多模态模型是以 LLMs 为核心, 具备整合和建模包括语言、声音和视觉信号在内的多模态信息能力的模型(Yin et al., 2024)。具身智能强调智能受到大脑、身体和环境紧密耦合的影响(Liu et al., 2025)。当前的 LLMs 以处理封闭问题为核心(李开阳, 2025), 难以应对现实世界这一开放环境, 进而难以形成类似人类的智能。为突破该发展瓶颈, 研究者正尝试构建能够与环境进行自主交互, 具有感知、决策、计划和执行能力的具身 AI (Ren et al., 2024)。当 LLMs 成为可有效处理多模态信息的具身 AI 系统, 可能真正实现对人类心理状态的理解。

尽管当前人工心理理论与心理理论间的类比存在局限, 但 LLMs 作为人类“思维黑箱”的技术表现形式或延续(张珺皓, 2025), 在发展模型人工心理理论的过程中有机会加深对心理理论的理解。探索 LLMs 的“算法黑箱”时应超越其本身, 回归到“人”身上。实现人工心理理论与心理理论的对齐, 有赖于在模拟人脑运作机制方面取得实质性进展, 从而为揭示人类心理理论的涌现及内部过程提供新思路(He et al., 2023)。

参考文献

李开阳. (2025). 弱具身、具身认知与具身 AI. *自然辩证法研究*, *41*(3), 81−87. https://doi.org/10.19484/j.cnki.1000-8934.2025.03.014

林曦. (2025, 3 月). 人工智能“幻觉”的存在主义阐释. *社会科学辑刊*, (2), 81−91.

舒文韬, 李睿潇, 孙天祥, 黄萱菁, 邱锡鹏. (2023). 大型语言模型: 原理、实现与发展. *计算机研究与发展*, *61*(2), 351−361.

张珺皓. (2025). 算法黑箱研究：基于认知科学的视角. *科学学研究*, *43*(9), 1872−1880. https://doi.org/10.16192/j.cnki.1003-2053.20250107.001

赵泽林, 程聪瑞. (2024). ChatGPT、人类心灵与人工心灵——科辛斯基 ChatGPT 实验的考察与反思. *江汉论坛*, (7), 99−105.

Amirizaniani, M., Martin, E., Sivachenko, M., Mashhadi, A., & Shah, C. (2024, October). Can LLMs reason like humans? Assessing theory of mind reasoning in LLMs for open-ended questions. *In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management* (pp. 34−44). Association for Computing Machinery. https://doi.org/10.1145/3627673.3679832

Aru, J. (2025). Artificial intelligence and the internal processes of creativity. *The Journal of Creative Behavior, 59*(2), e1530. https://doi.org/10.1002/jocb.1530

Aru, J., Larkum, M. E., & Shine, J. M. (2023). The feasibility of artificial consciousness through the lens of neuroscience. *Trends in Neurosciences*, *46*(12), 1008−1017. https://doi.org/10.1016/j.tins.2023.09.009

Attanasio, M., Mazza, M., Le Donne, I., Masedu, F., Greco, M. P., & Valenti, M. (2024). Does ChatGPT have a typical or atypical theory of mind? *Frontiers in Psychology*, *15*, 1488172. https://doi.org/10.3389/fpsyg.2024.1488172

Baker, C. L., Saxe, R. R., & Tenenbaum, J. B. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 33, No. 33, pp. 2469−2474). Cognitive Science Society.

Beaudoin, C., Leblanc, É., Gagner, C., & Beauchamp, M. H. (2020). Systematic review and inventory of theory of mind measures for young children. *Frontiers in Psychology*, *10*, 2905. https://doi.org/10.3389/fpsyg.2019.02905

Bendell, R., Williams, J., Fiore, S. M., & Jentsch, F. (2024). Individual and team profiling to support theory of mind in artificial social intelligence. *Scientific Reports*, *14*(1), 12635. https://doi.org/10.1038/s41598-024-63122-8

Blank, I. A. (2023). What are large language models supposed to model? *Trends in Cognitive Sciences*, *27*(11), 987−989. https://doi.org/10.1016/j.tics.2023.08.006

Bridgelall, R. (2024). Unraveling the mysteries of AI chatbots. *Artificial Intelligence Review*, *57*(4), 89. https://doi.org/10.1007/s10462-024-10720-7

Brunet-Gouet, E., Vidal, N., & Roux, P. (2023, September). Can a conversational agent pass theory-of-mind tasks? A case study of ChatGPT with the hinting, false beliefs, and strange stories paradigms. In: J. Baratgin, B. Jacquet, & H. Yama (Eds.), *Human and Artificial Rationalities (HAR 2023). Lecture Notes in Computer Science, vol 14522* (pp. 107-126). Springer, Cham. https://doi.org/10.1007/978-3-031-55245-8_7

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences*, *12*(5), 187−192. http://dx.doi.org/10.1016/j.tics.2008.02.010

Chang, E. Y. (2023, December). Examining GPT-4's capabilities and enhancement with SocraSynth. In *2023 International Conference on Computational Science and Computational Intelligence (CSCI)* (pp. 7−14). IEEE. https://doi.org/10.1109/CSCI62032.2023.00009

Chen, Z., Wu, J., Zhou, J., Wen, B., Bi, G., Jiang, G., ...

Huang, M. (2024, August). ToMBench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 15959−15983). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.847

Deshpande, M., & Magerko, B. (2024, May). Embracing embodied social cognition in AI: Moving away from computational theory of mind. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (pp. 1−7). Association for Computing Machinery. https://doi.org/10.1145/3613905.3650998

Devine, R. T., & Hughes, C. (2013). Silent films and strange stories: Theory of mind, gender, and social experiences in middle childhood. *Child Development*, *84*(3), 989−1003. https://doi.org/10.1111/cdev.12017

Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., ... Sui, Z. (2024, November). A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 1107−1128). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.64

Duan, Q., Jing, Z., Zou, X., Wang, Y., Yang, K., Zhang, T., ... Yang, Y. (2020). Spiking neurons with spatiotemporal dynamics and gain modulation for monolithically integrated memristive neural networks. *Nature Communications*, *11*(1), 3399. https://doi.org/10.1038/s41467-020-17215-3

Ebert, S. (2020). Theory of mind, language, and reading: Developmental relations from early childhood to early adolescence. *Journal of Experimental Child Psychology*, *191*, 104739. https://doi.org/10.1016/j.jecp.2019.104739

Feng, J., Jirsa, V., & Lu, W. (2024). Human brain computing and brain-inspired intelligence. *National Science Review*, *11*(5), nwae144. https://doi.org/10.1093/nsr/nwae144

Flavell, J. H. (1999). Cognitive development: Children's knowledge about the mind. *Annual Review of Psychology*, *50*(1), 21−45. https://doi.org/10.1146/annurev.psych.50.1.21

Frank, M. C. (2023). Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*, *27*(11), 990−992. https://doi.org/10.1016/j.tics.2023.08.007

Fu, I. N., Chen, K. L., Liu, M. R., Jiang, D. R., Hsieh, C. L., & Lee, S. C. (2023). A systematic review of measures of theory of mind for children. *Developmental Review*, *67*, 101061. https://doi.org/10.1016/j.dr.2022.101061

Gandhi, K., Fränken, J. P., Gerstenberg, T., & Goodman, N. D. (2023, December). Understanding social reasoning in language models with language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)* (pp. 13518-13529). Curran Associates Inc. https://doi.org/10.48550/arXiv.2306.15448

Griffiths, T. L., Zhu, J. Q., Grant, E., & Thomas McCoy, R. (2024). Bayes in the age of intelligent machines. *Current Directions in Psychological Science*, *33*(5), 283−291. https://doi.org/10.1177/09637214241262329

Hagendorff, T. (2024). Deception abilities emerged in large language models. *Proceedings of the National Academy of Sciences*, *121*(24), e2317967121. https://doi.org/10.1073/pnas.2317967121

He, Y., Wu, Y., Jia, Y., Mihalcea, R., Chen, Y., & Deng, N. (2023, December). Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 10691−10706). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.717

Hou, G., Zhang, W., Shen, Y., Wu, L., & Lu, W. (2024, August). TimeToM: Temporal space is the key to unlocking the door of large language models' theory-of-mind. In *Findings of the Association for Computational Linguistics ACL 2024* (pp. 11532−11547). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.685

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), 1−55. https://doi.org/10.1145/3703155

Huang, X. A., La Malfa, E., Marro, S., Asperti, A., Cohn, A. G., & Wooldridge, M. J. (2024, November). A notion of complexity for theory of mind via discrete world models. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 2964−2983). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.167

Huang, Y., Zhu, H., Guo, T., Jiao, J., Sojoudi, S., Jordan, M. I., ... Mei, S. (2025). Generalization or hallucination? Understanding out-of-context reasoning in transformers. *arxiv* preprint arxiv:2506.10887. https://doi.org/10.48550/arXiv.2506.10887

Hutchins, T. L., Prelock, P. A., & Bonazinga, L. (2012). Psychometric evaluation of the theory of mind inventory (ToMI): A study of typically developing children and children with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *42*(3), 327−341. https://doi.org/10.1007%2Fs10803-011-1244-7

Jamali, M., Grannan, B. L., Fedorenko, E., Saxe, R., Báez-Mendoza, R., & Williams, Z. M. (2021). Single-neuronal predictions of others' beliefs in humans. *Nature*, *591*(7851), 610−614. https://doi.org/10.1038/s41586-021-03184-0

Jiao, L., Ma, M., He, P., Geng, X., Liu, X., Liu, F., ... Tang, X. (2025). Brain-inspired learning, perception, and cognition: A comprehensive review. *IEEE Transactions on Neural Networks and Learning Systems*, *36*(4), 5921−5941. https://doi.org/10.1109/TNNLS.2024.3401711

Jin, C., Wu, Y., Cao, J., Xiang, J., Kuo, Y.-L., Hu, Z., ... Shu, T. (2024, August). MMToM-QA: Multimodal theory of mind question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 16077−16102). Association for Computational Linguistics. https://doi.org/

10.18653/v1/2024.acl-long.851

Jones, C. R., Trott, S., & Bergen, B. (2024). Comparing humans and large language models on an experimental protocol inventory for theory of mind evaluation (EPITOME). *Transactions of the Association for Computational Linguistics*, *12*, 803−819. https://doi.org/10.1162/tacl_a_ 00674

Kim, H., Sclar, M., Zhou, X., Bras, R. L., Kim, G., Choi, Y., & Sap, M. (2023). FANToM: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 14397−14413). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.890

Korman, J., & Malle, B. F. (2016). Grasping for traits or reasons? How people grapple with puzzling social behaviors. *Personality and Social Psychology Bulletin*, *42*(11), 1451−1465. https://doi.org/10.1177%2F0146167216663704

Kosinski, M. (2023). Theory of mind might have spontaneously emerged in large language models. *arxiv* preprint arxiv: 2302.02083v5. https://doi.org/10.48550/arXiv.2302.02083

Kosinski, M. (2024). Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, *121*(45), e2405460121. https://doi.org/10.1073/pnas.2405460121

Laskar, M. T. R., Alqahtani, S., Bari, M. S., Rahman, M., Khan, M. A. M., Khan, H., ... Huang, J. (2024, November). A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 13785−13816). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.764

Lebiere, C., Pirolli, P., Johnson, M., Martin, M., & Morrison, D. (2025). Cognitive models for machine theory of mind. *Topics in Cognitive Science*, *17*(2), 268−290. https://doi.org/10.1111/tops.12773

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436−444. https://doi.org/10.1038%2Fnature14539

Li, H., Chong, Y., Stepputtis, S., Campbell, J., Hughes, D., Lewis, C., & Sycara, K. (2023, December). Theory of mind for multi-agent collaboration via large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 180−192). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.13

Lin, Z. (2024). How to write effective prompts for large language models. *Nature Human Behaviour*, *8*(4), 611−615. https://doi.org/10.1038/s41562-024-01847-2

Liu, H., Guo, D., & Cangelosi, A. (2025). Embodied intelligence: A synergy of morphology, action, perception and learning. *ACM Computing Surveys*, *57*(7), 1−36. https://doi.org/10.1145/3717059

Lu, J. G., Song, L. L., & Zhang, L. D. (2025). Cultural tendencies in generative AI. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-025-02242-1

Lu, Y. L., Zhang, C., Song, J., Fan, L., & Wang, W. (2025). Do theory of mind benchmarks need explicit human-like reasoning in language models?. *arxiv* preprint arxiv:2504.01698. https://doi.org/10.48550/arXiv.2504.01698

Ma, X., Gao, L., & Xu, Q. (2023, December). ToMChallenges: A principle-guided dataset and diverse evaluation tasks for exploring theory of mind. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)* (pp. 15−26). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.conll-1.2

Mao, Y., Liu, S., Ni, Q., Lin, X., & He, L. (2024). A review on machine theory of mind. *IEEE Transactions on Computational Social Systems*, *11*(6), 7114−7132. https://doi.org/10.1109/TCSS.2024.3416707

Marchetti, A., Di Dio, C., Cangelosi, A., Manzi, F., & Massaro, D. (2023). Developing ChatGPT's theory of mind. *Frontiers in Robotics and AI*, *10*, 1189525. https://doi.org/10.3389/frobt.2023.1189525

Marchetti, A., Manzi, F., Riva, G., Gaggioli, A., & Massaro, D. (2025). Artificial intelligence and the illusion of understanding: A systematic review of theory of mind and large language models. *Cyberpsychology, Behavior, and Social Networking*, *28*(7), 505−514. http://dx.doi.org/10.1089/cyber.2024.0536

Meskó, B. (2023). Prompt engineering as an important emerging skill for medical professionals: Tutorial. *Journal of Medical Internet Research*, *25*, e50638. http://dx.doi.org/10.2196/50638

Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child Development*, *78*(2), 622−646. https://doi.org/10.1111/j.1467-8624.2007.01018.x

Nickel, C., Schrewe, L., & Flek, L. (2024). Probing the robustness of theory of mind in large language models. *arxiv* preprint arxiv:2410.06271. https://doi.org/10.48550/arXiv.2410.06271

O'Hare, A. E., Bremner, L., Nash, M., Happé, F., & Pettigrew, L. M. (2009). A clinical assessment tool for advanced theory of mind performance in 5 to 12 year olds. *Journal of Autism and Developmental Disorders*, *39*(6), 916−928. https://doi.org/10.1007%2Fs10803-009-0699-2

Osterhaus, C., Koerber, S., & Sodian, B. (2016). Scaling of advanced theory-of-mind tasks. *Child Development*, *87*(6), 1971−1991. https://doi.org/10.1111%2Fcdev.12566

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (pp. 27730−27744). Curran Associates Inc. https://doi.org/10.48550/arXiv.2203.02155

Paus, T. (2023). Tracking development of connectivity in the

human brain: Axons and dendrites. *Biological Psychiatry*, *93*(5), 455−463. https://doi.org/10.1016/j.biopsych.2022.08.019

Pei, Z., Yin, J., & Zhang, J. (2025). Language models for materials discovery and sustainability: Progress, challenges, and opportunities. *Progress in Materials Science*, *154*, 101495. https://doi.org/10.1016/j.pmatsci.2025.101495

Peng, Y., Han, J., Zhang, Z., Fan, L., Liu, T., Qi, S., ... Zhu, S. C. (2024). The tong test: Evaluating artificial general intelligence through dynamic embodied physical and social interactions. *Engineering*, *34*, 12−22. https://doi.org/10.1016/j.eng.2023.07.006

Pi, Z., Vadaparty, A., Bergen, B. K., & Jones, C. R. (2024). Dissecting the Ullman variations with a SCALPEL: Why do LLMs fail at trivial alterations to the false belief task?. *arxiv* preprint arxiv:2406.14737. https://doi.org/10.48550/arXiv.2406.14737

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515−526. https://doi.org/10.1017/S0140525X00076512

Press, O., Zhang, M., Min, S., Schmidt, L., Smith, N., & Lewis, M. (2023, December). Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 5687−5711). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.378

Quesque, F., & Rossetti, Y. (2020). What do theory-of-mind tasks actually measure? Theory and practice. *Perspectives on Psychological Science*, *15*(2), 384−396. https://doi.org/10.1177/1745691619896607

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., & Botvinick, M. (2018, July). Machine theory of mind. In *Proceedings of the 35th International Conference on Machine Learning* (pp. 4218−4227). PMLR. https://doi.org/10.48550/arXiv.1802.07740

Rakoczy, H. (2022). Foundations of theory of mind and its development in early childhood. *Nature Reviews Psychology*, *1*(4), 223−235. https://doi.org/10.1038/s44159-022-00037-z

Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C. E., & Van Bavel, J. J. (2024). GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, *121*(34), e2308950121. https://doi.org/10.1073/pnas.2308950121

Ren, L., Dong, J., Liu, S., Zhang, L., & Wang, L. (2024, August). Embodied intelligence toward future smart manufacturing in the era of AI foundation model. *IEEE/ASME Transactions on Mechatronics, 30*(4), 2632-2642. https://doi.org/10.1109/TMECH.2024.3456250

Richter, F. (2025). From human-system interaction to human-system co-action and back: Ethical assessment of generative AI and mutual theory of mind. *AI and Ethics*, *5*(1), 19−28. https://doi.org/10.1007/s43681-024-00626-z

Riva, G., Mantovani, F., Wiederhold, B. K., Marchetti, A., & Gaggioli, A. (2025). Psychomatics—A multidisciplinary framework for understanding artificial minds. *Cyberpsychology, Behavior, and Social Networking*, *28*(7), 515−523. http://dx.doi.org/10.1089/cyber.2024.0409

Ruffman, T. (2023). Belief it or not: How children construct a theory of mind. *Child Development Perspectives*, *17*(2), 106−112. https://doi.org/10.1111%2Fcdep.12483

Sap, M., Le Bras, R., Fried, D., & Choi, Y. (2022, December). Neural theory-of-mind? On the limits of social intelligence in large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 3762−3780). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.248

Sarıtaş, K., Tezören, K., & Durmazkeser, Y. (2025). A systematic review on the evaluation of large language models in theory of mind tasks. *arxiv* preprint arxiv:2502.08796/ https://doi.org/10.48550/arXiv.2502.08796

Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., & Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences*, *19*(2), 65−72. https://doi.org/10.1016/j.tics.2014.11.007

Schurz, M., Radua, J., Tholen, M. G., Maliske, L., Margulies, D. S., Mars, R. B., ... Kanske, P. (2021). Toward a hierarchical model of social cognition: A neuroimaging meta-analysis and integrative review of empathy and theory of mind. *Psychological Bulletin*, *147*(3), 293−327. http://dx.doi.org/10.1037/bul0000303

Sclar, M., Kumar, S., West, P., Suhr, A., Choi, Y., & Tsvetkov, Y. (2023, July). Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 13960−13980). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.780

Sclar, M., Neubig, G., & Bisk, Y. (2022, June). Symmetric machine theory of mind. In *Proceedings of the 39th International Conference on Machine Learning* (pp. 19450−19466). PMLR.

Seth, A. K. (2024). Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences*, 1-42. https://doi.org/10.1017/S0140525X25000032

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature*, *623*(7987), 493−498. https://doi.org/10.1038/s41586-023-06647-8

Shapira, N., Levy, M., Alavi, S. H., Zhou, X., Choi, Y., Goldberg, Y., ... Shwartz, V. (2024, March). Clever hans or neural theory of mind? Stress testing social reasoning in large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2257−2273). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.eacl-long.138

Shapira, N., Zwirn, G., & Goldberg, Y. (2023, July). How well do large language models perform on faux pas tests? In *Findings of the Association for Computational*

Linguistics: ACL 2023 (pp. 10438−10451). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.663

Sileo, D., & Lernould, A. (2023, December). MindGames: Targeting theory of mind in large language models with dynamic epistemic modal logic. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 4570−4577). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.303

Sodian, B., Kristen-Antonow, S., & Kloo, D. (2020). How does children's theory of mind become explicit? A review of longitudinal findings. *Child Development Perspectives*, *14*(3), 171−177. https://doi.org/10.1111%2Fcdep.12381

Strachan, J. W. A., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., ... Becchio, C. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*, *8*(7), 1285−1295. https://doi.org/10.1038/s41562-024-01882-z

Tait, I., Bensemann, J., & Wang, Z. (2024). Is GPT-4 conscious? *Journal of Artificial Intelligence and Consciousness*, *11*(1), 1−16. https://doi.org/10.1142/S270507852450005X

Trott, S., Jones, C., Chang, T., Michaelov, J., & Bergen, B. (2023). Do large language models know what humans know? *Cognitive Science*, *47*(7), e13309. https://doi.org/10.1111%2Fcogs.13309

Ullman, T. (2023). Large language models fail on trivial alterations to theory-of-mind tasks. *arxiv* preprint arxiv: 2302.08399. https://doi.org/10.48550/arXiv.2302.08399

Ünlütabak, B., & Bal, O. (2025). Theory of mind performance of large language models: A comparative analysis of Turkish and English. *Computer Speech & Language*, *89*, 101698. https://doi.org/10.1016/j.csl.2024.101698

Van Der Meer, L., Groenewold, N. A., Nolen, W. A., Pijnenborg, M., & Aleman, A. (2011). Inhibit yourself and understand the other: Neural basis of distinct processes underlying theory of mind. *NeuroImage*, *56*(4), 2364−2374. http://dx.doi.org/10.1016/j.neuroimage.2011.03.053

Van Duijn, M., Van Dijk, B., Kouwenhoven, T., De Valk, W., Spruit, M., & van der Putten, P. (2023, December). Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7-10 on advanced tests. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)* (pp. 389−402). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.conll-1.25

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017, December). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS '17)* (pp. 6000−6010). Curran Associates Inc. https://doi.org/10.48550/arXiv.1706.03762

Verma, M., Bhambri, S., & Kambhampati, S. (2024, March). Theory of mind abilities of large language models in human-robot interaction: An illusion?. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 36−45). Association for Computing Machinery. https://doi.org/10.1145/3610978.3640767

Wang, J., Zhang, C., Li, J., Ma, Y., Niu, L., Han, J., ... Fan, L. (2024, May). Evaluating and modeling social intelligence: A comparative study of human and ai capabilities. *In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 46)* (pp. 5983−5990). Cognitive Science Society. https://doi.org/10.48550/arXiv.2405.11841

Wang, Q., & Goel, A. K. (2024). Mutual theory of mind for human-AI communication. *arxiv* preprint arxiv:2210. 03842. https://doi.org/10.48550/arXiv.2210.03842

Wang, Q., Walsh, S., Si, M., Kephart, J., Weisz, J. D., & Goel, A. K. (2024, May). Theory of mind in human-AI interaction. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (pp. 1−6). Association for Computing Machinery. https://doi.org/10.1145/3613905.3636308

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... Fedus, W. (2022). Emergent abilities of large language models. *arxiv* preprint arxiv:2206.07682. https://doi.org/10.48550/arXiv.2206.07682

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)* (pp. 24824-24837). Curran Associates Inc. https://doi.org/10.48550/arXiv.2201.11903

Wellman, H. M. (2017). The development of theory of mind: Historical reflections. *Child Development Perspectives*, *11*(3), 207−214. https://doi.org/10.1111%2Fcdep.12236

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development*, *75*(2), 523−541. https://doi. org/10.1111/j.1467-8624.2004.00691.x

Westby, C., & Robinson, L. (2014). A developmental perspective for promoting theory of mind. *Topics in Language Disorders*, *34*(4), 362−382. https://doi.org/10.1097%2FTLD.0000000000000035

Wilf, A., Lee, S., Liang, P. P., & Morency, L. P. (2024, August). Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 8292−8308). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.451

Williams, J., Fiore, S. M., & Jentsch, F. (2022). Supporting artificial social intelligence with theory of mind. *Frontiers in Artificial Intelligence*, *5*, 750763. https://doi.org/10.3389/frai.2022.750763

Winfield, A. F. (2018). Experiments in artificial theory of mind: From safety to story-telling. *Frontiers in Robotics and AI*, *5*, 357467. https://doi.org/10.3389/frobt.2018.00075

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q. L., & Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, *10*(5), 1122−1136. https://doi.

org/10.1109/JAS.2023.123618

Xu, H., Zhao, R., Zhu, L., Du, J., & He, Y. (2024, August). OpenToM: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 8593−8623). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.466

Yildirim, I., & Paul, L. A. (2024). From task structures to world models: what do LLMs know? *Trends in Cognitive Sciences*, *28*(5), 404−415. https://doi.org/10.1016/j.tics.2024.02.008

Yıldız, T. (2025). The minds we make: A philosophical inquiry into theory of mind and artificial intelligence. *Integrative Psychological and Behavioral Science*, *59*(1), 10. https://doi.org/10.1007/s12124-024-09876-2

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. (2024). A survey on multimodal large language models. *National Science Review*, *11*(12), nwae403. https://doi.org/10.1093/nsr/nwae403

Yongsatianchot, N., Thejll-Madsen, T., & Marsella, S. (2024, September). Exploring theory of mind in large language models through multimodal negotiation. In *Proceedings of the ACM International Conference on Intelligent Virtual Agents* (pp. 1−9). Association for Computing Machinery. https://doi.org/10.1145/3652988.3673960

Yu, C. L., & Wellman, H. M. (2024). A meta-analysis of sequences in theory-of-mind understandings: Theory of mind scale findings across different cultural contexts. *Developmental Review*, *74*, 101162. https://doi.org/10.1016/j.dr.2024.101162

Yuan, L., Gao, X., Zheng, Z., Edmonds, M., Wu, Y. N., Rossano, F., ... Zhu, S. C. (2022). In situ bidirectional human-robot value alignment. *Science Robotics*, *7*(68), eabm4183. https://doi.org/10.1126/scirobotics.abm4183

Zelazo, P. D. (2004). The development of conscious control in childhood. *Trends in Cognitive Sciences*, *8*(1), 12−17. https://doi.org/10.1016/j.tics.2003.11.001

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... Wen, J. R. (2023). A survey of large language models. *arxiv* preprint arxiv:2303.18223. https://doi.org/10.48550/arXiv.2303.18223

Zhou, L., Schellaert, W., Martínez-Plumed, F., Moros-Daval, Y., Ferri, C., & Hernández-Orallo, J. (2024). Larger and more instructable language models become less reliable. *Nature*, *634*(8032), 61−68. https://doi.org/10.1038/s41586-024-07930-y

Zhou, P., Madaan, A., Potharaju, S. P., Gupta, A., McKee, K. R., Holtzman, A., ... Faruqui, M. (2023). How far are large language models from agents with theory-of-mind?. *arxiv* preprint arxiv:2310.03051. https://doi.org/10.48550/arXiv.2310.03051

Artificial theory of mind in large language models: Evidence, conceptualization, and challenges

DU Chuanchen[1], ZHENG Yuanxia[1,2], GUO Qianqian[1], LIU Guoxiong[1]

([1] *School of Psychology, Nanjing Normal University, Nanjing* 210097, *China*)

([2] *College of Teacher Education, Ningbo University, Ningbo* 315211, *China*)

Abstract: Traditionally, theory of mind has been regarded as a distinctive social cognitive ability exclusive to conscious beings. However, the rapidly developing large language models (LLMs) can solve various theory of mind tasks, which has provoked intense debate about whether LLMs possess theory of mind capabilities. Artificial theory of mind in LLMs exhibits similarities to theory of mind in performance but differs in internal processes. First, we systematically synthesize research on artificial theory of mind from the objects of evaluation and the characteristics of tasks. By comprehensively analyzing GPT-4's high accuracy on theory of mind tasks alongside intrinsic and extrinsic factors limiting its performance, we demonstrate that current models achieve performance similar to those of humans. Second, by comparing the neural foundations and developmental factors underpinning theory of mind and artificial theory of mind, we reveal essential distinctions in their internal processes, thereby refining the conceptual definition of artificial theory of mind. Future research should prioritize developing and using standardized evaluation protocols, investigating the mechanisms of artificial theory of mind within the mutual theory of mind framework, and aligning artificial theory of mind with its human counterpart.

Keywords: large language models, theory of mind, artificial theory of mind, artificial intelligence