arXiv:2305.03353v2 [cs.CL] 7 Nov 2023

MindGames：利用动态认知模态逻辑探索大语言模型中的心智理论

DamienSileo Antoine Lernould里尔大学、Inria、法国国家科学研究中心 里尔大学、里尔中央理工学院、UMR 9189 CRIStAL，F-59000里尔，法国 CRIStAL，F-59000 里尔，法国
damien.sileo@inria.fr

摘要

心智理论（ToM）是智能的重要组成部分，但其评估仍存在激烈争议。以往研究将人类心智理论评估应用于自然语言处理模型，采用的要么是由人类创建的标准化测试，要么是基于规则的模板。然而，这些方法主要关注过于简单的推理，仍需进一步验证。在本文中，我们利用动态认知逻辑，分离出心智理论的一个特定组成部分，并生成受控问题。我们还引入了新的语言表达技术，以英语自然语言表达这些问题。研究结果表明，某些语言模型规模的扩大（从 70M 到 6B，以及从 350M 到 174B）并不能始终取得优于随机概率的结果。尽管 GPT-4 展现出更强的认知推理能力，但仍有改进空间。我们的代码和数据集已公开[1]

1 引言

心智理论（心智理论）是指将信念、愿望和意图等心理状态归因于自己和他人的认知能力，使个体能够根据推断出的这些心理状态理解并预测行为。它是通用文本理解或人工智能的一项重要要求（Navarro 等人，2020年），但有关心智理论的论断容易受到人类预期的偏见影响（de Waal，2016年）。Kosinski (2023) 最近引发了一场争论：研究显示，扩大大型语言模型（LLM）的规模能够提升其在旨在衡量心智理论的标准化测试中的表现。然而，这些测试在学术研究中被广泛讨论，可能已泄露到 LLM 的训练语料库中。早期研究则改为生成合成示例，扩展了 bAbi（Weston 等人，2016年）框架。Nematzadeh 等人(2018年) 提出了一套基于固定模板的数据集，基于该模板的

> 莎莉-安妮问题（Baron-Cohen 等人，1985）：莎莉在安妮陪伴时把一颗弹珠放进盒子里。莎莉暂时离开，玛丽把弹珠放进篮子里。莎莉会去哪里找弹珠？ [答案=盒子]

Le 等人(2019年) 认为这些问题过于简单，并将其扩展到追踪二阶信念（例如，莎莉对安妮信念的信念）。

> 在我们的研究中，我们生成动态认知逻辑（DEL）问题，并设计相应的文字表述，将其转化为自然语言推理问题。动态认知逻辑是模态逻辑的一个分支，可以对个体关于特定事实或其他智能体知识的认知进行建模。动态认知逻辑还支持对连续公开公告影响的推理：爱丽丝和鲍勃的头上都沾有泥巴。他们的父亲说，他们至少有一个人头上沾了泥巴。他问爱丽丝和鲍勃头上是否沾了泥巴。爱丽丝和鲍勃知道自己头上沾了泥巴吗？ [AN- SWER=NO] 他们回答说不知道。现在爱丽丝和鲍勃知道自己头上沾了泥巴吗？ [ANSWER=YES]

如果爱丽丝没有弄脏，鲍勃就会对第一个问题回答“是”，因此，在鲍勃第一次作答后，爱丽丝可以知道自己弄脏了。[2] 动态认知逻辑（DEL）可以形式化某些心智理论（ToM）问题，因此为心智理论评估提供了一个有价值的视角。我们创建的问题可能要求追踪多个智能体的信念，并对高阶信念进行推理[3]。我们的数据集涵盖了多种泥孩子和饮酒逻辑学家问题变体（vanEijck，2014年）。这个受控测试平台为理解语言模型扩展带来了新的视角，并提供了首个复杂到足以挑战监督学习模型的数据集。该数据集及其生成脚本已公开提供[1]。

[1] [代码:GitHub][数据:HF-datasets]

[2] 如果把鲍勃和爱丽丝互换，结论也同样成立。[3] 例如，安妮对莎莉关于安妮对玛丽信念的信念的信念。

2 相关工作

自然语言处理中的逻辑推理逻辑与自然语言处理有着深刻的联系。早期系统以逻辑为基础构建，而近期方法则将逻辑推理融入神经网络（Hamilton 等人，2022年；Helwe 等人，2022年）。另一个与我们研究更为接近的研究方向，利用文本数据集以及通过逻辑推理工具生成的标签，考察自然语言处理模型的逻辑能力。RuleTaker（Clark 等人，2020年）以命题逻辑探索这一领域，而 LogicNLI 则关注一阶逻辑（Tian 等人，2021年）。Richardson 和 Sabharwal（2022年）研究自然语言中的可满足性问题。Sileo 和 Moens（2022年）则以概率逻辑为研究对象。我们的研究首次聚焦于自然语言中的模态逻辑，具体而言是认知逻辑。

自然语言处理中的心智理论 为衡量自然语言处理模型的心智理论能力，Nematzadeh 等人（2018年）使用莎莉—安妮模板构建了示例，Le 等人（2019年）则通过引入二阶知识增加了数据的复杂度。这两项研究都将示例设计为问答任务。Kosinski（2023）采用人工设计的测试来评估语言模型的下一个词预测能力。Ullman（2023）指出，在这些数据集上进行干预会暴露大语言模型的脆弱性；Ma 等人（2023）则将此前的数据集整合为一套有原则的评估套件。Social-IQA 数据集（Sap 等人，2019年）涵盖广泛的社会常识，包括心智理论的相关方面，以及理解愿望和情绪等挑战。Cohen（2021年）使用人工设计的模式，研究自然语言推理模型是否能通过 know 和 think 等认知动词捕捉真实性。该任务被纳入 BIG-Bench 框架（Srivastava 等人，2022年），成为认知推理任务，但它只针对认知推理中一个较为浅层的方面。Bara 等人（2021年）使用 Minecraft 数据集，在协作任务中进行实时信念推断。Shapira 等人（2023b）指出，大语言模型在失言测试中表现吃力。Shapira 等人（2023a）则对大语言模型的社会推理能力进行了压力测试。

认识论逻辑与心智理论 Bolander（2018年）表明，莎莉-安问题可以用认识论逻辑进行建模。Van Ditmarsch 和 Labuschagne (2007) 考察了动态认知逻辑与心智理论之间更普遍的联系，而 Dissing 和 Bolander (2020年) 证明了动态认知逻辑的适用性在机器人学中。VanDe Pol 等人（2018年）通过研究其理论计算可处理性，探讨了认识逻辑用于心智理论的合理性。

3 动态认知逻辑问题生成与表述

3.1 问题定义

我们的目标是同时生成动态认知逻辑问题及其对应的自然语言表述，并采用（前提、假设、标签）格式。

一个认知逻辑问题可以分解为以下几个组成部分：

智能体： 一组由 N 个体组成，每个个体都被赋予一个不同的任意名称。

谓词： 一组布尔谓词。在这里，我们使用 N 个谓词，每个智能体对应一个（例如，爱丽丝头上沾有泥巴）。

可观测性： 对每个智能体初始知晓谓词值情况的描述。我们用一个大小为$N \times N$的布尔矩阵 $\mathcal{O}$ 表示可观测性，其中 $\mathcal{O}_{i,j}=1$ 表示智能体 i 初始时知道谓词 j 是否为真。

公告： 向所有智能体共享的一系列表达式（谓词或智能体关于谓词的知识）。公告按顺序发布，每条新公告都可能改变智能体所知道的内容，即使同一条公告被重复发布两次也是如此。

假设： 在给定智能体、可观测性和公告并将其归入一个前提之后，公告发布完毕时可能包含谓词以及智能体关于特定表达式的知识的表达式。

3.2 设置：连接谓词与可观测性

谓词的具体选择决定了可观测性的结构。例如，谓词“爱丽丝头上有泥”可被爱丽丝以外的智能体观察到，而“爱丽丝手上有泥”则可能被所有人观察到。我们将谓词和可观测性组合成所谓的设置，以生成文本描述。我们定义以下设置：

> **额头泥设置**
> 谓词$_i$：<代理$_i$>的额头上沾满了泥。$\mathcal{O}$：一元N身份N

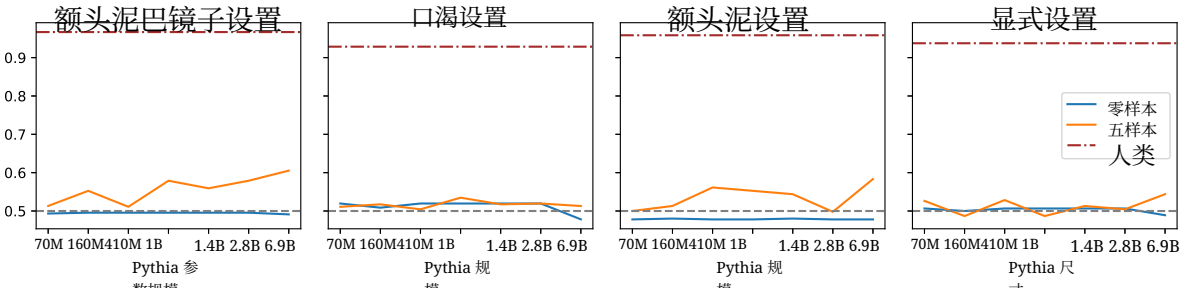


图 1：Pythia 语言模型在 MindGames 设置中的准确率。

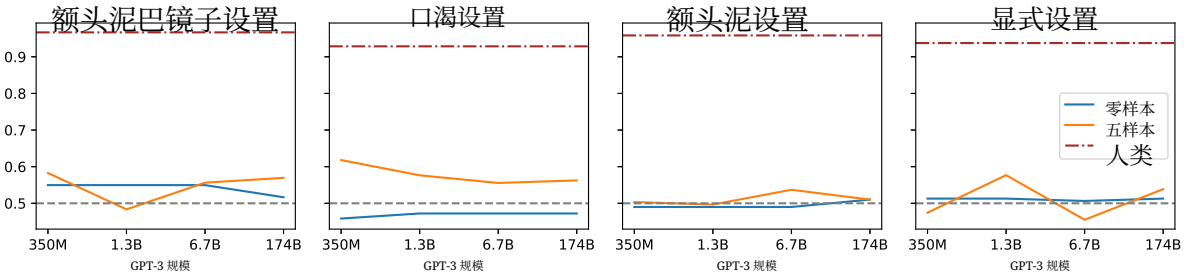


图 e 2：GPT-3 系列（ada、cabbage、curie、davinci）语言模型在 MindGames 设置中的准确率 s.

额头泥巴镜子设置
谓词$_i$： <代理$_i$>的额头上沾满了泥。$\mathcal{O}$：一元（N）观察：房间里有一面镜子。

口渴设置
P谓词$_i$： <A代理$_i$>'s 口渴了。$\mathcal{O}$： I身份(N)

显式设置
谓词$_i$: <智能体$_i$> 抽到了一张红牌。$\mathcal{O}$: RANDBOOL(N, N) $\mathbb{E}(sum(\mathcal{O}))=N$观察: 每个人抽一张牌，牌面朝下（红色或黑色）。 < <智能体$_j$>的牌向 <智能体$_i$>公开。对于所有 i, j其中$\mathcal{O}_{i,j}=1$>

3.3 问题表述

接着，我们根据给定设置，使用以下自然语言模板构建问题：

[前提] 有 <N> 个人。每个人都能被其他人看到。 <观察> 公开宣布有人 <谓词> <$[0-N]$公告>

[假设] <$[1-K]^{th}$ 信念顺序>

$[0-N]$ 表示从 0 到N的均匀采样。我们将公告限制为一阶信念。一阶信念具有以下结构： <智能体> （能够知道是否 |能够知道 | 无法知道 | 无法知道是否）（<谓词>|<否定谓词>），例如，爱丽丝无法知道鲍勃是否不浑浊。我们使用"能够"来表示某个智能体理论上可以推断出某件事，但实际上未能察觉到它。 K^{th}阶信念是关于（$K-1$）th阶信念的一阶信念。对于设置谓词，我们将所有人、并非所有人和无人视为可能的主体。主体从这些量词和智能体个体列表中均匀采样。我们将抽象的问题表示转换为自然语言和代码，以便输入模型检查器，从而确定假设是否由前提蕴含。我们使用 SMCDEL 模型检查器（Benthem 等，2018年），这是一种基于 S5（Lewis 等，1959年）模态逻辑的公告逻辑。截至 2023年4月，该实现是公开可用且被引用次数最多的认知逻辑。我们舍弃前提包含矛盾[4]的示例。为了生成多样且性别均衡的随机英语姓氏，我们使用 CensusName[5]（Qian 等，2022年）。

4 实验

4.1 问题生成参数

我们随机抽取 $N\in\{2,3,4\}$ 个智能体，因为我们观察到，即使只有三个智能体，问题也已足够具有挑战性——

[4]我们通过检查未使用的谓词是否可由前提推出，来识别矛盾。[5]https://pypi.org/project/censusname/

我们使用 $K=2$，原因相同。在公告中，我们有80%的时间使用知识谓词否定，以鼓励进行更丰富的推理（因为某个智能体不知道某事这一事实会向其他智能体传递信息）；其他情况下则有50%的时间使用这种否定。

4.2 控制样例难度

捷径，例如仅假设偏差（Gururangan 等人，2018年；Zhang 等人，2023），可能无需正确推理即可得出答案。为控制捷径的影响，我们在一个结合了所有设置的训练集上训练了一个相对浅层的监督模型（deberta-small（He 等人，2021年），6 层，4400 万个骨干参数），并确保其中没有重复样例，也没有任何同时出现在测试集中的样例。我们使用了 1.12 万个训练样例进行 3 个 epoch 的训练，学习率为 3e-5，并使用了 3730 个测试和验证样例。总体验证准确率为 83%。我们还尝试了更简单的词汇基线，例如 TF-IDF，但其对否定的捕捉能力不足。我们假设，那些被 deberta-small 以高置信度正确预测的样例包含捷径线索。我们将 deberta-small 的预测结果和置信度作为额外元数据。我们发现，所评估的语言模型在简单样例上已经表现不佳。因此，我们在实验中使用了验证子集和测试子集的随机子集；不过，用户可以根据所提供的置信度以及 deberta-small 的预测结果与真实标签之间的差异，按难度筛选我们的数据集。

我们将智能体数量限制为 3 个，并对问题去重后再进行欠采样，从而为每种设置生成 400 个测试案例，且真/假标签完全均衡。我们将所得数据集称为 MindGames。

4.3 扩展实验

我们使用一系列语言模型进行零样本实验和少样本实验。我们采用标准提示，遵循 Kosinski (2023)的设置。我们使用 lm-eval-harness 软件（Gao et al.,2021），在多项选择设置下测量语言模型的困惑度是否更倾向于正确的推理；所用的自然语言推理提示来自Brown 等人(2020年)： <前提>问题： <假设> 正确还是错误？"续写有两种可能的选项：正确和错误。我们评估两类语言模型：

人工评估我们向两位自然语言处理研究人员展示了每种设置下的 50 个测试样本，他们仅接受了执行蕴涵检测的指示。标注者间一致性为 0.89，平均准确率为 94%[6]。

Pythia 语言模型 我们在开源扩展实验中选择了 Pythia（Biderman 等人，2023）语言模型。我们使用在去重语料库（去重版）上训练的检查点，其规模分别为 70M、160M、410M、1B、1.4B、2.8B 和 6.9B。

OpenAI API 我们通过公共 API 评估 OpenAI GPT-3（Brown 等人，2020年）模型，具体包括 ada、babbage、curie 和 davinci 检查点。我们假设它们的模型规模依次为 350M、1.3B、6.7B 和 174B，并使用默认温度。

图1 展示了不同规模 Pythia 模型的结果。我们观察到，模型规模扩大提升了 5 次样本推理能力[7]，但对零样本推理没有影响。与 Kosinski (2023) 报告的涌现结果不同，图2并未显示 GPT-3 模型在 MindGames 数据上存在明确的规模效应，这表明他们观察到的涌现行为并非源于稳健的认知逻辑能力。

4.4 使用 ChatGPT 进行定性分析

截至 2023 年 5 月，我们还使用 GPT-3.5 和 GPT-4（OpenAI，2023）进行了简短的定性分析。在随机抽取的 20 道题中，我们发现 GPT3 的正确率为60%，GPT-4 的正确率为70%。我们对这两个模型进行了简短的定性分析。

如附录A所示，GPT3.5 往往会回答信息不足，只有在所需推理非常浅显时才能做出正确推断。GPT-4 能够解决这个具体例子。不过，如图3所示，有些问题仍然颇具挑战性。GPT-4 很少回答信息不足，其推理表面上看似认知推理，但偶尔会包含明显的错误。

5 结论

我们开发了一个新颖的数据集，旨在评估认知逻辑推理，并解决一个特定问题

[6]大多数错误源于未能区分是否知道与知道。[7]增加示例数量并未提高验证准确率。

用户：

[前提] 有两个人。每个人都能看见其他人。房间里有一面镜子。有人公开宣布，有人的额头沾了泥。有人公开宣布，并非所有人的额头都沾了泥。有人公开宣布，并非所有人的额头都沾了泥。 [假设]Robert 现在能否知道是否所有人的额头都沾了泥？该假设为真还是为假？

GPT-4：

错误。根据所提供的信息，Robert 无法确定所有人的额头是否都沾有泥。已知陈述如下：有两个人。每个人都能被其他人看见。房间里有一面镜子。有人额头沾泥这一事实已被公开宣布。并非所有人的额头都沾泥这一事实已被公开宣布。既然已经公开宣布并非所有人的额头都沾泥，Robert 可以推断至少有一个人的额头没有沾泥。然而，仅凭这一信息，Robert 仍无法明确判断自己的额头是否沾泥。

图 3：GPT-4 对话

心智理论的一个方面。我们的结果表明，这项任务仍然给当代大规模语言模型带来了挑战。当未来的模型能够为 2—3 个智能体解决 MindGames 时，只需增加智能体数量，就可以轻松提升难度。未来的研究可以更深入地探究人类在我们数据集上的表现，同时考虑年龄和教育背景等因素。此外，还可以进一步研究微调对其他下游任务的影响，并评估 Transformer 电路对表示模态逻辑问题的克里普克结构进行建模的能力。

6 局限性

心智理论是一个复杂的主题，而我们的研究有意采取了一个具体的视角，因此仍留下了多个有待解决的问题：

语言 我们的工作以英语为中心，该方法可以适用于其他采用主谓宾结构的语言。此外，我们将研究限制在了一些无法涵盖英语全部多样性的模板上。

提示结构与模型扩展我们关注零样本和少样本提示，这已足以（Kosinski，2023），不过，Moghaddam 和 Honey（2023）最近指出，更先进的提示方案会产生显著差异。此外，由于计算资源有限，我们未能探索全部 Pythia 模型。

任务复杂性、标注员差异 我们提出的任务相对复杂，并引发了这样的问题：什么样的标注员特征会与符号推理器的结果相匹配。动态认知逻辑本身的框架也可以为心智理论提供洞见，因为动态认知逻辑求解器能够完美地解决这一任务，尽管我们可能会对将心智理论归因于该求解器感到不适。我们或许可以认为，在简单的动态认知逻辑示例上失败就能证明其不具备心智理论，但正如上一段所述，要证明失败是很困难的。

7 伦理考量

这项工作涉及人工标注。不过，我们使用的是程序生成的数据，从而确保其中不含机密或有害内容。此外，标注工作是在研究人员的工作时间内完成的。基于这些原因，我们的机构审查委员会根据内部指南认定该研究可免于正式审查。

参考文献

Cristian-Paul Bara、Sky CH-Wang 和 Joyce Chai。2021年。MindCraft：协作任务中情境对话的心智理论建模。载于*2021*年自然语言处理实证方法会议论文集，第 1112–1125 页，线上会议及多米尼加共和国蓬塔卡纳。计算语言学协会。

Simon Baron-Cohen、Alan M Leslie 和 Uta Frith。1985年。自闭症儿童是否具有“心智理论”？认知，21(1):37–46。

Johan Benthem、Jan van Eijck、Malvin Gattinger 和 Kaile Su。2018年。动态认知逻辑的符号模型检验——S5 及更多*。逻辑与计算杂志，28:367–402。

Stella Biderman、Hailey Schoelkopf、Quentin Anthony、Herbie Bradley、Kyle O’Brien、Eric Hallahan、Mohammad Aflah Khan、Shivanshu Purohit、USVSN Sai Prashanth、Edward Raff 等。2023。Pythia：用于分析大型语言模型训练与扩展过程的一套工具。*arXiv* 预印本 *arXiv:2304.01373*。

Thomas Bolander。2018年。《眼见为实：在动态认知逻辑中形式化错误信念任务》。《Jaakko Hintikka论知识与博弈论语义学》，第207–236页。

Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell、Sandhini Agarwal、Ariel Herbert-Voss、Gretchen Krueger、Tom Henighan、Rewon Child、Aditya Ramesh、Daniel Ziegler、Jeffrey Wu、Clemens Winter、Chris Hesse、Mark Chen、Eric Sigler、Mateusz Litwin、Scott Gray、Benjamin Chess、Jack Clark、Christopher Berner、Sam McCandlish、Alec Radford、Ilya Sutskever 和 Dario Amodei。2020年。语言模型是少样本学习者。载于神经信息处理系统进展，第 33 卷，第 1877–1901 页。Curran Associates, Inc.

Peter Clark、Oyvind Tafjord 和 Kyle Richardson。2020年。作为语言软推理器的 Transformer。见 第二十九届国际人工智能联合会议论文集，*IJCAI-20*，第 3882–3890 页。国际人工智能联合会。主会场。

Michael Cohen。2021年。通过文本蕴含探索 RoBERTa 的心智理论。*PhilArchive*。

F. de Waal。2016年。我们够聪明，能知道动物有多聪明吗？W. W. Norton。

Lasse Dissing 和 Thomas Bolander。2020年。使用动态认知逻辑在机器人上实现心智理论。见 第二十九届国际人工智能联合会议论文集，*IJCAI-20*，第 1615–1621 页。国际人工智能联合会。主会场。

Leo Gao、Jonathan Tow、Stella Biderman、Sid Black、Anthony DiPofi、Charles Foster、Laurence Golding、Jeffrey Hsu、Kyle McDonell、Niklas Muennighoff、JasonPhang、Laria Reynolds、Eric Tang、Anish Thite、Ben Wang、Kevin Wang 和 Andy Zou。2021年。少样本语言模型评估框架。

Suchin Gururangan、Swabha Swayamdipta、Omer Levy、Roy Schwartz、Samuel Bowman 和 Noah A. Smith。2018年。自然语言推理数据中的标注伪影。载于计算语言学协会北美分会 *2018*年会议论文集：人类语言技术，第*2*卷（短文），第107–112页，美国路易斯安那州新奥尔良。计算语言学协会。

KyleHamilton、AparnaNayak、BojanBožić 和 Luca Longo。2022年。神经符号人工智能是否兑现了其在自然语言处理中的承诺？一篇结构化综述。语义网，第 1–42 页。

Pengcheng He、Jianfeng Gao 和 Weizhu Chen。2021年。DeBERTaV3：通过采用 ELECTRA 风格的预训练和梯度解耦嵌入共享来改进 DeBERTa。*arXiv* 预印本 *arXiv:2111.09543*。

Chadi Helwe、Chloé Clavel 和 Fabian Suchanek。2022年。Logitorch：一个基于 PyTorch、用于自然语言逻辑推理的库。载于*2022*年会议 自然语言处理实证方法：系统演示。

Michal Kosinski。2023。心智理论可能已在大语言模型中自发涌现。*arXiv* 预印本 *arXiv:2302.02083*。

Matthew Le、Y-Lan Boureau 和 Maximilian Nickel。2019年。通过问答重新审视心智理论的评估。载于*2019*年自然语言处理经验方法会议和第九届国际自然语言处理联合会议（*EMNLP-IJCNLP*）论文集，第5872–5877页，中国香港。计算语言学协会。

Clarence Irving Lewis、Cooper Harold Langford 和 P Lamprecht。1959 年。符号逻辑，第 170 卷。Dover 出版社，纽约。

Xiaomeng Ma、Lingyu Gao 和 Qihui Xu。2023。Tom- challenges：用于探索心智理论的原则指导型数据集和多样化评估任务。*arXiv* 预印本 *arXiv:2305.15068*。

Shima Rahimi Moghaddam 和 Christopher J Honey。2023。通过提示提升大语言模型的心智理论表现。*arXiv* 预印本 *arXiv:2304.11490*。

Ester Navarro、Sara Anne Goring 和 Andrew R. A. Conway。2020年。心智理论与智力之间的关系：一种形成性 g 方法。智力杂志，9。

Aida Nematzadeh、Kaylee Burns、Erin Grant、Alison Gopnik 和 Tom Griffiths。2018年。在问答中评估心智理论。载于*2018*年自然语言处理实证方法会议论文集，第2392–2400页，比利时布鲁塞尔。计算语言学协会。

OpenAI。2023。GPT-4技术报告。

Rebecca Qian、Candace Ross、Jude Fernandes、Eric Smith、Douwe Kiela 和 Adina Williams。2022。通过扰动增强实现更公平的自然语言处理。*arXiv* 预印本 *arXiv:2205.12586*。

Kyle Richardson 和 Ashish Sabharwal。2022年。《通过自然语言可满足性拓展 Transformer 中规则推理的极限》。载于人工智能 *AAAI* 会议论文集，第 36 卷，第 11209–11219 页。

Maarten Sap、Hannah Rashkin、Derek Chen、Ronan Le Bras 和 Yejin Choi。2019年。Social IQa：关于社会互动的常识推理。载于*2019*年自然语言处理实证方法会议论文集和第九届国际自然语言处理联合会议（*EMNLP-IJCNLP*），第 4463–4473 页，中国香港。计算语言学协会。

Natalie Shapira、Mosh Levy、Seyed Hossein Alavi、XuhuiZhou、Yejin Choi、YoavGoldberg、Maarten Sap 和 Vered Shwartz。2023a。聪明的汉斯，还是神经心智理论？对大语言模型中的社会推理进行压力测试。*arXiv* 预印本 *arXiv:2305.14763*。

Natalie Shapira、Guy Zwirn 和 Yoav Goldberg。2023b。大语言模型在失言测试中的表现如何？载于计算语言学协会：*ACL 2023* 论文集，第 10438–10451 页，加拿大多伦多。计算语言学协会。

Damien Sileo 和 Marie-Francine Moens。2022年。探测神经语言模型对概率估计词语的理解能力。*arXiv* 预印本 *arXiv:2211.03358*。

Aarohi Srivastava、Abhinav Rastogi、Abhishek Rao、Abu Awal Md Shoeb、Abubakar Abid、Adam Fisch、Adam R Brown、Adam Santoro、Aditya Gupta、Adrià Garriga-Alonso 等。2022年。超越模仿游戏：量化并外推语言模型的能力。*arXiv* 预印本 *arXiv:2206.04615*。

Jidong Tian、Yitian Li、Wenqing Chen、Liqiang Xiao、Hao He 和 Yaohui Jin。2021年。通过 LogicNLI 诊断一阶逻辑推理能力。载于*2021*年自然语言处理经验方法会议论文集，第 3738–3747 页，在线出版，地点为多米尼加共和国蓬塔卡纳。计算语言学协会。

Tomer Ullman。2023。大语言模型在心智理论任务的细微改动上表现不佳。*arXiv* 预印本 *arXiv:2302.08399*。

Iris Van De Pol、Iris Van Rooij 和 Jakub Szymanik。2018年。动态认知逻辑中心智理论推理的参数化复杂度。逻辑、语言与信息杂志，27：255–294。

Hans Van Ditmarsch 和 Willem Labuschagne。2007。关于你的信念的我的信念：心智理论与认知逻辑案例研究。*Synthese*，155：191–209。

Jan van Eijck。2014年。动态认知逻辑。*Johan van Benthem* 论逻辑与信息动态，第175–202页。

Jason Weston、Antoine Bordes、Sumit Chopra 和 Tomás Mikolov。2016年。迈向人工智能完备的问答：一组先修玩具任务。载于第*4*届国际学习表征会议（*ICLR 2016*），*2016*年*5*月*2*日至*4*日，波多黎各圣胡安，会议论文集。

Honghua Zhang、Liunian Harold Li、Tao Meng、Kai- Wei Chang 和 Guy Van den Broeck。2023。论从数据中学习推理的悖论。载于*IJCAI23*，第3365–3373页。国际人工智能联合会议组织。主会场。

使用 GPT3.5 的一个定性示例

用户：
[前提] 有三个人。每个人都能被其他人看见。公开地，存在一个——
宣布有人额头上沾了泥。现公开宣布 Her-
伯特知道某人的额头是否沾有泥巴。 [假设]Herbert 可以
现在知道 Herbert 的额头是脏的。这个假设是真的还是假的?

GPT-3.5：
无法根据给定的前提确定假设的真值。
虽然已公开宣布某人的额头沾有泥巴，但尚不清楚这片
沾有泥巴的额头属于谁。因此，仅凭这些信息无法确定 Herbert 的
额头是否沾有泥巴。然而，已说明 Herbert 知道某人的额头是否沾有泥巴，因此 Herbert-
bert 可能知道他们的
根据前提中未提供的额外信息，判断自己的额头是否沾有泥巴。

图4：GPT-3.5 对话