对 *Strachan 等(2024)*研究结果的计算可重复性检验

小组成员分工

组长	平航		
组员	常霖、苏欣、李佳		
分工			
数据分析	平航 60%、李佳 20%、常霖 20%	PPT 制作	苏欣 70%、平航 15%、常霖 15%
文字报告制作	李佳 30%、常霖 30%、苏欣 20%、平航 20%	PPT 展示	苏欣

* 同一名同学可负责多个部分；如同一内容由多位同学负责，可按百分比注明贡献占比。

摘要 我们作为人类的核心特征之一是心理理论，即理解并追踪他人心理状态的能力。而近年来，大语言模型（LLMs）的发展引发了关于这些模型在理解心理理论任务中是否能与人类表现无法区分的激烈讨论。方法：研究采用了一系列不同心理理论测试（理解错误信念、解释间接请求、察觉失言等），比较两种 LLMs（GPT 和 LLaMA2）与 1907 名人类参与者的表现。结果：结果显示，在理解间接请求、错误信念和误导方面，GPT-4 模型的表现达到甚至有时超过了人类水平；然而，在识别失言方面，GPT 表现较差。相比之下，LLaMA2 在识别失言测试中表现优于人类，但后续的信念可能性调查显示，LLaMA2 的优势可能是一种倾向于归因于无知的假象。研究还发现，GPT 表现不佳源于对结论的过度保守，而非推理失败。结论：这些发现不仅证明了 LLMs 表现出的行为与人类心理推理的结果是一致的，而且还强调了进行系统测试的重要性，以确保对人工智能和人类之间进行非表面化的比较。

关键词 心理理论；大语言模型 (LLMs)；人类；比较测试；计算可重复性

1 引言

1.1 所选文献

文献：Testing theory of mind in large language models and humans

数据及代码：https://osf.io/fwj6v/

1.2 文献介绍

我们作为人类的核心定义之一是心理理论的概念，即理解和追踪他人心理状态的能力。最近，大语言模型（LLMs），例如 ChatGPT 的发展，引发了关于这些模型在心理理论任务中表现是否能与人类行为难以区分的激烈讨论。

为此，研究者们对比了人类和 LLMs 在一系列不同心理理论能力测量中的表现，这些能力包括理解错误信念、解释间接请求（暗示任务）、识别讽刺、奇怪故事和察觉失言等。研

究者们反复测试了两个 LLMs 家族（GPT 和 LLaMA2），并将其表现与 1907 名人类参与者的表现进行了比较。

在一系列心理理论测试中，我们发现 GPT-4 模型在解释间接请求、错误信念和误导方面的表现达到了或甚至超过了人类水平，但在察觉失言方面却表现不佳。而在察觉失言测试中，LLaMA2 表现优于人类。然而，后续对信念可能性的进一步研究表明，LLaMA2 的优越表现可能是表象的，可能反映了其倾向于归因于无知的偏好。相比之下，GPT 表现不佳的原因并非是推理失败，而是因为对做出结论过于保守。

这些发现不仅表明 LLMs 在行为上表现出与人类心理推理输出一致的特征，还强调了系统测试的重要性，以确保对人类和人工智能的比较更加深入和准确。

2 方法

2.1 原研究方法简介

2.1.1 原研究设计

原研究选取人类被试和两种 GPT 模型以及三种 LLaMA2 模型作为被试，进行一系列心理理论系列测试，包括理解错误信念、识别讽刺、察觉失言、暗示任务、奇怪故事等，其中察觉失言测试在研究者分析不同模型表现水平后进行了失言可能性测试、信念似然测试的后续研究。

原研究中的数据分析和绘图使用了 R 4.1.2 和 RStudio 2024.04.0-daily+368 “Chocolate Cosmos” Daily，贝叶斯分析使用了 JASP 0.18.3。研究者们通过反应编码的方式将被试表现转化为分数，使用并通过经 Holm 校正的两独立样本 Wilcoxon 检验来进行统计分析。

2.1.2 小组复现内容

本小组的复现工作主要有以下两部分：

一，对比了人类和三种大语言模型（LLM）在错误信念（false belief）、讽刺（irony）、暗示任务（hinting）、察觉失言任务（faux pas）和奇怪故事任务（strange stories）这五个心理理论测试中的成绩。

二，对比了人类和三种大语言模型（LLM）在错误信念（false belief）、暗示任务（hinting）、察觉失言任务（faux pas）奇怪故事任务（strange stories）中新旧项目测试的成绩。

本小组选用的数据为“scores_gpt.csv”和“scores_human.csv”，是 GPT - 4、GPT - 3.5、LLaMA2 和人类分别在错误信念（false belief）、讽刺（irony）、暗示任务（hinting）、察觉失言任务（faux pas）和奇怪故事任务（strange stories）这五个心理理论测试的新旧题项上的得分。

在计算可重复性检验过程中，我们使用了用于数据操作的 dplyr，主要使用的函数有创建新变量或修改现有变量的 mutate、计算分组数据的汇总统计量的 summarise、按行进行操作的 rowwise；用于数据整理的 tidyr，主要使用的函数有将数据从宽格式转换为长格式的 gather；用于数据可视化的 ggplot2，主要使用的函数有用于绘制小提琴图的 geom_violin、

用于设置图形主题和样式的 theme_classic 和 theme、用于添加统计汇总的 stat_summary、用于创建分面图的 facet_grid；用于统计分析和添加比较的 rstatix，主要使用的函数有进行 Wilcoxon 秩和检验的 wilcox_test、进行多重比较校正的 adjust_pvalue；用于创建增强版的表格的 kableExtra，主要使用的函数有创建一个基本表格的 kable、对表格进行美化的 kable_styling；用于创建美观表格的 gt，主要使用的函数有设置表格标题、列标签和数字格式的 tab_header、cols_label、fmt_number；用于引导法计算的 boot，主要使用的函数有进行引导法计算的 boot、计算引导法置信区间的 boot.ci；用于非参数统计的 coin，主要使用的函数有进行 Wilcoxon 秩和检验 wilcox_test。

2.2 复现思路与 R 包

2.2.1 准备工作

获取必要的软件和工具：安装并设置好 R、RStudio 和 JASP。获取数据和代码：载数据集和评分文件：从 OSF 库下载，访问 https://osf.io/dbn92。下载数据分析代码：访问 https://osf.io/j3vhq。

2.2.2 数据理解和预处理

了解数据结构：仔细阅读研究论文的“方法”部分，了解数据的来源、格式和变量定义。打开下载的数据文件，查看数据的整体结构和内容，确保对每个变量的含义有清晰的理解。

2.2.3 设置分析环境

在 RStudio 中设置项目：打开 RStudio，创建一个新的项目，并将下载的数据和代码文件导入项目目录中。安装必要的 R 包：在 RStudio 的控制台输入并运行以下代码，安装所有必要的 R 包：dplyr、tidyr、ggplot2、ggpubr、rstatix、kableExtra、gt、boot、coin。

2.2.4 运行数据分析代码

打开并运行 RMarkdown 文件。在 RStudio 中打开下载的 RMarkdown 文件（.Rmd），逐块运行代码，或点击“Run”按钮运行整个 RMarkdown 文件。确保每个代码块都成功运行，并记录下每步的输出结果。

2.2.5 结果对比和验证

对照三人的输出结果。将分析结果与研究论文中报告的结果进行详细对比，确保各项结果一致，并根据原代码复现图表，验证其准确性。

调试和修正。如果发现任何不一致，仔细检查数据预处理、代码实现和分析步骤，找出可能的原因并进行修正。

2.2.6 修改与验证

为了验证研究中提到的数据的可重复性，我们更换了一些 R 包和函数，具体包括计算 Z 值、*p* 值、r 值，以及 95%的置信区间。这些修改旨在确保分析过程的准确性和可靠性。通过使用不同的 R 包和函数，我们能够进一步确认计算的可重复性。

3 结果

3.1 描述性统计

由于原文未提及描述性统计结果，因此未对该部分进行复现。

3.2 推断性统计

报告对原文献推断性统计进行重复的结果，并汇总表格：

3.2.1 研究一

对比了人类和三种大语言模型（GPT - 4、GPT - 3.5、LLaMA2）在错误信念任务（False belief）、讽刺任务（Irony）、失言识别任务（Faux pas）、暗示任务（Hinting）和奇怪故事任务（Strange stories）这五个心理理论测试中的成绩，见下表：

其中错误信念任务（False belief）产生天花板效应，原文未展示数据。

表 1-1 讽刺任务的复现结果推断性统计

	讽刺		
	GPT-4 vs. Human	GPT-3.5 vs. Human	LLaMA2 vs. Human
	P	*P*	*P*
原文数据	0.040	2.37×10^{-6}	2.39×10^{-7}
复现数据	0.040	2.37×10^{-6}	2.39×10^{-7}
δ	0%	0%	0%
level	完全一致	完全一致	完全一致

表 1-2 失言识别任务的复现结果推断性统计

	失言识别		
	GPT-4 vs. Human	GPT-3.5 vs. Human	LLaMA2 vs. Human
	P	*P*	*P*
原文数据	5.42×10^{-5}	5.95×10^{-8}	0.002
复现数据	5.42×10^{-5}	5.95×10^{-8}	0.002
δ	0%	0%	0%
level	完全一致	完全一致	完全一致

表 1-3 暗示任务的复现结果推断性统计

	暗示		
	GPT-4 vs. Human	GPT-3.5 vs. Human	LLaMA2 vs. Human
	P	*P*	*P*
原文数据	0.040	0.626	5.42×10^{-5}
复现数据	0.040	0.626	5.42×10^{-5}
δ	0%	0%	0%
level	完全一致	完全一致	完全一致

表 1-4 奇怪故事任务的复现结果推断性统计

	奇怪故事		
	GPT-4 vs. Human	GPT-3.5 vs. Human	LLaMA2 vs. Human
	P	*P*	*P*
原文数据	1.04×10^{-5}	0.110	0.005
复现数据	1.04×10^{-5}	0.110	0.005
δ	0%	0%	0%
level	完全一致	完全一致	完全一致

3.2.2 研究二

对比了人类和三种大语言模型（LLMs）在错误信念任务（False belief）、失言识别任务（Faux pas）、暗示任务（Hinting）和奇怪故事任务（Strange stories）中新旧项目测试的成绩，见下表：

表 2-1 失言识别任务复现结果推断性统计

	失言识别							
	Human				GPT3.5			
	Z	*P*	*r*	CI	*Z*	*P*	*r*	CI
原文数据	-0.10	0.029	0.29	0.10-0.50	0.10	0.002	0.69	0.49-0.88
复现数据	-0.10	0.03	0.29	0.10-0.50	0.10	0.00	0.69	0.49-0.88
δ	0%	0%	0%	0%	0%	0%	0%	0%
level	完全一致	完全一致	完全一致	完全一致	完全一致	完全一致	完全一致	完全一致

表 2-2 暗示任务复现结果推断性统计

	暗示任务															
	Human				GPT3.5				GPT4				LLaMA2			
	Z	*P*	*r*	CI	*Z*	*P*	*r*	CI	*Z*	*P*	*r*	CI	*Z*	*P*	*r*	CI
原文数据	-0.10	0.008	0.34	0.14-0.53	-0.03	0.955	0.24	0.02-0.59	-0.10	0.123	0.44	0.07-0.75	-0.20	9.18×10^{-4}	0.73	0.50-0.87
复现数据	-0.10	0.008	0.34	0.14-0.53	-0.03	0.955	0.24	0.02-0.59	-0.10	0.123	0.44	0.07-0.75	-0.20	0.001	0.73	0.50-0.87
δ	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
level	完全一致	完全一致	完全一致	完全一致	完全一致	完全一致	完全一致	完全一致	完全一致	完全一致	完全一致	完全一致	完全一致	完全一致	完全一致	完全一致

3.3 对原文计算可重复性进行评估

报告对原文献推断性统计进行重复的结果，并汇总表格：

报告原文献的值的评级分布情况，整理成表格，如下表所示：

表 3 计算可重复性的评估表

可重复性情况	数量及占比	
	N	%
完全一致($\delta = 0\%$)	18	100
偏差较小($0\% < \delta < 10\%$)	0	0
偏差较小($\delta > 10\%$)	0	0
因舍入导致的偏差	0	0

* 这里的 *N* 指的在重复分析中，对重复分析结果与原结果进行配对比较的次数。例如，原文仅进行了一个 *t* 检验，则 *N*＝1；如果原文进行了一个 2*2 的方差分析，并进行简单效应分析，则有可能有 7 个统计检验的数值：两个主效应，一个交互作用，四个可能的效应效应分析的 *t* 或者 *F* 值，因此 *N*＝7。

4 讨论

4.1 计算可重复性检验结果分析

一，未提供原始数据，仅提供了清洗过的数据。

二，为了确保模型不仅仅是复现训练集中的数据，研究者为每个已发表的测试生成了新的测试项。研究假设生成的新测试项未被 LLM 学习过，以此来比较新旧项目之间的差异。然而，不能完全保证生成的新测试项未被 LLM 学习过。

三，复现结果与文中结果的微小差异，是由于保留小数位的不同而产生的。

四，能完全复现结果，是因为作者公开了清洗好的数据集以及完整的代码。

五，文中未提供描述性统计数据，因而无法对这部分进行复现。

表 4 计算上可重复的原因分析表

	可能原因	研究一	研究二
一般开放获取性问题	几个结果的微小差异，可能是由于分析中使用了没有设置固定种子的随机数；	无	无
	个别结果的微小差异，可能是由于印刷或复制粘贴错误；	无	无
	文章文本中程序报告不明确，包括纳入亚组的标准、缺乏或不正确报告用于回归模型的变量、以及未报告的单侧分析；	无	无
	在文章的开放实践声明中对研究的模糊标记。	无	无
OSF 开放获取特定问题	OSF 中缺乏对数据和/或代码内容进行说明的文档(readme 文档)；	无	无
	OSF 上的数据与代码文件不一致，如代码中对部分数据进行了操作，但这部分数据在数据文件中无对应；	无	无
	OSF 上的数据存储问题，包括文件损坏或无法下载。	无	无
数据开放获取特定问题	没有提供原始数据；	存在。但不影响本研究复现。	存在。但不影响本研究复现。
	没有提供处理后的数据；	无	无
	没有提供数据处理过程的描述或代码。	无	无
代码开放获取特定问题	缺乏共享的分析代码或建模代码；	无	无
	软件包或软件版本的问题。	无	无
其他可能因素	文章出版年代；	无	无
	重复者此前是否有过 R 语言使用经验；	存在。R 语言使用经验缺乏者，对于选哪个R包和函数以及函数中的代码怎么写会有困惑。	存在。R 语言使用经验缺乏者，对于选哪个R包和函数以及函数中的代码怎么写会有困惑。
	重复者对关于 R 的知识或操作上存在漏洞，较难理解原文章中的部分操作(可做简单说明)。	无	无

4.2 其他思考

首先，通过这次作业，我们深刻认识到导入、清理和转换数据的重要性。这要求我们熟练掌握数据框的操作，了解如何处理缺失值和异常值。我们知道了数据质量对研究结果至关重要，也意识到数据采集和预处理的重要性，包括数据清洗等工作。尽管本文提供的数据已经经过清洗，但我们在此过程中缺少了对数据处理的训练实践，这也是未来需要进一步强化的部分。

其次，在解释结果时，我们学会了谨慎对待，避免过度推断。文中对于 LLaMA2 在失言识别任务上的完美表现并未简单推断，而是进一步思考了可能存在的更深层次原因。这种不满足于表面结论的态度值得在今后的研究设计中借鉴和学习。

另外，通过小组合作完成这次复现顶刊数据分析的作业，我们感受到了团队的力量。作业准备的前期，每个人都在独自摸索，感觉这个作业非常困难，没有清晰的思路和具体的步骤。后来，我们开始进行小组讨论，确定具体的作业实施方案。团队的思想碰撞激发了很多好点子，每个人可以将自己的想法说出来，大家一起讨论完善，集思广益，很快就确定了复现的思路和分工。作业的顺利完成离不开小组合作的功劳。

还有，查找文献和数据分析的过程，也是对我们之前学习的 R 语言知识的检验。这次作业让我们能够独立实践并巩固了所学内容，因此是一次很好的锻炼和提升机会。

最后，建议老师和助教能及时提供关于作业的反馈和指导，帮助大家更好地改进和提高。

参考文献

Strachan, J. W. A., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., Graziano, M. S. A., & Becchio, C. (2024). Testing theory of mind in large language models and humans. *Nature human behaviour*, 10.1038/s41562-024-01882-z. Advance online publication. https://doi.org/10.1038/s41562-024-01882-z