自动发现的思维链提示可泛化至新模型和数据集

Konstantin Hebenstreit[1, 2, *], **Robert Praas**[1, 3, *], **Louis P Kiesewetter**[4],**Matthias Samwald**[1, §]人工智能研究所，维也纳医科大学，奥地利[2]林茨约翰·开普勒大学，奥地利[3] 瑞典皇家理工学院，斯德哥尔摩，瑞典[4]柏林洪堡大学，德国[*]共同贡献者， [§]通讯作者 matthias.samwald[@]meduniwien.ac.at

arXiv:2305.02897v2 [cs.CL] 3 Aug 2023

摘要

大语言模型（LLMs）涌现出的思维链（CoT）推理能力有望提升其性能和可解释性。然而，对于为先前模型代际制定的推理策略如何泛化到新的模型代际和不同数据集，人们仍存在诸多疑问。在这项小规模研究中，我们比较了通过零样本提示在六个近期发布的大语言模型（Davinci-002、Davinci-003、GPT-3.5-turbo、GPT-4、Flan-T5-xxl 和 Cohere command-xlarge）上诱导出的不同推理策略；实验使用了由六个问答数据集组成的混合数据集，其中包括来自科学和医学领域的数据集。我们的研究结果表明，尽管有效性存在一些差异，但思维链推理策略带来的收益在不同模型和数据集之间仍然稳健。GPT-4 从当前最先进的推理策略中获益最多，并且通过应用一个此前通过自动发现得到的提示，展现出最佳性能。

1 动机

大语言模型（LLMs）涌现出的思维链（CoT）推理能力，有望在处理复杂任务时提升模型的预测性能和可解释性（Wei 等人，2021年）。虽然针对具体任务选择合适示例进行少样本上下文提示工程可以取得良好效果，但零样本提示无需根据任务选择示例（Kojima 等人，2022年）。近期，面向对话优化的模型（如 GPT-3.5）取得成功，进一步提升了人们的预期：模型能够通过临时采用的推理策略实现稳健的性能，并且较少受到细微变化的影响。本研究通过实证评估，考察先前发现的零样本思维链提示风格能否泛化到新一代模型和数据集，以及它们与新开发的推理策略相比如何。我们在六个复杂程度各异的问答数据集上开展评估，其中包括科学和医学领域的数据集。

2 方法

2.1 数据集

我们的研究采用了 ThoughtSource 框架（Ott 等人，2023年），该框架提供了一个全面的元数据集和软件库，旨在简化思维链（CoT）推理的生成、评估和标注。我们选取了六个数据集的子样本，涵盖常识（Talmor 等人，2019年；Geva 等人，2021年）、科学（Xie 等人，2020；Mihaylov 等人，2018年）和医学领域（Jin 等人，2021年；Pal 等人，2022年）（表1）。所有这些问答数据集均为多项选择题，每题包含两个至五个答案选项，且只有一个正确答案。

2.2 提示

我们整理了一组包含十种零样本推理策略的提示（表5和附录A），包括一种基线、两种已有设计和七种新设计：

1. 直接提示：不使用特定触发语，作为比较基线。2. Kojima：广为人知的思维链提示词：“让我们一步一步地思考。”(Kojima等人，2022年) 3. Zhou：通过自动化提示词工程创建的增强版本：“让我们一步一步地解决这个问题，以确保答案正确。”(Zhou等人，2023年) 4. 我们设计的七种原创推理策略，灵感来自各种公开资源(OpenAI，2023a；Schulhoff，2022年)，并通过基于……的迭代适应加以改进

数据集	描述
CommonsenseQA	语义复杂性较高的通用领域众包问题，需要运用先验知识。
StrategyQA	需要隐式推理的通用领域众包问题，以及多步作答策略。是/否答案。
WorldTree v2	面向三至五年级水平的小学科学问题，结合领域特定知识和世界知识。
OpenBookQA	科学和广泛常识问题，需要多步推理和丰富的文本理解能力。
MedQA	医学资格考试中的问题。我们仅使用了来自美国的示例（美国执业医师资格考试子集）。
MedMCQA	现实世界中的医学入学考试题。

表 1：各种数据集的描述。

<escape>分析输出。其中一种策略采用了自我批评策略，要求模型先给出初步答案，对其进行批评，然后提出修订后的回答（Madaan 等人，2023年； Saunders 等人，2022年）。</escape>

2.3 模型

我们根据相关报告中模型在思维链推理方面的能力，纳入了六个指令微调模型：davinci-002（Brown 等人，2020）、davinci-003（Ouyang 等人，2022年）、GPT-3.5-turbo（OpenAI，2022年）和 GPT-4（OpenAI，2023b），它们均来自 OpenAI；Flan-T5-xxl 来自谷歌（Chung 等人，2022年）；command-xlarge-nightly 来自 Cohere（Cohere.ai，2023年）。我们使用 LangChain 框架（Chase，2022年）访问多个应用程序接口。在 2023年2月至4月期间，我们进行了 11,880 次实验，所有实验均将模型温度设为 0，最大令牌长度设为 512。

2.4 评估

我们选择 Krippendorff α系数作为评估指标（Krippendorff，2011）。它通过校正相应的基准概率率，使我们能够合并来自答案选项数量不同的子数据集的结果。Krippendorff α系数以从零（随机机会）到一（完全一致）的尺度测量评定者间信度，并用于将模型预测结果与金标准答案进行比较（Castro，2017）。

为确定合适的样本量，我们进行了功效分析，显著性水平设为 0.05，Krippendorff α系数取中等值 0.8，基线正确概率为 0.2，并考虑了子数据集中最多有五个答案选项（附录 B）。分析得出的所需样本量为164个条目，我们将其增加至198个条目，并划分为六个子数据集，每个子数据集包含33个条目。我们采用自助法（r = 1000）计算生成结果的均值和置信区间。为确保 Krippendorff 评分的准确性（该评分取决于选项的数量），我们在必要时分别对每个子数据集进行自助抽样，并通过合并标准差来计算置信区间。

3 结果

本文中的所有分数均以 95% 置信区间（置信区间）显示。

提示	GPT-4 α (置信区间) 每个提示的数量 = 198	模型平均值 α (置信区间) 每个提示的 n = 1188
Zhou	.83 (.77, .90)	.53 (.50, .57)
Kojima	.80 (.73, .87)	.51 (.47, .55)
Zhou-instr.	.79 (.72, .86)	.50 (.46, .54)
清晰表达	.79 (.71, .86)	.52 (.48, .56)
改述	.78 (.71, .85)	.54 (.51, .58)
Plan	.77 (.71, .84)	.50 (.46, .54)
详细阐述	.77 (.70, .84)	.51 (.47, .55)
自我批评	.76 (.69, .84)	.49 (.45, .53)
对话	.74 (.66, .81)	.47 (.43, .51)
直接	.71 (.64, .79)	.49 (.45, .52)

表 2：提示在各数据集上的 Krippendorff α系数（α）性能。平均值仅针对 GPT-4 和总计六个模型计算，n 总计为 = 11880。

尽管许多提示在所有数据集上的平均性能相当接近，但我们发现，采用推理策略的表现优于直接提示。进一步审视结果可

得自最新模型 GPT-4 的结果，凸显了使用特定提示的优势（表2）。结果显示，Zhou等人（2023年）自动发现的提示，其性能得以保持；该提示在跨模型平均得分方面也表现不俗。有趣的是，自我批判提示的得分相对较低。此外，在多个观察到的实例中，它还导致生成了多个答案，这些答案未纳入评分过程。将Zhou的推理提示置于问题之前而非之后来构建指令提示，并未带来更好的结果。

性能更强的模型发现 WorldTree v2 和 CommonsenseQA 越来越容易（表9），而 StrategyQA 则受到一些特殊题目的影响。这凸显了开发更加精细的通识知识数据集或采用领域专用数据集的必要性，例如两个医学数据集（表3）。

数据集	α（置信区间） 每个数据集的 n 值 = 1980
WorldTree v2	.83 (.81, .85)
CommonsenseQA	.71 (.68, .73)
OpenBookQA	.65 (.63, .68)
StrategyQA	.31 (.27, .36)
MedMCQA	.31 (.28, .34)
MedQA	.21 (.19, .24)

表 3：在数据集上对模型和提示取平均后的 Krippendorff α系数（α）表现。总计 N = 11880。

FLAN-T5 对于其规模而言表现良好，但其结果可能受到数据污染的影响：它使用子数据集 CommonsenseQA 和 StrategyQA 进行了指令微调。GPT-3.5-turbo 和 GPT-4 表现最佳（表 4），也是唯一在医学数据集上表现尚可的模型（表9）。

模型	α（置信区间） 每个模型 = 1980
GPT-4	.78 (.76, .81)
GPT-3.5-turbo	.62 (.59, .65)
Davinci-003	.47 (.45, .50)
Flan-T5-XXL	.45 (.42, .47)
Davinci-002	.41 (.38, .44)
Command-XL	.32 (.29, .35)

表 4：模型在各数据集和提示上的 Krippendorff α系数（α）表现的平均值。总计 N = 11880。

更多详细结果以及以准确率值报告的结果，请参阅附录。

4 局限性

这项工作存在一些局限性。我们的研究旨在预算受限的情况下，测试各种提示、数据集和模型组合。因此，我们根据统计功效分析对数据集进行了子抽样。这限制了我们的结果与完整基准测试集评估结果之间的直接比较。在检查通过众包生成的一些学术基准数据集的结果时，我们发现相当一部分样例的质量并不理想。我们发现的一个常见问题是，问题和答案选项无法明确地选出最佳答案。更先进的模型往往会在推理回答中正确指出这类问题，并避免只选择一个答案选项。我们没有采用自一致性等方法（Wang etal.,2022年），因为这类方法以牺牲实际可解释性为代价来最大化最终准确率；换言之，我们关注的是用户期待获得一条高质量、易于理解的推理链，而非一组嘈杂推理链的情形。将提示与集成方法结合使用时，所取得的结果可能会有所不同。我们的研究纳入了最先进的闭源模型，而这些模型不断变化，因此难以进行复现和跨时间比较。我们通过公开实验期间模型生成的所有数据，部分回应了这一顾虑。闭源模型缺乏文档记录，也让人担忧训练数据可能受到基准数据集的污染。虽然不同提示之间的比较不会受到严重影响，但出于这一原因，我们提醒读者不要过度解读不同模型之间的结果。我们注意到，Flan-T5（Longpreet al.,2023年）在 CommonsenseQA 和 StrategyQA 的子集上经过指令微调后，在 CommonsenseQA 上的表现优于 GPT-3.5-turbo。

5 讨论

相关工作。 一些相关研究评估了零样本提示的表现。一个值得注意的例子是，Liévin（Liévin等人，2022年）开展了类似的零样本思维链评估，重点关注医学数据集。更早的工作评估了多种

模型和数据集的零样本表现，包括常识数据（Zhou等人，2020）以及对T0在多项选择任务上的表现进行评估（Or-lanski， 2022年）。HELM（Liang等人，2022年）涵盖了广泛的模型比较。我们的研究通过聚焦于寻找简单且通用的思维链提示方法，为现有知识作出了补充；这些方法能够适用于一系列问答数据集和模型。

未来工作。 当前研究可以通过使用更多模型评估提示和数据集来扩展，尤其是大量公开可用的大语言模型，例如 LLaMA、Pythia 套件、对话调优模型 Alpaca（Touvron 等，2023年；Biderman 等，2023年；Taori 等，2023年）、StableLM（Stability AI，2023年）以及 OpenAssistant（LAION，2023年）。最后，还需要开展用户评估，以考察不同提示和模型生成的推理链的质量及其解释效用。

6 致谢

感谢 Cohere 团队提供定制的应用程序接口访问权限，使我们能够更快地进行推理，并且不受限制地分析医疗问题数据集；这些数据集偶尔会被标准应用程序接口标记。

参考文献

Stella Biderman、Hailey Schoelkopf、Quentin Anthony、Herbie Bradley、Kyle O'Brien、Eric Hallahan、Mohammad AflahKhan、ShivanshuPurohit、USVSN Sai Prashanth、Edward Raff、Aviya Skowron、Lintang Sutawika 和 Oskar van der Wal。2023年。 Pythia：用于分析训练和扩展过程中的大语言模型的套件。 *arXiv*。

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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。*arXiv*。

Santiago Castro。2017。Fast{K}rippendorff：快速计算{K}rippendorff 的 alpha 一致性度量。

Harrison Chase。2022年。Langchain。

Hyung Won Chung、Le Hou、Shayne Longpre、Barret Zoph、Yi Tay、William Fedus、Eric Li、XuezhiWang、Mostafa Dehghani、Siddhartha Brahma、Albert Webson、Shixiang Shane Gu、Zhuyun Dai、Mirac Suz- gun、Xinyun Chen、Aakanksha Chowdhery、Sharan Narang、Gaurav Mishra、Adams Yu、Vincent Zhao、Yanping Huang、Andrew Dai、Hongkun Yu、Slav Petrov、Ed H. Chi、Jeff Dean、Jacob Devlin、Adam Roberts、Denny Zhou、Quoc V. Le 和 Jason Wei。2022年。指令微调语言模型的扩展。*arXiv*。

Cohere.ai。2023年。Command nightly。

Mor Geva、DanielKhashabi、Elad Segal、Tushar Khot、Dan Roth 和 Jonathan Berant。2021年。亚里士多德用过笔记本电脑吗？一种采用隐式推理策略的问答基准。计算语言学协会会刊，9：346–361。

Di Jin、Eileen Pan、Nassim Oufattole、Wei-Hung Weng、Hanyi Fang 和 Peter Szolovits。2021年。这位患者患有什么疾病？一项基于医学考试的大规模开放域问答数据集。应用科学，11(14)：6421。

Takeshi Kojima、Shixiang Shane Gu、Machel Reid、Yu- takaMatsuo 和 YusukeIwasawa。2022年。大型语言模型是零样本推理器。*arXiv*。

K Krippendorff。2011年。Krippendorff α系数信度的计算。

LAION. 2023年。Open Assistant |Open Assistant。

Percy Liang、Rishi Bommasani、Tony Lee、Dimitris Tsipras、Dilara Soylu、Michihiro Yasunaga、Yian Zhang、Deepak Narayanan、Yuhuai Wu、Ananya Kumar、Benjamin Newman、Binhang Yuan、Bobby Yan、Ce Zhang、Christian Cosgrove、Christopher D. Manning、Christopher Ré、Diana Acosta-Navas、Drew A. Hudson、Eric Zelikman、Esin Durmus、Faisal Ladhak、Frieda Rong、Hongyu Ren、Huaxiu Yao、Jue Wang、Keshav Santhanam、Laurel Orr、Lucia Zheng、Mert Yuksekgonul、Mirac Suzgun、Nathan Kim、Neel Guha、Niladri Chatterji、Omar Khattab、Peter Henderson、Qian Huang、Ryan Chi、Sang Michael Xie、Shibani Santurkar、Surya Ganguli、Tatsunori Hashimoto、Thomas Icard、Tianyi Zhang、Vishrav Chaudhary、William Wang、Xuechen Li、Yifan Mai、Yuhui Zhang 和 Yuta Koreeda。2022年。语言模型的全面评估。*arXiv*。

Valentin Liévin、Christoffer Egeberg Hother 和 Ole Winther。2022年。大语言模型能推理医学问题吗？*arXiv*。

Shayne Longpre、Le Hou、Tu Vu、Albert Webson、Hyung Won Chung、Yi Tay、Denny Zhou、QuocV.Le、Barret Zoph、Jason Wei 和 Adam Roberts。2023年。Flan Collection：为有效的指令微调设计数据和方法。。

Aman Madaan、Niket Tandon、Prakhar Gupta、Skyler Hallinan、Luyu Gao、Sarah Wiegreffe、Uri Alon、Nouha Dziri、Shrimai Prabhumoye、Yiming Yang、Sean Welleck、Bodhisattwa Prasad Majumder、

<escape>Shashank Gupta、Amir Yazdanbakhsh 和 Peter Clark。2023年 Self-refine：借助自我反馈进行迭代式优化。 *arXiv*。</escape>

Todor Mihaylov、Peter Clark、Tushar Khot 和 Ashish Sabharwal。2018年。盔甲能导电吗？一个用于开放式阅读理解问答的新数据集。*arXiv* 预印本 *arXiv:1809.02789*。

OpenAI。2022年。介绍 ChatGPT。

OpenAI。2023a。GPT-4 开发者直播。

OpenAI。2023b。GPT-4 技术报告。*arXiv*。

Gabriel Orlanski。2022年。在零样本设置下评估多项选择任务中的提示。*arXiv*。

Simon Ott、Konstantin Hebenstreit、Valentin Liévin、Christoffer Egeberg Hother、Milad Moradi、Maximilian Mayrhauser、Robert Praas、Ole Winther 和 Matthias Samwald。2023年。ThoughtSource：大语言模型推理数据的中心枢纽。

Long Ouyang、Jeff Wu、Xu Jiang、Diogo Almeida、Carroll L. Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agarwal、Katarina Slama、Alex Ray、John Schulman、Jacob Hilton、Fraser Kelton、Luke Miller、Maddie Simens、Amanda Askell、Peter Welinder、Paul Christiano、Jan Leike 和 Ryan Lowe。2022年。通过人工反馈训练语言模型遵循指令。*arXiv*。

Ankit Pal、Logesh Kumar Umapathi 和 Malaikannan Sankarasubbu。2022年。Medmcqa：一个用于医学领域问答的大规模多学科多项选择数据集。在健康、推理与学习会议，第248–260页。PMLR。

William Saunders、Catherine Yeh、Jeff Wu、Steven Bills、Long Ouyang、Jonathan Ward 和 Jan Leike。2022年。用于协助人工评估者的自我批评模型。*arXiv*。

Sander Schulhoff。2022年。学习提示工程。

Stability AI。2023年。Stability AI 推出其 StableLM 语言模型套件的首个模型——Stability AI。

Alon Talmor、Jonathan Herzig、Nicholas Lourie 和 Jonathan Berant。2019年。CommonsenseQA：一项以常识知识为目标的问答挑战。

R Taori、I Gulrajani、T Zhang、Y Dubois、X Li、C Guestrin、P Liang 和 TB Hashimoto。2023年。斯坦福 Alpaca：一种指令遵循型 Llama 模型。

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, BaptisteRozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave 和 Guillaume Lample。2023年。LLaMA：开放且高效的基础语言模型。*arXiv*。

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery 和 DennyZhou。2022年。自一致性提升语言模型中的思维链推理。*arXiv*。

Jason Wei、Maarten Bosma、Vincent Y. Zhao、Kelvin Guu、Adams Wei Yu、Brian Lester、Nan Du、Andrew M. Dai 和 Quoc V. Le。2021年。微调后的语言模型是零样本学习器。*arXiv*。

Zhengnan Xie、Sebastian Thiem、Jaycie Martin、Elizabeth Wainwright、Steven Marmorstein 和 Peter Jansen。2020。WorldTree v2：一个科学领域结构化解释及推理模式语料库，支持多跳推理。载于第十二届语言资源与评测会议论文集，第5456–5473页。

Xuhui Zhou、Yue Zhang、Leyang Cui 和 Dandan Huang。2020。评估预训练语言模型中的常识。*AAAI* 人工智能会议论文集，34(05):9733– 9740。

Yongchao Zhou、Andrei Ioan Muresanu、Ziwen Han、Keiran Paster、Silviu Pitis、Harris Chan 和 Jimmy Ba。2023年。大语言模型是人类水平的提示工程师。载于第十一届国际学习表征会议。

A 模型输入模板

本附录概述了我们研究中使用的文本模板和提示结构。
模型输入的结构如下：

{instruction}
{question} {answer_选
项} {cot_触发器}

占位符描述：

- {问题}：模型预期需要回答的多项选择问题。

- {answer_选项}：为多项选择问题提供的选项。

采用两种提示方法（一次仅使用一种）：

- {instruction}：置于问题和答案选项之前。

- {cot_trigger}：置于问题之后。

表5中每个提示的括号内都附有其类型。

B 功率计算公式

$$T\left(P_c, \alpha_{\text{min}}, p\right) =$$

$$= 2z_p^2 \left(\frac{\left(1 + \alpha_{\text{min}}\right)\left(3 - \alpha_{\text{min}}\right)}{4\left(1 - \alpha_{\text{min}}\right) P_c \left(1 - P_c\right)} - \alpha_{\text{min}} \right)$$

其中：

$P_c =$ 值的概率 $c\alpha_{\text{min}} =$ 最小 α 可接受为可靠编码的$p =$ 显著性水平$z_p =$ 在 p处的标准化 z统计量

C 表格

有些提示在特定数据集上的效果尤其好。改写提示似乎有助于处理我们发现 StrategyQA 数据集中普遍存在的模糊问题。

ID	提示名称	Text
None	直接	“直接提示”。不使用特定提示。只】【。 问题和答案选项是.....的输入。 模型。”
kojima-01 （cot_触发器）	*Kojima*	“答案：让我们一步一步地思考。”
zhou-01 （cot_触发器）	*Zhou*	“答案：让我们一步一步地解决这个问题， 确保我们有正确的答案。
zhou-01-ins (指令)	*Zhou*指令	“让我们一步一步地推导，确保无误。 我们有正确答案。
qa-10（指令）	*Plan*	“首先逐步思考——描述你的计划，说明如何 以极其详尽的方式逐步写出得出正确答案的过程。 然后回答这个问题。
qa-12（指令）	清晰表达	"仔细阅读问题，并逐步推导 以确保得出正确答案。 务必详细说明你的想法和推理过程， 以便任何人都能验证。把所有内容都 一丝不苟地详细写出，不要跳过任何步骤！"
qa-13（指令）	改述	“指令：首先，让我们改述一下问题，确保我们正确理解了它。 其次，让我们逐步推演，详细阐明我们的想法和推理过程， 以便任何人都能验证。第三，确保 我们的答案正确。” 确定我们得出了正确答案。”
qa-16（指令）	详细阐述	“通过仔细的、连贯的.....回答以下问题。 精确的逐步推理。首先，补充..... 提供有用知识和重要广告的問題 附加事实。其次，生成子问题，这些子问题..... 需要回答原问题，回答 它们，直到你能回答原问题为止。”
qa-17（指令）	对话	“创建一段教授与学生之间的对话。 学生针对该问题提出子问题。 教授会逐步解答这些问题，并且 确保学生理解了他们是如何得出 正确答案的。”
refl-01（指令）	自我批评	“回答问题，然后批评答案。 根据批评意见，重新考虑另一个答案 选项，并给出一个最终答案。”

表 5：使用的提示及其对应的编号和文本。

提示	模型平均准确率 (CI) 每个提示的数量 = 1188
Zhou	.68 (.65, .70)
清晰表达	.67 (.64, .70)
改述	.67 (.64, .69)
详细阐述	.66 (.63, .69)
Zhou指令	.65 (.63, .68)
Plan	.65 (.62, .68)
Kojima	.64 (.62, .67)
直接	.64 (.61, .67)
自我批评	.64 (.61, .67)
对话	.64 (.61, .66)

<escape>表6：各数据集上的提示准确率平均值。表 5 中可以找到与提示名称对应的文本。对全部六个模型取平均值。总计 = 11880。</escape>

数据集	准确率 （置信区间） 每个数据集的数量 = 1980	基准率
WorldTree v2	.88 (.86, .89)	.25
CommonsenseQA	.77 (.75, .79)	.2
OpenBookQA	.74 (.72, .76)	.25
StrategyQA	.67 (.65, .69)	.5
MedMCQA	.49 (.46, .51)	.25
MedQA	.38 (.36, .40)	.2

表 7：各数据集的准确率，按模型和提示取平均值。随机猜测的基准率取决于数据集中的答案选项数量。总数 N = 11880

模型	准确率 (置信区间) 每个模型的 n = 1980
GPT-4	.85 (.83, .86)
GPT-3.5-turbo	.74 (.72, .76)
Davinci-003	.63 (.61, .65)
Flan-T5-XXL	.61 (.59, .63)
Davinci-002	.59 (.56, .61)
Command-XL	.52 (.50, .55)

表 8：模型在各数据集和提示上的平均准确率。总计 N = 11880。

模型 数据集	Command-XL	Flan-T5-XXL	GPT-3.5-turbo	GPT-4	Davinci-002	Davinci-003
CommonsenseQA	.57 (.50, .64)	.81 (.75, .85)	.70 (.64, .76)	.82 (.76, .87)	.68 (.62, .74)	.68 (.62, .74)
MedQA	.06 (.01, .13)	.02 (.00, .07)	.40 (.32, .47)	.55 (.47, .61)	.09 (.03, .15)	.17 (.11, .24)
MedMCQA	.08 (.01, .14)	.10 (.03, .17)	.51 (.44, .58)	.73 (.67, .79)	.20 (.13, .27)	.21 (.14, .28)
OpenBookQA	.43 (.36, .50)	.69 (.63, .76)	.77 (.71, .83)	.91（.87, .95）	.45 (.37, .52)	.66（.59、.72）
StrategyQA	.10 (.00, .21)	.23 (.12, .34)	.44 (.33, .55)	.69 (.61, .76)	.20 (.09, .32)	.22 (.12, .31)
WorldTree v2	.67 (.61, .73)	.77 (.72, .83)	.89 (.85, .93)	.97 (.95, .99)	.84 (.79, .89)	.84 (.80, .89)

表 9：各模型在不同数据集上的 Krippendorff α系数（α）表现（对提示取平均）。Flan-T5-XXL 在 MedQA 上的低分说明，Krippendorff α系数（α）对表 10 中的准确率进行了校正，以反映表 7 中的基准率。

模型 数据集	Command-XL	Flan-T5-XXL	GPT-3.5-turbo	GPT-4	Davinci-002	Davinci-003
CommonsenseQA	.66 (.61, .71)	.85 (.81, .89)	.76 (.71, .81)	.85 (.81, .90)	.75 (.70, .79)	.75 (.70, .80)
MedQA	.27 (.22, .32)	.22 (.17, .26)	.53 (.47, .58)	.65 (.60, .70)	.28 (.23, .33)	.35 (.30, .40)
MedMCQA	.31 (.26, .36)	.35 (.30, .40)	.63 (.58, .69)	.80 (.76, .85)	.41 (.35, .46)	.41 (.36, .47)
OpenBookQA	.58 (.52, .63)	.78 (.73, .82)	.83 (.79, .88)	.93 (.91, .96)	.59 (.54, .65)	.75 (.70, .80)
StrategyQA	.57 (.51, .62)	.62 (.57, .68)	.73 (.68, .79)	.85 (.81, .89)	.63 (.57, .68)	.63 (.58, .69)
WorldTree v2	.75 (.71, .80)	.83 (.79, .87)	.92 (.89, .95)	.98 (.96, .99)	.88 (.85, .92)	.88 (.85, .92)

表 10：各数据集上模型的准确率（按提示取平均值）。

数据集 提示	CommonsenseQA	MedQA	MedMCQA	OpenBookQA	StrategyQA	WorldTree v2
直接	.68 (.60, .76)	.21 (.12, .30)	.28 (.18, .37)	.65 (.56, .73)	.24 (.10, .38)	.84 (.77, .90)
Kojima	.69 (.61, .77)	.22 (.14, .31)	.25 (.16, .35)	.61 (.52, .70)	.46 (.32, .59)	.79 (.72, .86)
Zhou	.72 (.64, .79)	.23 (.14, .32)	.37 (.27, .46)	.74 (.66, .81)	.32 (.19, .44)	.83 (.77, .89)
Plan	.73 (.65, .80)	.19 (.11, .28)	.30 (.21, .40)	.65 (.56, .73)	.27 (.12, .42)	.82 (.75, .88)
清晰表达	.72 (.64, .80)	.22 (.14, .31)	.35 (.25, .45)	.67 (.59, .75)	.27 (.13, .40)	.88 (.83, .93)
改述	.75 (.68, .82)	.21 (.13, .29)	.31 (.22, .41)	.61 (.51, .70)	.42 (.30, .55)	.87 (.82, .92)
详细阐述	.68（.60、.76）	.25（.17、.34）	.36（.25、.45）	.64（.56、.72）	.33（.20、.47）	.82（.75、.88）
对话	.63 (.55, .72)	.20 (.12, .29)	.32 (.23, .41)	.63 (.55, .72)	.30 (.16, .43)	.78 (.71, .85)
自我批评	.73 (.65, .80)	.19 (.11, .27)	.25 (.16, .34)	.66 (.56, .74)	.23 (.09, .37)	.82 (.75, .88)
Zhou指令	.73 (.66, .81)	.19 (.11, .28)	.26 (.17, .36)	.65 (.57, .74)	.28 (.14, .42)	.86 (.80, .92)

表 11：各数据集上提示的 Krippendorff α系数（α）表现，按模型取平均。

数据集 提示	CommonsenseQA	MedQA	MedMCQA	OpenBookQA	StrategyQA	WorldTree v2
直接	.74 (.68, .81)	.38 (.31, .45)	.46 (.39, .53)	.74 (.68, .80)	.64 (.57, .71)	.88 (.84, .93)
Kojima	.75 (.69, .81)	.39 (.32, .46)	.44（.37、.51）	.71 (.65, .78)	.73 (.67, .80)	.85 (.79, .90)
Zhou	.78 (.72, .84)	.40 (.33, .47)	.54 (.47, .61)	.81 (.75, .87)	.67（.60、.74）	.87 (.83, .92)
Plan	.78 (.73, .84)	.36 (.30, .43)	.48 (.41, .56)	.74 (.68, .80)	.66 (.59, .73)	.87 (.82, .91)
清晰表达	.78 (.72, .84)	.39 (.32, .46)	.52 (.44, .59)	.76 (.70, .82)	.66 (.59, .73)	.91（.87, .95）
改述	.80 (.75, .86)	.38 (.31, .45)	.49 (.42, .56)	.71 (.64, .77)	.72 (.65, .78)	.91（.87, .95）
详细阐述	.75 (.68, .81)	.41 (.34, .48)	.53 (.46, .60)	.74 (.67, .80)	.68 (.61, .75)	.87 (.82, .92)
对话	.71 (.64, .78)	.37 (.30, .44)	.51 (.43, .58)	.73 (.67, .79)	.66 (.59, .73)	.84 (.79, .89)
自我批评	.79 (.73, .84)	.37 (.30, .43)	.44（.37、.51）	.75 (.69, .81)	.62 (.55, .69)	.87 (.82, .92)
Zhou指令	.79 (.73, .85)	.37 (.30, .44)	.45 (.38, .52)	.74 (.68, .81)	.66 (.59, .73)	.90 (.86, .94)

表12：各数据集的提示准确率在不同模型上的平均值。

模型 提示	Command-XL	Flan-T5-XXL	GPT-3.5-turbo	GPT-4	Davinci-002	Davinci-003
直接	.26 (.18, .33)	.49 (.41, .58)	.61 (.53, .69)	.71 (.64、.79)	.41 (.31、.50)	.44 (.35、.53)
Kojima	.25 (.16、.34)	.46 (.38、.55)	.66 (.57, .75)	.80 (.73, .87)	.42 (.33, .51)	.45 (.36, .54)
Zhou	.35 (.27, .43)	.44 (.37, .51)	.62 (.53, .71)	.83 (.77, .90)	.53 (.45, .62)	.50 (.41, .59)
Plan	.34 (.25, .42)	.45 (.37, .53)	.61 (.52, .70)	.77 (.71, .84)	.37 (.30, .45)	.46 (.37, .55)
清晰表达	.33 (.26, .40)	.50 (.42, .58)	.59 (.49, .68)	.79 (.71, .86)	.44 (.35, .53)	.52 (.43, .60)
改述	.42 (.33, .51)	.46 (.38, .54)	.61 (.52, .70)	.78 (.71, .85)	.44 (.35, .53)	.46 (.37, .55)
详细阐述	.34 (.26, .42)	.42 (.33, .51)	.61 (.51, .70)	.77 (.70, .84)	.51 (.42, .60)	.43 (.35, .51)
对话	.31 (.22, .40)	.44 (.35, .52)	.58 (.49, .67)	.74 (.66, .81)	.35 (.26, .43)	.46 (.38, .54)
自我批评	.32 (.26, .39)	.41 (.35, .47)	.58 (.49, .68)	.76 (.69, .84)	.38 (.30, .47)	.48 (.39、.57)
Zhou指令	.38 (.30, .46)	.43 (.35, .51)	.64 (.54, .73)	.79 (.72, .86)	.33 (.26, .40)	.49 (.41, .58)

表 13：按模型计算、跨数据集平均的提示表现 Krippendorff α系数（α）。

模型 提示	Command-XL	Flan-T5-XXL	GPT-3.5-turbo	GPT-4	Davinci-002	Davinci-003
直接	.47 (.40, .54)	.62 (.55, .69)	.75 (.69, .81)	.81 (.76, .87)	.59 (.52, .66)	.61 (.54, .68)
Kojima	.45 (.38, .52)	.62 (.55, .69)	.76 (.70, .82)	.86 (.81, .91)	.56 (.49, .63)	.61 (.54, .68)
Zhou	.55 (.48, .62)	.58 (.51, .65)	.75 (.69, .81)	.89 (.84, .93)	.65 (.58, .71)	.66 (.59, .73)
Plan	.53 (.46, .60)	.62 (.55, .69)	.73 (.66, .79)	.84 (.79, .89)	.55 (.48, .62)	.63 (.56, .70)
清晰表达	.53 (.46, .60)	.65 (.58, .71)	.74 (.67, .80)	.86 (.81, .91)	.61 (.54, .69)	.65 (.58, .72)
改述	.58 (.51, .65)	.62 (.55, .69)	.73 (.66, .79)	.84 (.79, .89)	.61 (.54, .68)	.63 (.56, .70)
详细阐述	.55 (.48, .62)	.59 (.52, .66)	.73 (.66, .79)	.84 (.79, .89)	.65 (.58, .72)	.60 (.53, .67)
对话	.52 (.45, .59)	.61 (.55, .68)	.71 (.65, .78)	.81 (.76, .87)	.53 (.46, .61)	.62 (.55, .69)
自我批评	.51 (.43, .58)	.58 (.51, .65)	.72 (.65, .78)	.84 (.79, .89)	.57 (.50, .64)	.64 (.57, .71)
Zhou指令	.56 (.49, .63)	.59 (.52, .66)	.75 (.69, .82)	.85 (.80, .90)	.54 (.47, .61)	.65 (.58, .71)

表14：各模型的提示准确率在不同数据集上的平均值。