Evaluating the Zero-shot Robustness of Instruction-tuned Language Models

Jiuding Sun
Khoury College of Computer Sciences
Northeastern University
sun.jiu@northeastern.edu

Chantal Shaib
Khoury College of Computer Sciences
Northeastern University
shaib.c@northeastern.edu

Byron C. Wallace
Khoury College of Computer Sciences
Northeastern University
b.wallace@northeastern.edu

Abstract

*警告：该PDF由GPT-Academic开源项目调用大语言模型+Latex翻译插件一键生成，版权归原文作者所有。翻译内容可靠性无保障，请仔细鉴别并以原文为准。项目Github地址 https://github.com/binary-husky/gpt_academic/。当前大语言模型: gpt-4，当前语言模型温度设定: 1。为了防止大语言模型的意外谬误产生扩散影响，禁止移除或修改此警告。

指令微调近期已经作为一种改善大型语言模型（LLMs）在新任务的零样本能力的有效手段崭露头角。此技术在提高适度大小的LLMs的性能方面显示出特别的优势，有时候甚至可以产生与更大型的模型变体相竞争的性能。在本篇论文中，我们提出两个问题：（1）微调模型对于指令表述的特定词句有多敏感？和（2）我们如何使他们对于这样的自然语言变化有更强的鲁棒性？为了回答前者，我们收集了319份由NLP从业者手动编写用于广泛使用的基准测试中的超过80个独特任务的指令，我们评估这些指令与在指令微调过程中观测到的指令表述的平均性能和差异性。我们发现使用新颖的（未被观察到的）但恰当的指令表述会持续降低模型的性能，有时候甚至会大幅度降低。此外，尽管它们在语义上等价，这些自然的指令产生了广泛的下游性能差异。换句话说，经过指令微调的模型对指令的重新表述并不特别鲁棒。我们提出一种简单的方法来减轻这个问题，通过引入“软提示”嵌入参数，并优化这些参数，最大化语义等效指令的表示之间的相似性。我们证明了这种方法可以持续改善经过指令微调的模型的鲁棒性。[1]

1 Introduction

大型语言模型（LLMs）在某种程度上由于它们可以通过提示 [3; 4; 10; 37]实现对新任务的零射击和少射击适应，从而主导了自然语言处理（NLP）领域。最近的工作已经展示了利用自

[1]代码和指令已在以下地址公开:https://github.com/jiudingsun01/InstructionEval

Preprint. Under review.

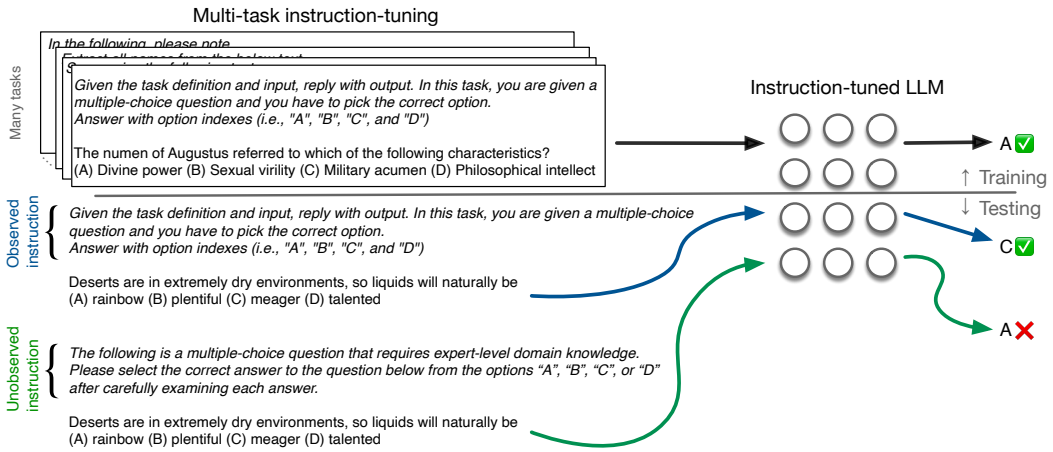


图 1: 模型在指导调整数据集上训练有多好地泛化到新的指令（在训练中未观察到）呢？我们的分析表明，它们并没有做得非常好。上面我们展示了一个案例，其中将一个示例与观察到的指令配对会产生正确的输出，而提供一个不同但语义等价的指令会产生一个错误的响应。我们提出并评估了一种简单的方法来改善这种情况。

然语言指令对这些模型进行微调的前景。这种指令调优改进了LLM在零射击和少射击设置下的性能，有时会有显著的提升，尤其是对于"中等大小"的模型 [5; 22]。例如，在某些基准测试中，指令调优的Flan-T5-XL (3B参数) [5]胜过了GPT-3 (175B)，尽管它的体积明显更小。此外，LLaMa-7B [27] — 在Alpaca [26]指令集的大规模语料库上进行细微调整后 — 在一系列NLP基准测试中均优于GPT-3。

这些经验性的成功促使了人们进行指令增强任务收集的努力，用于元学习 [31; 33; 33]，以及改进指令调优的研究 [17; 34; 24]。在这项工作中，我们调查了指令调优模型的鲁棒性。更具体地说，我们的问题是：在测试时，指令调优的语言模型对指令措词的变动有多敏感？这点尤其重要，因为指令调优的主要动机是通过自然语言指令来促进零射击适应：如果模型对任务指令的特定措词过于敏感，那么它们在实践中的效用可能会大大受限。

在2部分详细审查的早期工作已经证明，LLM似乎并不直观地“理解”提示[32; 12; 38]，但这些努力并未特别考虑经过指令调优的模型。我们的最新研究[8]调查了指令调优模型的鲁棒性，发现在少射击设置中，指令调优的T5 [23]对指令扰动具有鲁棒性，但在零射击应用中鲁棒性较差。我们提供了对这种现象进行更深入分析的贡献，涵盖了大量的指令调优模型和基准测试。我们还介绍并评估了一种改进这些模型鲁棒性的方法，得到了有希望的结果。

更具体地说，我们收集了一组由NLP研究人员手动编写的任务指令；这些都是有效的指令，但与Flan集合中的指令不同。然后，我们评估了在给定这些新指令的情况下，对Flan集合指令集进行细微调整的LLMs在两个基准：MMLU [9] 和 BBL [25]中的表现。我们发现，在零射击应用中使用新的指令会大大降低精度（图 1 描述了这一点）。例如，比较Flan-T5 XXL在使用（a）训练中看到的指令和（b）在训练中未观察到但在语义上等效的指令时，我们在大型基准测试中观察到平均6.9个百分点的绝对性能下降。

我们的**主要贡献**如下所示：(1)我们对三个这样的模型（Flan-T5 [33]， Alpaca [26]，和 T0 [24]）的"家族"中的指令调谐LLMs实施了全面且深入的鲁棒性分析，使用大型基准测试[9; 25]。我们收集了由NLP研究人员手动编写的大量新任务指令；我们将发布这个数据集以便于进行更多关于指令鲁棒性的工作。在使用"新的"（在训练中未见过的）指令时，我们观察

到了大量的性能下降。 (2) 我们提出了一种简单的方法来通过强制LLMs对语义上等效的指令产生相似的表示来提高鲁棒性。我们发现这种方法在使用新的，但合适的任务指令时始终能够提高实现的性能。

2 Related Work

多任务学习与指令调优 对于至少半个十年来的自然语言处理(NLP)领域，能回应任意查询的单一文本与文本模型的训练一直是一个抱负。例如，在现代提示和指令策略之前，有努力统一不同的任务，通过将这些任务重新定位为一般问题解答 [18; 14; 13]的实例。最近的努力更侧重于编译和微调在自然语言指令下的大语言模型(LLMs)的各个任务[33; 20; 24]，我们将这种策略称为指令调优。这种策略的一个例子是`Super-NaturalInstructions` [31], 它汇总了超过1600个任务，并丰富了这些任务的指令和负面示例。类似地，最近发布的 OPT-IML Bench [11] 汇集了2000个 NLP 任务。 Flan 2022 任务集合[17] 还额外包含了 *Chain-of-Thought* (CoT) 风格的 "推理" 链条在指令模板中，作者们表明，在指令微调过程中包含这些（以及零样本示例和 "输入反转"）可以在保留任务上取得改进。

这些元资源—指令、任务和样本的集合—有助于训练经过指令调优的模型家族，如 Flan-T5、Flan-PaLM [5] 和 OPT-IML [11]。[2] 结果令人鼓舞；对 LLMs 进行指令微调可以提供明确和持续的收益，并且，也许最令人兴奋的是，使得相对 "小" (∼10B) 的 LLMs 能够达到与巨型 (∼175B) 模型相当的近似最先进的性能 [26]。这激发了人们对描述指令是如何帮助模型的兴趣，并且开发了用来进一步改善指令调优的技术；我们在下面回顾了与这两个研究主题相关的近期努力。

评估提示与指令能力 指令可以被视为一种特殊类型的模型提示，最近有一些努力对其进行了批判性的评估。例如，Webson 和 Pavlick 提出模型是否有意义地 "理解" 提示的问题 [32]，他们发现大多数情况下并非如此：当提供无关的和误导性的提示时，性能通常并未受到影响。在跟进的工作中，Jang 等人 [12]评估了在否定式提示下的性能，发现其中有一个 "反比例" 现象，即在这种情况下，较大的模型的性能反而更差。

其他研究试图描绘出在上下文学习（ICL）—即在提示中包括一些例子—是如何以及何时工作的 [19; 29; 6; 1; 36]。 ICL 是一种与当前努力正交的提示形式，因为我们主要关注经过指令调优的 LLMs 的零样本适应性。

在我们对同时进行的研究中，Gu 等人 [8] 研究了指令调优模型对于指令扰动（如，删除词语）和释义的稳健性。他们发现，当给定例子时（即，在少样本环境下），模型相对较为稳健，但在零样本使用时相当敏感；这与我们的发现在定性上是一致的。我们的工作与这项同期研究有很大的不同：(1) 我们提供了对稳健性的更全面评估；Gu 等人只考虑了仅针对单个指令数据集进行 T5 指令调优，而我们评估了三种 LLMs（以及各种大小）使用的五种指令调优数据集，并且我们在所有的 80 多个测试任务中进行了评估（Gu 等人只考虑了12个）。(2) 我们提出并评估了一种新的方法来提高指令调优模型的稳健性；Gu 等人并未提供任何改善稳健性的机制。

改进指令调整 过去的工作也寻求通过各种方式改进指令调整。做到这一点的一种方法是根据人类反馈进行指令调整[22; 7; 2; 21; 39]。这倾向于改进开放式模型的回应，但会降低

[2]有点令人困惑的是,在 FLAN 和 OPT 的情况下,语料库（即任务和指令组成的基准）和使用它们进行微调的 LLMs 都用相关的首字母缩写词作为前缀来表示:例如,Flan-T5 指的是一个用 Flan 集合进行微调的 T5 [23] 变体。

对下游任务的性能。另一种策略是利用现有资源大规模自动生成指令调整数据集。例如，Wang *et al.* [30] 使用LLMs生成指令、输入和输出，然后利用这些来提高他们自己的指令遵循能力。同样地，Zhou和他的同事们 [40] 提出使用LLMs来设计提示。最后，Ye *et al.* [35] 提出"翻转"标准任务，让LLMs在给定输入和标签的情况下生成指令。

3 Instruction Datasets

3.1 Evaluation Benchmarks

我们在两个大型测试基准上评估了一组指令调优模型：MMLU [9]和BIG-BENCH [25]。MMLU是一个多项选择题答题测试基准，包含57个需要专家知识的任务。BIG-BENCH是一个由协作构建的测试基准，包含来自各个领域的204个多样化任务；在此，我们考虑BIG-BENCH LITE子集，并且我们只包括QA、多类别和二元分类任务，总共产生了18个任务。

3.2 Collecting New Instructions from NLP Researchers

我们的目标是评估当提供的指示在语义上等同于，但在表面上不同于训练时所使用的指示时，对指示调整模型的效果进行评估。为此，我们招募了自然语言处理（NLP）研究人员（研究生）来为所考虑的任务编写新颖的指示；因此，这些特定的指示措词在指示微调过程中是未被观察到的。

更具体地说，我们招募了36位在自然语言处理领域工作的研究生。所有人都至少有一些与指示调整模型以及评估基准中包含的下游任务的经验。对于BBL中的18项任务和MMLU中的所有任务，我们要求12位研究生写出他们在使用指示调整模型进行零次推断时会使用的一个（不同的）指示。我们在附录A中提供了关于这个指示收集过程的详细信息。我们将发布所有为这项工作获取的319条指示，以确保这项工作的可重复性，并促进对指示调整模型鲁棒性的进一步研究。

4 Evaluating the Robustness of Instruction-tuned LLMs

4.1 Models and Data

我们针对三个指令集合进行实验，这三个集合分别训练了模型的各种变体（它们提供了观察到的任务指令）：P3 [24]、Flan-2022 [5]和Alpaca [26]。为方便我们的分析，我们手动识别出所有对应于 (a) 多选问题回答（QA），(b) 二元分类（BC），或要求"是"或"否"响应的任务，以及 (c) 多类别分类（MC）的指令，这需要将输入分为一组有限的类别。

为了评估模型对于指令措辞的鲁棒性，我们使用了两个基准测试：MMLU [9] 和 BIG-BENCH LITE (BBL) [25]，同时还使用了在第 3.2 剖述的新指令集。我们包括了所有的57个来自MMLU的任务，和24个来自BBL的任务中的14个。在后者中，我们排除了两个依赖生成度量的任务，四个使用精确匹配的任务，以及四个包含T5和/或LLaMa标记化器无法识别的代币的任务（例如，一项任务中的输入是表情符号）。

我们对同一类别的所有任务使用相同的指示，这些指示取自与每个模型相关的已发布的指示调优数据集。这些指示是通用的，例如，在分类的情况下，他们要求模型考虑一个与分类标准和由实例提供的标签空间相对应的示例，并选择一个合适的类别（参见表格 1 的例子）。只要这些指示符合任务类型，就可以"混合和匹配"。

QA	In this task, you are given a multiple-choice question and you have to pick the correct option. Answer with option indexes (i.e., "A", "B", "C", and "D"). Q: {question} A. {choiceA} B. {choiceB} C. {choiceC} D. {choiceD}
MC	Pick one category for the following text. The options are - {options} {text}
BC	{paragraph} Choose your answer: According to the above paragraph, the question "{question}" is "{response}"?

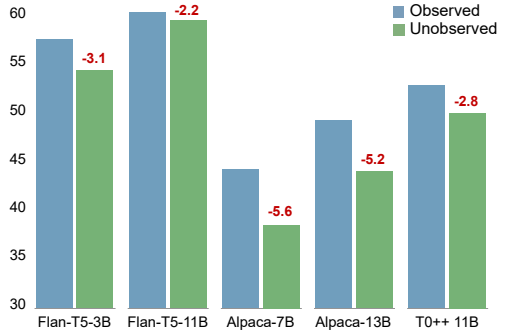
表 1: 我们收集了三种常见任务类型的观察指令示例。

Observed Instructions				Unobserved Instructions		
Instruction Type	QA	MC	BC	Number of tasks	1	14
Flan	50	35	18	Instructions per task	20	10
Alpaca	20	20	11	Total instructions	20	140
P3	13	8	7			

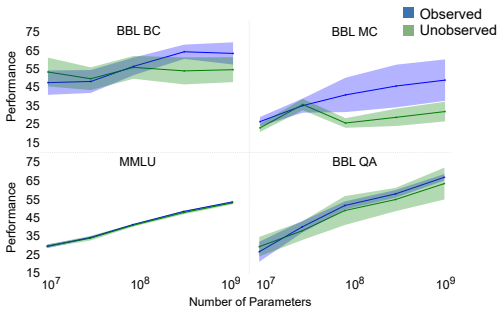
表 2: 我们用于评估的指令措辞（未观察到和已观察到）的计数。

4.2 Results

我们在图2和表3中呈现了主要的聚合分析结果。这里的关键信息是，使用在训练中未被观察到的指令——但是手动为手头的任务编写的，因此在语义上是恰当的——会导致性能的显著下降：平均来说，未观察到的指令会使模型的准确性降低五个百分点以上。表3报告的结果是按任务类型进行分解的；我们观察到，使用新颖指令对分类任务的影响最大。我们在附录中提供了更详细（数据集级别）的结果。



(a) 使用观察到的和未观察到的指示时，所有任务的平均零射击性能。



(b) 使用观寔和未观察指令作为模型大小函数的Flan-T5性能表现。

图 2: 在测试时使用新颖但有效的指令（在训练中未观察到的措辞）会持续地降低指令调整后的LLMs的性能（a）。规模并不一定能解决这个问题（b）。

4.3 A Closer Look at Instruction Robustness

在上文中，我们使用了一些通用的指令来让模型执行任务（表1）。这里我们更深入地研究了在使用新颖指令时观察到的性能下降。我们报告了一个令人好奇的结果，突显模型依赖于之前观测到的指令的程度：不正确但被观测到的指令的表现优于适当但未被观测到的指令（图 3）。

Model	MMLU		BBL-QA		BBL-BC		BBL-MC		**Overall**	
	Avg.	Std.	Avg.	Std.	Avg.	Std.	Avg.	Std.	Avg.	Std.
Flan-T5-3B										
OBSERVED	**48.1**	(±0.3)	**59.0**	(±2.1)	**66.5**	(±3.8)	**55.6**	(±0.7)	**57.3**	(±1.7)
UNOBSERVED	47.5	(±0.9)	56.0	(±7.3)	61.1	(±6.9)	52.1	(±5.4)	54.2	(±5.1)
Performance Δ	↓ 0.6		↓ 3.0		↓ 5.5		↓ 3.5		↓ 3.1	
Alpaca-7B										
OBSERVED	**41.9**	(±0.6)	**48.6**	(±2.8)	**53.8**	(±3.4)	**32.1**	(±2.2)	**44.1**	(±2.3)
UNOBSERVED	39.7	(±2.2)	45.3	(±6.5)	52.4	(±6.5)	16.4	(±3.5)	38.5	(±4.7)
Performance Δ	↓ 2.2		↓ 3.3		↓ 1.4		↓ 15.7		↓ 5.6	
T0++ 11B										
OBSERVED	48.3	(±0.9)	54.1	(±4.1)	**66.1**	(±2.1)	**42.0**	(±2.1)	**52.6**	(±2.3)
UNOBSERVED	**48.5**	(±0.9)	**54.7**	(±3.7)	54.7	(±4.3)	41.4	(±2.4)	49.8	(±2.8)
Performance Δ	↑ 0.2		↑ 0.7		↓ 11.4		↓ 0.6		↓ 2.8	
Flan-T5-11B										
OBSERVED	**53.2**	(±0.2)	**67.9**	(±1.8)	**65.6**	(±6.0)	**58.7**	(±0.5)	**61.4**	(±2.1)
UNOBSERVED	52.7	(±0.8)	64.6	(±8.5)	63.6	(±6.1)	55.9	(±5.5)	59.2	(±5.2)
Performance Δ	↓ 0.5		↓ 3.4		↓ 2.0		↓ 2.8		↓ 2.2	
Alpaca-13B										
OBSERVED	**47.8**	(±0.5)	**53.9**	(±2.2)	**57.9**	(±4.8)	**36.7**	(±1.8)	**49.1**	(±2.3)
UNOBSERVED	47.0	(±0.8)	51.7	(±5.7)	54.1	(±5.6)	22.7	(±7.5)	43.9	(±14.0)
Performance Δ	↓ 0.9		↓ 2.2		↓ 3.8		↓ 14.0		↓ 5.2	

表 3: 在基准任务中使用观察和未观察的指令的结果（按类型分组）。当使用(UNOBSERVED)指令时，性能会降低——有时降低10+点——这表明经过指令调整的模型的鲁棒性不强。BC、MC和QA分别代表二进制分类、多类分类和问答。

我们通过评估 Flan-T5-XXL（11B）在使用六种指令类型处理来自 BIG-BENCH 的七个数据集的性能，来得出这个观察结果。特别的，这包括在训练中观察到的两个指令的（变体）：**最接近** 是最相似任务中指令调优集的指令；**不正确** 是一个完全不同且不适用的任务的观察指令（但有相同的期望输出格式，例如，分类）——直观地说，这些不应产生期望的行为；**否定**和**最接近**一样，但我们否定指令以表示它不应执行任务。

对于未观察到的指令，我们考虑：**任务设计者**，也就是在BIG-BENCH中任务的作者提供的指令（任务前缀）；以及**新收集**，或者是从NLP研究生那里收集的新的指令，如上所述。作为对照参考，我们也考虑了**无意义**，它是完全与任何任务无关的随机“指令”。

图 3 报告了这些变体的平均结果。与我们的发现一致，使用在训练中未观察到的指导降低了性能。引人注目的是，我们这里也发现，使用一个不适当但观察到的指导比使用适当但未观察到的指导表现更好。这指示了指导调整的模型——或至少我们在这里评估的适度大小的这些模型——可能在某种程度上过度依赖在训练中观察到的指导，并且并未像我们希望的那样推广到新的指导和措辞。我们在附录中提供了所有的指导和结果。

4.4 Scaling

指令健壮性是否随着规模的扩大而开始显现？为了尝试回答这个问题，我们用Flan-T5模型（参数范围从小（80M）到XXL（11B）），重复了表3中的所有实验。我们在图2b中观察到，

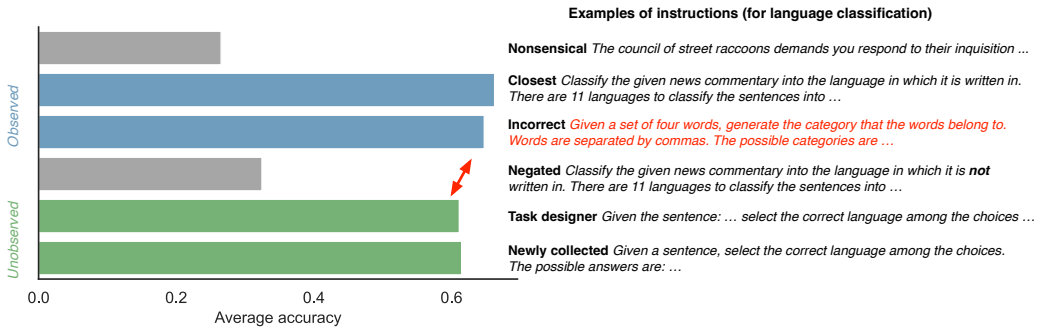


图 3: **不正确但被观察到的指令的平均表现优于正确但未被观察到的指令**。我们报告了基准测试的平均结果，但在右侧为一个具体的，说明性的任务展示了示例指令。我们在附录中提供了所有指令。

与未观察的指令相比，观察到的指令所得结果的差异**并未**随着模型规模的增大而减少，至少到目前为止是这样。也就是说，大规模模型（175B+）可能会提供更大的健壮性。然而，我们要重申的是，关于指令调整的许多兴奋之处在于，这种技术似乎允许规模较小的模型达到与大规模模型竞争的结果。

4.5 Robustness with Semantic Distance

在4.2中的一个观察是，使用未观察到的指令对MMLU的性能影响较小。 MMLU是一个包含57个关于不同知识领域的 QA 任务的基准测试，这些任务都有着相似的输入-输出（问题，四个选项→ 答案）形式。在收集指令过程中，我们将MMLU中的所有任务都视为一般的 QA 任务，并要求 NLP 研究人员编写一般的 QA 指令。因此，我们假设这些指令与观察到的指令相比较为类似，这也解释了在这种情况下的相对鲁棒性。

我们在图4和表4中经验性地验证了这一点。对于每个实例（指令加示例），我们提取了第一个解码标记在倒数第二层的表示。我们使用 tSNE [28]来可视化在MMLU和BBL的实例中观察到和未观察到的指令的这些表示。图4显示，在MMLU的情况下，我们收集到的未观察到的指令非常类似于观察到的，而在BBL中，未观察到的指令和观察到的指令之间有更大的差距。我们还在表4中提供了这一现象的数值测量。我们报告了未观察到的指令表示与最近观察到的对应表示之间的平均 $\ell 2$距离。我们发现MMLU的未观察到的指令平均而言更接近最近的观察到的指令，这与观察到的性能下降较低有关。这些发现与假设一致，即MMLU的未观察到的指令与此数据集的观察到的指令更为相似，这很可能解释了在这种情况下明显的鲁棒性。

我们将平均性能降级（以%为单位）作为平均相似度的函数绘制出来，平均相似度是指第一次解码的标记（根据未观察到的指令）的相似度和最相似的已观察到指令的相似度。负斜率暗示了直观的关系：在模型表示方面不相似的指令往往导致性能较差。然而，这种关系相对较弱，得到的截距估计是-0.8，斜率是-0.2（p =0.08）。

4.6 Robustness Under In-Context Learning (ICL)

以前的研究 [8] 已经表明，当上下文中提供了少量样本时，LLMs 对提示/指令的变化不太敏感。尽管我们主要关注零样本能力，但为了完整性，我们在少样本设置中重新进行了所有实验。我们在 C 中报告了这些结果。主要的发现是，尽管仍存在一些差异，但总的来说 ICL

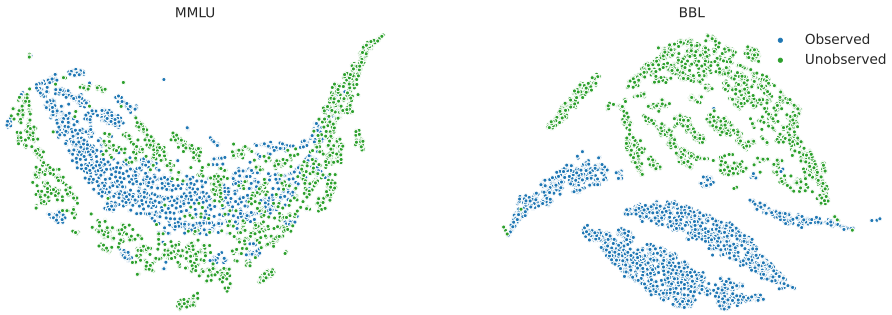


图 4: 对来自MMLU和BBL中通过Flan-T5 (XXL)解码的首个token的300个随机样本进行tSNE图的表示。观察到的和未观察到的MMLU指令的嵌入相似，而BBL的则大不相同。这个结果在大多数但并非所有考虑的模型中都成立：请参见D获得所有模型的可视化。

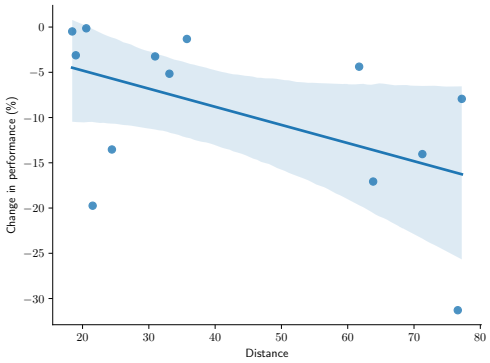


图 5: 使用未观察指令时，平均性能降级与语义距离的关系图。

Dataset	Avg. $\Delta\ell 2$	Avg. Δ Acc.
MMLU	**19.8**	**-0.5**
BBL-QA	37.9	-3.4
BBL-BC	25.3	-2.0
BBL-MC	26.1	-2.8

表 4: 四个类别的平均性能下降。可以看出，MMLU 的平均距离最小，这表明其分布偏移较小，因此导致最小的性能下降。

稍微减少了模型对未观察到的指令使用的敏感性。这是直观的，因为样本本身很可能暗示了期望的任务，并可能影响分布。

5 Aligning Equivalent Instructions

我们现在引入一种简单、轻量级且有效的方法来提高指令调优LLMs的稳健性。直观的想法是在目标中引入一个项，明确鼓励模型对于由不同但语义相等的指令提供的同一输入做出类似的预测（因此生成类似的表示）。

更具体地说，我们的目标是在模型诱导的空间中对齐语义等价的指示。为了此目的，我们引入了具有维度$\mathbb{R}^{d\times n}$的软嵌入参数；这等同于添加n个新的令牌（具有嵌入维度d）作为输入的前缀（先于指令）。直观的想法是推动语义等价任务的表示接近。为此，我们在损失中添加额外的项：给定任务的参考指令和复述（语义相等）版本之间的输出概率的KL散度$\mathcal{L}_{\text{KL}}$。我们将这个和标准交叉熵损失结合起来，并在这个目标下微调*only*引入的软提示参数（图7）。这里λ是一个损失加权超参数，$\hat{y}_i^{(j)}$ 和 $\hat{y}_r^{(j)}$ 是由模型诱导出的具有复述指令i和参考指令r在令牌位置j的词汇表$\mathcal{V}$上的分布。[3]

[3]我们填充实例使得在给定批次中的长度有效地相等；因此,总和是从1到与当前批次相关的长度,为了简化,我们省略这一点。

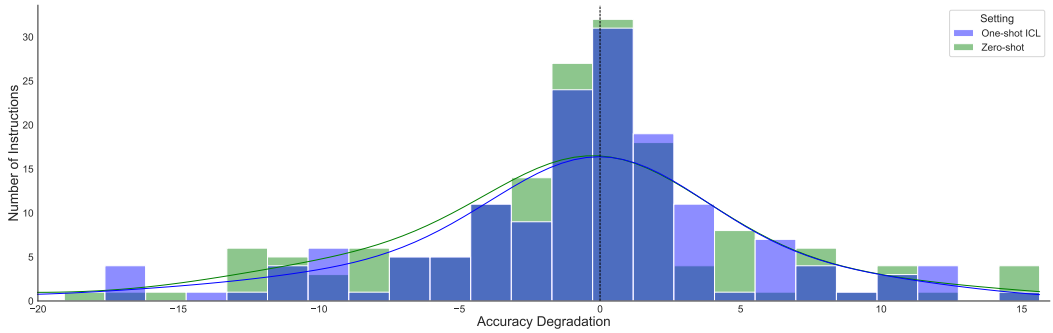
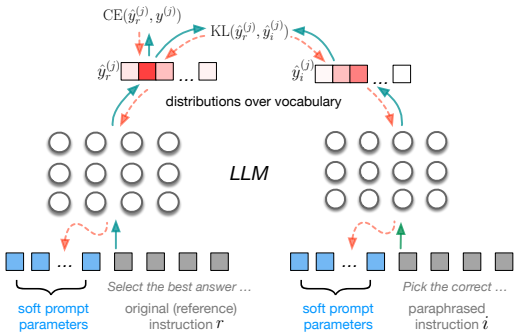


图 6: 在使用BBL和MMLU中的未观测指令时，Flan-T5-XXL的性能降低。我们绘制了所有未观测指令与观测指令的平均准确度对比的准确度降低情况。可以看出，在一次性情境下的学习中，模型稍微更加稳健，因为性能差异趋近于0.



$$\mathcal{L} = (1-\lambda)\mathcal{L}_{\text{CE}} + \lambda\mathcal{L}_{\text{KL}}$$

$$\mathcal{L}_{\text{KL}} = \frac{1}{N-1}\sum_{i\neq r}^{N}\sum_{j}\text{KL}(\hat{y}_i^{(j)}||\hat{y}_r^{(j)})$$

$$\hat{y}_i^{(j)} = \text{Softmax}(p_i^{(j)}),\ p_i^{(j)} \in \mathbb{R}^{|\mathcal{V}|}$$

图 7: 所提出的指令对齐方法的示意图（左）和相关的损失项（右）。虚线（红色）指示反向传播；我们只更新软提示参数，我们展示了这带来的性能优于微调所有模型参数。

为了优化上述目标，我们需要为训练数据中的每个任务生成经过释义的指令i，我们按照以下方式自动生成这些指令。对于指令调优数据集，我们从训练数据中抽取一小部分用于对齐。我们使用GPT-4对这些参考指令进行改述。对于Alpaca集合，我们随机抽取了1000个任务，并通过三个提示对其进行了改述，然后在温度0.5下收集了排名前三的候选项。对于Flan集合，我们从混合中随机抽取了986个实例，并用贪婪解码器对其进行了3个提示的改写。

针对微调，我们然后为每个示例创建实例，将它们与相应任务的所有不同指令配对。我们通过包含一个带有原始指令的实例，其余实例均采用改写指令来形成批处理。对于前缀的实现，我们遵循[16]的设置，即冻结模型参数并仅训练前缀嵌入与MLP层。

6 Results

我们使用两种代表性的指令调整LLM进行了实验：Flan-XL (3B) 和 Alpaca (7B)。我们比较了用常规方式训练的这些模型的规范版本(与表 3 中评估的相同)以及使用我们提出的方法进行微调的变体。我们消减我们方法的组成部分，以揭示数据和目标的贡献。

具体来说，我们考虑了以下几种变体：在额外的，自动产生的指令释义上微调所有的模型参数(FT)；施加新的KL损失项（再次微调所有的模型参数； FT+KL)；引入额外的柔性提示

参数并在释义实例上进行微调，但没有KL (PT)；然后是完全提出的策略，它引入了柔性提示参数，并优化了他们增加了KL项的损失(**PT+KL**)。

	MMLU			BBL		
Model	OBS.	UNOBS.	Avg.	OBS.	UNOBS.	Avg.
FLAN-T5-3B	48.1	47.5	47.8	**56.1**	51.9	54.0
FT	39.4 **(-8.7)**	40.1 **(-7.4)**	39.8 **(-8.0)**	48.2 **(-7.9)**	42.3 **(-9.2)**	45.3 **(-8.7)**
FT+KL	41.8 **(-6.3)**	43.6 **(-3.9)**	45.9 **(-1.9)**	47.7 **(-8.4)**	43.1 **(-8.8)**	45.4 **(-8.6)**
PT	48.1 **(+0.0)**	47.6 **(+0.1)**	47.9 **(+0.1)**	55.9 **(-0.2)**	52.1 **(+0.2)**	54.0 **(+0.0)**
PT+KL	**48.1** **(+0.1)**	**47.9** **(+0.4)**	**48.0** **(+0.2)**	55.9 **(-0.2)**	**53.7** **(+1.8)**	**54.8** **(+0.8)**
ALPACA-7B	41.9	39.7	40.8	47.6	42.9	45.3
FT	40.3 **(-1.6)**	39.1 **(-0.6)**	39.7 **(-1.1)**	44.4 **(-3.2)**	42.1 **(-0.8)**	43.4 **(-2.0)**
FT+KL	39.7 **(-2.2)**	40.2 **(+0.5)**	40.0 **(-0.8)**	45.6 **(-2.0)**	42.8 **(-0.1)**	44.2 **(-1.1)**
PT	42.1 **(+0.2)**	40.0 **(+0.3)**	41.1 **(+0.3)**	47.5 **(-0.1)**	43.0 **(+0.1)**	45.3 **(+0.0)**
PT+KL	**42.4** **(+0.5)**	**41.8** **(+2.1)**	**42.1** **(+1.3)**	**47.9** **(+0.3)**	**46.6** **(+3.7)**	**47.3** **(+2.0)**

表 5: 提出的软提示对齐方法的结果和消融。所有消融版本都使用自动改写的说明的增强集。FT参考的是在这个额外的数据上进行简单的微调（使用强制教学）；PT表示前缀调整（即，引入软提示参数）；KL参考的是我们上面提出的对齐目标。将所有这些组件同时使用可以得到最佳性能，特别是在未观察到的指令上。

我们在表 5 中报告结果。两个观察结果：（1）提出的软提示对齐策略（**PT+KL**）在考虑的任务和模型中带来了一致的改进，特别是在未观察到的指令上，如预期的那样，性能有所提高。（2）只有当所有组件—— 额外的自动化释义训练指令，软提示参数，以及额外的 KL 损失项—— 都到位时，才能实现该方法的全部效益。

Dataset	**Closest Distance Before**	**Closest Distance After**	**Δ Acc. Improvement** (%)
MMLU	22.2	21.3	+ 0.3%
BBL QA	22.4	23.0	+ 0.4%
BBL BC	30.1	**27.9**	**+ 4.2%**
BBL MC	26.0	24.6	+ 0.3%

表 6: Flan-T5-XL的软提示对齐前后的平均距离。

根据我们在4.5中的方法，我们取观察到的指令与未观察到的指令在对齐前后的平均距离。表6显示，我们的方法将观察到的和未观察到的指令表示法拉近了。在精度增益最大的情况下，相似性最大的增加，进一步表明了软提示对齐提供的改进机制。

7 Conclusions

指令调制的LLMs已经成为实现零次学习的有前途的方法，其使用较小的模型达到的性能有竞争力，甚至有时候比使用大得多的LLMs观察到的结果还要好 [17; 26]。在这项工作中，我们实证研究了这类模型对于指令改写的鲁棒性。具体来说，我们从36名NLP研究生中收集了75个任务的手工制作的指令，评估了不同类型的指令调整LLMs (Flan, Alpaca, 和 T0)在提供观察和未观察指令（分别在训练中看到和没有看到）时的表现。我们发现，使用后者会导致模型性能一致性下降，表明模型对于指令表述过于敏感。

然后，我们提出了一种旨在提高指令调整LLMs的鲁棒性的简单机制。这种方法包括引入一个额外的损失项，惩罚模型在使用（a）改写的指令而不是（b）同一任务的参考指令时，导致输出令牌分布的不相似度。我们发现，在此目标下进行训练一致（尽管适度）改善了结果，特别是缓解了使用以前未观察到的指令时观察到的性能降低。

8 Limitations

这项工作有重要的限制：例如，我们只对“中等大小”的模型（<20B参数）进行了评估，我们的发现是否能推广到更大的指令调整模型尚不清楚。（然而，我们注意到对于较小的模型，指令调整最有前景。）我们还将评估限制在三种任务类型中：问答，多类和二元分类。

伦理道德 这项工作并没有明确的伦理维度，但我们承认所有的LLMs都可能编码出存在问题的偏见；指令调整如何与这些偏见相互作用尚不清楚。

9 Acknowledgments

本工作得到了美国国家科学基金(NSF) 1901117号资助。

我们感谢Jay DeYoung和Alberto Mario Ceballos Arroyo对本文的建议和反馈。我们还要感谢Alberto Mario Ceballos Arroyo，Arnab Sen Sharma，Bowen Zhao，Eric Todd，Hanming Li，Hiba Ahsan，Hye Sun Yun，Shulin Cao，Jay DeYoung，Jered McInerney，Ji Qi，Jifan Yu，Jize Jiang，Kaisheng Zeng，Koyena Pal，Kundan Krishna，Linxiao Nie，Hailong Jin，Jinxin Matthew Liu，Millicent Li，Monica Munnangi，Nikhil Prakash，Pouya Pezeshpour，Sanjana Ramprasad，Sarthak Jain，Shangqing Tu，Somin Wadhwa，Tingjian Zhang，Hao Wesley Peng，Xiaozhi Wang，Xingyu Lu，Xin Lv，Zijun Yao提供手写指令。

参考文献

[1] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*, 2022.

[2] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[4] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[5] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[6] Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. Why can gpt learn in-context? language models secretly perform gradient descent as meta optimizers. *arXiv preprint arXiv:2212.10559*, 2022.

[7] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022.

[8] Jiasheng Gu, Hanzi Xu, Liangyu Nie, and Wenpeng Yin. Robustness of learning from task instructions. 2023.

[9] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[10] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[11] Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Dániel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. OPT-IML: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*, 2022.

[12] Joel Jang, Seonghyeon Ye, and Minjoon Seo. Can large language models truly understand prompts? a case study with negated prompts. In *Transfer Learning for Natural Language Processing Workshop*, pages 52–62. PMLR, 2023.

[13] Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. Unifying question answering, text classification, and regression via span extraction. *arXiv preprint arXiv:1904.09286*, 2019.

[14] Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. Unifiedqa: Crossing format boundaries with a single qa system. *arXiv preprint arXiv:2005.00700*, 2020.

[15] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.

[16] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

[17] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*, 2023.

[18] Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*, 2018.

[19] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022.

[20] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*, 2021.

[21] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

[22] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

[23] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[24] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021.

[25] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

[26] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpaca: A strong, replicable instruction-following model, 2023.

[27] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[28] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[29] Xinyi Wang, Wanrong Zhu, and William Yang Wang. Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning. *arXiv preprint arXiv:2301.11916*, 2023.

[30] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

[31] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. Benchmarking generalization via in-context instructions on 1,600+ language tasks. *arXiv preprint arXiv:2204.07705*, 2022.

[32] Albert Webson and Ellie Pavlick. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States, July 2022. Association for Computational Linguistics.

[33] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

[34] Zhiyang Xu, Ying Shen, and Lifu Huang. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. *arXiv preprint arXiv:2212.10773*, 2022.

[35] Seonghyeon Ye, Doyoung Kim, Joel Jang, Joongbo Shin, and Minjoon Seo. Guess the instruction! making language models stronger zero-shot learners. *arXiv preprint arXiv:2210.02969*, 2022.

[36] Ping Yu, Tianlu Wang, Olga Golovneva, Badr Alkhamissy, Gargi Ghosh, Mona Diab, and Asli Celikyilmaz. Alert: Adapting language models to reasoning tasks. *arXiv preprint arXiv:2212.08286*, 2022.

[37] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022.

[38] Kai Zhang, Bernal Jiménez Gutiérrez, and Yu Su. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. *arXiv preprint arXiv:2305.11159*, 2023.

[39] Tianjun Zhang, Fangchen Liu, Justin Wong, Pieter Abbeel, and Joseph E Gonzalez. The wisdom of hindsight makes language models better instruction followers. *arXiv preprint arXiv:2302.05206*, 2023.

[40] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*, 2022.

附录

Table of Contents

A Experimental Setup Details

为确保可重复性，我们提供所有关于我们评估指令调谐LLMs稳健性的详细信息。

A.1 Evaluation Protocols

LLM有时会生成正确但与（自然语言）目标不同的输出。因此，我们根据BIG-BENCH建议的“多项选择”评分来预测答案，我们取logits分数并在所有可能的选择中取argmax以获得预测。在大多数情况下，这种方法产生的准确性与使用精确匹配进行评估的准确性相同。以下是我们评估的所有模型的配置。

Models	Node Type	Precision	Batch Size	Hours	CO_2 emission (KG)
Inference					
Flan-T5-Small	V100-SXM2-32G	FP16	128	64	4.0
Flan-T5-Base	V100-SXM2-32G	FP16	128	128	8.1
Flan-T5-Large	V100-SXM2-32G	FP16	32	256	16.2
Flan-T5-XL	V100-SXM2-32G	FP16	32	512	32.3
Flan-T5-XXL	RTX-A6000-46G	BF16	8	600	37.8
T0++	RTX-A6000-46G	BF16	2	128	8.1
Alpaca-7B	A100-SXM4-80G	BF16	16	160	13.4
Alpaca-13B	A100-SXM4-80G	BF16	8	192	16.1
Training					
Flan-T5-XL	A100-SXM4-80G	BF16	256	256	21.5
Alpaca-7B	A100-SXM4-80G	BF16	128	80	6.7
Estimated Total CO_2 Emission (KG)				164.2	

表 7: 用于评估不同指令调优语言模型的配置。CO_2排放量是由[15]估计的。总的排放量估计相当于一辆平均内燃机汽车驾驶679公里产生的排放量。

A.2 Hyperparameters

我们在具有80GB内存的8个A100上进行所有的训练和消融研究。我们将KL-Loss权重保持在0.8。我们同时训练Flan-T5-XL和Alpaca-7B，批量大小为4。权重衰减设定为$1e-5$。实验的学习率为$5e-4$。

B Disaggregated Results

B.1 Main Results and Scaling Results

在主论文中，我们报告了在基准语料库上的汇总结果。在这里，我们报告了对BBL的单个数据集的结果。对于MMLU，我们一起评估所有的57个数据集，因为这些都是QA任务（我们期望QA模型能够回答跨越多样化领域集的问题）。我们在表9中报告了在所有指示中 achieved 的准确率的均值和标准偏差。设定左侧的数字 suggets 使用的指令数量。我们还分享了更细粒度的结果 — 报告每个指令的性能 — 在补充材料中提供的CSV文件中。

	MMLU				
Model	Flan-T5-XL	Flan-T5-XXL	T0pp-11B	Alpaca-7B	Alpaca-13B
MMLU					
OBSERVED	**48.1 (± 0.3)**	**53.2 (± 0.2)**	48.3 (± 0.9)	**41.9 (± 0.6)**	**47.8 (± 0.5)**
UNOBSERVED	47.5 (± 0.9)	52.7 (± 0.8)	**48.5 (± 0.9)**	39.7 (± 2.2)	47.0 (± 0.8)

表 8: 对于表3中的每个MMLU数据集的粒子结果。我们将MMLU中的所有任务平等地视为一般QA，并计算了总体准确度。

	MMLU				
Size Variance	Small (80M)	Base (250M)	Large (780M)	XL (3B)	XXL (11B)
MMLU					
OBSERVED	29.4 (± 1.0)	**34.1 (± 0.4)**	**41.1 (± 0.2)**	**48.1 (± 0.3)**	**53.2 (± 0.2)**
UNOBSERVED	**29.6 (± 0.9)**	33.8 (± 1.2)	40.7 (± 0.7)	47.5 (± 0.9)	52.7 (± 0.8)

表 9: 针对MMLU中的每个数据集，图2b展示了粒度结果。我们将MMLU中的所有任务平等地视为一般的QA，并计算了总体准确度。

BBL-QA					
Model	Flan-T5-XL	Flan-T5-XXL	T0pp-11B	Alpaca-7B	Alpaca-13B
BBQ Lite					
OBSERVED	66.5 (± 1.5)	**77.4 (± 2.4)**	**51.8 (± 5.3)**	32.6 (± 1.0)	43.5 (± 1.4)
UNOBSERVED	**67.0 (± 7.0)**	73.7 (± 11.4)	51.6 (± 3.0)	**33.1 (± 1.3)**	**45.5 (± 2.9)**
Code Desc.					
OBSERVED	**73.6 (± 3.4)**	**83.6 (± 1.7)**	70.3 (± 3.0)	**70.2 (± 2.5)**	**85.2 (± 2.4)**
UNOBSERVED	69.7 (± 12.4)	72.9 (± 22.2)	**70.5 (± 3.7)**	67.5 (± 11.3)	82.2 (± 8.5)
Hindu Know.					
OBSERVED	**52.4 (± 1.6)**	53.9 (± 1.8)	**57.1 (± 2.5)**	50.9 (± **2.1**)	63.8 (± 0.7)
UNOBSERVED	47.1 (± 5.4)	**56.5 (± 3.5)**	53.2 (± 3.0)	49.8 (± 5.1)	**63.9 (± 1.1)**
Known Unk.					
OBSERVED	**79.3 (± 2.5)**	**84.7 (± 2.1)**	70.9 (± 10.2)	**75.2 (± 4.7)**	**81.9 (± 4.3)**
UNOBSERVED	69.0 (± 6.7)	80.6 (± 8.1)	**76.1 (± 5.9)**	60.9 (± 11.2)	71.1 (± 16.3)
Logical Ded.					
OBSERVED	**52.5 (± 1.0)**	**58.0 (± 0.7)**	**45.5 (± 0.8)**	**25.5 (± 1.1)**	**29.2 (± 1.3)**
UNOBSERVED	52.1 (± 1.1)	57.8 (± 0.6)	45.3 (± 1.2)	24.5 (± 2.3)	28.0 (± 1.6)
Novel Conc.					
OBSERVED	29.8 (± 2.4)	**50.1 (± 1.9)**	28.8 (± 2.9)	**37.2 (± 5.2)**	**20.0 (± 3.1)**
UNOBSERVED	**31.2 (± 5.0)**	46.0 (± 5.4)	**31.5 (± 5.1)**	36.1 (± 7.6)	19.6 (± 4.0)
Logic Grid					
OBSERVED	**41.8 (± 1.1)**	**43.2 (± 1.8)**	**37.6 (± 1.7)**	24.4 (± 1.9)	**29.3 (± 0.8)**
UNOBSERVED	38.6 (± 5.4)	39.6 (± 4.4)	36.4 (± 3.6)	**25.4 (± 1.1)**	28.7 (± 1.2)
Conc. Com.					
OBSERVED	**75.9 (± 1.8)**	**75.0 (± 2.6)**	73.3 (± 2.3)	**58.7 (± 4.0)**	**63.0 (± 2.2)**
UNOBSERVED	75.0 (± 1.9)	73.6 (± 4.8)	**74.2 (± 2.6)**	55.9 (± 6.2)	61.1 (± 3.3)

表 10: 针对类别BBL-QA中的每个数据集，表3列出了细致的结果

BBL-QA					
Size Variance	Small (80M)	Base (250M)	Large (780M)	XL (3B)	XXL (11B)
BBQ Lite					
OBSERVED	28.3 (± 1.3)	**51.5 (± 1.4)**	56.6 (± 2.0)	66.5 (± 1.5)	**77.4 (± 2.4)**
UNOBSERVED	**28.6 (± 4.3)**	50.5 (± 4.1)	**56.7 (± 4.7)**	**67.0 (± 7.0)**	73.7 (± 11.4)
Code Desc.					
OBSERVED	22.0 (± 4.0)	**55.7 (± 3.3)**	**72.4 (± 3.2)**	**73.6 (± 3.4)**	**83.6 (± 1.7)**
UNOBSERVED	**32.1 (± 7.2)**	48.6 (± 7.2)	63.3 (± 14.2)	69.7 (± 12.4)	72.9 (± 22.2)
Hindu Know.					
OBSERVED	25.1 (± 15.2)	**30.7 (± 2.6)**	34.9 (± 0.9)	**52.4 (± 1.6)**	53.9 (± 1.8)
UNOBSERVED	**31.6 (± 10.7)**	26.9 (± 4.4)	**37.5 (± 7.0)**	47.1 (± 5.4)	**56.5 (± 3.5)**
Known Unk.					
OBSERVED	49.9 (± 1.9)	**66.9 (± 4.7)**	**76.2 (± 3.5)**	**79.3 (± 2.5)**	**84.7 (± 2.1)**
UNOBSERVED	**52.8 (± 5.2)**	63.8 (± 7.3)	68.4 (± 11.1)	69.0 (± 6.7)	80.6 (± 8.1)
Logical Ded.					
OBSERVED	19.8 (± 0.7)	27.1 (± 1.3)	45.9 (± 1.1)	**52.5 (± 1.0)**	**58.0 (± 0.7)**
UNOBSERVED	**19.9 (± 0.4)**	**28.9 (± 2.8)**	**46.4 (± 2.6)**	52.1 (± 1.1)	57.8 (± 0.6)
Novel Conc.					
OBSERVED	**22.9 (± 9.2)**	15.9 (± 5.4)	**31.0 (± 2.3)**	29.8 (± 2.4)	**50.1 (± 1.9)**
UNOBSERVED	19.3 (± 3.6)	**16.8 (± 4.0)**	28.4 (± 6.9)	**31.2 (± 5.0)**	46.0 (± 5.4)
Logic Grid					
OBSERVED	22.3 (± 4.0)	**31.7 (± 0.8)**	32.6 (± 2.1)	**41.8 (± 1.1)**	**43.2 (± 1.8)**
UNOBSERVED	**28.8 (± 3.1)**	29.4 (± 5.1)	**34.1 (± 2.8)**	38.6 (± 5.4)	39.6 (± 4.4)
Conc. Com.					
OBSERVED	30.4 (± 10.6)	**55.6 (± 5.1)**	**64.2 (± 1.9)**	**75.9 (± 1.8)**	**75.0 (± 2.6)**
UNOBSERVED	**32.2 (± 16.7)**	54.5 (± 9.4)	58.1 (± 11.6)	75.0 (± 1.9)	73.6 (± 4.8)

表 11: 关于类别BBL-QA的每个数据集的图 2b 的粒度化结果

BBL-BC					
Model	Flan-T5-XL	Flan-T5-XXL	T0pp-11B	Alpaca-7B	Alpaca-13B
Play Dialog					
OBSERVED	**61.6 (± 5.8)**	51.8 (± 9.5)	**62.7 (± 0.4)**	**45.0 (± 2.0)**	**53.4 (± 5.8)**
UNOBSERVED	53.0 (± 6.9)	**58.1 (± 4.4)**	55.2 (± 8.1)	42.9 (± 7.9)	42.9 (± 8.8)
Strat. QA					
OBSERVED	58.7 (± 3.3)	**64.2 (± 3.0)**	51.0 (± 1.8)	53.0 (± 2.1)	56.7 (± 3.8)
UNOBSERVED	**60.7 (± 7.5)**	59.3 (± 6.1)	**54.5 (± 0.9)**	**53.3 (± 4.1)**	**61.0 (± 1.9)**
Strange St.					
OBSERVED	69.3 (± 4.4)	71.0 (± 7.3)	**51.2 (± 5.1)**	**67.0 (± 4.7)**	**69.8 (± 5.0)**
UNOBSERVED	**70.5 (± 7.0)**	**77.4 (± 6.1)**	48.4 (± 3.1)	59.9 (± 9.4)	57.5 (± 5.6)
Winowhy					
OBSERVED	**76.5 (± 1.9)**	**75.6 (± 4.0)**	**99.6 (± 1.0)**	50.1 (± 4.8)	51.9 (± 4.6)
UNOBSERVED	60.2 (± 6.2)	59.7 (± 7.7)	60.9 (± 5.1)	**53.4 (± 4.6)**	**55.2 (± 6.0)**

表 12: 针对类别 BBL-BC 的每个数据集，表 3 的细粒度结果

BBL-BC					
Size Variance	Small (80M)	Base (250M)	Large (780M)	XL (3B)	XXL (11B)
Play Dialog					
OBSERVED	51.6 (± 13.3)	54.6 (± 10.7)	**59.0 (± 6.7)**	**61.6 (± 5.8)**	51.8 (± 9.5)
UNOBSERVED	**61.6 (± 4.6)**	**56.3 (± 10.9)**	57.3 (± 7.8)	53.0 (± 6.9)	**58.1 (± 4.4)**
Strat. QA					
OBSERVED	**52.3 (± 1.0)**	48.9 (± 2.1)	**60.9 (± 1.3)**	58.7 (± 3.3)	**64.2 (± 3.0)**
UNOBSERVED	51.5 (± 2.7)	**52.9 (± 1.3)**	53.9 (± 3.8)	**60.7 (± 7.5)**	59.3 (± 6.1)
Strange St.					
OBSERVED	41.3 (± 10.3)	**43.1 (± 4.2)**	54.4 (± 1.2)	69.3 (± 4.4)	71.0 (± 7.3)
UNOBSERVED	**55.9 (± 18.5)**	42.0 (± 5.5)	**67.9 (± 8.0)**	**70.5 (± 7.0)**	**77.4 (± 6.1)**
Winowhy					
OBSERVED	**54.8 (± 1.6)**	55.9 (± 7.6)	**60.4 (± 9.8)**	**76.5 (± 1.9)**	**75.6 (± 4.0)**
UNOBSERVED	53.7 (± 5.1)	**57.1 (± 6.7)**	53.5 (± 4.9)	60.2 (± 6.2)	59.7 (± 7.7)

表 13: 对于类别BBL-BC的每个数据集上，图2b的颗粒结果

	BBL-MC				
Model	Flan-T5-XL	Flan-T5-XXL	T0pp-11B	Alpaca-7B	Alpaca-13B
Language ID					
OBSERVED	**32.6 (± 0.2)**	**38.9 (± 0.3)**	**15.7 (± 3.0)**	12.9 (± 0.7)	18.5 (± 0.7)
UNOBSERVED	25.5 (± 7.3)	31.6 (± 9.4)	14.3 (± 2.4)	**14.7 (± 1.7)**	**21.7 (± 0.7)**
Vitamin C					
OBSERVED	78.6 (± 1.1)	78.5 (± 0.7)	68.3 (± 1.1)	**51.4 (± 3.6)**	**54.9 (± 2.9)**
UNOBSERVED	78.6 (± 3.6)	**80.2 (± 1.6)**	**68.5 (± 2.4)**	18.1 (± 5.3)	23.6 (± 14.2)

表 14: 针对BBL-MC类别的每个数据集，表3的粒度结果

	BBL-MC				
Size Variance	Small (80M)	Base (250M)	Large (780M)	XL (3B)	XXL (11B)
Language ID					
OBSERVED	**11.9 (± 0.2)**	**17.0 (± 0.3)**	**25.8 (± 0.3)**	**32.6 (± 0.2)**	**38.9 (± 0.3)**
UNOBSERVED	9.5 (± 0.2)	12.4 (± 1.5)	19.2 (± 4.4)	25.5 (± 7.3)	31.6 (± 9.4)
Vitamin C					
OBSERVED	**46.6 (± 4.0)**	60.7 (± 5.6)	**72.6 (± 1.5)**	78.6 (± 1.1)	78.5 (± 0.7)
UNOBSERVED	40.8 (± 4.2)	**63.0 (± 4.6)**	36.4 (± 0.8)	78.6 (± 3.6)	**80.2 (± 1.6)**

表 15: 对于类别BBL-MC的每个数据集上，图2b的粒度结果

B.2 "Closer Look" Experiment Results

在这里，我们提供了我们在3中报告的详细结果。

Dataset	Observed		Unobserved		Control	
	Closest	Incorrect	Collected	Task Designer	Negated	Nonsensical
Intent Recognition	93.6 ±0.3	93.1 ±1.0	94.1 ±0.6	**94.66** -	28.0 ±6.5	40.7 ±7.2
Empirical Judgments	39.2 ±0.8	**41.62** ±6.3	37.6 ±1.7	37.4 -	28.1 ±3.5	30.9 ±2.5
Conceptual Combinations	78.0 ±1.6	**78.92** ±0.5	75.3 ±3.3	58.3 -	11.2 ±2.1	63.7 ±2.6
Language Identification	**38.94** ±0.3	29.3 ±5.0	28.8 ±5.8	27.6 -	36.9 ±0.5	12.4 ±0.5
Logical Sequence	**56.92** ±6.6	49.4 ±6.1	52.8 ±5.3	53.8 -	11.8 ±6.9	34.4 ±5.9
Crash Blossom	53.6 ±2.8	50.0 ±5.3	50.5 ±2.2	**63.16** -	28.6 ±6.2	43.7 ±1.4
Epistemic Reasoning	62.8 ±2.9	59.3 ±3.4	58.1 ±1.7	60.2 -	**65.49** ±4.6	49.5 ±1.3
Overall	**60.45** ±2.1	57.4 ±2.2	56.8 ±1.8	56.4 -	30.0 ±2.2	39.3 ±2.3

表 16: “深入研究”的详细结果。我们在E.3部分提供了所有选取的指令及其来源。可以看出，在大多数情况下，“错误”但观察到的指令优于未观察到的正确指令（“收集”和“任务设计者”）。

C Instruction Robustness with In-context Learning

我们一直关注的是零样本设置，但在这里我们也报告在背景学习（ICL）下取得的结果。考虑到Flan-T5的上下文窗口限制无法提供额外的背景样本，我们采用了单样本ICL。我们重复了主论文中的实验，如下所示的一次ICL实验结果。我们（任意地）从每个数据集中选取了第一个实例作为样本。由结果可以看出，提供了背景示例后，观察到的和未观察到的指令之间的性能差距已显著缩小。

	MMLU **(One-shot)**				
Flan-T5	Small (80M)	Base (250M)	Large (780M)	XL (3B)	XXL (11B)
MMLU					
OBSERVED	29.3 (± 0.9)	**33.9 (± 0.5)**	**40.7 (± 0.2)**	**47.5 (± 0.2)**	52.7 (± 0.2)
UNOBSERVED	**29.6 (± 0.6)**	33.8 (± 1.0)	40.4 (± 0.9)	47.5 (± 0.7)	**52.8 (± 1.0)**

表 17: 对于Flan-T5模型和一次性ICL，我们在MMLU的每个数据集上的粒度结果。我们将MMLU中的所有QA任务分组，并报告这些任务的总体准确率。

BBL-QA **(One-shot)**					
Flan-T5	Small (80M)	Base (250M)	Large (780M)	XL (3B)	XXL (11B)
BBQ Lite					
OBSERVED	**29.6 (± 2.2)**	50.5 (± 1.7)	57.0 (± 2.0)	66.7 (± 1.6)	77.2 (± 2.7)
UNOBSERVED	28.9 (± 1.4)	**51.0 (± 3.6)**	**58.0 (± 2.7)**	**69.0 (± 5.9)**	**77.6 (± 5.5)**
Code Desc.					
OBSERVED	20.5 (± 3.1)	**56.9 (± 5.1)**	**76.0 (± 2.2)**	**73.6 (± 1.8)**	**85.3 (± 1.6)**
UNOBSERVED	**25.6 (± 8.7)**	43.6 (± 8.4)	53.5 (± 16.0)	65.8 (± 15.3)	75.0 (± 12.3)
Hindu Know.					
OBSERVED	**24.5 (± 12.7)**	**30.5 (± 3.6)**	36.2 (± 1.7)	**50.9 (± 1.3)**	52.9 (± 1.9)
UNOBSERVED	23.0 (± 5.8)	22.4 (± 5.8)	**37.7 (± 4.7)**	50.2 (± 5.8)	**54.6 (± 3.2)**
Known Unk.					
OBSERVED	**49.2 (± 3.8)**	**66.7 (± 8.3)**	**73.6 (± 2.7)**	**76.3 (± 2.0)**	**84.7 (± 3.8)**
UNOBSERVED	49.1 (± 4.1)	60.2 (± 7.3)	67.7 (± 12.9)	63.7 (± 9.8)	76.0 (± 12.7)
Logical Ded.					
OBSERVED	20.2 (± 0.4)	26.8 (± 1.0)	**45.9 (± 1.1)**	**53.0 (± 0.7)**	**58.2 (± 0.5)**
UNOBSERVED	**20.2 (± 0.8)**	**27.8 (± 3.5)**	44.3 (± 8.6)	48.6 (± 10.1)	55.0 (± 10.5)
Novel Conc.					
OBSERVED	**21.3 (± 4.3)**	**14.8 (± 4.9)**	**29.1 (± 4.0)**	31.2 (± 2.0)	**47.7 (± 3.1)**
UNOBSERVED	17.2 (± 6.8)	14.7 (± 9.0)	24.7 (± 6.2)	**35.1 (± 5.2)**	42.5 (± 9.0)
Conc. Com.					
OBSERVED	28.0 (± 3.6)	**38.5 (± 2.3)**	**65.0 (± 1.8)**	**77.7 (± 1.6)**	**77.1 (± 2.2)**
UNOBSERVED	**28.6 (± 4.5)**	36.3 (± 6.4)	58.2 (± 10.7)	74.8 (± 11.9)	75.2 (± 2.8)

表 18: 在类别BBL-QA的每个数据集上的颗粒化结果，使用一次性的情境学习。

D Representational Similarity and Model Performance

我们为所有的T5模型大小重新生成了图4中的可视化图像，如图 . 定性结果——遵循观察到的指令的令牌表示通常与遵循未观察到的指令的表示不同——在很大程度上保持一致，尽管差异不太明显，例如，对于XL（特别是在这里，MMLU样本并不像我们所期望的那样混乱）。

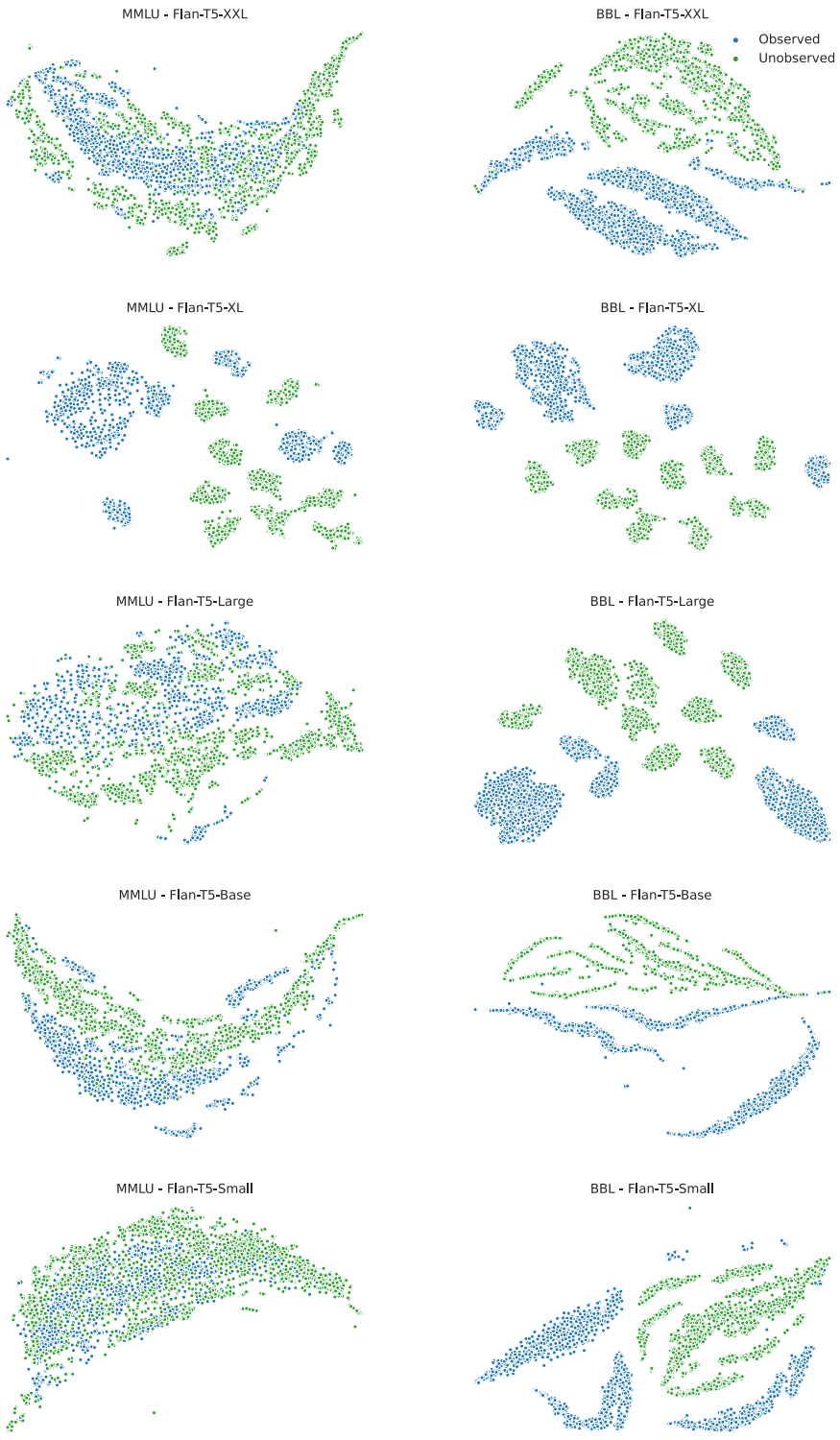


图 8: 我们在所有T5模型大小上复制了主论文中的图4。

BBL-BC **(One-shot)**					
Flan-T5	Small (80M)	Base (250M)	Large (780M)	XL (3B)	XXL (11B)
Play Dialog					
OBSERVED	49.8 (± 11.9)	**54.4 (± 10.3)**	**58.1 (± 5.7)**	**57.8 (± 3.3)**	43.9 (± 4.2)
UNOBSERVED	**55.8 (± 10.2)**	52.5 (± 10.7)	48.6 (± 8.7)	46.3 (± 3.9)	**51.3 (± 3.3)**
Strat. QA					
OBSERVED	52.5 (± 0.8)	49.3 (± 3.1)	**60.6 (± 1.5)**	60.6 (± 6.2)	**66.2 (± 2.6)**
UNOBSERVED	**53.2 (± 0.0)**	**53.3 (± 0.8)**	55.9 (± 4.3)	**61.2 (± 6.0)**	62.2 (± 5.5)
Strange St.					
OBSERVED	40.7 (± 12.3)	**41.8 (± 2.3)**	51.7 (± 1.6)	74.4 (± 3.6)	78.5 (± 2.1)
UNOBSERVED	**46.7 (± 5.1)**	37.9 (± 8.6)	**56.3 (± 3.0)**	**78.7 (± 3.2)**	**83.2 (± 7.6)**
Winowhy					
OBSERVED	**52.7 (± 2.8)**	57.3 (± 6.2)	**62.1 (± 5.9)**	**77.2 (± 0.6)**	**76.7 (± 1.0)**
UNOBSERVED	48.1 (± 4.2)	**58.3 (± 8.1)**	57.1 (± 8.8)	66.3 (± 8.8)	65.5 (± 9.9)

表 19: 在类别BBL-BC的每个数据集上的颗粒结果，使用一次性上下文学习。

BBL-MC **(One-shot)**					
Flan-T5	Small (80M)	Base (250M)	Large (780M)	XL (3B)	XXL (11B)
Language ID					
OBSERVED	**11.7 (± 0.2)**	**13.5 (± 2.8)**	**25.6 (± 0.4)**	**31.8 (± 0.3)**	**38.7 (± 0.5)**
UNOBSERVED	9.7 (± 0.9)	11.6 (± 1.5)	16.9 (± 5.4)	20.3 (± 7.5)	28.4 (± 10.3)
Vitamin C					
OBSERVED	**50.9 (± 1.6)**	60.9 (± 5.4)	**73.3 (± 0.8)**	**78.6 (± 1.4)**	80.4 (± 0.5)
UNOBSERVED	50.1 (± 0.9)	**64.5 (± 1.8)**	73.2 (± 1.5)	77.5 (± 9.3)	**80.7 (± 3.6)**

表 20: 在类别BBL-MC的每个数据集上得到的粒度结果，利用一次性情境学习。

E Instruction Collection

在本节中，我们详细报告了我们如何收集用于评估指令调整类型长期记忆模型（LLM）领域外（O.O.D.）鲁棒性的指令。

E.1 Observed Instructions

我们手动审查了用于训练我们评估的三个指令调整LLM的集合：Flan [5]，Alpaca [26]和P3 [24]。我们考虑足够通用的指令，它们可以“混合搭配”用于三种任务类型：问答，二元分类，多类别分类。

下面我们为所有(观察到的)我们汇总的指令和它们在集合中的来源提供指令模板：

Flan

为了简化，我们提供了我们从Flan集收集的观察到的指令模板的任务和指数。我们用于处理数据的确切代码可以在公开可用的Flan存储库 [4] 中找到

QA - 01 来源：NIV2 - 任务73 - 模板1

输入： {question}，{options}

模板：

你被给予一个问题和一些回答选项（分别与"A", "B", "C", "D"关联）。你应该根据常识选择正确的答案。避免根据关联性回答问题，答案集是为了捕捉超出关联性的常识而有意选择的。除了以下字符之一，不要生成任何其他东西：{options letter}，并且每个问题只给出一个答案。

{question} {options}

QA - 02 来源：NIV2 - 任务73 - 模板2

输入： {question}，{options}

模板：

首先，你会得到一个任务的定义，然后是一些任务的输入。
你被给予一个问题和一些回答选项（分别与"A", "B", "C", "D"关联）。你应该根据常识选择正确的答案。避免根据关联性回答问题，答案集是为了捕捉超出关联性的常识而有意选择的。除了以下字符之一，不要生成任何其他东西：{options letter}，并且每个问题只给出一个答案。

{question} {options}
输出：

QA - 03 来源：NIV2 - 任务73 - 模板3

输入： {question}，{options}

模板：

定义：你被给予一个问题和一些回答选项（分别与"A", "B", "C", "D"关联）。你应该根据常

[4] https://github.com/google-research/FLAN

识选择正确的答案。避免根据关联性回答问题，答案集是为了捕捉超出关联性的常识而有意选择的。除了以下字符之一，不要生成任何其他东西：{options letter}，并且每个问题只给出一个答案。
输入：{question} {options}
输出：

QA - 04 来源：NIV2 - 任务73 - 模板4

输入：{question}，{options}

模板：

指导：你被给予一个问题和一些回答选项（分别与"A", "B", "C", "D"关联）。你应该根据常识选择正确的答案。避免根据关联性回答问题，答案集是为了捕捉超出关联性的常识而有意选择的。除了以下字符之一，不要生成任何其他东西：{options letter}，并且每个问题只给出一个答案。
输入：{question} {options}
输出：

QA - 05 来源：NIV2 - 任务 73 - 模板 5

输入：{question}, {options}

模板：

提供一个问题和一些答案选项（分别与"A"，"B"，"C"，"D"相关联）。你应该根据常识选择正确答案。避免基于关联性来回答问题，答案集灵活选择以捕获超越关联性的常识。除了下列字符之一：{options letter}，不要生成任何其他东西，并且每道题只给出一个答案。
Q：{question} {options}
A:

QA - 06 来源：NIV2 - 任务 73 - 模板 6

输入：{question}, {options}

模板：

根据任务定义和输入，回复输出。你被给予一个问题和一些答案选项（分别与"A"，"B"，"C"，"D"相关联）。你应该根据常识选择正确答案。避免基于关联性来回答问题，答案集灵活选择以捕获超越关联性的常识。除了下列字符之一：{options letter}，不要生成任何其他东西，并且每道题只给出一个答案。

{question} {options}

QA - 07 来源：NIV2 - 任务 73 - 模板 7

输入：{question}, {options}

模板：

教师：你被给予一个问题和一些答案选项（分别与"A"，"B"，"C"，"D"相关联）。你应该根据常识选择正确答案。避免基于关联性来回答问题，答案集灵活选择以捕获超越关联性的常识。除了下列字符之一：{options letter}，不要生成任何其他东西，并且每道题只给出一个答案。

教师：现在，理解问题了吗？解决这个实例：{question} {options}
学生：

QA - 08 来源：NIV2 - 任务 73 - 模板 8

输入： {question}, {options}

模板：

Q：给你一个问题和一些答案选项（分别与"A"，"B"，"C"，"D"相关联）。你应该根据常识选择正确答案。避免基于关联性来回答问题，答案集灵活选择以捕获超越关联性的常识。除了下列字符之一：{options letter}，不要生成任何其他东西，并且每道题只给出一个答案。
{question} {options}
A:

QA - 09 来源：NIV2 - 任务 73 - 模板 9

输入： {问题}, {选项}

模板：

详细指导：给你一个问题和一些答案选择（与"A"，"B"，"C"，"D"相关联）。你应该根据常识选择正确的答案。避免根据关联来回答问题，答案的选择是有意识地抓取超越关联的常识。除了以下字符之一外，不要生成其他内容：{选项字母}，并且每个问题只给出一个答案。
问题：{问题} {选项}
解决方案：

QA - 10 来源：NIV2 - 任务 73 - 模板 10

输入： {问题}, {选项}

模板：

详细指导：给你一个问题和一些答案选择（与"A"，"B"，"C"，"D"相关联）。你应该根据常识选择正确的答案。避免根据关联来回答问题，答案的选择是有意识地抓取超越关联的常识。除了以下字符之一外，不要生成其他内容：{选项字母}，并且每个问题只给出一个答案。
Q：{问题} {选项}
A：

QA - 11 来源：NIV2 - 任务 1420 - 模板 1

输入： {问题}, {选项}

模板：

在这个任务中，你需要从提供的选项中提供给定问题的正确选项。

问题：{问题}
{选项}

QA - 12 来源：NIV2 - 任务 1420 - 模板 2

输入： {问题}, {选项}

模板：

你将首先得到一个任务的定义，然后是任务的一些输入。
在这个任务中，你需要从提供的选择中为给定的问题提供正确的选项。

问题：{问题}
{选项}
输出：

QA - 13 来源：NIV2 - 任务 1420 - 模板 3

输入： {问题}, {选项}

模板：

定义：在这个任务中，你需要从提供的选项中，为给定的问题提供正确的选项。
输入：问题：{问题}
{选项}
输出：

QA - 14 来源：NIV2 - 任务 1420 - 模板 4

输入： {问题}, {选项}

模板：

指导：在这个任务中，你需要从提供的选项中为给定的问题提供正确的选项。
输入：问题：{问题}
{选项}
输出：

QA - 15 来源：NIV2 - 任务1420 - 模板5

输入： {问题}, {选项}

模板：

在这个任务中，你需要从给出的选项中提供给定问题的正确选项。
Q: 问题:{问题}
{选项}
A:

QA - 16 来源：NIV2 - 任务1420 - 模板6

输入： {问题}, {选项}

模板：

根据任务定义和输入，回复输出。在此任务中，你需要为给定问题从提供的选项中提供正确选项。

问题:{问题}
{选项}

QA - 17 来源：NIV2 - 任务1420 - 模板7

输入： {问题}, {选项}

模板：

老师：在这个任务中，你需要从给出的选项中提供给定问题的正确选项。
老师：现在，理解这个问题了吗？解答这个实例：问题:{问题}
{选项}
学生:

QA - 18 来源：NIV2 - 任务1420 - 模板8

输入： {问题}, {选项}

模板：

Q: 在这个任务中，你需要从给出的选项中提供给定问题的正确选项。
问题:{问题}
{选项}
A:

QA - 19 来源：NIV2 - 任务1420 - 模板9

输入： {问题}, {选项}

模板：

详细说明：在此任务中，您需要从提供的选项中为给定问题提供正确选项。
问题：问题:{问题}
{选项}
解决方案:

QA - 20 来源：NIV2 - 任务1420 - 模板10

输入： {问题}, {选项}

模板：

详细说明：在此任务中，您需要从提供的选项中为给定问题提供正确选项。
Q: 问题:{问题}
{选项}
A:

QA - 21 来源：NIV2 - 任务1286 - 模板1

输入： {问题}, {选项}

模板：

在此任务中，你会收到一个多项选择题，你必须选择错的选项。用选项索引答题（即，{选项字母}）。

{问题} {选项}

QA - 22 来源：NIV2 - 任务 1286 - 模板 2

输入： {question}, {options}

模板：

首先，你将获得一个任务的定义，然后是任务的一些输入。
在这个任务中，给你一个多选题，你需要选出不正确的选项。用选项索引进行回答（即，

{options letter}）。

{question} {options}
输出：

QA - 23 来源：NIV2 - 任务 1286 - 模板 3

输入： {question}, {options}

模板：

定义：在这个任务中，给你一个多选题，你需要选出不正确的选项。用选项索引进行回答（即，{options letter}）。
输入：{question} {options}
输出：

QA - 24 来源：NIV2 - 任务 1286 - 模板 4

输入： {question}, {options}

模板：

指导：在这个任务中，给你一个多选题，你需要选出不正确的选项。用选项索引进行回答（即，{options letter}）。
输入：{question} {options}
输出：

QA - 25 来源：NIV2 - 任务 1286 - 模板 5

输入： {question}, {options}

模板：

在这个任务中，给你一个多选题，你需要选出不正确的选项。用选项索引进行回答（即，{options letter}）。
问：{question} {options}
答：

QA - 26 来源：NIV2 - 任务 1286 - 模板 6

输入： {question}, {options}

模板：

按照任务定义和输入给出输出。在这个任务中，给你一个多选题，你需要选出不正确的选项。用选项索引进行回答（即，{options letter}）。

{question} {options}

QA - 27 来源：NIV2 - 任务 1286 - 模板 7

输入： {question}, {options}

模板：

教师：在这个任务中，给你一个多选题，你需要选出不正确的选项。用选项索引进行回答（即，{options letter}）。

教师：现在，理解了问题吗？解决这个实例：{question} {options}
学生：

QA - 28 来源: NIV2 - 任务 1286 - 模板 8

输入: {问题}, {选项}

模板:

Q: 在此任务中，你会得到一个多选题，并需要挑选出错误的选项。请使用选项的索引（例如，{选项字母}）回答问题。
{问题} {选项}
A:

QA - 29 来源: NIV2 - 任务 1286 - 模板 9

输入: {问题}, {选项}

模板:

详细说明: 在此任务中，你会得到一个多选题，并需要挑选出错误的选项。请使用选项的索引（例如，{选项字母}）回答问题。
问题:{问题} {选项}
解答:

QA - 30 来源: NIV2 - 任务 1286 - 模板 10

输入: {问题}, {选项}

模板:

详细说明: 在此任务中，你会得到一个多选题，并需要挑选出错误的选项。请使用选项的索引（例如，{选项字母}）回答问题。
Q: {问题} {选项}
A:

QA - 31 来源: NIV2 - 任务 1565 - 模板 1

输入: {问题}, {选项}

模板:

本任务涉及提出一个问题，并提供一系列{选项长度}的选项。你需要选择最适合该问题的答案。输出形式是{选项字母}，对应于选择的选项。

{问题}, 选项: [{选项}]

QA - 32 来源: NIV2 - 任务 1565 - 模板 2

输入: {问题}, {选项}

模板:

首先你将获得一项任务的定义，然后是一些任务的输入内容。
本任务涉及提出一个问题，并提供一系列{选项长度}的选项。你需要选择最适合该问题的答案。输出形式是{选项字母}，对应于选择的选项。

{问题}, 选项: [{选项}]
输出：

QA - 33 来源: NIV2 - 任务 1565 - 模板 3

输入: {问题}, {选项}

模板:

定义: 本任务涉及提出一个问题，并提供一系列{选项长度}的选项。你需要选择最适合该问题的答案。输出形式是{选项字母}，对应于选择的选项。
输入: {问题}, 选项: [{选项}]
输出:

QA - 34 来源：NIV2 - 任务 1565 - 模板 4

输入：{问题}, {选项}

模板：

指导：本任务涉及提出一个问题，提供一组{选项长度}选项。您需要选择问题的最佳答案。输出将以{选项字母}的形式，对应所选的选项。
输入：{问题}, 选项: [{选项}]
输出：

QA - 35 来源：NIV2 - 任务 1565 - 模板 5

输入：{问题}, {选项}

模板：

本任务涉及提出一个问题，提供一组{选项长度}选项。您需要选择问题的最佳答案。输出将以{选项字母}的形式，对应所选的选项。
Q：{问题}, 选项: [{选项}]
A：

QA - 36 来源：NIV2 - 任务 1565 - 模板 6

输入：{问题}, {选项}

模板：

根据任务定义和输入，回复输出。此任务涉及提出一个问题，提供一组{选项长度}选项。你需要选择问题的最佳答案。输出将以{选项字母}的形式，对应所选的选项。

{问题}, 选项: [{选项}]

QA - 37 来源：NIV2 - 任务 1565 - 模板 7

输入：{问题}, {选项}

模板：

教师：此任务涉及提出一个问题，提供一组{选项长度}选项。你需要选择问题的最佳答案。输出将以{选项字母}的形式，对应所选的选项。
教师：现在，理解问题了吗？解决这个实例：{问题}, 选项: [{选项}]
学生：

QA - 38 来源：NIV2 - 任务 1565 - 模板 8

输入： {问题}, {选项}

模板：

Q：此任务涉及提出一个问题，提供一组{选项长度}选项。你需要选择问题的最佳答案。输出将以{选项字母}的形式，对应所选的选项。
{问题}, 选项: [{选项}]
A：

QA - 39 来源：NIV2 - 任务 1565 - 模板 9

输入： {问题}, {选项}

模板：

详细说明：此项任务包括提问和提供一套{选项长度}选项。你需要选择对问题的最佳答案。输出将以{选项字母}的形式，对应选择的选项。
问题：{问题}, 选项：[{选项}]
解答：

QA - 40 来源：NIV2 - 任务 1565 - 模板 10

输入： {问题}, {选项}

模板：

详细说明：此项任务包括提问和提供一套{选项长度}选项。你需要选择对问题的最佳答案。输出将以{选项字母}的形式，对应选择的选项。
Q：{问题}, 选项：[{选项}]
A：

QA - 41 来源：NIV2 - 任务 229 - 模板 1

输入： {问题}, {选项}

模板：

你会得到一个科学问题（高级别）和{选项长度}个答案选项（与{选项字母}关联）。你的任务是根据科学事实、知识和推理找出正确答案。除了以下字符之一：{选项字母}之外，不要生成任何其他内容。每个问题只有一个正确答案。

{问题} {选项}

QA - 42 来源：NIV2 - 任务 229 - 模板 2

输入： {问题}, {选项}

模板：

首先，你会得到一个任务的定义，然后是一些任务的输入。
你会得到一个科学问题（高级别）和{选项长度}个答案选项（与{选项字母}关联）。你的任务是根据科学事实、知识和推理找出正确答案。除了以下字符之一：{选项字母}之外，不要生成任何其他内容。每个问题只有一个正确答案。

{问题} {选项}
输出：

QA - 43 来源：NIV2 - 任务 229 - 模板 3

输入： {问题}, {选项}

模板：

定义：你会得到一个科学问题（高级别）和{选项长度}个答案选项（与{选项字母}关联）。你的任务是根据科学事实、知识和推理找出正确答案。除了以下字符之一：{选项字母}之外，不要生成任何其他内容。每个问题只有一个正确答案。
输入：{问题} {选项}
输出：

QA - 44 Source: NIV2 - Task 229 - Template 4

输入： {问题}, {选项}

模板：

说明：您将获得一道科学问题（高难度级别）和{选项数量}个答案选项（与{选项字母}相关）。您的任务是根据科学事实、知识和推理找出正确答案。除了下列字符之一：{选项字母}，不要生成其他任何内容。每个问题只有一个正确答案。
输入：{问题} {选项}
输出：

QA - 45 Source: NIV2 - Task 229 - Template 5

输入： {问题}, {选项}

模板：

您将获得一道科学问题（高难度级别）和{选项数量}个答案选项（与{选项字母}相关）。您的任务是根据科学事实、知识和推理找出正确答案。除了下列字符之一：{选项字母}，不要生成其他任何内容。每个问题只有一个正确答案。
问：{问题} {选项}
答：

QA - 46 Source: NIV2 - Task 229 - Template 6

输入： {问题}, {选项}

模板：

根据任务定义和输入，用输出回答。您将获得一道科学问题（高难度级别）和{选项数量}个答案选项（与{选项字母}相关）。您的任务是根据科学事实、知识和推理找出正确答案。除了下列字符之一：{选项字母}，不要生成其他任何内容。每个问题只有一个正确答案。

{问题} {选项}

QA - 47 Source: NIV2 - Task 229 - Template 7

输入： {问题}, {选项}

模板：

教师：您将获得一道科学问题（高难度级别）和{选项数量}个答案选项（与{选项字母}相关）。您的任务是根据科学事实、知识和推理找出正确答案。除了下列字符之一：{选项字母}，不要生成其他任何内容。每个问题只有一个正确答案。

教师：现在，理解问题了吗？解决这个实例：{问题} {选项}
学生：

QA - 48 来源：NIV2 - 任务 229 - 模板 8

输入： {问题}, {选项}

模板：

问：你给出了一个科学问题（高难度）和 {选项长度}个答案选项（与{选项字母} 相关联）。你的任务是根据科学事实，知识和推理找出正确答案。不要生成除以下字符之外的任何其他内容：{选项字母}。每个问题只有一个正确答案。
{问题} {选项}
答：

QA - 49 来源：NIV2 - 任务 229 - 模板 9

输入： {问题}, {选项}

模板：

详细指导：你给出了一个科学问题（高难度）和 {选项长度}个答案选项（与{选项字母} 相关联）。你的任务是根据科学事实，知识和推理找出正确答案。不要生成除以下字符之外的任何其他内容：{选项字母}。每个问题只有一个正确答案。
问题： {问题} {选项}
解决方案：

QA - 50 来源：NIV2 - 任务 229 - 模板 10

输入： {问题}, {选项}

模板：

详细指导：你给出了一个科学问题（高难度）和 {选项长度}个答案选项（与{选项字母} 相关联）。你的任务是根据科学事实，知识和推理找出正确答案。不要生成除以下字符之外的任何其他内容：{选项字母}。每个问题只有一个正确答案。
问： {问题} {选项}
答：

MC - 01 来源：NIV2 - 任务 1135 - 模板 1

输入： {问题}, {选项}

模板：

在这个任务中，你将被提出一个有多个可能答案的问题。你应该根据你的常识知识选择 {选项字母}中最合适的选项。

{问题} 选项： {选项}

MC - 02 来源：NIV2 - 任务 1135 - 模板 2

输入： {问题}, {选项}

模板：

你将首先得到一个任务的定义，然后是一些任务的输入。
在这个任务中，你将被提出一个有多个可能答案的问题。你应该根据你的常识知识选择 {选

项字母}中最合适的选项。

{问题} 选项： {选项}
输出：

MC - 03 来源：NIV2 - 任务 1135 - 模板 3

输入： {问题}, {选项}

模板：

定义：在这个任务中，你会面对一个有多个可能答案的问题。你应该根据你的常识知识，从{选项字母}中选择最合适的选项。
输入：{问题} 选项：{选项}
输出：

MC - 04 来源：NIV2 - 任务 1135 - 模板 4

输入： {问题}, {选项}

模板：

说明：在这个任务中，你会面对一个有多个可能答案的问题。你应该根据你的常识知识，从{选项字母}中选择最合适的选项。
输入：{问题} 选项：{选项}
输出：

MC - 05 来源：NIV2 - 任务 1135 - 模板 5

输入： {问题}, {选项}

模板：

在这个任务中，你会面对一个有多个可能答案的问题。你应该根据你的常识知识，从 {选项字母} 中选择最合适的选项。
Q：{问题} 选项：{选项}
A：

MC - 06 来源：NIV2 - 任务 1135 - 模板 6

输入： {问题}, {选项}

模板：

根据任务定义和输入，进行回答。在这个任务中，你会面对一个有多个可能答案的问题。你应该根据你的常识知识，从 {选项字母} 中选择最合适的选项。

{问题} 选项：{选项}

MC - 07 来源：NIV2 - 任务 1135 - 模板 7

输入： {问题}, {选项}

模板：

教师：在这个任务中，你会面对一个有多个可能答案的问题。你应该根据你的常识知识，从{选项字母}中选择最合适的选项。

教师：现在，理解问题了吗？解决这个实例：{问题} 选项：{选项}
学生：

MC - 08 来源：NIV2 - 任务 1135 - 模板 8

输入： {问题}, {选项}

模板：

Q：在这个任务中，你会面对一个有多个可能答案的问题。你应该根据你的常识知识，从{选项字母}中选择最合适的选项。
{问题} 选项：{选项}
A：

MC - 09 来源：NIV2 - 任务 1135 - 模板 9

输入： {问题}, {选项}

模板：

详细说明：在此任务中，您将面对一个有多个可能答案的问题。您应根据您的常识知识，从{选项字母}中选择最合适的选项。
问题:{问题} 选项: {选项}
解答：

MC - 10 来源：NIV2 - 任务 1135 - 模板 10

输入： {问题}, {选项}

模板：

详细说明：在此任务中，您将面对一个有多个可能答案的问题。您应根据您的常识知识，从{选项字母}中选择最合适的选项。
问题：{问题} 选项：{选项}
答案：

MC - 11 来源：NIV2 - 任务 900 - 模板 1

输入： {文本}, {选项}

模板：

给定一个小常识问题，请从这个列表中分类出一个大的主题类别：{选项}。

{问题}

MC - 12 来源：NIV2 - 任务 900 - 模板 2

输入： {文本}, {选项}

模板：

首先，您将得到一个任务的定义，然后是一些任务的输入。
给定一个小常识问题，请从这个列表中分类出一个大的主题类别：{选项}。

{问题}
输出：

MC - 13 来源：NIV2 - 任务 900 - 模板 3

输入：{文本}, {选项}

模板：

定义：给定一个小常识问题，请从这个列表中分类出一个大的主题类别：{选项}。
输入：{问题}
输出：

MC - 14 来源：NIV2 - 任务 900 - 模板 4

输入：{文本}, {选项}

模板：

说明：给定一个小常识问题，请从这个列表中分类出一个大的主题类别：{选项}。
输入：{问题}
输出：

MC - 15 来源：NIV2 - 任务 900 - 模板 5

输入：{文本}, {选项}

模板：

给定一个小常识问题，请从这个列表中分类出一个大的主题类别：{选项}。
问题：{问题}
答案：

MC - 16 来源：NIV2 - 任务 900 - 模板 6

输入：{文本}, {选项}

模板：

根据任务定义和输入，回复输出。给出一个小常识问题，从这个列表中分类广泛的主题种类：{选项}。

{问题}

MC - 17 来源：NIV2 - 任务 900 - 模板 7

输入：{文本}, {选项}

模板：

老师：给出一个小常识问题，从这个列表中分类广泛的主题种类：{选项}。
老师：现在，理解问题了吗？解答这个实例：{问题}
学生：

MC - 18 来源：NIV2 - 任务 900 - 模板 8

输入：{文本}, {选项}

模板：

Q：给出一个小常识问题，从这个列表中分类广泛的主题种类：{选项}。
{问题}
A：

MC - 19 来源：NIV2 - 任务 900 - 模板 9

输入：{文本}, {选项}

模板：

详细指南：给出一个小常识问题，从这个列表中分类广泛的主题种类：{选项}。
问题：{问题}
解答：

MC - 20 来源：NIV2 - 任务 900 - 模板 10

输入：{文本}, {选项}

模板：

详细指南：给出一个小常识问题，从这个列表中分类广泛的主题种类：{选项}。
Q：{问题}
A：

MC - 21 来源：Flan2021 - ARC - 模板 1

输入：{问题}, {选项}

模板：

{文本}
选项：
{选项}

MC - 22 来源：Flan2021 - ARC - 模板 2

输入：{问题}, {选项}

模板：

问题：{文本}?
选项：{选项}
答案：

MC - 23 来源：Flan2021 - ARC - 模板 3

输入：{问题}, {选项}

模板：

问题：{文本}

从以下选择中，问题的正确答案是什么？
选项：{选项}

MC - 24 来源: Flan2021 - ARC - Template 4

输入: {问题}, {选项}

模板:

问题: {文本}
这个问题的正确答案是什么？
选项:{选项}...A:

MC - 25 来源: Flan2021 - ARC - Template 5

输入: {问题}, {选项}

模板:

选择你的答案?

{文本}

选项:{选项}

MC - 26 来源: Flan2021 - ARC - Template 6

输入: {问题}, {选项}

模板:

回答问题

{文本}
选项:{选项}

MC - 27 来源: Flan2021 - ARC - Template 7

输入: {问题}, {选项}

模板:

{文本}

从这些选项中选择答案

选项:{选项}

MC - 28 来源: Flan2021 - CosmosQA - Template 1

输入: {背景}, {问题}, {选项}

模板:

{背景}

带有选项的问题: {问题}
选项:{选项}

MC - 29 来源: Flan2021 - CosmosQA - Template 2

输入: {背景}, {问题}, {选项}

模板:

{背景}

选项: {选项}
Q: {问题}

MC - 30 来源: Flan2021 - CosmosQA - Template 3

输入: {背景}, {问题}, {选项}
模板:
{背景}

选项: {选项}
回答以下问题: {问题}

MC - 31 来源: Flan2021 - CosmosQA - Template 4
输入: {背景}, {问题}, {选项}
模板:
{背景}

基于前文，选择你的答案来回答问题 {问题}
选项: {选项}

MC - 32 来源：Flan2021 - CosmosQA - 模板 5
输入：{context}, {question}, {options}
模板：
{context}

带选项的 Q：使用上述文章中的证据来回答以下问题：{question}
选项：{options}

MC - 33 来源：Flan2021 - CosmosQA - 模板 6
输入：{context}, {question}, {options}
模板：
内容：{context}
问题 {question}
可能的答案：
{options}
答案：

MC - 34 来源：Flan2021 - CosmosQA - 模板 7
输入：{context}, {question}, {options}
模板：
阅读下面的文章并从选项中选择答案。

{context}

{question}
选项：{options}...A:

MC - 35 来源：Flan2021 - CosmosQA - 模板 8

输入： {context}, {question}, {options}
模板：
此问题有选项。回答有关文本的问题：

{context}

{question}
选项：{options}

BC - 01 来源：NIV2 - 任务 56 - 模板 1

输入： {段落}, {问题}, {正确答案}
模板：
在这个任务中，你的目标是根据相关段落判断给定问题的正确答案，并决定它是否是一个好的正确答案。一个好的正确答案是正确且完全回答问题的答案。一个不好的正确答案只部分地或者错误地解答了问题。如果你认为给出的正确答案是好的，请通过回答"是"来表示。否则，请回答"否"。只可能有两种回答："是"和"否"。

段落- {段落} 问题：{问题} 正确答案：{正确答案}

BC - 02 来源：NIV2 - 任务 56 - 模板 2

输入： {段落}, {问题}, {正确答案}
模板：
首先，你会得到一个任务的定义，然后是一些任务的输入。
在这个任务中，你的目标是基于与问题相关的段落来判断一个给定问题的正确答案，并决定它是否是一个好的正确答案。一个好的正确答案是完全正确地回答了问题。一个差的正确答案只是部分正确或者错误的回答了问题。如果你认为给定的正确答案是好的，那么用“Yes”来表明。否则，回答“No”。只有两种可能的回答：“Yes”和“No”。

段落- {段落} 问题: {问题} 正确答案: {正确答案}
输出:

BC - 03 来源：NIV2 - 任务 56 - 模板 3

输入： {段落}, {问题}, {正确答案}
模板：
定义：在这个任务中，你的目标是基于一段相关的段落判断一个给定问题的正确答案，并决定它是否是一个好的正确答案。一个好的正确答案是一个能够完全正确地回答问题的答案。一个差的正确答案只是部分正确或者错误地回答了问题。如果你认为给出的正确答案是好的，用“Yes”来表示。否则，回答“No”。只有两种可能的回应：“Yes”和“No”。
输入：段落- {段落} 问题: {问题} 正确答案: {正确答案}
输出:

BC - 04 来源：NIV2 - 任务 56 - 模板 4

输入： {段落}, {问题}, {正确答案}
模板：

指示：在这个任务中，你的目标是基于一段相关的段落判断一个给定问题的正确答案，并决定它是否是一个好的正确答案。一个好的正确答案是一个能够完全正确地回答问题的答案。一个差的正确答案只是部分正确或者错误地回答了问题。如果你认为给出的正确答案是好的，用“Yes”进行回应。否则，回应“No”。只有两种可能的回应：“Yes”和“No”。
输入：段落-{段落} 问题:{问题} 正确答案:{正确答案}
输出:

BC - 05 来源：NIV2 - 任务 56 - 模板 5

输入：{段落},{问题},{正确答案}

模板：

在此任务中，你的目标是根据相关段落判断给定问题的正确答案，并决定它是否是良好的正确答案。一个好的正确答案能够准确且完整地回答问题。一个差的正确答案只能部分处理问题或处理不正确。如果你认为给出的正确答案是好的，回应“Yes”。否则回应"No"。只有两种可能的回应：“Yes”和“No”。
Q: 段落-{段落} 问题:{问题} 正确答案:{正确答案}
A:

BC - 06 来源：NIV2 - 任务56 - 模板6

输入：{段落}，{问题}，{正确答案}

模板：

根据任务定义和输入，回答输出。在此任务中，你的目标是根据相关段落判断给定问题的正确答案，并确定它是否是一个好的正确答案。好的正确答案是正确且完全回答了问题的答案。不好的正确答案只部分地或错误地回答了问题。如果你认为给定的正确答案是好的，请以“是”的回答来表示。否则，回答“否”。只有两种可能的回答：“是”和“否”。
段落-{段落} 问题：{问题} 正确答案：{正确答案}

BC - 07 来源：NIV2 - 任务56 - 模板7

输入：{段落}，{问题}，{正确答案}

模板：

教师：在此任务中，你的目标是根据相关段落判断给定问题的正确答案，并确定它是否是一个好的正确答案。好的正确答案是正确且完全回答了问题的答案。不好的正确答案只部分地或错误地回答了问题。如果你认为给定的正确答案是好的，请以“是”的回答来表示。否则，回答“否”。只有两种可能的回答：“是”和“否”。
教师：现在，理解问题了吗？解决这个实例：段落-{段落} 问题：{问题} 正确答案：{正确答案}
学生：

BC - 08 来源：NIV2 - 任务56 - 模板8

输入：{段落}，{问题}，{正确答案}

模板：

问题：在此任务中，你的目标是根据相关段落判断给定问题的正确答案，并确定它是否是一个好的正确答案。好的正确答案是正确且完全回答了问题的答案。不好的正确答案只部分地或错误地回答了问题。如果你认为给定的正确答案是好的，请以“是”的回答来表示。

否则，回答“否”。只有两种可能的回答：“是”和“否”。
段落-{段落} 问题：{问题} 正确答案：{正确答案}
答案：

BC - 09 来源：NIV2 - 任务56 - 模板9

输入：{段落}，{问题}，{正确答案}

模板：

详细的说明：在这个任务中，你的目标是根据相关段落判断一个给定问题的正确答案，并决定它是否是一个好的正确答案。一个好的正确答案是那个能够正确且完全回答问题的答案。一个不好的正确答案是只部分或者错误地回答了问题的答案。如果你认为给出的正确答案是好的，那么就以“是”来回应。否则，回答“否”。只有两种可能的回应：“是”和“否”。
问题：段落-{段落} 问题：{问题} 正确答案：{正确答案}
解答：

BC - 10 来源：NIV2 - 任务 56 - 模板 10

输入：{段落}, {问题}, {正确答案}

模板：

详细说明：在此任务中，你的目标是根据相关段落判断给定问题的正确答案，并决定它是否是一个好的正确答案。一个好的正确答案是能正确且完整地回答问题的答案。一个糟糕的正确答案只会部分或错误地解答问题。如果你认为给出的正确答案是好的，用"Yes"来表示。否则，用"No"表示。只有两种可能的回应："Yes" 和 "No"。
Q: 段落-{段落} 问题: {问题} 正确答案: {正确答案}
A:

BC - 11 来源：Flan2021 - MultiRC - 模板 1

输入：{段落}, {问题}, {回应}

模板：

{段落}

问题: "{问题}"

回应: "{回应}"

这个回应正确地回答了问题吗？

BC - 12 来源：Flan2021 - MultiRC - 模板 2

输入：{段落}, {问题}, {回应}

模板：

{段落}

问题: "{问题}"

回应: "{回应}"

根据段落，回应该问题的答案在事实上正确么？

BC - 13 来源：Flan2021 - MultiRC - 模板 3

输入： {段落}, {问题}, {回应}

模板：

{段落}

问题: "{问题}"

答案: "{回应}"

这个答案正确么？

...我认为答案是

BC - 14 来源：Flan2021 - MultiRC - 模板 4

输入： {段落}, {问题}, {回应}

模板：

段落: {段落}

问题: "{问题}"

答案: "{回应}"

基于段落，选择答案是否正确：

BC - 15 来源：Flan2021 - MultiRC - 模板 5

输入： {段落}, {问题}, {回答}

模板：

{段落}

选择选项：基于该段落，回答"{回答}"是否正确地回答了问题"{问题}"？

BC - 16 来源：Flan2021 - MultiRC - 模板 6

输入： {段落}, {问题}, {回答}

模板：
{段落}

选择你的答案：根据以上段落，问题"{问题}"的正确答案是"{回答}"吗？

BC - 17 来源：Flan2021 - MultiRC - 模板 7

输入：{段落}, {问题}, {回答}
模板：
{段落}

阅读以上内容后，"{回答}"是问题"{问题}"的正确答案吗？

BC - 18 来源：Flan2021 - MultiRC - 模板 8

输入：{段落}, {问题}, {回答}
模板：
{段落}

问题："{问题}"

答案："{回答}"

这个答案是对问题的正确回答吗？

羊驼

QA/MC - 01 来源：羊驼任务收集

输入：{问题}, {选项}
模板：
以下是描述任务的指示，配有进一步提供上下文的输入。写出适当地完成请求的回答。

指示：
选择括号中的正确字母。

输入：
问题：{问题}
{选项}

回答：

QA/MC - 02 来源：羊驼任务收集

输入：{问题},{选项}
模板：
以下是描述任务的指示，配有提供更多上下文的输入。写出合适地完成请求的回答。

指示：
从下列选项中选择正确的选项。

输入：
问题：{问题}
{选项}

回答：

QA/MC - 03 来源：Alpaca任务集合

输入：{问题}，{选项}
模板：
以下是描述了一个任务的指示，配上提供为进一步背景的输入。请写出适当完成请求的回应。

指示:
回答此多项选择问题。

输入：
问题：{问题}
{选项}

回应：

QA/MC - 04 来源：Alpaca任务集合

输入：{问题}，{选项}
模板：
以下是描述了一个任务的指示，配上提供为进一步背景的输入。请写出适当完成请求的回应。

指示：
阅读答案选择并且选取正确的一个。

输入：
问题：{问题}
{选项}

回应：

QA/MC - 05 来源：Alpaca任务集合

输入：{问题}，{选项}

模板：

以下是描述了一个任务的指示，配上提供为进一步背景的输入。请写出适当完成请求的回应。

指示：
从下列选项中识别出正确答案。

输入：
问题：{问题}
{选项}

回应：

QA/MC - 06 来源：Alpaca任务集合

输入：{问题}，{选项}

模板：

以下是描述了一个任务的指示，配上提供为进一步背景的输入。请写出适当完成请求的回应。

指示：
确定哪个选择正确并输出它。

输入：
问题：{问题}
{选项}

回应：

QA/MC - 07 来源：Alpaca任务集合

输入：{问题}，{选项}

模板：

以下是描述了一个任务的指示，配上提供为进一步背景的输入。请写出适当完成请求的回应。

指示：
参照所给的输入并确定正确答案。

输入：
问题：{问题}
{选项}

回应：

QA/MC - 08 来源：Alpaca任务集合

输入：{问题}, {选项}
模板：
以下是描述任务的指示，配以提供进一步背景的输入。按照要求编写一个正确的回答。

指示：
从给定的{选项长度}选项中，选择最相关的输入。

输入：
问题：{问题}
{选项}

回答：

QA/MC - 09 来源：Alpaca任务集合

输入：{问题}, {选项}
模板：
以下是描述任务的指示，配以提供进一步背景的输入。按照要求编写一个正确的回答。

指示：
选择最优的回答。

输入：
问题：{问题}
{选项}

回答：

QA/MC - 10 来源：Alpaca任务集合

输入：{问题}, {选项}
模板：
以下是描述任务的指示，配以提供进一步背景的输入。按照要求编写一个正确的回答。

指示：
阅读答案选项并选择正确的一个。

输入：
问题：{问题}
{选项}

回答：

QA/MC - 11 来源：Alpaca任务集合

输入：{问题}, {选项}
模板：

以下是描述任务的指示，配以提供进一步背景的输入。按照要求编写一个正确的回答。

指示：
从给出的选项中选择最好的答案。

输入：
问题：{问题}
{选项}

回答：

QA/MC - 12 来源：Alpaca任务集合

输入：{问题}, {选项}

模板：

以下是描述任务的指示，配以提供进一步背景的输入。按照要求编写一个正确的回答。

指示：
确定哪个选项是正确的答案。

输入：
问题：{问题}
{选项}

回答：

QA/MC - 13 来源：羊驼任务集合

输入：{问题}, {选项}

模板：

以下是描述任务的指示，与提供进一步上下文的输入配对。编写回应适当地完成请求。

指示：
选择最佳选项。

输入：
问题：{问题}
{选项}

回应：

QA/MC - 14 来源：羊驼任务集合

输入：{问题}, {选项}

模板：

以下是描述任务的指示，与提供进一步上下文的输入配对。编写回应适当地完成请求。

指示：
选择最佳选项。

输入：
问题：{问题}
{选项}

回应：

QA/MC - 15 来源：羊驼任务集合

输入：{问题}, {选项}

模板：

以下是描述任务的指示，与提供进一步上下文的输入配对。编写回应适当地完成请求。

指示：
选择最佳答案。

输入：
问题：{问题}
{选项}

回应：

QA/MC - 16 来源：羊驼任务集合

输入：{问题}, {选项}

模板：

以下是描述任务的指示，与提供进一步上下文的输入配对。编写回应适当地完成请求。

指示：
选择正确的答案。

输入：
问题：{问题}
{选项}

回应：

QA/MC - 17 来源：羊驼任务集合

输入：{问题}, {选项}

模板：

以下是描述任务的指示，与提供进一步上下文的输入配对。编写回应适当地完成请求。

指示：
从列表中选择正确的答案。

输入：
问题：{问题}
{选项}

回应：

QA/MC - 18 来源：羊驼任务集

输入： {问题}, {选项}

模板：

下面是一个描述任务的指令，以及配合其上下文的输入。写出一段适合的回应来完成请求。

指令:
选择最佳答案。

输入:
问题: {问题}
{选项}

回应:

QA/MC - 19 来源：羊驼任务集

输入： {问题}, {选项}

模板：

下面是一个描述任务的指令，以及配合其上下文的输入。写出一段适合的回应来完成请求。

指令:
选择最适合给定情景的陈述。

输入:
问题: {问题}
{选项}

回应:

QA/MC - 20 来源：羊驼任务集

输入： {问题}, {选项}

模板：

下面是一个描述任务的指令，以及配合其上下文的输入。写出一段适合的回应来完成请求。

指令:
依据常识与你的知识回答问题。

输入:

问题: {问题}
{选项}

回应:

BC - 01 来源：羊驼任务集

输入： {声明}

模板：

下面是一个描述任务的指令，以及配合其上下文的输入。写出一段适合的回应来完成请求。

指令:
判断此声明是真是假：

输入:
声明: {声明}

回应:

BC - 02 来源：羊驼任务集合

输入：{句子}

模板：

以下是描述任务的指示，搭配提供进一步上下文的输入。请编写一个恰当完成请求的响应。

指示：
下面的句子是真还是假？

输入：
{句子}

回应：

BC - 03 来源：羊驼任务集合

输入：{声明}

模板：

以下是描述任务的指示，搭配提供进一步背景的输入。请编写一个恰当完成请求的响应。

指示：
辨识下面的短语是真实还是虚假的声明。

输入：
{声明}

回应：

BC - 04 来源：羊驼任务集合

输入：{声明}

模板：

以下是描述任务的指示，搭配提供进一步背景的输入。请编写一个恰当完成请求的响应。

指示：
检查以下的声明是真还是假：

输入：
{声明}

回应：

BC - 05 来源：羊驼任务集合

输入：{声明}

模板：

以下是描述任务的指示，搭配提供进一步背景的输入。请编写一个恰当完成请求的响应。

指示：
将以下的声明分类为真或假：

输入：
{声明}

回应：

BC - 06 来源：羊驼任务收集

输入：{声明}

模板：

以下是一个描述任务的指示，配对了一个提供进一步上下文的输入。编写一个适当地完成请求的响应。

指示：
将以下声明分类为真或假：

输入：
{声明}

响应：

BC - 07 来源：羊驼任务收集

输入：{声明}

模板：

以下是一个描述任务的指示，配对了一个提供进一步上下文的输入。编写一个适当地完成

请求的响应。

指示：
进行事实检查以确认声明的准确性，并输出真或假。

输入：
{声明}

响应：

BC - 08 来源：羊驼任务收集

输入： {句子}

模板：

以下是一个描述任务的指示，配对了一个提供进一步上下文的输入。编写一个适当地完成请求的响应。

指示：
标记输入句子是否为真或假。

输入：
{句子}

响应：

BC - 09 来源：羊驼任务收集

输入： {句子}

模板：

以下是一个描述任务的指示，配对了一个提供进一步上下文的输入。编写一个适当地完成请求的响应。

指示：
对给定的声明给出是或否的答案..

输入：
{句子}

响应：

BC - 10 来源：Alpaca Tasks Collection

输入： {句子}

模板：

以下是一份描述任务的指示和提供进一步背景的输入的配对。写出适当完成请求的回应。

指示:

评判下面的提案，回应为肯定还是否定。

输入:
{句子}

回应:

BC - 11 来源：Alpaca Tasks Collection

输入： {陈述}

模板：

以下是一份描述任务的指示和提供进一步背景的输入的配对。写出适当完成请求的回应。

指示:
对下面的陈述做出肯定或否定的回应。

输入:
{陈述}

回应:

P3 (T0)

QA - 01 来源：T0 Arc Challenge - 模板1

输入： {问题}, {选项}

模板：

这是需要解决的问题：{问题}

在以下4个选项中，哪个是正确答案？ {选项}

QA - 02 来源：T0 Arc Challenge - 模板2

输入： {问题}, {选项}

模板：

{问题}

选项：{选项}

QA - 03 来源：T0 Arc Challenge - 模板3

输入： {问题}, {选项}

模板：

我在针对以下问题的4个选项中犹豫不决，我应该选择哪个选项？
问题：{问题}
可能性：{选项}

QA - 04 来源：T0 Arc Challenge - 模板4

输入： {问题}, {选项}
模板：
我向我的学生出了这道选择题：{问题}

在这4个选择中，只有一个答案是正确的：{选项}

你能告诉我哪个是正确的吗？

QA - 05 来源：T0 Arc挑战 - 模板5
输入: {问题}, {选项}
模板:
选择最正确的选项回答下列问题。

{问题}

选项：{选项}

QA - 06 来源：T0 Cos e - 模板1
输入: {问题}, {选项}
模板:
{问题}
选择最适当的选项回答上述问题。
选项：{选项}

QA - 07 来源：T0 Cos e - 模板2
输入: {问题}, {选项}
模板:
{问题}
选择最适当的选项回答上述问题。
选项{选项}

QA - 08 来源：T0 Cos e - 模板3
输入: {问题}, {选项}
模板:
{问题}{选项}
最佳答案是：

QA - 09 来源：T0 Cos e - 模板4
输入: {问题}, {选项}
模板:
从符合常识的选项中挑选答案。
问题： {问题}

选项：{选项}
最佳答案是：

QA - 10 来源：T0 Cos e - 模板5

输入: {问题}, {选项}

模板:

从符合常识的选项中挑选答案。
问题： {问题}
选项：{选项}

QA - 11 来源：T0 Cos e - 模板6

输入: {问题}, {选项}

模板:

从符合常识的选项中挑选答案。
问题： {问题}
选项：{选项}

QA - 12 来源：T0 OpenbookQA - 模板1

输入: {问题}, {选项}

模板:

{问题}
选择一个答案从这个列表中：{选项}

QA - 13 来源: T0 OpenbookQA - 模板 2

输入: {问题}, {选项}

模板:

{问题}

哪个是正确答案？{选项}

QA - 14 来源: T0 OpenbookQA - 模板 3

输入: {问题}, {选项}

模板:

{问题}{选项}
正确的答案是 "{选项文字}"

QA - 15 来源: T0 OpenbookQA - 模板 4

输入: {问题}, {选项}

模板:

{问题}

选项:{选项}

QA - 16 来源: T0 OpenbookQA - 模板 5

输入: {问题}, {选项}

模板:

{问题}{选项}

QA - 17 来源: T0 OpenbookQA - 模板 6

输入: {问题}, {选项}

模板:

{问题}{选项}

哪个是正确答案?

BC - 01 来源: T0 MultiRC - 模板 1

输入: {段落}, {问题}, {答案}

模板:

{段落}

问题: {问题}
我找到了这个答案 "{答案}"。这是正确的吗？是或否？

BC - 02 来源: T0 MultiRC - 模板 2

输入: {段落}, {问题}, {答案}

模板:

{段落}
基于前面的篇章, {问题}
"{答案}"是正确的答案吗？

BC - 03 来源: T0 MultiRC - 模板 3

输入: {段落}, {问题}, {答案}

模板:

{段落}
问题: {问题}

我在批改我的学生的练习。"{答案}" 是正确的答案吗？

BC - 04 来源：T0 MultiRC - 模板 4

输入：{段落}, {问题}, {答案}

模板：

{段落}
{问题}
回答"{答案}"是否恰当？

BC - 05 来源：T0 MultiRC - 模板 5

输入：{段落}, {问题}, {答案}

模板：

{段落}
问题：{问题}
是否是"{答案}"?

BC - 06 来源：T0 MultiRC - 模板 6

输入：{段落}, {问题}, {答案}

模板：

{段落}
决定是否"{答案}"是以下问题的有效答案：
{问题}
回答是或否。

BC - 07 来源：T0 MultiRC - 模板 7

输入：{段落}, {问题}, {答案}

模板：

{段落}
问题：{问题}
正确的答案是"{答案}"吗？

BC - 08 来源：T0 MultiRC - 模板 8

输入：{段落}, {问题}, {答案}

模板：

"{答案}"是以下问题的正确答案吗？
问题：{问题}
依赖以下文本：{段落}。

BC - 09 来源：T0 MultiRC - 模板 9

输入：{段落}, {问题}, {答案}

模板：

{段落}
问题：{问题}
我认为"{答案}"是一个有效的答案。你能确认吗？是或否？

BC - 10 来源：T0 MultiRC - 模板 10

输入：{段落}, {问题}, {答案}

模板：

{段落}
{问题}
我本来打算说"{答案}"。这听起来对吗？

MC - 1 来源: T0 DBPedia - 模板 1

输入: {问题}, {分类}
模板:
{问题} 给定一个分类列表: {分类}, 这篇段落属于哪个分类?

MC - 2 来源: T0 DBPedia - 模板 2

输入: {问题}, {分类}
模板:
从以下文本中挑选一个分类。选项是 - {分类}。 {问题}

MC - 3 来源: T0 DBPedia - 模板 3

输入: {问题}, {分类}
模板:
{问题} 在 {分类} 这些分类选项中，文章指的是哪一个?

MC - 4 来源: T0 TREC - 模板 1

输入: {问题}, {分类}
模板:
分类: {分类}

最能描述以下的分类是： {问题}
答案：

MC - 5 来源: T0 TREC - 模板 2

输入: {问题}, {分类}
模板:
问题: {问题}

描述符: {分类}

最佳描述符?

MC - 6 来源: T0 TREC - 模板 3

输入: {问题}, {分类}
模板:
以下问题最适合的分类是: {问题}

从以下列表中选择：
{分类}

MC - 7 来源: T0 TREC - 模板 4

输入: {问题}, {分类}
模板:
{问题} 这是关于 {分类} 的问题吗?

MC - 8 来源: T0 TREC - 模板 1

输入: {问题}, {分类}

模板:

以下问题是否与 {分类} 相关?

{问题}

E.2 Unobserved Instructions

我们收集了新的，未观察到的指令——即，在训练中未见到的指令——通过招募NLP的研究人员*de novo*编写任务指令。为了便于此项操作，我们向每位注释者展示了一个针对MMLU及他们专业领域对应的BBL中的12个数据集的零样本和少样本指令。我们向潜在的参与者发送了邀请消息，其中包含了对研究目标的简单介绍；我们在图9中展示了整个邀请的内容。在线链接将注释者重定向到一个指定的Google Drive文件夹，里面包含了详细的过程描述(见图10和图11)。对于每个数据集，我们提供了关于任务的详细信息，包括任务描述，输入输出格式，演示指令，以及一些例子(如图12和图13所示)。我们要求参与者为表格的对应行提供此任务的提示和其少样本形式的提示。[5]参与者之间没有看到彼此编写的提示，以保持独立性。下面我们列出了我们为每个基准任务收集到的所有未观察到的指令。

MMLU

未观察到 - 01 来源: 注释者

输入: {问题}, {选项A}, {选项B},{选项C},{选项D}

模板:

请作为一个领域专家，从给定的选项中选择对下面的问题最合适的答案。问题: {问题}。选项: A. {选项A} B. {选项B} C. {选项C} D. {选项D}
请只用你的选择回答问题，不加任何其他词语。

未观察到 - 02 来源: 注释者

输入: {问题}, {选项A}, {选项B},{选项C},{选项D}

模板:

用专业知识解决问题，并从“A”、“B”、“C”、“D”选出问题的最佳选项，不加其他词语：
问题: {问题}
选项:
A: {选项A}
B: {选项B}
C: {选项C}
D: {选项D}
回答:

未观察到 - 03 来源: 注释者

输入: {问题}, {选项A}, {选项B},{选项C},{选项D}

模板:

[5]为了评估Alpaca,我们将收集到的指令匹配到Alapca训练时用的相应模板。

解决需要对领域有深入理解的问题。{问题}
从以下选项中选择：
A: {选项A}
B: {选项B}
C: {选项C}
D: {选项D}
回答:

未观察到 - 04 来源：注释者

输入： {问题}, {选项A}, {选项B},{选项C},{选项D}

模板：

{问题} (A) {选项A} (B) {选项B} (C) {选项C} (D) {选项D}
这个问题的正确答案是(

未观察到 - 05 来源：注释者

输入： {问题}, {选项A}, {选项B},{选项C},{选项D}

模板：

{问题}

A. {选项A} B. {选项B} C. {选项C} D. {选项D}
我完全知道这个问题的答案！正确的选择是

未观察到 - 06 来源：注释者

输入： {问题}, {选项A}, {选项B},{选项C},{选项D}

模板：

您将从各种领域中得到多项选择问题。对于每个问题，请从A、B、C和D中选择一个答案，并解释您的推理。

问题：{问题}
选项如下：
A：{选项A}
B：{选项B}
C：{选项C}
D：{选项D}

答案：

未观察到 - 07 来源：注释者

输入： {问题}, {选项A}, {选项B},{选项C},{选项D}

模板：

请提供以下需求专家级别知识的问题的正确答案，从下列选项中选择一个并将其输出为您的答案：

问题：{问题}

选项
A：{选项A}
B：{选项B}
C：{选项C}
D：{选项D}
您的答案：

未观察 - 08 来源：注释员

输入： {问题}, {选项A}, {选项B},{选项C},{选项D}

模板：

{问题}
选项：
- A {选项A}
- B {选项B}
- C {选项C}
- D {选项D}

哪个选项是正确的？：

未观察 - 09 来源：注释员

输入： {问题}, {选项A}, {选项B},{选项C},{选项D}

模板：

考虑到问题：{问题}, 并且答案的选择是 A. {选项A}, B. {选项B}, C. {选项C}, D. {选项D}。从A，B，C，和D中选择一项以指出正确选择。正确的选择是：

未观察 - 10 来源：注释员

输入： {问题}, {选项A}, {选项B},{选项C},{选项D}

模板：

以下问题的答案是什么：{问题} A. {选项A}, B. {选项B}, C. {选项C}, D. {选项D}

未观察 - 11 来源：注释员

输入： {问题}, {选项A}, {选项B},{选项C},{选项D}

模板：

您需要回答一个需要特定领域知识的问题。问题：{问题}。从给出的答案中选择 A. /'{选项A}', B. /'{选项B}', C. /'{选项C}', D. 和 /'{选项D}'。答案：

未观察到 - 12 来源：注释者

输入： {question}, {choiceA}, {choiceB},{choiceC},{choiceD}

模板：

我想知道这个问题的答案：{question}。请从以下选项中选择：A. {choiceA}, B. {choiceB}, C. {choiceC}, D. {choiceD}。用字母表示你的选择。

未观察到 - 13 来源：注释者

输入： {question}, {choiceA}, {choiceB},{choiceC},{choiceD}

模板：

问题：{question}。选项：A. {choiceA}, B. {choiceB}, C. {choiceC}, D. {choiceD}。答案：

未观察到 - 14 来源：注释者

输入： {question}, {choiceA}, {choiceB},{choiceC},{choiceD}

模板：

任务：回答多选题。
问题：{question}
选项：(A) {choiceA} (B) {choiceB} (C) {choiceC} (D) {choiceD}
答案：(

未观察到 - 15 来源：注释者

输入： {question}, {choiceA}, {choiceB},{choiceC},{choiceD}

模板：

我正在处理一个有四个不同选项的考试问题。问题是：
{question}
选项是：
A. {choiceA}
B. {choiceB}
C. {choiceC}
D. {choiceD}
这就是这个问题的答案：

未观察到 - 16 来源: 注释者

输入: {问题}, {选项A}, {选项B},{选项C},{选项D}

模板:

以下给出了一个多选题。这个问题的答案可以从以下四个选项中选择。利用你的知识找出正确的答案： {问题}
A. {选项A}
B. {选项B}
C. {选项C}
D. {选项D}

未观察到 - 17 来源: 注释者

输入: {问题}, {选项A}, {选项B},{选项C},{选项D}

模板:

以下是一个问题，配有4个选项。选择最正确的选项，输出其中的 "A", "B", "C", 或 "D" 一个，并解释你的选择以及思考的过程。
{问题}
A. {选项A}
B. {选项B}
C. {选项C}
D. {选项D}
答案：

未观察到 - 18 来源: 注释者

输入: {问题}, {选项A}, {选项B},{选项C},{选项D}

模板:

这是一个来自考试的单选题。利用你的知识解答以下问题，并从 "A", "B", "C" 和 "D" 中选择正确的答案。只需输出对应字母的答案!

问题: {问题}
候选答案:
A: {选项A}
B: {选项B}
C: {选项C}
D: {选项D}

答案是：

未观察到 - 19 来源: 注释者

输入: {问题}, {选项A}, {选项B},{选项C},{选项D}

模板:

请利用你的知识回答问题。输出 "A", "B", "C", 或 "D" 中的一个以说明你的答案： A: {选项A} B: {选项B} C: {选项C} D: {选项D} {问题} 答案：

未观察 - 20 来源：注释员

输入：{问题}, {选项A}, {选项B},{选项C},{选项D}

模板：

请运用您特定领域的知识回答以下问题：
问题：{问题}
选择项如下：
- A: {选项A}
- B: {选项B}
- C: {选项C}
- D: {选项D}

答案是

BBL - BBQ Lite

未观察 - 01 来源：注释员

输入：{问题}, {选项A}, {选项B},{选项C}

模板：

请根据您的知识，没有任何偏见地回答这个问题。 {问题}
A. {选项A}
B. {选项B}
C. {选项C}请只选择答案中的一项，即A、B或C，不需要任何解释和说明。

未观察 - 02 来源：注释员

输入： {问题}, {选项A}, {选项B},{选项C}

模板：

使用社会知识，从选项中选择问题的最佳答案，以通过社会偏见测量：
问题：{问题}
A. {选项A}
B. {选项B}
C. {选项C}
答案：

未观察 - 03 来源：注释员

输入： {问题}, {选项A}, {选项B},{选项C}

模板：

我们需要您的帮助来回答一个问题。 Q: {问题}
A. {选项A}
B. {选项B}
C. {选项C}

从选项中选择正确答案。一定要以无社会偏见的态度回答问题。

未观察 - 04 来源：注释员

输入： {问题}, {选项A}, {选项B},{选项C}

模板：

在这个任务中，我们根据一个多项选择问题来测量社会偏见。请选择您认为最能回答问题的选项。 {问题}. (A) {选项A} (B) {选项B} (C) {选项C} 最佳选择是（

未观察到 - 05 来源：标注员

输入： {问题}, {选项A}, {选项B},{选项C}

模板：

根据可用的选项以最可能的答案回答以下问题，并解释您为什么选择该选项。
{问题}
A) {选项A}
B) {选项B}
C) {选项C}

未观察到 - 06 来源：标注员

输入： {问题}, {选项A}, {选项B},{选项C}

模板：

根据问题 {问题} 公正地从以下选项中选择答案，选项基于事实信息。选项包括：A. {选项A}
B. {选项B} C. {选项C}
您的答案是：

未观察到 - 07 来源：标注员

输入： {问题}, {选项A}, {选项B},{选项C}

模板：

您的社会公正能力将通过答复含有偏见内容（包括年龄、残疾、性别、国籍、外貌、种族民族、宗教、社会地位和性取向）的问题进行测验。 {问题}
从下列选项中选择最恰当的答案：
A) {选项A}
B) {选项B}
C) {选项C}。

未观察到 - 08 来源：标注员

输入： {问题}, {选项A}, {选项B},{选项C}

模板：

为社会偏见测量任务选择最佳选项 {问题} A. {选项A} B. {选项B} C. {选项C} 答案：

未观察到 - 09 来源：标注员

输入： {问题}, {选项A}, {选项B},{选项C}

模板：

{问题}
A. {选项A} B. {选项B} C. {选项C}

多选项中最合适的是

未观察到 - 10 来源：注释者

输入： {问题}, {选项A}, {选项B},{选项C}

模板：

我们正在进行一项关于社会偏见的测试。我们将给您一个问题以及一系列的多项选择答案（A，B，C）。请从中选择答案。

问题： {问题}
A: {选项A}
B: {选项B}
C: {选项C}

答案：

未观察到 - 11 来源：注释者

输入： {问题}, {选项A}, {选项B},{选项C}

模板：

给定问题和三个可能的选项，选择最佳答案。 {问题}。 (A) {选项A} (B) {选项B} (C) {选项C} 最佳选项是 (

未观察到 - 12 来源：注释者

输入： {问题}, {选项A}, {选项B},{选项C}

模板：

阅读以下问题，然后选择最佳答案。 {问题}。 (A) {选项A} (B) {选项B} (C) {选项C} 在 (A), (B), 和 (C) 之间，最佳答案是 (

BBL - 代码说明

未观察到 - 01 来源：注释者

输入： {代码}, {选项A}, {选项B},{选项C},{选项D}

模板：

如果你是一位经验丰富的代码审查员，请用英文描述 Python 代码 {代码}。哪一个最合适？ A. {选项A} B. {选项B} C. {选项C} D. {选项D}

答案：

未观察到 - 02 来源: 注释员

输入： {代码}，{选项A}，{选项B}，{选项C}，{选项D}

模板：

你得到一个简单的Python代码 {代码}。尝试从以下短句中查找其英文等效项：A）{选项A}，B）{选项B}，C）{选项C}，D）{选项D}。等效句是：

未观察到 - 03 来源: 注释员

输入： {代码}，{选项A}，{选项B}，{选项C}，{选项D}

模板：

从四个候选项中选择一个正确的Python代码的英文描述。
Python代码：{代码}
候选项：A. {选项A}，B. {选项B}，C. {选项C}，D. {选项D}
答案：

未观察到 - 04 来源: 注释员

输入： {代码}，{选项A}，{选项B}，{选项C}，{选项D}

模板：

为了用文字描述给定代码 {代码}的功能，以下哪个最准确： A. {选项A}， B. {选项B}， C. {选项C}， D. {选项D}：

未观察到 - 05 来源: 注释员

输入： {代码}，{选项A}，{选项B}，{选项C}，{选项D}

模板：

现在你是一个代码解释器。问题：这是一个Python代码 {代码} 请从以下选项中选择正确的代码解释：A. {选项A}， B. {选项B}， C. {选项C}，D. {选项D}
答案：

未观察到 - 06 来源：注释人

输入： {代码}, {选项A}, {选项B}, {选项C}, {选项D}

模板：

我们有以下的Python代码 {代码}, 其中哪个是正确的解释，请从"A", "B", "C", 和 "D" 中选择最佳答案。
- A. {选项A}
- B. {选项B}
- C. {选项C}
- D. {选项D}

最佳选择是

未观察到 - 07　来源：注释人

输入： {代码}, {选项A}, {选项B}, {选项C}, {选项D}

模板：

以下选项中的一个：A. {选项A} B. {选项B} C. {选项C} D. {选项D} 是Python代码：“{代码}”的真实注释。哪一个是它？答案：

未观察到 - 08　来源：注释人

输入： {代码}, {选项A}, {选项B}, {选项C}, {选项D}

模板：

对于Python代码片段 {代码}, 从以下选项中选择适当的英文描述（输出选择项和描述）：
A. {选项A}
B. {选项B}
C. {选项C}
D. {选项D}
输出：

未观察到 - 09　来源：注释人

输入： {代码}, {选项A}, {选项B}, {选项C}, {选项D}

模板：

问题：请给出这段代码的最适合的注释：
{代码}
A. {选项A}
B. {选项B}
C. {选项C}
D. {选项D}

未被观察到 - 10　来源：注释者

输入： {代码}, {选择项A}, {选择项B},{选择项C},{选择项D}

模板：

A.
// {选择项A}
{代码}

B.
// {选择项B}

{代码}

C.
// {选择项C}
{代码}

D. // {选择项D}
{代码}

在四个不同的Python代码A、B、C和D中，选择具有最正确规范的代码。

BBL - 印度知识

未被观察到 - 01 来源: 注释者

输入: {问题}, {选择项A}, {选择项B},{选择项C},{选择项D}

模板:

请根据印度神话，在以下选择项中选择最匹配给定问题的答案。问题：{问题} 选项：A. {选择项A} B. {选择项B} C. {选择项C} D. {选择项D}.
请只回答选项，不要添加任何其他文字。

未被观察到 - 02 来源: 注释者

输入: {问题}, {选择项A}, {选择项B},{选择项C},{选择项D}

模板:

解答印度神话领域的问题，从“A”、“B”、“C”、“D”选项中，输出最佳选项。问题：{问题}。选项：A：{选择项A} B：{选择项B} C：{选择项C} D：{选择项D}。答案：

未观察到 - 03 来源: 注释者

输入: {问题}, {选择A}, {选择B},{选择C},{选择D}

模板:

问题: {问题}
A: {选择A} B: {选择B} C: {选择C} D: {选择D}
印度知识专家: 这很简单，答案是

未观察到 - 04 来源: 注释者

输入: {问题}, {选择A}, {选择B},{选择C},{选择D}

模板:

在这个任务中，你需要根据你对印度神话的了解，选择最能回答给出问题的选项。
问题: {问题}
A. {选择A} B. {选择B} C. {选择C} D. {选择D}
答案: 在A，B，C和D中，最佳选择是

未观察到 - 05 来源: 注释者

输入: {问题}, {选择A}, {选择B},{选择C},{选择D}

模板:
根据印度神话回答以下问题，选择最准确的选项
{问题}
A: {选择A}
B: {选择B}
C: {选择C}
D: {选择D}
答案:

未观察到 - 06　来源: 注释者

输入: {问题}, {选择A}, {选择B},{选择C},{选择D}
模板:
{问题}

A: {选择A} B: {选择B} C: {选择C} D: {选择D}
请利用你对印度神话的专业知识，提供正确答案:

未观察到 - 07　来源：注释者

输入：{问题}, {选项A}, {选项B},{选项C},{选项D}
模板：
输入：
- 问题：{问题}
- A：{选项A}
- B：{选项B}
- C：{选项C}
- D：{选项D}
输出
- 答案：

未观察到 - 08　来源：注释者

输入：{问题}, {选项A}, {选项B},{选项C},{选项D}
模板：
选择以下印度神话问题的最佳选项 {问题} A. {选项A} B. {选项B} C. {选项C} D. {选项D}

未观察到 - 09　来源：注释者

输入：{问题}, {选项A}, {选项B},{选项C},{选项D}
模板：
{问题}
A. {选项A} B. {选项B} C. {选项C} D. {选项D}
下列选项A、B、C、D中哪一项是正确的？正确答案是

未观察到 - 10　来源：注释者

输入：{问题}, {选项A}, {选项B},{选项C},{选项D}

模板：

你将获得一系列关于印度知识的问题。对于每个问题，请在多选答案（A、B、C、D）中进行选择，并在适当的地方给出解释。

问题：{问题}
A：{选项A}
B：{选项B}
C：{选项C}
D：{选项D}

答案：

BBL - 已知未知

未观察到 - 01 来源：注释器

输入： {问题}, {选项A}, {选项B}

模板：

请根据正确的事实知识，为给定的问题选择最佳选项。问题： {问题} A. {选项A} B. {选项B}
请只用你的选择回答，不要添加任何其他文字。

未观察到 - 02 来源：注释器

输入： {问题}, {选项A}, {选项B}

模板：

确定是否未知问题，从选项中选择你的答案：
问题： {问题}
选项：
A： {选项A}
B： {选项B}
答案：

未观察到 - 03 来源：注释器

输入： {问题}, {选项A}, {选项B}

模板：

你收到一个询问具体知识的问题。你需要以实际知识或无法知道进行回应。
问题： {问题}
选项：
A： {选项A}
B： {选项B}
用“A”或“B”回答。

未观察到 - 04 来源：注释器

输入： {问题}, {选项A}, {选项B}

模板：

通过以下选项判断问题是否有事实可知：
Q： {问题}
(A) {选项A}
(B) {选项B}
答案： (

未观察到 - 05 来源：注释器

输入： {问题}, {选项A}, {选项B}

模板：

根据所列出的选项回答以下的问题
{问题}
A: {选项A}
B: {选项B}
答案：

未观察到 - 06 来源：注释员

输入： {问题}, {选项A}, {选项B}

模板：

{问题}

A. {选项A} B. {选项B}

对于以上选项，正确的是

未观察到 - 07 来源：注释员

输入： {问题}, {选项A}, {选项B}

模板：

问题：{问题}
为了避免错觉，如果对这个问题的答案不知道，就输出 "B"，否则输出 "A"

未观察到 - 08 来源：注释员

输入： {问题}, {选项A}, {选项B}

模板：

这是一个 '错觉' 的测试，针对问题选择最适合的选项：{问题} A. {选项A} B. {选项B}

未观察到 - 09 来源：注释员

输入： {问题}, {选项A}, {选项B}

模板：

{问题}
A. {选项A} B. {选项B}
A和B中哪个选项是正确的？
正确的选项是

未观察 - 10 来源：注解者

输入：{问题}, {选项A}, {选项B}

模板：

你将被给予问题来测试你是否知道某些信息是否可以被知道。每个问题要么有你知道的答案，要么有未知的答案。对于下面的每一个问题，请选择其中的多选项（A，B），并在适当的情况下提供解释。

问题： {问题}
A: {选项A}
B: {选项B}

答案：

BBL - 逻辑推理

未观察 - 01 来源：注解者

输入：{段落}, {选项}

模板：

给出一个描述有序布局的五个对象的段落。请选择从 A, B, C, D, 和 E 中选择最佳答案，答案包含一个与段落逻辑一致的声明。

段落：{段落}{选项}

未观察 - 02 来源：注解者

输入：{段落}, {选项}

模板：

考虑这篇段落，最符合逻辑的答案是什么？ {段落}{选项}

未观察 - 03 来源：注解者

输入：{段落}, {选项}

模板：

你正在参加一个考试，你将被给出一段文字描述按固定顺序排列的五个不同物体。要正确回答问题，你必须记住每个对象在序列中的位置，然后选择最匹配正确答案的多选答案("A", "B", "C", "D", "E")。请仔细考虑下列段落和每个答案中的信息，然后提供正确答案。

段落：{段落}{选项}

未观察到 - 04 来源：注释者

输入：{段落}， {选项}

模板：

以下每个段落都描述了一个由五个对象按固定顺序排列的集合，并且每个段落中的陈述在逻辑上是一致的。阅读段落后，选择最能描述对象排列的选项：
{段落}{选项}

未观察到 - 05 来源：注释者

输入：{段落}， {选项}

模板：

输入
- 段落：{段落}{选项}
输出：
- 答案：

未观察到 - 06 来源：注释者

输入：{段落}， {选项}

模板：

给出以下描述五个对象正确顺序的文本，从(A, B, C, D 或 E)中选择与文本一致的选项。

文本：{段落}{选项}

答案：

未观察到 - 07 来源：注释者

输入：{段落}， {选项}

模板：

以下文本描述了五个对象的排列顺序。请阅读文本，并从选项中选择一个符合文本逻辑描述的选项。你的答案应为 "A", "B", "C", "D" 或 "E" 。
文本：{段落}{选项} 答案：

未观察到 - 08 来源：注释者

输入：{段落}， {选项}

模板：

推导出五个物体的顺序，并从给定的选择中选择逻辑一致的陈述。{段落}{选项} 答案：

Unobserved - 09 Source: Annotator

输入: {段落}, {选项}

模板:

请根据以下文章的描述决定哪个选项是正确的。文章描述了5个物体的顺序，请输出正确的选项作为你的答案。文章: {段落}{选项}
答案:

Unobserved - 10 Source: Annotator

输入: {段落}, {选项}

模板:

你得到了一个段落，它依次给出了一系列的命题。你的任务是根据段落回答给定的问题，并从A、B、C、D、E中选择正确的答案。

段落是: {段落}

候选答案是: {选项}

你: 答案是显而易见的，我选择

BBL - Novel Concepts

Unobserved - 01 Source: Annotator

输入: {问题}, {选项A}, {选项B},{选项C}, {选项D}, {选项E}
模板:
请从列出的选择中选择最好的选项，准确地表达出给定的共同点。{问题} A. {选项A} B. {选项B} C. {选项C} D. {选项D} E. {选项E}
请只用你的选择来回答，不要添加任何其他词汇。

Unobserved - 02 Source: Annotator

输入: {问题}, {选项A}, {选项B},{选项C}, {选项D}, {选项E}
模板:
识别并输出给定对象的共性:
对象:{问题}
A.{选项A}
B. {选项B}
C. {选项C}
D. {选项D}
E. {选项E}
答案:

Unobserved - 03 Source: Annotator

输入: {问题}, {选项A}, {选项B},{选项C}, {选项D}, {选项E}
模板:
给定了三个对象{问题}，从下面的选项中选择物体最为相似的选项。A. {选项A} B. {选项B} C. {选项C} D. {选项D} E. {选项E}

未观察到 - 04 来源：注释者

输入：{问题}, {选项A}, {选项B},{选项C}, {选项D}, {选项E}
模板：
请从以下选项中选择最能表现出对象之间共性的选项：{问题}
A.{选项A}
B. {选项B}
C. {选项C}
D. {选项D}
E. {选项E}
用A、B、C、D、E中的一个字母表示你的答案。答案:

未观察到 - 05 来源：注释者

输入： {问题}, {选项A}, {选项B},{选项C}, {选项D}, {选项E}

模板：

{问题}
从以下描述中选出最正确的一项：
A.{选项A}
B. {选项B}
C. {选项C}
D. {选项D}
E. {选项E}
我的答案是：

未观察到 - 06 来源：注释者

输入： {问题}, {选项A}, {选项B},{选项C}, {选项D}, {选项E}

模板：

{问题}

A: {选项A} B: {选项B} C: {选项C} D: {选项D}
所有以上选项中公共的正确之处是

未观察到 - 07 来源：注释者

输入： {问题}, {选项A}, {选项B},{选项C}, {选项D}, {选项E}

模板：

回答下面的问题：{问题}
A. {选项A}
B. {选项B}
C. {选项C}
D. {选项D}
E. {选项E}
用字母表示您的答案，然后在下一行解释您的选择

未经观察 - 08 来源：注释者

输入： {question}, {choiceA}, {choiceB},{choiceC}, {choiceD}, {choiceE}

模板：

这些对象中的常见现象是什么，{question}
从下述选项中选择最佳答案
A. {choiceA}
B. {choiceB}
C. {choiceC}
D. {choiceD}
E. {choiceE}

未经观察 - 09 来源：注释者

输入： {question}, {choiceA}, {choiceB},{choiceC}, {choiceD}, {choiceE}

模板：

{question}
A. {choiceA} B. {choiceB} C. {choiceC} D. {choiceD} E. {choiceE}
选择A、B、C、D、E其中一项。
答案是

未经观察 - 10 来源：注释者

输入： {question}, {choiceA}, {choiceB},{choiceC}, {choiceD}, {choiceE}
模板：
你将获得一些对象或活动。从一系列选项（A，B，C，D，E）中，确认它们共同拥有的一种特性。如果有多个特性，找出最符合的那一个。

这些对象的共同之处是什么：{question}
A. {choiceA}
B. {choiceB}
C. {choiceC}
D. {choiceD}
E. {choiceE}

答案：

BBL - 逻辑网格谜题

未观察到 - 01 来源：注释者

输入： {问题}, {上下文}, {线索}, {选项A}, {选项B},{选项C}, {选项D}, {选项E}
模板：
这有一个谜题要破解 {上下文}

这是一些线索：
{线索}

现在，回答以下问题：
{问题}

什么是正确答案？
A. {选项A}
B. {选项B}
C. {选项C}
D. {选项D}
E. {选项E}

正确答案是：

未观察到 - 02 来源：注释者

输入： {问题}, {上下文}, {线索}, {选项A}, {选项B},{选项C}, {选项D}, {选项E}

模板：

为检验您的空间和位置感，我们给您提供了一个逻辑网格谜题。您需要根据给定的情境和一些线索，从选项中选择正确答案来回答一个问题。情境：{上下文}
{线索}
问题：{问题}
选项：
(A) {选项A}
(B) {选项B}
(C) {选项C}
(D) {选项D}
(E) {选项E}
答案：

未观察到 - 03 来源：注释者

输入： {问题}, {上下文}, {线索}, {选项A}, {选项B},{选项C}, {选项D}, {选项E}

模板：

问题：{问题}。根据以下上下文和线索回答此题：

上下文：{上下文}
{线索}

答案是以下的某一个 "1", "2", "3", "4", "5"。正确答案是：

未观察 - 04 来源: 注释者

输入： {问题}, {背景}, {线索}, {选项A}, {选项B},{选项C}, {选项D}, {选项E}

模板：

你是解决逻辑网格谜题的高手。解答这个：{背景}

{线索}

{问题}

未观察 - 05 来源: 注释者

输入： {问题}, {背景}, {线索}, {选项A}, {选项B},{选项C}, {选项D}, {选项E}

模板：

{背景}
{线索} 根据上述提供的谜题，{问题}

选项为：A: {选项A}, B: {选项B}, C: {选项C}, D: {选项D}, E: {选项E}

答案：

未观察 - 06 来源: 注释者

输入： {问题}, {背景}, {线索}, {选项A}, {选项B},{选项C}, {选项D}, {选项E}

模板：

任务是解决一个逻辑网格谜题。你将获得问题的背景和一些解决谜题的线索。
背景：{背景} {线索}
问题是：{问题}
选项：
A. {选项A}
B. {选项B}
C. {选项C}
D. {选项D}
E. {选项E}
用"A"，"B"，"C"，"D"，"E"中的一个输出你的答案。

未观察到 - 07 来源：注释员

输入： {问题}, {上下文}, {线索}, {选项A}, {选项B}, {选项C}, {选项D}, {选项E}

模版：

这是一个复杂的谜题。利用你的技能，通过逻辑网格表解决这个谜题：{上下文}
{线索}
{问题}

未观察到 - 08 来源：注释员

输入： {问题}, {上下文}, {线索}, {选项A}, {选项B}, {选项C}, {选项D}, {选项E}

模版：

{问题}
这个问题只有在你完全理解内容的情况下才能回答：{上下文}
{线索}。你的选择是：
(A) {选项A}
(B) {选项B}
(C) {选项C}
(D) {选项D}
(E) {选项E}

答案：

未观察到 - 09 来源：注释员

输入： {问题}, {上下文}, {线索}, {选项A}, {选项B}, {选项C}, {选项D}, {选项E}

模板：

测试你解答逻辑网格问题的能力。{问题}
{线索}
{问题}
答案总是'1'、'2'、'3'、'4'或'5'之一。输出你的答案并给出解释

未观察 - 10 源自：批注者

输入：{问题}, {背景}, {线索}, {选项A}, {选项B}, {选项C}, {选项D}, {选项E}
模版：
背景：{背景}

以下线索始终为真：{线索}

现在， 推断出对这个问题：“{问题}” 的答案， 并从选项A) {选项A} B) {选项B} C) {选项C} D) {选项D} E) {选项E} 中选择正确的答案

答案：

BBL - 概念组合

未观察 - 01 源自：批注者

输入：{问题}, {背景}, {选项A}, {选项B}, {选项C}, {选项D}
模版：
{背景} 问题：{问题}
以下是选项：
A. {选项A}
B. {选项B}
C. {选项C}
D. {选项D}
用你的常识来选择答案，并输出字母"A"，"B"，"C" 或 "D"。

未观察 - 02 源自：批注者

输入：{问题}, {背景}, {选项A}, {选项B}, {选项C}, {选项D}
模版：
给出一个概念或事实情境，根据情境内容回答选择题，并从所提供的选项中做出选择。
背景：{背景}
问题：{问题}
选择：
A. {选项A}
B. {选项B}
C. {选项C}
D. {选项D}
答案：

未观察到 - 03 来源：注释者

输入：{问题}, {语境}, {选项A}, {选项B}, {选项C}, {选项D}
模板：
你是语言专家，掌握了大部分概念和词语组合。现在，请回答以下问题：{语境} 问题：{问题} (A) {选项A} (B) {选项B} (C) {选项C} (D) {选项D}
你的答案是：

未观察到 - 04 来源：注释者

输入：{问题}, {语境}, {选项A}, {选项B}, {选项C}, {选项D}

模板：

回答关于概念组合的问题。具体来说，你需要考虑到矛盾，突现性质，奇特的虚构组合，同音词，创造新词，以及令人惊讶的非常规组合。 {语境} 问题：{问题} (A) {选项A} (B) {选项B} (C) {选项C} (D) {选项D}
你的答案是：

未观察到 - 05 来源：注释者

输入：{问题}, {语境}, {选项A}, {选项B}, {选项C}, {选项D}

模板：

问题：{问题}
选项如下：
A.{选项A}
B. {选项B}
C. {选项C}
D. {选项D}
这是一个帮助你回答问题的语境：{语境}。从 "A", "B", "C", "D"中选择最佳答案。

未观察到 - 06 来源：注释者

输入：{问题}, {语境}, {选项A}, {选项B}, {选项C}, {选项D}

模板：

以下是一个关于词语概念含义的多选题问题。你应根据语境选择最能回答问题的答案。 {语境} 问题：{问题}
选项如下：
A. {选项A}
B. {选项B}
C. {选项C}
D. {选项D}
答案：

未观察到 - 07 来源：注释者

输入：{问题}，{上下文}，{选项A}，{选项B}，{选项C}，{选项D}

模板：

上下文：{上下文}。理解上下文，并回答以下问题：{问题}选项：
(A) {选项A}
(B) {选项B}
(C) {选项C}
(D) {选项D}
答案：

未观察到 - 08 来源：注释者

输入：{问题}，{上下文}，{选项A}，{选项B}，{选项C}，{选项D}

模板：

语言学教授：{上下文} {问题}
学生：你能提供一下选项吗？
语言学教授：选择有 A) {选项A} B) {选项B} C) {选项C} D) {选项D}
学生：我明白了。答案是

未观察到 - 09 来源：注释者

输入：{问题}，{上下文}，{选项A}，{选项B}，{选项C}，{选项D}

模板：

任务是回答关于概念组合的语言学问题。上下文：{上下文}

问题：{问题}

选项：
A. {选项A}
B. {选项B}
C. {选项C}
D. {选项D}

答案：

未观察到 - 10 来源：注释者

输入：{问题}，{上下文}，{选项A}，{选项B}，{选项C}，{选项D}

模板：

{问题}
选项：A. {选项A}, B. {选项B}, C. {选项C}, 或 D. {选项D}。这个概念组合问题的正确答案是什么？基于上下文“{上下文}”，我认为最准确的答案是

BBL - 戏剧对话

Unobserved - 01 来源: 注释者

输入: {play}, {line1}, {line2}

模板:

以下对话的剧本选自莎士比亚的戏剧，但剧本未明示哪段话由哪个人说出。根据这些内容和风格，你的任务是判断问题中的句子是否由同一人或不同人说出。
{play}
从上段对话中，句子 {line1} 和 {line2} 是由同一人或不同人说出的？

答案:

Unobserved - 02 来源: 注释者

输入: {play}, {line1}, {line2}

模板:

以下是来自莎士比亚戏剧的对话剧本。
{play}

请确定这两段台词是否由同一人说出。回答是或否。
台词1: {line1}
台词2: {line2}
答案:

Unobserved - 03 来源: 注释者

输入: {play}, {line1}, {line2}

模板:

你已经阅读了所有的莎士比亚戏剧。你肯定认出了这段对话：{play}
现在，这两句台词是由同一角色还是不同角色说的？从“相同”或“不同”中选择答案。
台词1: {line1}
台词2: {line2}
你的答案:

Unobserved - 04 来源: 注释者

输入: {play}, {line1}, {line2}

模板:

下段文字是来自莎士比亚某部戏剧的对话，但没有相应说话角色的信息。你需要决定我给你的这两句台词是否为同一个角色的台词。
{play}
从上段对话，台词 {line1} 和 {line2} 是由同一角色或不同角色说出的？
答案:

Unobserved - 05 来源: 注释者

输入: {play}, {line1}, {line2}

模板:

现在你是一个剧作家。下列对话的剧本选自莎士比亚的戏剧，但剧本上并没有标明哪句话是由哪个人说出的。你的任务是确定问题中的句子是否由同一人或者不同人说出。这就是戏剧：
{play}
问题：在前述对话中，台词 {line1} 和 {line2} 是由同一个人还是不同人说出的？请给出简短的答案：相同或不同。

你的答案:

未观察到 - 06 数据源：注释者

输入：{play}，{line1}，{line2}

模板：

以下对话记录取自莎士比亚的戏剧，但记录中并未说明是谁说的这些话。我们选择了两句话，从记录中，对话描绘为 {play}，请判断 {line1} 和 {line2} 两句是否由同一个人说出。

未观察到 - 07 数据源：注释者

输入：{play}，{line1}，{line2}

模板：

对话：{play}
从上述对话中，判断 {line1} 和 {line2} 是否由同一角色说出，或是由不同角色说出。
回答 "相同" 或 "不同"。
回答：

未观察到 - 08 数据源：注释者

输入：{play}，{line1}，{line2}

模板：

在莎士比亚剧目 {play}的背景之下检查给定对话的录音记录。确定 {line1} 和 {line2} 是被一个人还是不同的人说出的。
回答：

未观察到 - 09 数据源：注释者

输入：{play}，{line1}，{line2}

模板：

剧目：{play}
在这部由莎士比亚创作的戏剧中，判断
角色 A：{line1}
角色 B：{line2} 它们是否由同一个角色说出？只需回答“是”或“否”。

未观察到 - 10 数据源：注释者

输入：{play}，{line1}，{line2}

模板：

上下文：{play}
问题：阅读这段由莎士比亚创作的剧剧中的选段，判断 {line1} 和 {line2} 是否来自同一个角色？
选项：
A) 是
B) 否。回答：

BBL - 策略 QA

未观察到 - 01 来源：注释者

输入：{问题}

模板：

你将获得一个问题，该问题需要你进行暗示在问题中的推理步骤。请选择"是"或"否"中的最佳答案，并提供解释。

问题：{问题}
答案和解释：

未观察到 - 02 来源：注释者

输入：{问题}

模板：

关于问题的答案进行推理。{问题}

未观察到 - 03 来源：注释者

输入：{问题}

模板：

你正在参加一个考试，其中每个问题都需要进行推理步骤来回答。答案总是"是"或"否"。请仔细考虑以下问题，其含义，以及你可能需要提供的任何相关信息以提供正确答案。
问题：{问题}
答案：

未观察到 - 04 来源：注释者

输入：{问题}

模板：

回答假设问题提示中含有隐含推理步骤的问题：{问题}

未观察到 - 05 来源：注释者

输入：{问题}

模板：

输入：
- 问题：{问题}
输出：
- 答案：

Unobserved - 06 来源：注释器

输入：{question}

模板：

用逻辑推理来回答以下问题，答案应为“是”或“否”。
问题：{question}
答案:

Unobserved - 07 来源：注释器

输入：{question}

模板：

请回答以下问题，你应该逐步思考，但请用“是”或“否”来回答。
问题：{question}
答案：

Unobserved - 08 来源：注释器

输入：{question}

模板：

回答问题中所需的推理步骤隐含在问题中。请先回答“是”或“否”，然后输出你的解释。
{question} 答案：

Unobserved - 09 来源：注释器

输入： {question}

模板：

请对以下问题给出你的答案，应该回答是或否。这个问题可能需要你进行隐含的多跳推理。
问题：{question}
答案：

Unobserved - 10 来源：注释器

输入： {question}

模板：

这个问题需要通过将其分解成多个子问题并比较子问题的结果来解决。
问题是 {question}
分解问题后，我们发现答案是

BBL - 奇怪的故事

未观察到 - 01 来源：注释者

输入： {问题}, {背景}

模板：

你获得了一个心理学问题，要求你阅读一篇短篇故事后提供一个社会智能化的回答。请对给出的问题回答"是"或"否"。

{背景}

问题：{问题}
答案：

未观察到 - 02 来源：注释者

输入： {问题}, {背景}

模板：

给定一个故事，回答问题是否为真或假。
{背景}
问题：{问题}
答案：

未观察到 - 03 来源：注释者

输入： {问题}, {背景}

模板：

你正在进行阅读理解测试。你将被给出一个故事，并被问到一个与故事相关的问题。问题的答案要么是"是"，要么是"否"。请在选择你的答案之前仔细思考以下的故事。
故事：{背景}
问题：{问题}
答案：

未观察到 - 04 来源：注释者

输入： {问题}, {背景}

模板：

一个通过自然主义短篇故事测量社会智能的心理测试。布尔选项。 {背景}
问题：{问题} 答案:

未观察到 - 05 来源：注释者

输入: {问题}, {背景}

模板：

故事： {背景}
问题： {问题}
输出：

未观察到 - 06 来源：注释者

输入： {question}, {context}

模板：

根据以下文本，用“是”或“否”回答问题：

文本：{context}
问题：{question}
回答：

未观察到 - 07 来源：注释者

输入： {question}, {context}

模板：

请阅读以下文本并根据文本内容回答问题，您的答案应为“是”或“否”。
文本：{context}
问题：{question}
回答：

未观察到 - 08 来源：注释者

输入： {question}, {context}

模板：

设想你正在参加一场心理测试。请阅读给定的故事并回答问题。请回答“是”或“否”。
故事：{context}
问题：{question}
回答：

未观察到 - 09 来源：注释者

输入： {question}, {context}

模板：

请根据故事酌情回答以下问题，答案应为“是”或“否”。
故事：{context}

问题：{question}
回答：

未观察到 - 10 来源：注释者

输入：{问题}, {背景}

模板：

以下故事与一个问题相关，其答案为“是”或“否”。

故事：{背景}
问题：{问题}

根据故事，答案是

BBL - Winowhy

未观察到 - 01 来源：注释者

输入：{问题}

模板：

在句子中：{问题}。代词推理是否正确？请以“正确”或“不正确”回答。不要包含其他任何词语。

未观察到 - 02 来源：注释者

输入：{问题}

模板：

验证在给定的单词中，关于某些代词指代哪些单词的推理是否正确，从“正确”和“不正确”中选择一个答案：
推理：{问题}
答案：

未观察到 - 03 来源：注释者

输入：{问题}

模板：

给定的背景：{问题}，确定共指解析和解释是否正确，输出“正确”或“不正确”。答案：

未观察到 - 04 来源：注释员

输入: {问题}

模板:

背景: {问题}
问题：代词是否指向了正确的对象？以"是"或"否"回答。

未观察到 - 05 来源：注释员

输入: {问题}

模板:

判断代词理解的正确性：
{问题}
以"正确"或"错误"作为您的回答。您的答案是：

未观察到 - 06 来源：注释员

输入: {问题}

模板:

您要在代词理解上接受测试。这是一个句子，后面是解释：{问题}
如果你认为解释是正确的，输出"正确"；如果解释是错误的，输出"错误"。

未观察到 - 07 来源：注释员

输入: {问题}

模板:

阅读以下有关特定代词指向谁的推理：{问题}
推理是否正确？

Unobserved - 08 来源：注释者

输入：{question}

模板：

阅读下面的推理，并回答其是否正确。{question}

Unobserved - 09 来源：注释者

输入：{question}

模板：

{question} 上述推理是“正确的”还是“错误的”？这是

Unobserved - 10 来源：注释者

输入：{context}, {explanation}

模板：

你将会看到一个句子，然后是对该句子中代词使用的解释。请回答这个解释是对的还是错的。

句子：{context}
解释：{explanation}
答案：

BBL - 语言ID

未观察到 - 01 来源：注释员

输入：{句子}, {选择A}, {选择B}, {选择C}, {选择D}, {选择E}, {选择F}, {选择G}, {选择H}, {选择I}, {选择J}, {选择K}

模板：

识别给定句子的正确语言。请从A、B、C、D、E、F、G、H、I、J和K中选择最佳答案。

句子：{句子}
A: {选择A}
B: {选择B}
C: {选择C}
D: {选择D}
E: {选择E}
F: {选择F}
G: {选择G}
H: {选择H}
I: {选择I}
J: {选择J}
K: {选择K}
答案：

未观察到 - 02 来源：注释员

输入： {句子}, {选择A}, {选择B}, {选择C}, {选择D}, {选择E}, {选择F}, {选择G}, {选择H}, {选择I}, {选择J}, {选择K}

模板：

{句子}
上述语言是什么语言？ A: {选择A} B: {选择B} C: {选择C} D: {选择D} E: {选择E} F: {选择F} G: {选择G} H: {选择H} I: {选择I} J: {选择J} K: {选择K}

未观察到 - 03 来源：注释员

输入： {句子}, {选项A}, {选项B}, {选项C}, {选项D}, {选项E}, {选项F}, {选项G},{选项H}, {选项I}, {选项J}, {选项K}

模板：

你正在参加一个测试，需要你识别出给定句子的语言。为了帮助你缩小选择范围，我们把这个问题设为了多项选择题。在仔细考察句子和以下每个答案之后，请从"A", "B", "C", "D", "E", "F", "G", "H", "I", "J" 或 "K" 中选择句子的正确语言。
句子: {句子}
- A: {选项A}
- B: {选项B}
- C: {选项C}
- D: {选项D}
- E: {选项E}
- F: {选项F}
- G: {选项G}
- H: {选项H}
- I: {选项I}
- J: {选项J}
- K: {选项K}
答案：

未观察到 - 04　来源：注释员

输入： {句子}, {选项A}, {选项B}, {选项C}, {选项D}, {选项E}, {选项F}, {选项G},{选项H}, {选项I}, {选项J}, {选项K}

模板：

请从下列选项中选择与所给句子正确对应的语言：
句子: {句子}
选项：
A: {选项A}
B: {选项B}
C: {选项C}
D: {选项D}
E: {选项E}
F: {选项F}
G: {选项G}
H: {选项H}
I: {选项I}
J: {选项J}
K: {选项K}
你的答案：

未观察到 - 05　来源：注释者

输入： {句子}, {选项A}, {选项B}, {选项C}, {选项D}, {选项E}, {选项F}, {选项G}, {选项H}, {选项I}, {选项J}, {选项K}

模板：

输入
- 句子：{句子}
- A: {选项A}
- B: {选项B}
- C: {选项C}
- D: {选项D}
- E: {选项E}
- F: {选项F}
- G: {选项G}
- H: {选项H}
- I: {选项I}
- J: {选项J}
- K: {选项K}

输出
- 答案：

未观察到 - 06　来源：注释者

输入： {句子}, {选项A}, {选项B}, {选项C}, {选项D}, {选项E}, {选项F}, {选项G}, {选项H}, {选项I}, {选项J}, {选项K}

模板：

给定以下文本，通过选择列表中的一个选项（A，B，C，D，E，F，G，H，I，J，K）确定正确的语言：

文本：{句子}

A: {选项A}
B: {选项B}
C: {选项C}
D: {选项D}
E: {选项E}
F: {选项F}
G: {选项G}
H: {选项H}
I: {选项I}
J: {选项J}
K: {选项K}

答案：

未观察到 - 07 来源：注释者

输入： {句子}, {选项A}, {选项B}, {选项C}, {选项D}, {选项E}, {选项F}, {选项G},{选项H}, {选项I}, {选项J}, {选项K}

模板：

请阅读以下句子， 然后从选项中选择你认为最有可能来自哪种语言。 你的答案应为"A"，"B"，"C"，"D"，"E"，"F"，"G"，"H"，"I"，"J"或"K"
句子：{句子}
选项：
A: {选项A}
B: {选项B}
C: {选项C}
D: {选项D}
E: {选项E}
F: {选项F}
G: {选项G}
H: {选项H}
I: {选项I}
J: {选项J}
K: {选项K}
答案：

未观察到 - 08 来源：注释者

输入： {句子}, {选项A}, {选项B}, {选项C}, {选项D}, {选项E}, {选项F}, {选项G},{选项H}, {选项I}, {选项J}, {选项K}

模板：

请给出以下句子所用的语言。每个句子都会给出五个选项，请输出相应的选项 (即 A, B, C,

D, E, F, G, H, I, J, 或 K) 以表示相应的答案。
句子：{句子}
选项:
A: {选项A}
B: {选项B}
C: {选项C}
D: {选项D}
E: {选项E}
F: {选项F}
G: {选项G}
H: {选项H}
I: {选项I}
J: {选项J}
K: {选项K}
答案：

未观察到 - 09 来源：注解器

输入： {句子}，{选项A}，{选项B}，{选项C}，{选项D}，{选项E}，{选项F}，{选项G}，{选项H}，{选项I}，{选项J}，{选项K}

模板：

给定句子：{句子}，请在以下选项中选择正确的语言 A. {选项A} B. {选项B} C. {选项C} D. {选项D} E. {选项E} F. {选项F} G. {选项G} H. {选项H} I. {选项I} J. {选项J} K. {选项K}
- A: {选项A}
- B: {选项B}
- C: {选项C}
- D: {选项D}
- E: {选项E}
- F: {选项F}
- G: {选项G}
- H: {选项H}
- I: {选项I}
- J: {选项J}
- K: {选项K}

语言：

未观察到 - 10 来源：注解器

输入： {句子}，{选项A}，{选项B}，{选项C}，{选项D}，{选项E}，{选项F}，{选项G}，{选项H}，{选项I}，{选项J}，{选项K}

模板：

{句子}

这是一段以{选项A}，{选项B}，{选项C}，{选项D}，{选项E}，{选项F}，{选项G}，{选项H}，{选项I}，{选项J}，{选项K}中的一种语言写就的句子。根据单词和语言结构，我可以判断这个语言是：

BBL - 维他命C

未观察到 - 01 来源：标注员

输入： {context}, {claim}

模板：

您现在是一位非常有经验的法官。仅根据维基百科引述的简短信息，回答相关声明是否为真、假或无法判断。当维基百科引述的内容没有提供解决问题所需的必要信息时，使用“无法判断”。

{context}
声明：{claim}
这是真的、假的，还是无法判断?

未观察到 - 02 来源：标注员

输入： {context}, {claim}

模板：

现在你是一名维生素事实验证者。仅根据维基百科引述的简短信息，回答相关声明是否为真、假或无法判断。当维基百科引述的内容没有提供解决问题所需的必要信息时，使用“无法判断”。

{context}
声明：{claim}
问题：这是真的、假的，还是无法判断？
您的回答：

未观察到 - 03 来源：标注员

输入： {context}, {claim}

模板：

{context}
阅读上述段落，并回答以下声明 {claim}。回答真、假或无法判断。无法判断意味着维基百科引述的内容没有提供解决问题所需的必要信息。答案：

未观察到 - 04 来源：标注员

输入： {context}, {claim}

模板：

给定一个声明和来自维基百科的相关信息环境，确定该声明是真、假还是无法判断。无法判断意味着给定的信息不足以判断该声明是真或假，大致相当于不确定。

情境：{context}
声明：{claim}
是真的、假的，还是无法判断?

未观察到 - 05 来源：标注员

输入： {context}, {claim}

模板：

{context} 声明：{claim}
根据上述情况， 声明是真的吗？ 是假的吗？ 还是无法判断？ 请给出您的答案， 是"真"、"假"还是"无法判断"

未观察到 - 06 来源：标注员

输入： {context}, {claim}

模板：

仅根据给定情境中的信息，请判断相关声明是真的、假的还是无法判断。

{context}
声明：{claim}
是真的、假的，还是无法判断？

Unobserved - 07 来源：注释者

输入： {context}, {claim}

模板：

维基百科：{context}
有人云：基于给定的情境，{claim} 是正确、错误，还是无法确定？

Unobserved - 08 来源：注释者

输入： {context}, {claim}

模板：

仅根据维基百科摘录中给出的信息，评估相关声明是正确、错误，还是无法确定。如果摘录没有提供足够的信息来解答问题，那么选择无法确定。
{context}
声明：{claim}
答案（正确、错误，还是无法确定）：

Unobserved - 09 来源：注释者

输入： {context}, {claim}

模板：

输入：{claim}
基于以下情境，验证声明的真实性
{context}
如果声明在事实上正确，为 "True"
如果声明在事实上错误，为 "False"
如果事实性无法确定，则为 "Neither"。用 "True"，"False" 或 "Neither" 中的一个输出你的答案。答案：

Unobserved - 10 来源：注释者

输入： {context}, {claim}

模板：

情境：{context}
现在将此声明分类为 'True'，'False' 或 'Neither' 中的一个。
{claim}

E.3 Granular Experiment Instructions

在本节中，我们提供了我们用于所有6种设置的指令，按数据集分类。

BBH - 意图识别

最接近 - 1 来源: NIV2 任务 163 OpenPI 分类 - 模板 2

输入： {passage}

模板：

首先，您将获得一个任务的定义，然后是一些任务的输入。
给定一个段落作为输入，回答这个段落属于哪一个类别。这里有分类 - {options}。答案应该是基于段落中的词汇密切关联的类别中的一个。

{passage}
输出:

最接近 - 2 来源：NIV2 任务 163 OpenPI 分类 - 模板 4

输入： {passage}

模板：

指令：给定一个段落作为输入，回答这个段落属于哪一个类别。这里有分类 - {options}。答案应该是基于段落中的词汇密切关联的类别中的一个。
输入：{passage}
输出：

最接近 - 3 来源：NIV2 任务 163 OpenPI 分类 - 模板 6

输入： {passage}

模板：

根据任务定义和输入，回复输出。给定一个段落作为输入，回答这个段落属于哪一个类别。这里有分类 - {options}。答案应该是基于段落中的词汇密切关联的类别中的一个。

{passage}

最接近 - 4 来源：NIV2 任务 163 OpenPI 分类 - 模板 8

输入： {passage}

模板：

Q：给定一个段落作为输入，回答这个段落属于哪一个类别。这里有分类 - {options}。答案应该是基于段落中的词汇密切关联的类别中的一个。
{passage}
A:

最接近 - 5 来源：NIV2 任务 163 OpenPI 分类 - 模板 10

输入： {passage}

模板：

详细指令：给定一个段落作为输入，回答这个段落属于哪一个类别。这里有分类 - {options}。答案应该是基于段落中的词汇密切关联的类别中的一个。
Q：{passage}
A:

不正确 - 1 来源：NIV2 任务 562 语言识别 - 模板 10

输入： {文本}, {选项}

模板：

详细指导：在这项任务中，给出了一句可能是在{选项}语言中的输入句。一共有{选项长度}种语言。您的任务是识别输入句的语言。输入句只能是提供的{选项长度}种语言中的任何一种。
Q: {文本}
A:

不正确 - 2 来源：NIV2 任务 1588 Tecla 分类 - 模板 10

输入： {文本}, {选项}

模板：

详细指导：在这个任务中，您得到的是一段加泰罗尼亚语文本。您的任务是将其分类到{选项长度}个不同的给定主题中。所有类别的名称都是{选项}。
Q: {文本}
A:

不正确 - 3 来源：NIV2 任务 564 DiscoFuse 分类 - 模板 10

输入： {文本}, {选项}

模板：

详细指导：在这个任务中，你得到的是两个英语句子，你的任务是将它们分类到它们的一种话语类型中。话语类型是一种分类标准，根据上下文和相关情境对给定的两个句子进行分类。总共有{选项长度}种话语类型，它们是{选项}。
Q: {文本}
A:

不正确 - 4 来源：NIV2 任务 1193 课程分类 - 模板 10

输入： {文本}, {选项}

模板：

详细指导：在这个任务中，你得到的是一个印度菜肴的名称。你需要将这道菜分类为{选项}。
Q: {文本}
A:

不正确 - 5 来源：Flan Sentiment140 - 模板 2

输入： {文本}, {选项}
模板：
{文本}

这条推特的情感如何描述？
{选项}

收集 - 1 来源：注释者

输入： {text}, {options}
模板：
您将得到一组要进行预测的意图：{options}。选择最合适的一个来描述以下话语：{text}。
意图：

收集 - 2 来源：注释者

输入： {text}, {options}
模板：
您是一位识别和分类用户意图的对话助手。
始终使用ĂL'项中的一个： [{options}] 来指示意图。
话语：{text}
意图：

收集 - 3 来源：注释者

输入： {text}, {options}
模板：
任务是将话语：'{text}' 的意图分类到以下内容之一：{options}。您的答案是：

收集 - 4 来源：注释者

输入： {text}, {options}
模板：
给定标签空间：{options}, 对给定话语的意图进行分类。
{text}
意图：

收集 - 5 来源：注释者

输入： {text}, {options}
模板：
从列表中输出话语的意图：{options}。准确输出相应的字词或短语。 {text}

任务设计者 请参见 BIG-BENCH 评估文件。

否定 - 1 来源：NIV2 任务 163 OpenPI 分类 - 模板 2

输入： {段落}
模板：

你将首先给出一个任务的定义，然后是一些任务的输入。
给定一个段落作为输入，回答段落 不 属于哪个分类。这里有几个分类 - {选项}。答案应该是基于段落中的词语 不 属于该类别的一个类别。

{段落}
输出：

否定 - 2 来源：NIV2 任务 163 OpenPI 分类 - 模板 4

输入： {段落}

模板：

说明：给定一个段落作为输入，回答段落 不 属于哪个分类。这里有几个分类 - {选项}。答案应该是基于段落中的词语 不 属于该类别的一个类别。
输入：{段落}
输出：

否定 - 3 来源：NIV2 任务 163 OpenPI 分类 - 模板 6

输入： {段落}

模板：

给定任务定义和输入，请回答输出。给定一个段落作为输入，回答段落 不 属于哪个分类。这里有几个分类 - {选项}。答案应该是基于段落中的词语 不 属于该类别的一个类别。

{段落}

否定 - 4 来源：NIV2 任务 163 OpenPI 分类 - 模板 8

输入： {段落}

模板：

Q：给定一个段落作为输入，回答段落 不 属于哪个分类。这里有几个分类 - {选项}。答案应该是基于段落中的词语 不 属于该类别的一个类别。
{段落}
A：

否定 - 5 来源：NIV2 任务 163 OpenPI 分类 - 模板 10

输入： {段落}

模板：

详细说明：给定一个段落作为输入，回答段落 不 属于哪个分类。这里有几个分类 - {选项}。答案应该是基于段落中的词语 不 属于该类别的一个类别。
Q：{段落}
A：

荒诞 - 1 来源：注释员

输入： {文字}

模板：

街道浣熊议会要求你回答他们的询问。{文字}

荒诞 - 2 来源: 注释员

输入: {文本}
模板:
监视鸟类向您询问关于种子的知识。{文本}

荒诞 - 3 来源: 注释员

输入: {文本}
模板:
达斯·维德要求你向黑暗面回答。{文本}

荒诞 - 4 来源: 注释员

输入: {文本}
模板:
回应火星工作海豚联盟的要求。{文本}

荒诞 - 5 来源: 注释员

输入: {文本}
模板:
你正在接受当地松鼠法庭的询问。{文本}

BBH - 实证判断

最接近 - 1 来源: NIV2 任务 163 OpenPI 分类 - 模板 2

输入: {段落}
模板:
首先会给出一个任务的定义，然后是一些任务的输入。
给定一个段落作为输入，回答出这个段落属于哪个类别。这有几种类别 - {选项}。答案应该是基于段落中与类别紧密相关的词语来确定的其中一个类别。

{段落}
输出:

最接近 - 2 来源: NIV2 任务 163 OpenPI 分类 - 模板 4

输入: {段落}
模板:
说明：给定一个段落作为输入，回答这个段落属于哪个类别。这里有类别 - {选项}。答案应该是根据段落中紧密关联到类别的词汇选择的类别之一。
输入：{段落}
输出：

最接近 - 3 来源: NIV2 任务 163 OpenPI 分类 - 模板 6

输入: {段落}

模板:

根据任务定义和输入，给出输出。给定一个段落作为输入，回答这个段落属于哪个类别。这里有类别 - {选项}。答案应该是根据段落中紧密关联到类别的词汇选择的类别之一。

{段落}

最接近 - 4 来源: NIV2 任务 163 OpenPI 分类 - 模板 8

输入: {段落}

模板:

Q: 给定一个段落作为输入，回答这个段落属于哪个类别。这里有类别 - {选项}。答案应该是根据段落中紧密关联到类别的词汇选择的类别之一。
{段落}
A:

最接近 - 5 来源: NIV2 任务 163 OpenPI 分类 - 模板 10

输入: {段落}

模板:

详细说明：给定一个段落作为输入，回答这个段落属于哪个类别。这里有类别 - {选项}。答案应该是根据段落中紧密关联到类别的词汇选择的类别之一。
Q: {段落}
A:

不正确 - 1 来源: NIV2 任务 143 Odd Man Out分类 - 模板 10

输入: {输入}, {类别}

模板:

详细说明：给定一组四个单词，生成单词所属的类别。单词由逗号分隔。可能的类别是 {类别}
Q: {输入}
A:

不正确 - 2 来源: NIV2 任务 137 Newscomm分类 - 模板 10

输入: {输入}, {选项}

模板:

详细说明：将给定的新闻评论分类到其所写的语言中。这里有 {选项的长度}种语言可以将句子分类为 {选项}
Q: {输入}
A:

不正确 - 3 来源：Flan2021 - Sentiment140 - 模板 1

输入：{input}，{options}

模板：

{text}
从选项中选择你的答案。这条推文的情感是什么？
选项：{options}...我认为答案是

不正确 - 4 来源：Flan2021 - Sentiment140 - 模板 6

输入：{input}，{options}
模板：
从选项中选择你的答案。如何描述这条推文的情绪？
{text}
{options}

不正确 - 5 来源：NIV2 任务 1422 MathQA 物理 - 模板 10

输入：{input}，{options}
模板：
详细说明：在这个任务中，你需要回答关于物理学的给定多项选择题。将你的答案分类为{letter length}
Q：问题：{input}
{options}
A：

Collected - 1 来源：注释员

输入：{events}
模板：
下面的句子描述了两个事件：{events}
判断事件之间的关系是 '因果'，'相关' 或者 '中性'。

Collected - 2 来源：注释员

输入：{events}
模板：
因果关系：如果一事件导致另一事件发生，那么两者之间就有因果关系。
相关关系：如果两个事件之间没有明显的因果关系，但它们是相关的，那么它们之间就有相关关系。
中性关系：如果两个事件之间没有明显的相关性，那么它们之间就有中性关系。

{events} 句子中描述的事件有因果关系、相关关系还是中性关系？

Collected - 3 来源：注释员

输入：{events}
模板：
句子：{events} 对句子中事件的关系做出判断。可能的关系包括："因果"，"相关"，"中性"

收集 - 4 来源：注释员

输入：{事件}

模板：

这些事件：{事件}之间的关系是什么？
将它分类为“因果”，“相关”，“中立”

收集 - 5 来源：注释员

输入：{事件}

模板：

现在，你被给予一句描述两个或更多事件的句子。现在，请将关系分类为"因果"，"相关"，"中立"之一。
句子："{事件}"
答案：

任务设计者 参见 BIG-BENCH 评估文件。

否定 - 1 来源：NIV2任务163 OpenPI分类 - 模板2

输入：{段落}

模板：

首先，你会被给一个任务的定义，然后是任务的一些输入。
给定一个段落作为输入，回答一下段落不属于哪个类别。有这些类别 - {选项}。答案应该是基于段落中的词，这些词不属于该类别的一个类别。

{段落}
输出：

否定 - 2 来源：NIV2任务163 OpenPI分类 - 模板4

输入：{段落}

模板：

说明：给定一个段落作为输入，回答这个段落不属于哪个类别。有这些类别 - {选项}。答案应该是基于段落中的词，这些词不属于该类别的一种类别。
输入：{段落}
输出：

否定 - 3 来源：NIV2任务163 OpenPI分类 - 模板6

输入：{段落}

模板：

给定任务定义和输入，回复输出。将一段文本作为输入，回答此文本不属于哪个类别。有如下类别 - {选项}。答案应该是基于段落中的真正不属于该类别领域的词汇的那个类别。

{段落}

否定 - 4 来源：NIV2 Task 163 OpenPI 分类 - 模板 8

输入：{段落}

模板：

Q：给定一个输入段落，答案应为该段落不属于的类别。这里有以下这些类别 - {选项}。答案应根据段落中的词汇选择一个不属于的类别。
{段落}
A:

否定 - 5　来源：NIV2 Task 163 OpenPI 分类 - 模板 10

输入：{段落}

模板：

详细说明：给定一个输入段落，答案应为该段落不属于的类别。这里有以下这些类别 - {选项}。答案应根据段落中的词汇选择一个不属于的类别。
Q：{段落}
A:

无意义 - 1　来源：注释员

输入：{文本}

模板：

街头浣熊委员会要求你回答他们的质询。 {文本}

无意义 - 2　来源：注释员

输入：{文本}

模板：

监控鸟类询问你对种子的了解。 {文本}

无意义 - 3　来源：注释员

输入：{文本}

模板：

达斯·维达要求你回答黑暗面 {文本}

无意义 - 4　来源：注释员

输入：{文本}

模板：

回应火星工作海豚工会的要求。 {文本}

无意义 - 5　来源：标注员

输入：{text}

模板：

你正在接受当地松鼠法庭的审问。 {text}

BBL - 概念结合

最接近 - 1　来源：Flan2021 - CosmosQA - 模板1

输入：{context}, {question}, {options}

模板：

{context}

以下问题要从选项中选择: {question}
选项：{options}

最接近 - 2 来源：Flan2021 - CosmosQA - 模板2

输入： {context}, {question}, {options}

模板：

{context}

选项: {options}
问: {question}

最接近 - 3 来源：Flan2021 - CosmosQA - 模板3

输入： {context}, {question}, {options}

模板：

{context}

选项: {options}
回答以下问题: {question}

最接近 - 4 来源：Flan2021 - CosmosQA - 模板4

输入： {context}, {question}, {options}

模板：

{context}

根据前述段落，为问题选择你的答案 {question}
选项: {options}

最接近 - 5 来源: Flan2021 - CosmosQA - 模板 5

输入: {context}, {question}, {options}

模板:

{context}

有选项的问题：使用上述段落中的证据回答以下问题：{question}
选项：{options}

最接近 - 6 来源: Flan2021 - CosmosQA - 模板 6

输入: {context}, {question}, {options}

模板:

上下文：{context}

问题 {question}
可能的答案：
{options}
答案：

最接近 - 7 来源: Flan2021 - CosmosQA - 模板 7

输入: {context}, {question}, {options}

模板:

阅读以下文章并通过选择选项来回答问题。

{context}

{question}
选项：{options}...A：

最接近 - 8 来源: Flan2021 - CosmosQA - 模板 8

输入: {context}, {question}, {options}

模板:

此问题有选项。回答关于文字的问题：

{context},

{question}
选项：{options}

错误 - 1 来源: Flan2021 WSC273 - 模板 2

输入: {context}, {options}

模板:

完成文章。

{context}
选项： {options}

错误 - 2 来源: Flan2021 WSC273 - 模板 9

输入: {context}, {options}

模板:

在选项中列出的下一个事件是正确的吗？

{context}
选项： {options}
A:

错误 - 3 来源: Flan2021 Winograde - 模板 3

输入: {context}, {options}

模板:
选择你的故事来继续下面的故事。

{context}

{options}

错误 - 4 来源: Flan2021 Story Cloze - 模板 1

输入: {context}, {options}
模板:
{context}
{options}
哪个选项是下一句话?

错误 - 5 来源: Flan2021 Sentiment140 - 模板 1

输入: {text}, {options}
模板:
{text}
从选项中选择你的答案。这条推文的情绪是什么?
{options}...我认为答案是

错误 - 6 来源: NIV2 任务 1422 MathQA 物理 - 模板 10

输入: {input}, {options}
模板:
详细的指令：在这个任务中，你需要回答物理学的给定的多选题问题。将你的答案分类为{letter length}
Q: 问题: {input}
{options}
A:

错误 - 7 来源: NIV2 任务 1297 QASC - 模板 10

输入: {input}, {options}
模板:
详细的指令：在这个任务中，你被给予两个事实和一个多选题。根据给定的事实，用正确选项的索引来回答问题（例如，"A"）。
Q: {input} {options}
A:

收集 - 01 来源：注解者

输入：{question}, {context}, {choiceA}, {choiceB},{choiceC},{choiceD}
模版：
{context} 问题：{question}
下面是选项：

A. {choiceA}
B. {choiceB}
C. {choiceC}
D. {choiceD}
根据您的常识选择一个"A", "B", "C", 或者 "D"作为您的答案。

收集 - 02 来源：注解者

输入： {question}, {context}, {choiceA}, {choiceB},{choiceC},{choiceD}

模版：

您收到了一个概念或实际情境。基于这个情境，选择所提供的选项来回答多选题
情境：{context}
问题：{question}
选项：
A. {choiceA}
B. {choiceB}
C. {choiceC}
D. {choiceD}
答案：

收集 - 03 来源：注解者

输入： {question}, {context}, {choiceA}, {choiceB},{choiceC},{choiceD}

模版：

您是一位语言学专家，能掌握大多数概念和词语的组合。现在，请回答以下问题：{context}
问题：{question} (A) {choiceA} (B) {choiceB} (C) {choiceC} (D) {choiceD}
您的答案是：

收集 - 04 来源：注解者

输入： {question}, {context}, {choiceA}, {choiceB},{choiceC},{choiceD}

模版：

回答关于概念组合的问题。特别是，您需要考虑矛盾、新出现的特性、奇特的虚构组合、同义词、创新的词汇和出乎意料的罕见组合。{context} 问题：{question} (A) {choiceA} (B) {choiceB} (C) {choiceC} (D) {choiceD}
您的答案是：

已收集 - 05 来源: 注释者

输入： {问题}, {背景}, {选项A}, {选项B},{选项C},{选项D}

模板：

问题：{问题}
可选项有：
A. {选项A}
B. {选项B}
C. {选项C}
D. {选项D}

下面是一个背景来帮助你回答这个问题：{背景}。从“A”，“B”，“C”，“D”中选择最佳答案。

已收集 - 06 来源: 注释者

输入： {问题}, {背景}, {选项A}, {选项B},{选项C},{选项D}

模板：

以下是一个关于词语概念意义的多选问题解答问题。你应该根据上下文选择最佳答案。 {背景} 问题：{问题}
可选项有：
A. {选项A}
B. {选项B}
C. {选项C}
D. {选项D}
答案：

已收集 - 07 来源: 注释者

输入： {问题}, {背景}, {选项A}, {选项B},{选项C},{选项D}

模板：

背景：{背景}。理解这个背景，并回答以下问题：{问题} 选项：
(A) {选项A}
(B) {选项B}
(C) {选项C}
(D) {选项D}
答案：

已收集 - 08 来源: 注释者

输入： {问题}, {背景}, {选项A}, {选项B},{选项C},{选项D}

模板：

语言学教授：{背景} {问题}
学生：你能提供选项吗？
语言学教授：选项有 A) {选项A} B) {选项B} C) {选项C} D) {选项D}
学生：我明白了。答案是

收集 - 09 来源：注解者

输入： {问题}, {内容}, {选项A}, {选项B},{选项C},{选项D}

模板：

任务是回答有关概念组合的语言问题。内容：{内容}

问题：{问题}

选项：
A. {选项A}
B. {选项B}
C. {选项C}

D. {选项D}

答案：

收集 - 10 来源：注解者

输入： {问题}, {内容}, {选项A}, {选项B},{选项C},{选项D}
模板：
{问题}
选项： A. {选项A}, B. {选项B}, C. {选项C}, 或 D. {选项D}。这个概念组合问题的正确答案是什么呢？基于 "{内容}" 的上下文，我认为最准确的答案是

任务设计者 请参看 BIG-BENCH 评估文件。

否定 - 1 来源：Flan2021 - CosmosQA - 模板 1

输入： {内容}, {问题}, {选项}
模板：
{内容}

有选项可供选择的问题：{问题}
选项：{选项} 答案不是：

否定 - 2 来源：Flan2021 - CosmosQA - 模板 2

输入： {内容}, {问题}, {选项}
模板：
{内容}

选项：{选项}
问题：{问题} 答案不是：

否定 - 3 来源：Flan2021 - CosmosQA - 模板 3

输入： {内容}, {问题}, {选项}
模板：
{内容}

选项：{选项}
回答如下问题：{问题}
答案不是：

否定 - 4 来源：Flan2021 - CosmosQA - 模板 4

输入： {context}, {question}, {options}
模板：
{context}

基于前述段落，回答问题 {question}
选择项： {options}
答案不是:

否定 - 5 来源：Flan2021 - CosmosQA - 模板 5

输入： {context}, {question}, {options}

模板：

{context}

问题及选项：使用上述段落中的证据回答以下问题： {question}
选择项: {options}
答案不是:

否定 - 6 来源：Flan2021 - CosmosQA - 模板 6

输入： {context}, {question}, {options}

模板：

情境：{context}
问题 {question}
可能的答案：
{options}
答案不是:

否定 - 7 来源：Flan2021 - CosmosQA - 模板 7

输入： {context}, {question}, {options}

模板：

阅读以下文章并从选项中回答问题。

{context}

{question}
选择项： {options}...答案不是:

否定 - 8 来源：Flan2021 - CosmosQA - 模板 8

输入： {context}, {question}, {options}

模板：

这个问题有选项。回答关于文本的问题：

{context}

{question}
选项：{options}
答案不是：

荒谬 - 1 来源：注释者

输入：{text}
模板：
街头浣熊议会要求你回答他们的质询。{text}

Nonsensical - 2 来源：注释者

输入：{text}
模板：
监视鸟类询问你对种子的知识。 {text}

Nonsensical - 3 来源：注释者

输入：{text}
模板：
达斯·维达要求你回答黑暗面 {text}

Nonsensical - 4 来源：注释者

输入：{text}
模板：
回应火星工作海豚联盟的要求。 {text}

Nonsensical - 5 来源：注释者

输入：{text}
模板：
你正在接受当地松鼠法庭的询问。 {text}

BBL - 语言识别

Closest - 1 来源：NIV2 任务 137 新闻评论分类 - 模板 2

输入：{passage}
模板：
首先，您将被给予一个任务的定义，然后是一些任务的输入。
将给定的新闻评论分类为其所写的语言。这里有 {option length}种语言可以将句子分类为 {options}

{sentence}
输出：

最接近 - 2 来源：NIV2 任务 137 Newscomm分类 - 模板 4

输入：{文章}
模板：
说明：将给定的新闻评论分类为其所写的语言。有 {option length} 种语言可以将句子分类到 {options}
输入：{句子}
输出：

最接近 - 3 来源：NIV2 任务 137 Newscomm分类 - 模板 6

输入：{文章}

模板：

基于任务定义和输入， 回复输出。 将给定的新闻评论分类为其所写的语言。 有 {option length} 种语言可以将句子分类到 {options}

{句子}

最接近 - 4 来源：NIV2 任务 137 Newscomm分类 - 模板 8

输入：{文章}

模板：

问：将给定的新闻评论分类为其所写的语言。有 {option length} 种语言可以将句子分类到 {options}
{句子}
答：

最接近 - 5 来源：NIV2 任务 137 Newscomm分类 - 模板 10

输入：{文章}

模板：

详细说明：将提供的新闻评论内容分类为其所写的语言。有 {option length} 种语言可供句子分类 {options}
问：{句子}
答：

错误 - 1 来源：NIV2任务143 Odd Man Out分类 - 模板10

输入：{输入}, {类别}

模板：

详细说明：给定四个单词，生成单词所属的类别。单词由逗号隔开。可能的类别是 {类别}
问：{输入}
答：

错误 - 2 来源：NIV2 任务 1322 政府类型分类 - 模板 10

输入：{输入}, {选项}

模板：

详细说明：在此任务中，您将得到一个国家名称，您需要回答该国2015年的政府类型。以下是被认为是有效答案的可能的政府类型：{选项}
问：{输入}
答：

错误 - 3 来源：NIV2 任务 1422 MathQA 物理 - 模板 10

输入：{输入}, {选项}

模板：

详细说明：在此任务中，您需要回答关于物理的给定多项选择题。请将您的答案分类为{字母长度}
问：问题：{输入}
{选项}
答：

错误 - 4 来源：NIV2 任务 154 HateXPlain 分类 - 模板 10

输入：{输入}, {标签}

模板：

详细说明：输入是一条可能是恶意言论、冒犯或正常的推文。恶意言论和冒犯的推文将目标定向于一个社区。给出这样的推文，输出推文中的目标社区。社区将是以下九个值之一：{标签}。如果推文没有针对任何社区，输出'无'。一条推文只针对一个社区。
问：{输入}
答：

收集 - 01 来源：注释者

输入：{句子}, {选项A}, {选项B}, {选项C}, {选项D}, {选项E}, {选项F}, {选项G},{选项H}, {选项I}, {选项J}, {选项K}

模板：

请确定给定句子的正确语言。请从 A、B、C、D、E、F、G、H、I、J 以及 K 中选择最佳答案。

句子：{句子}
A：{选项A}
B：{选项B}
C：{选项C}
D：{选项D}
E：{选项E}
F：{选项F}
G：{选项G}
H：{选项H}
I：{选项I}
J：{选项J}
K：{选项K}
答案：

收集 - 02 来源：注释员

输入： {sentence}, {choiceA}, {choiceB}, {choiceC}, {choiceD}, {choiceE}, {choiceF}, {choiceG},{choiceH}, {choiceI}, {choiceJ}, {choiceK}

模板：

{sentence}
上述语言是什么语言？ A：{choiceA} B：{choiceB} C：{choiceC} D：{choiceD} E：{choiceE} F：{choiceF} G：{choiceG} H：{choiceH} I：{choiceI} J：{choiceJ} K：{choiceK}

收集 - 03 来源：注释员

输入： {sentence}, {choiceA}, {choiceB}, {choiceC}, {choiceD}, {choiceE}, {choiceF}, {choiceG},{choiceH}, {choiceI}, {choiceJ}, {choiceK}

模板：

你正在参加一个要求你识别给定句子所用语言的测试。为了帮助你缩小选择范围，我们已经把这个问题变成了多选题。请仔细阅读句子和以下的每一个答案，然后从"A"，"B"，"C"，"D"，"E"，"F"，"G"，"H"，"I"，"J"或"K"中选择句子的正确语言。
句子：{sentence}
- A：{choiceA}
- B：{choiceB}
- C：{choiceC}
- D：{choiceD}
- E：{choiceE}
- F：{choiceF}
- G：{choiceG}
- H：{choiceH}
- I：{choiceI}
- J：{choiceJ}
- K：{choiceK}

答案:

收集 - 04 来源：注释员

输入： {sentence}，{choiceA}，{choiceB}，{choiceC}，{choiceD}，{choiceE}，{choiceF}，{choiceG}，{choiceH}，{choiceI}，{choiceJ}，{choiceK}

模板：

请从以下选项中选择一个正确对应提供的句子的语言:
句子: {sentence}
选项:
A: {choiceA}
B: {choiceB}
C: {choiceC}
D: {choiceD}
E: {choiceE}
F: {choiceF}
G: {choiceG}
H: {choiceH}
I: {choiceI}
J: {choiceJ}
K: {choiceK}
你的答案：

收集 - 05 来源：注释员

输入： {sentence}，{choiceA}，{choiceB}，{choiceC}，{choiceD}，{choiceE}，{choiceF}，{choiceG}，{choiceH}，{choiceI}，{choiceJ}，{choiceK}

模板：

输入
- 句子: {sentence}
- A: {choiceA}
- B: {choiceB}
- C: {choiceC}
- D: {choiceD}
- E: {choiceE}
- F: {choiceF}
- G: {choiceG}
- H: {choiceH}
- I: {choiceI}
- J: {choiceJ}
- K: {choiceK}

输出
- 答案：

收集 - 06　来源：注释者

输入： {句子}, {选项A}, {选项B}, {选项C}, {选项D}, {选项E}, {选项F}, {选项G},{选项H}, {选项I}, {选项J}, {选项K}

模板：

阅读下面的文本，通过选择列表中的一种选项（A、B、C、D、E、F、G、H、I、J、K）来确定正确的语言：

文本：{句子}

A: {选项A}
B: {选项B}
C: {选项C}
D: {选项D}
E: {选项E}
F: {选项F}
G: {选项G}
H: {选项H}
I: {选项I}
J: {选项J}
K: {选项K}

回答：

收集 - 07　来源：注释者

输入： {句子}, {选项A}, {选项B}, {选项C}, {选项D}, {选项E}, {选项F}, {选项G},{选项H}, {选项I}, {选项J}, {选项K}

模板：

请读取下方的句子，然后选择你认为最可能出自哪种语言的选项。你的答案应是 "A"，"B"，"C"，"D"，"E"，"F"，"G"，"H"，"I"，"J" 或 "K"
句子：{句子}
选项：
A: {选项A}
B: {选项B}
C: {选项C}
D: {选项D}
E: {选项E}
F: {选项F}
G: {选项G}
H: {选项H}
I: {选项I}
J: {选项J}
K: {选项K}
回答：

收集 - 08 来源：注释者

输入：{句子}, {选项A}, {选项B}, {选项C}, {选项D}, {选项E}, {选项F}, {选项G},{选项H}, {选项I}, {选项J}, {选项K}

模板：

请给出以下句子所使用的语言。每个句子都会给出五个选项，请输出相应的选项（即 A、B、C、D、E、F、G、H、I、J 或 K）以代表相应的答案。

句子: {句子}
选项::
A: {选项A}
B: {选项B}
C: {选项C}
D: {选项D}
E: {选项E}
F: {选项F}
G: {选项G}
H: {选项H}
I: {选项I}
J: {选项J}
K: {选项K}
答案：

Collected - 09 来源：注释器

输入：{sentence}，{choiceA}，{choiceB}，{choiceC}，{choiceD}，{choiceE}，{choiceF}，{choiceG}，{choiceH}，{choiceI}，{choiceJ}，{choiceK}

模板：

给定句子：{sentence}，从选项中选择正确的语言 A. {choiceA} B. {choiceB} C. {choiceC} D. {choiceD} E. {choiceE} F. {choiceF} G. {choiceG} H. {choiceH} I. {choiceI} J. {choiceJ} K.

{choiceK}
- A: {choiceA}
- B: {choiceB}
- C: {choiceC}
- D: {choiceD}
- E: {choiceE}
- F: {choiceF}
- G: {choiceG}
- H: {choiceH}
- I: {choiceI}
- J: {choiceJ}
- K: {choiceK}
语言：

Collected - 10 来源：注释器

输入： {sentence}，{choiceA}，{choiceB}，{choiceC}，{choiceD}，{choiceE}，{choiceF}，{choiceG}，{choiceH}，{choiceI}，{choiceJ}，{choiceK}

模板：

{sentence}

这是一个用 {choiceA}， {choiceB}， {choiceC}， {choiceD}， {choiceE}， {choiceF}，{choiceG}，{choiceH}，{choiceI}，{choiceJ}，{choiceK}中的一种语言编写的句子。根据单词和语言结构，我可以判断语言为：

任务设计师 参见 BIG-BENCH 评估文件。

否定 - 1 来源： NIV2 任务 137 新闻评论分类 - 模板 2

输入： {passage}

模板：

首先，你将得到一个任务的定义，然后是一些任务的输入。
将给出的新闻评论分类到它不是用来写作的语言中。这有 {option length} 种语言可以将句子分类到 {options}

{sentence}
输出：

否定 - 2 来源：NIV2 任务 137 新闻评论分类 - 模板 4

输入： {passage}

模板：

指令：将给出的新闻评论归类到它不是用来编写的语言中。这有 {option length} 种语言可以将句子分类到 {options}
输入： {sentence}
输出：

否定 - 3 来源：NIV2 任务 137 新闻评论分类 - 模板 6

输入： {passage}
模板：
给出任务定义和输入，回答输出。将给出的新闻评论归类到它不是用来编写的语言中。这有 {option length} 种语言可以将句子分类到 {options}

{sentence}

否定 - 4 来源：NIV2 任务 137 新闻评论分类 - 模板 8

输入： {passage}
模板：
Q: 将给出的新闻评论归类到它不是用来编写的语言中。这有 {option length} 种语言可以将句子分类到 {options}
{sentence}
A:

否定 - 5 来源：NIV2 任务 137 新闻评论分类 - 模板 10

输入： {passage}
模板：
详细说明：将给出的新闻评论归类到它不是用来编写的语言中。这有 {option length} 种语言可以将句子分类到 {options}
Q: {sentence}
A:

荒谬的 - 1 来源：注解者

输入： {text}
模板：
街头浣熊委员会要求你回答他们的讯问。 {text}

荒谬的 - 2 来源：注解者

输入： {text}
模板：
监视鸟类询问你对种子的知识。 {text}

Nonsensical - 3 来源：注释员

输入： {text}
模板：
达斯·维达要求你回答黑暗面 {text}

Nonsensical - 4 来源：注释员

输入： {text}
模板：
回应火星工作海豚联盟的要求。 {text}

Nonsensical - 5 来源：注释员

输入： {text}

模板：

你正在接受当地松鼠法庭的询问。 {text}

BBH - 认知推理

Closest - 1 来源：FLAN2021 RTE - Template 1

输入： {premise}, {hypothesis}, {options}

模板：

{premise}

基于上述段落，我们可以得出 "{hypothesis}" 的结论吗？

选项： {options}

Closest - 2 来源：FLAN2021 RTE - Template 2

输入： {premise}, {hypothesis}, {options}

模板：

{premise}

基于该段落，我们能否得出以下的句子是真实的？
{hypothesis}

选项： {options}

Closest - 3 来源：FLAN2021 RTE - Template 3

输入： {premise}, {hypothesis}, {options}

模板：

{premise}

请问我们是否能够得出以下结论？
{hypothesis}

选项： {options}

最接近 - 4 来源：FLAN2021 RTE - 模板 4

输入： {前提}, {假设}, {选项}

模板：

{前提}
考虑到先前的文本，这句话是否跟随？
{假设}
OPTIONS: {选项}

最接近 - 5　来源：FLAN2021 RTE - 模板 5

输入：{前提}, {假设}, {选项}

模板：

{前提}
OPTIONS: {选项}
问题：我们可以推断以下内容吗？
{假设}

最接近 - 6　来源：FLAN2021 RTE - 模板 6

输入：{前提}, {假设}, {选项}

模板：

阅读下面的段落并确定假设是否正确。从最后的选项中选择：
{前提}
假设：{假设}
OPTIONS: {选项}
答案是

最接近 - 7　来源：FLAN2021 RTE - 模板 7

输入：{前提}, {假设}, {选项}

模板：

阅读文本并确定句子是否正确：
{前提}
句子：{假设}
OPTIONS: {选项}
A:

最接近 - 8　来源：FLAN2021 RTE - 模板 8

输入：{前提}, {假设}, {选项}

模板：

带有选项的问题：我们可以从上下文中得出以下假设吗？
上下文：
{前提}
假设：{假设}
OPTIONS: {选项}
A:

错误 - 1　来源：NIV2 任务 143 异类分类 - 模板 10

输入：{输入}, {类别}

模板：

详细说明：给出四个词的集合，生成词所属的类别。词由逗号分隔。可能的类别是 {类别}
问题：{输入}
答案：

错误 - 2 来源：NIV2 任务 137 新闻评论分类 - 模板 10

输入： {输入}, {选项}

模板：

详细说明：将给定的新闻评论归类为其所用的语言。有{选项长度}种语言可以将句子分类为{选项}
Q：{输入}
A：

错误 - 3 来源：Flan2021 - Sentiment140 - 模板 1

输入： {输入}, {选项}

模板：

{文本}
从选项中选择你的答案。这条推特的情绪是什么？
选项：{选项}...我认为答案是

错误 - 4 来源：Flan2021 - Sentiment140 - 模板 6

输入： {输入}, {选项}

模板：

从选项中选择你的答案。如何描述这条推特的情绪？
{文本}
{选项}

错误 - 5 来源：NIV2 任务 1422 MathQA 物理 - 模板 10

输入： {输入}, {选项}

模板：

详细说明：在此任务中，你需要回答给出的关于物理的多项选择问题。将你的答案分类为{字母长度}
Q: 问题：{输入}
{选项}
A：

错误 - 6 来源：NIV2 任务 562 语言识别 - 模板 10

输入： {文本}, {选项}

模板：

详细说明：在此任务中，给出了一个输入句子，该句子可以用{选项}种语言。总共有{选项长度}种语言。你的任务是识别输入句子的语言。输入句子只能用所提供的{选项长度}种语言中的任意一种。
Q：{文本}
A：

错误 - 7 来源：NIV2 任务 1193 课程分类 - 模板 10

输入： {文本}, {选项}

模板：

详细说明：在此任务中，你将获得一道印度食品的名称。你需要将食品分类为{选项}
Q：{输入}
A：

不正确 - 8 来源： NIV2 - 任务 56 - 模板 10

输入： {段落}, {问题}, {正确答案}

模板：

详细说明：在此任务中，你的目标是根据关联的段落，判断一个给定问题的正确答案是否是一个好的正确答案。一个好的正确答案是能够正确并完整地回答问题的答案。一个不好的正确答案只能部分或错误地回答问题。如果你认为给定的正确答案是好的，那么请以"是"回答。否则，请以"否"回答。只有两种可能的回应："是"和"否"。
Q: 段落- {段落} 问题： {问题} 正确答案： {正确答案}
A:

收集到的 - 01 来源：标注者

输入： {前提}, {假设}

模板：

请判断两个句子之间是否存在蕴含关系。如果存在则输出"是"，不存在则输出"否"。
前提：{前提}
假设：{假设}
答案：

Collect - 02 来源：标注者

输入： {前提}, {假设}

模板：

给定的两个句子之间的关系是什么？从'蕴含'和'非蕴含'中选择一种。

句子1：{前提}
句子2：{假设}
关系：

Collect - 03 来源：标注者

输入： {前提}, {假设}

模板：

前提： {前提}
前提的真实性是否蕴含以下假设？
假设：{假设}。

Collect - 04 来源：标注者

输入： {前提}, {假设}

模板：

给定前提'{前提}'，你能否推断出'{假设}'？以'是'或'否'作为你的答案。答案：

Collect - 05 来源：标注者

输入： {前提}, {假设}
模板：
我认为"{前提}"蕴含"{假设}"。

我对吗？

任务设计者 参见 BIG-BENCH 评估文件。

否定 - 1 来源：FLAN2021 RTE - 模板 1
输入： {前提}, {假设}, {选项}
模板：
{前提}

带选项的问题：基于上面的段落，我们能否得出 "{假设}" 的结论？
给出正确答案的相反答案。
选项：{选项}

否定 - 2 来源：FLAN2021 RTE - 模板 2
输入： {前提}, {假设}, {选项}
模板：
{前提}
基于该段落，我们能否得出下面的句子是真的？ 给出正确答案的相反答案。
{假设}

选项：{选项}

否定 - 3 来源：FLAN2021 RTE - 模板 3
输入： {前提}, {假设}, {选项}
模板：
{前提}
带选项的问题：我们能否得出以下结论？ 给出正确答案的反面答案。
{假设}

选项：{选项}

否定 - 4 来源：FLAN2021 RTE - 模板 4
输入： {前提}, {假设}, {选项}
模板：
{前提}
根据前面的文本，这个下面的句子是跟随推理的吗？ 给出正确答案的反面答案。
{假设}

选项：{选项}

否定 - 5 来源：FLAN2021 RTE - 模板 5

输入： {前提}, {假设}, {选项}

模板：

{前提}
选项：{选项}
问题：我们能否推出以下内容？ 给出正确答案的反面答案。
{假设}

否定 - 6 来源：FLAN2021 RTE - 模板 6

输入： {前提}, {假设}, {选项}

模板：

阅读下列段落，并确定假设是否正确。请从最后的选项中选择。回答与正确答案相反的答案:

{前提}

假设：{假设}
选项：{选项}
答案是

否定 - 7 来源：FLAN2021 RTE - 模板 7

输入： {前提}, {假设}, {选项}

模板：

阅读文本，并判断句子是否真实。回答与正确答案相反的答案：

{前提}

句子：{假设}
选项：{选项}
A:

否定 - 8 来源：FLAN2021 RTE - 模板 8

输入： {前提}, {假设}, {选项}

模板：

有选项的问题：我们能否从上下文中得出以下假设？回答与正确答案相反的答案。

上下文：

{前提}

假设：{假设}
选项：{选项}
A:

无意义 - 1 来源：注释者

输入： {文本}

模板：

街头浣熊议会要求你回答他们的询问。{文本}

无意义 - 2 来源：注释者

输入： {文本}

模板：

监视鸟类询问你对种子的知识。{文本}

无意义 - 3 来源：注释者

输入： {文本}

模板：

达斯维达要求你对黑暗面进行回答。{文本}

无意义 - 4 来源：注解员

输入： {text}

模板：

回应火星工作海豚联盟的要求。 {text}

无意义 - 5 来源：注解员

输入： {text}

模板：

你正在接受当地松鼠法庭的询问。 {text}

BBH - Crash Blossom

最接近 - 1 来源：Flan2021 - CosmosQA - Template 1

输入： {context}, {question}, {options}

模板：

{context}

可选问题： {question}
选项：{options}

最接近 - 2 来源：Flan2021 - CosmosQA - Template 2

输入： {context}, {question}, {options}

模板：

{context}

选项： {options}
Q: {question}

最接近 - 3 来源：Flan2021 - CosmosQA - Template 3

输入： {context}, {question}, {options}

模板：

{context}

选项： {options}
回答以下问题： {question}

最接近 - 4 来源：Flan2021 - CosmosQA - Template 4

输入： {context}, {question}, {options}

模板：

{context}

根据前面的段落，对问题 {question} 选择你的答案
选项： {options}

最接近 - 5 来源: Flan2021 - CosmosQA - 模板 5

输入： {上下文}, {问题}, {选项}

模板：

{上下文}

附带选项问答：请根据上述段落给出以下问题的答案：{问题}
选项：{选项}

最接近 - 6 来源: Flan2021 - CosmosQA - 模板 6

输入： {上下文}, {问题}, {选项}

模板：

上下文：{上下文}
问题：{问题}
可能的答案：
{选项}
答案：

最接近 - 7 来源: Flan2021 - CosmosQA - 模板 7

输入： {上下文}, {问题}, {选项}

模板：

阅读以下文章并通过选择选项来回答问题。

{上下文}

{问题}
选项：{选项}...答案：

最接近 - 8 来源: Flan2021 - CosmosQA - 模板 8

输入： {上下文}, {问题}, {选项}

模板：

这个问题有选项。回答有关文本的问题：

{上下文}

{问题}
选项：{选项}

错误 - 1 来源: NIV2 任务 143 分类 - 模板 10

输入： {输入}, {类别}

模板：

详细说明：给定四个词，生成这些词所属的类别。词语由逗号分隔。可能的类别有 {类别}
问题：{输入}
答案：

错误 - 2 来源: NIV2 任务 137 新闻评论分类 - 模板 10

输入： {输入}, {选项}

模板：

详细说明：将给定的新闻评论分类为其所写的语言。有{选项长度}种语言可以将句子分类到{选项}
问题：{输入}
答案：

错误 - 3 来源: Flan2021 - Sentiment140 - 模板 1

输入： {输入}, {选项}

模板：

{文本}
从选项中选择你的回答。这条推文的情绪是什么？
选项：{选项}...我认为答案是

错误 - 4 来源：Flan2021 - Sentiment140 - 模板 6

输入： {输入}, {选项}

模板：

从选项中选择你的答案。这条推文的情绪应该如何描述？
{文本}
{选项}

错误 - 5 来源：NIV2 任务 1422 MathQA 物理 - 模板 10

输入： {输入}, {选项}

模板：

详细说明：在此任务中，你需要回答关于物理的给定选择题。将你的答案分类为{字母长度}

Q：问题：{输入}
{选项}
A：

收集 - 01 来源：注释员

输入：{词}, {句子}, {选项}

模板：

将下列句子中的"{词}"归类为何种词性：{句子}。选项为：{选项}
回答：

收集 - 02 来源：注释员

输入：{词}, {句子}, {选项}

模板：

句子：{句子}
请确定句子中{词}的词性。从{选项}中选择你的答案，并输出最佳选择。

收集 - 03 来源：注释员

输入：{词}, {句子}, {选项}

模板：

'{词}'在'{句子}'中的词性是什么？你只能从以下选项中选择：{选项}。你的答案是：

收集 - 04 来源：注释员

输入：{词}, {句子}, {选项}

模板：

给定一个句子以及句子中含有的一个词，输出该词的词性。
词：{词}
句子：{句子}
选项：{选项}

回答：

收集 - 05 来源：注释员

输入：{词}, {句子}, {选项}

模板：

确定单词的词性。问题：在'{句子}'中'{词}'是{选项}中的哪一个？答案：

任务设计者 参见 BIG-BENCH 评估文件。

否定 - 1 源自: Flan2021 - CosmosQA - 模板 1

输入: {context}, {question}, {options}

模板:

{context}

含有选项的问题: {question}
选项: {options} 答案不是:

否定 - 2 源自: Flan2021 - CosmosQA - 模板 2

输入: {context}, {question}, {options}

模板:

{context}

选项: {options}
问题: {question} 答案不是:

否定 - 3 源自: Flan2021 - CosmosQA - 模板 3

输入: {context}, {question}, {options}

模板:

{context}

选项: {options}
回答以下问题: {question}
答案不是:

否定 - 4 源自: Flan2021 - CosmosQA - 模板 4

输入: {context}, {question}, {options}

模板:

{context}

根据前述段落，为问题 {question} 选择你的答案
选项: {options} 答案不是:

否定 - 5 源自: Flan2021 - CosmosQA - 模板 5

输入: {context}, {question}, {options}

模板:

{context}

带选项的问题：根据上述段落的内容回答以下问题: {question}
选项:{options} 答案不是:

否定 - 6 源自: Flan2021 - CosmosQA - 模板 6

输入: {context}, {question}, {options}

模板:

上下文: {context}
问题 {question}
可能的答案:

{options}
答案不是:

否定 - 7 来源：Flan2021 - CosmosQA - Template 7

输入： {context}, {question}, {options}

模板：

阅读以下文章，从选项中选择答案来回答问题。

{context}

{question}
选项：{options}...答案不是：:

否定 - 8 来源：Flan2021 - CosmosQA - Template 8

输入： {context}, {question}, {options}

模板：

此问题有选项。请回答有关文章的问题：

{context}

{question}
选项：{options} 答案不是：

无意义 - 1 来源：注释者

输入： {text}

模板：

街头浣熊议会要求你回答他们的调查。 {text}

无意义 - 2 来源：注释者

输入： {text}

模板：

监视鸟类询问你关于种子的知识。 {text}

无意义 - 3 来源：注释者

输入： {text}

模板：

达斯·维达要求你回答来自黑暗面的问题 {text}

无意义 - 4 来源：注释者

输入： {text}

模板：

回应火星工作海豚联盟的需求。 {text}

无意义 - 5 来源：注释者

输入： {text}

模板：

你正在经历当地松鼠法庭的询问。 {text}

BBH - 逻辑顺序

Closest - 1 来源: NIV2 - 任务 73 - 模板 2

输入: {问题}, {选项}

模板:

首先，会提供一个任务的定义，然后提供一些任务的输入。
你会收到一个问题和一些答案选项(与 "A", "B", "C", "D" 相关联)。你应该根据常识知识选择正确的答案。避免根据关联性回答问题，答案选项被特意选择来捕捉超越关联性的常识。不要生成除以下字符之外的任何其他内容：{选项字母} 并且每个问题只需提供一个答案。

{问题} {选项}
输出:

Closest - 2 来源: NIV2 - 任务 73 - 模板 4

输入: {问题}, {选项}

模板:

指示：你会收到一个问题和一些答案选项(与 "A", "B", "C", "D" 相关联)。你应该根据常识知识选择正确的答案。避免根据关联性回答问题，答案选项被特意选择来捕捉超越关联性的常识。不要生成除以下字符之外的任何其他内容：{选项字母} 并且每个问题只需提供一个答案。
输入: {问题} {选项}
输出:

Closest - 3 来源: NIV2 - 任务 73 - 模板 6

输入: {问题}, {选项}

模板:

给定任务定义和输入，使用输出来回答。你会收到一个问题和一些答案选项(与 "A", "B", "C", "D" 相关联)。你应该根据常识知识选择正确的答案。避免根据关联性回答问题，答案选项被特意选择来捕捉超越关联性的常识。不要生成除以下字符之外的任何其他内容：{选项字母} 并且每个问题只需提供一个答案。

{问题} {选项}

Closest - 4 来源: NIV2 - 任务 73 - 模板 8

输入: {问题}, {选项}

模板:

问题: 你会收到一个问题和一些答案选项(与 "A", "B", "C", "D" 相关联)。你应该根据常识知识选择正确的答案。避免根据关联性回答问题，答案选项被特意选择来捕捉超越关联性的

常识。不要生成除以下字符之外的任何其他内容：{选项字母} 并且每个问题只需提供一个答案。
{问题} {选项}
答案:

最接近 - 5 来源：NIV2 - 任务73 - 模板10

输入： {问题}, {选项}

模板：

详细指南：给你一个问题和一些答案选项（与“A”、“B”、“C”、“D”相关联）。你应该根据常识选择正确的答案。避免根据关联来回答问题，答案集是为了捕捉超越关联的常识而特意选择的。不要产生除了下列字符之一：{选项字母}，并且每个问题只给出一个答案。
Q：{问题} {选项}
A：

错误 - 1 来源：NIV2 任务1421 MathQA General - 模板10

输入： {输入}, {选项}

模板：

详细指南：在此任务中，你需要回答给定的普通数学多选题。将你的答案分类为 {选项字母}
Q: 问题：{输入}
{选项}
A:

错误 - 2 来源：NIV2 任务1422 MathQA 物理 - 模板10

输入： {输入}, {选项}

模板：

详细指南：在此任务中，你需要回答给定的物理多选题。将你的答案分类为 {字母长度}
Q: 问题：{输入}
{选项}
A:

错误 - 3 来源：Flan2021 - WSC273 - 模板1

输入： {语境}, {选项}

模板：

多选问题：{语境}
{选项}

错误 - 4 来源：Flan2021 - TREC - 模板1

输入： {文本}, {选项}

模板：

问题"{文本}"是关于什么类型的事物？

{选项}
答案：

错误 - 5 来源：Flan2021 - PIQA - 模板1

输入： {输入}, {选项}
模板：
这里有一个目标：{目标}

你将如何实现这个目标？

{选项}

Collected - 1 来源：注释者

输入： {listA}, {listB}, {listC}, {listD}
模板：
四个项目自然是按照顺序或时间顺序排列的。现在，从下列选项中选择这些项目的正确顺序：
A. {listA}
B. {listB}
C. {listC}
D. {listD}

Collected - 2 来源：注释者

输入： {listA}, {listB}, {listC}, {listD}
模板：
你将获得四个包含相同对象但顺序不同的列表。以下哪个列表按照时间顺序正确排列了？

列表：
A. {listA}
B. {listB}
C. {listC}
D. {listD}

Collected - 3 来源：注释者

输入： {listA}, {listB}, {listC}, {listD}
模板：
选择最佳答案，该答案能描述一个按时间顺序排列的序列. 选项：A: {listA}, B: {listB}, C: {listC}, D: {listD}

答案：

Collected - 4 来源：注释者

输入： {listA}, {listB}, {listC}, {listD}

模板：

在这项任务中，选择那些按时间顺序排序得最正确的物品列表。从以下选项中选择，并将输出的相应字母作为“A”、“B”、“C”或“D”之一。
A. {listA}
B. {listB}
C. {listC}
D. {listD}

Collected - 5 来源：注释者

输入： {listA}, {listB}, {listC}, {listD}

模板：

问题：以下哪个列表按时间顺序正确排序了？
从列表中选择正确的顺序：A. {listA}, B. {listB}, C. {listC}, D. {listD}. 答案：

任务设计者 参见 BIG-BENCH 评估文件。

否定 - 1 来源：NIV2 - 任务 73 - 模板 2

输入： {问题}, {选项}

模板：

首先，给你一个任务的定义，然后是该任务的一些输入。
你会收到一个问题和一些答案选项（与“A”，“B”，“C”，“D”相关联）。你需要根据常识知识，选择错误的答案。避免根据相关性回答问题，这些答案是刻意选择的，以获取超越关联的常识。除了以下字符之一外，不要生成其他任何东西：{选项字符}并且每个问题只给出一个答案。

{问题} {选项}
输出：

否定 - 2 来源：NIV2 - 任务 73 - 模板 4

输入： {问题}, {选项}

模板：

指令：你会收到一个问题和一些答案选项（与“A”，“B”，“C”，“D”相关联）。你需要根据常识知识，选择错误的答案。避免根据相关性回答问题，这些答案是刻意选择的，以获取超越关联的常识。除了以下字符之一外，不要生成其他任何东西：{选项字符}并且每个问题只给出一个答案。
输入：{问题} {选项}
输出：

否定 - 3 来源：NIV2 - 任务 73 - 模板 6

输入： {问题}, {选项}

模板：

根据任务定义和输入，以输出形式回答。你会收到一个问题和一些答案选项（与"A", "B", "C", "D"相关联）。你需要根据常识知识，选择错误的答案。避免根据相关性回答问题，这些答案是刻意选择的，以获取超越关联的常识。除了以下字符之一外，不要生成其他任何

东西： {选项字符}且每个问题只提供一个答案。

{问题} {选项}

否定 - 4 来源：NIV2 - 任务 73 - 模板 8

输入： {问题}, {选项}

模板：

Q: 你会收到一个问题和一些答案选项（与"A", "B", "C", "D"相关联）。你需要根据常识知识，选择错误的答案。避免根据相关性回答问题，这些答案是刻意选择的，以获取超越关联的常识。除了以下字符之一外，不要生成其他任何东西： {选项字符}且每个问题只提供一个答案。
{问题} {选项}
A:

否定 - 5 来源：NIV2 - 任务 73 - 模板 10

输入： {question}, {options}

模板：

详细指令：你会收到一个问题和一些答案选项（关联到“A”、“B”、“C”、“D”）。你应该根据常识选择错误的答案。请避免根据关联性来回答问题，答案集是精心选择的，以捕获普遍的常识知识。除了以下字符之一，不要生成其他任何内容：{options letter}，并且每个问题只给出一个答案。
Q：{question} {options}
A:

无意义 - 1 来源：注释者

输入： {listA}, {listB}, {listC}, {listD}

模板：

街头浣熊议会要求你回应他们的询问。
A. {listA}
B. {listB}
C. {listC}
D. {listD}

无意义 - 2 来源：注释者

输入： {listA}, {listB}, {listC}, {listD}

模板：

监控鸟类询问你对种子的知识。
A. {listA}
B. {listB}
C. {listC}
D. {listD}

无意义 - 3 来源：注释者

输入： {listA}, {listB}, {listC}, {listD}

模板：

达斯·维达要求你回应黑暗面。

A. {listA}
B. {listB}
C. {listC}
D. {listD}

无意义 - 4 来源：注释者

输入： {listA}, {listB}, {listC}, {listD}

模板：

回应火星工作海豚工会的要求。

A. {listA}
B. {listB}
C. {listC}
D. {listD}

无意义 - 5 来源：注释者

输入： {listA}, {listB}, {listC}, {listD}

模板：

你正在接受地方松鼠法庭的询问。

A. {listA}
B. {listB}
C. {listC}
D. {listD}

E.4 Paraphrased Instructions

在这里， 我们提供了用于自动生成5中的改写指令的提示。 我们还提供了所有改写指令的JSON文件。

Alpaca 为了生成Alpaca集合中观察到的指令的改写，我们从52002个Alpaca任务中随机独立地抽取了1000个样本，并使用了以下的提示，利用GPT-4生成了指令的改写。

- “改述这句话：\n\n{instruction}改述后的句子：\n\n”
- “将这条指令解释为一个较长的句子\n\n\n{instruction}新的句子:\n”
- “您将收到一条指令：\n\n{指令}现在， 请将它改述为一条具有相同意义的新指令：\n\n”

法兰 我们首先使用此管道[6]复制了Flan-T5的持有指令调谐集。我们从生成的数据中随机抽取了986个数据样本，按照[5]报告的B部分比例。我们使用以下提示使用GPT-4生成所选数据的释义。

- “这是一个输入话语：\n\n{instruction}\n \n \n 现在，你的任务是通过只改变指令内容但保持其它所有内容不变来对输入进行改述。\n 这是新的话语：\n\n”
- “给定的语句是任务指示和实际输入的组合。您的工作是对任务指示进行释义，并保持输入不变。以下是要进行释义的语句：\n\n\n{instruction}\n\n\n 现在，生成新的表述：\n\n\n”
- “您需要提供一个特定任务的表述，并且我需要您对其进行释义。任务中的实际输入、问题和例子都不应被改变。你只需要释义指示。任务:\n\n\n {instruction}\n\n\n\n释义后的表述:\n\n\n”

[6]https://github.com/google-research/FLAN

F Procedures and Surveys

Hi XXX!

Thank you for helping me with my research! It may take up to 30 minutes of your time, and your participation is deeply thanked and will be acknowledged in the final research paper.
The tasks/settings/ID of your instructions are:

MMLU General - Zero-shot - 1
MMLU General - Few-shot - 1
Hindu Knowledge - Zero-shot - 1
Hindu Knowledge - Few-shot - 1
Known Unknowns - Zero-shot - 1
Known Unknowns - Few-shot - 1
Novel Concepts - Zero-shot - 1
Novel Concepts - Few-shot - 1
winowhy - Zero-shot - 1
winowhy - Few-shot - 1
BBQ-Lite - Zero-shot - 1
BBQ-Lite - Few-shot - 1
Strange Stories - Zero-shot - 1
Strange Stories - Few-shot - 1
Emoji Movie - Zero-shot - 1
Emoji Movie - Few-shot - 1

To enter the google docs, click on this link: http://xxx.com/xxx
Be sure to read the instructions.docx on the front page for detailed instructions!
It is optimal that you can get it done before May.1st. If you have any questions regarding any of the procedures, please feel free to text me anytime for clarification!

Thank you!

图 9: Invitation note send to participant

First, I would like to express my appreciation for helping with my research project again!

Background

This research aims to evaluate the robustness of the instruction-tuned Language Models (LMs) with respect to the variation of instructions in zero-shot or few-shot settings. It is commonly acknowledged that multitask instruction tuning on a language model improves its zero-shot and few-shot ability. The model can understand and generalize to unseen instructions that users provide at inference time.

For instance, I use this instruction (prompt) as the Prefix:

"Complete this code written in Java SE11 …"

to the actual code, I want to complete, the LM can understand the task and perform inference accordingly.

Goal

As an NLP practitioner and expert, you can provide **instructions** that will prompt the instruction-tuned LMs well for **the given tasks**. The models in which the instructions might be evaluated are:

- GPT-4 / ChatGPT
- Text Davinci
- Flan-PaLM
- Flan-T5
- T0++
- mT0
- MetaIICL
- OPT-IML
- ChatGLM
- Alpaca

You are very well come to use your experience on these models to come up with the instruction you think will **perform the best**.

Tasks

The participation will take approximately **30 minutes**. You will be given **10-15** tasks/settings. For each task/setting, you are going to put your instruction in the row indicated by the order number. For instance:

"Auto Debugging - Zero-Shot - 5"

means that you are assigned to write an instruction on the task **"Auto Debugging"** with the setting **"Zero-shot,"** and you are putting your answer in the row with ID **5**.

图 10: 给注释者的指示的第一页

For each task, there is a folder with the exact same name. In the folder, there is a .docx file. Open the file, you will see detailed information about the task, including **input-output format, task description,** and **examples.** There will be two main tables - one to record instructions in **Zero-shot** and one to record instructions in **Few-shot**s. In each table, there will be an example provided. **Be sure to put your instruction under the correct table and follow the format of the example!**

If you want to see more examples, the "task.json" all the examples in the test set so you may have a better idea

Data source
The tasks given to you are sampled from the benchmarks BBH-Lite and MMLU.

Thank you!

图 11: 给标注者的指南的第二页

Thank you for helping me on this research project! The goal is to gather instructions from experienced **NLP researchers** on various downstream tasks incorporated in the benchmark *BBH*. Your task is to:

- Write down the instruction (prompt) for this task that you think will **work the best** for this task on **instruction-tuned** Seq2Seq LMs (Flan-T5-XXL, Davinci-text-003, OPT-IML, etc.) at zero-shot and few-shots (in-context learning).

- Please put your instruction in the **corresponding row** in the tables. The few-shots table is one page below the zero-shot table. Please use {…} to denote corresponding information. Note: you **do not** need to use all the information if you think some are distractions.

- For multiple choice tasks, you may either formulate the instruction to let the model output the **exact text** or **number/letter** of the text. Same goes for classification task.

- **Task Information** provides an overview of the task, including its input, output, and task description; **Example** provides an example to the test set so you may have a better grasp of the nature of the task; the tables of **Zero-shot Instruction** and **Few-shots Instruction** are in the following pages.

- Instead of using "\n" or "\t", you may directly use enter or tab.

- The given example also represents the **average** length of the input/output for this task. You may assume the maximum token length of the LM is **4096**

Task Information

Dataset	BIG-Bench
Task	Code Line Description
Metric	Accuracy
Task description	Give an English language description of Python code
Input	program, choiceA, choiceB, choiceC, choiceD
Output	answer

图 12: The first page of the dataset information

Example:

Input
- **program:** for i in range(23):\n\t print(i)
- **choiceA:** prints values from 0 to 22
- **choiceB:** computes first 10 prime numbers
- **choiceC:** prints values from 1 to 10
- **choiceD:** prints 'hello world' to the terminal

Output
- **answer:** prints values from 0 to 22 / **A**

Zero-shot Instruction:

You are given:
- {program}: the text sequence of the input code
- {choiceA}, {choiceB}, {choiceC}, {choiceD}: choices of the interpretation

ID	Instruction
Example	Give an English language description of Python code{program} A. {choiceA} B. {choiceB} C. {choiceC} D. {choiceD} English language description:
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

图 13: 数据集信息的第二页