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

Instruction fine-tuning has recently emerged as a promising approach for improving the zero-shot capabilities of Large Language Models (LLMs) on new tasks. This technique has shown particular strength in improving the performance of modestly sized LLMs, sometimes inducing performance competitive with much larger model variants. In this paper we ask two questions: (1) How sensitive are instruction-tuned models to the particular phrasings of instructions, and, (2) How can we make them more robust to such natural language variation? To answer the former, we collect a set of 319 instructions manually written by NLP practitioners for over 80 unique tasks included in widely used benchmarks, and we evaluate the variance and average performance of these instructions as compared to instruction phrasings observed during instruction fine-tuning. We find that using novel (unobserved) but appropriate instruction phrasings consistently degrades model performance, sometimes substantially so. Further, such natural instructions yield a wide variance in downstream performance, despite their semantic equivalence. Put another way, instruction-tuned models are not especially robust to instruction re-phrasings. We propose a simple method to mitigate this issue by introducing "soft prompt" embedding parameters and optimizing these to maximize the similarity between representations of semantically equivalent instructions. We show that this method consistently improves the robustness of instruction-tuned models. [1]

1 Introduction

Large Language Models (LLMs) have come to dominate NLP, in part because they enable zero- and few-shot adaptation to new tasks via *prompting* [3; 4; 10; 37]. Recent work has demonstrated the promise of fine-tuning such models with natural language instructions. Such *instruction-tuning* improves LLM performance in zero- and few-shot settings, sometimes dramatically, especially for "mid-sized" models [5; 22]. For example, on some benchmarks the instruction-tuned Flan-T5-XL (3B parameters) [5] outperforms GPT-3 (175B), despite being dramatically smaller. Furthermore, LLaMa-7B [27]—after being fine-tuned on large-scale corpora on the Alpaca [26] instruction set—outperforms GPT-3 across a range of NLP benchmarks.

[1] The code and instructions are publicly available at: https://github.com/jiudingsun01/InstructionEval

Preprint. Under review.

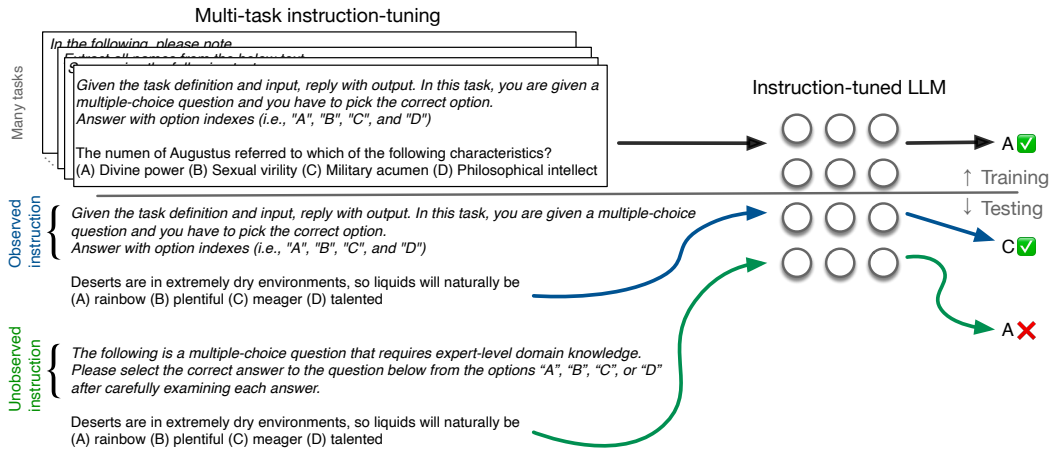


Figure 1: How well do models trained on instruction-tuning datasets generalize to novel instructions (unobserved in training)? Our analysis suggests that they do not do so very well. Above we show a case where pairing an example with an observed instruction yields the correct output, while providing a distinct but semantically equivalent instruction produces an incorrect response. We propose and evaluate a simple method that improves this.

These empirical successes have motivated efforts to curate instruction-augmented task collections for meta-learning [31; 33; 33], and research into improving instruction-tuning [17; 34; 24]. In this work we investigate how robust instruction-tuned models are. More specifically, we ask: How sensitive are instruction-tuned LMs to shifts in instruction phrasings at test time? This is particularly important given that the primary motivation of instruction tuning is to facilitate zero-shot adaptation via natural language instruction: If models are overly sensitive to the particular phrasing of a task instruction it may greatly limit their utility in practice.

Prior work—reviewed at length in Section 2—has established that LLMs do not seem to intuitively "understand" prompts [32; 12; 38], but these efforts did not consider instruction-tuned models specifically. Recent, contemporaneous work to ours [8] investigated the robustness of instruction-tuned models, and found that instruction-tuned T5 [23] is robust to instruction perturbations in few-shot settings, but less so in zero-shot application. We contribute a more in-depth analysis of this phenomena across a much wider set of instruction-tuned models and benchmarks. We also introduce and evaluate a method for improving the robustness of such models, with promising results.

More specifically, we collect a relatively large set of task instructions manually composed by NLP researchers; these are valid instructions but distinct from those found in the Flan collection. We then assess the performance of LLMs fine-tuned on the Flan collection instruction set when given these novel instructions on two benchmarks: MMLU [9] and BBL [25]. We find that using novel instructions in zero-shot application degrades accuracy considerably (Figure 1 illustrates this). For example, comparing the performance of Flan-T5 XXL when using (a) instructions that were seen in training to (b) semantically equivalent but unobserved in training, we observe a 6.9 point drop in absolute performance on average across large benchmarks.

Our **main contributions** are summarized as follows. (1) We perform a comprehensive and in-depth analysis of the robustness of instruction-tuned LLMs across three "families" of such models (Flan-T5 [33], Alpaca [26], and T0 [24]) using large benchmarks [9; 25]. For this we collect a large set of new task instructions manually composed by researchers in NLP; we will release this dataset to facilitate additional work on instruction robustness. We observe substantial performance degradation when using "novel" (unseen in training) instructions. (2) We propose a simple method to improve robustness by imposing an objective encouraging LLMs to induce similar representations for semantically equivalent instructions. We find that this consistently improves the performance realized when using novel but appropriate task instructions.

2 Related Work

Multitask learning and instruction-tuning Training a single text-to-text model capable of providing responses to arbitrary queries has been an aspiration in NLP for at least half a decade. For example, prior to modern prompting and instructing strategies, there were efforts to unify disparate tasks by reframing them as instances of general *question answering* [18; 14; 13]. More recent efforts have focussed on compiling and fine-tuning LLMs on corpora comprising diverse tasks with associated natural language instructions [33; 20; 24]; we refer to this strategy as instruction-tuning. One example of this is `Super-NaturalInstructions` [31], which compiles over 1600 tasks and enriches these with both instructions and negative examples. Similarly, the recently released OPT-IML Bench [11] comprises 2000 NLP tasks. The Flan 2022 task collection [17] additionally features *Chain-of-Thought* (CoT) style "reasoning" chains in instruction templates; the authors show that including these (as well as zero-shot examples and "input inversions") during instruction fine-tuning yields improvements on held-out tasks.

These meta-resources—collections of instructions, tasks, and samples—have facilitated the training of instruction-tuned model families such as Flan-T5, Flan-PaLM [5], and OPT-IML [11].[2] Results have been encouraging; fine-tuning LLMs to follow instructions provides clear and consistent gains across models, and, perhaps most exciting, enables relatively "small" ($\sim$10B) LLMs to achieve near SOTA performance comparable to massive ($\sim$175B) models [26]. This has motivated interest in characterizing how instructions help models, and developing techniques to further improve instruction-tuning; we review recent efforts related to these two research threads below.

Evaluating prompting and instruction capabilities Instructions may be seen as a special sort of model prompting, which a few recent efforts have critically evaluated. For example, Webson and Pavlick ask whether models meaningfully "understand" prompts [32], finding that they largely do not: Performance is often unaffected when irrelevant and misleading prompts are provided. In follow up work, Jang *et al.* [12] evaluates performance on negated prompts, observing an "inverse-scaling" phenomenon in which larger models perform worse in this case.

Other work has attempted to characterize how and when *in-context learning* (ICL)—i.e., including a few examples in prompts—works [19; 29; 6; 1; 36]. ICL is a form of prompting orthogonal to the present effort, as we are primarily interested in the zero-shot adaptability of instruction-tuned LLMs.

In work contemporaneous to ours, Gu *et al.* [8] investigated how robust instruction-tuned models are to instruction perturbations (e.g., dropping words) and paraphrasings. They found that models are relatively robust when given examples (i.e., in few-shot settings), but quite sensitive when used zero-shot; this is qualitatively in line with our findings. Our work differs in important way from this coincident research: (1) We provide a much more comprehensive analysis of robustness; Gu *et al.* considered *only* T5 instruction-tuned on a single instruction dataset, whereas we evaluate three LLMs (and different sizes of each) using five instruction tuning datasets, and we evaluate using over 80 test tasks in all (Gu *et al.* considered only 12). (2) We propose and evaluate a new approach to *improving* the robustness of instruction-tuned models; Gu *et al.* offered no mechanism to improve robustness.

Improving instruction-tuning Past work has also sought to improve instruction-tuning in various ways. One means to do so is to instruction tune based on human feedback [22; 7; 2; 21; 39]. This tends to improve open-ended model responses but degrade performance on downstream tasks. Another strategy is to leverage existing resources to automatically generate instruction-tuning datasets at scale. For example, Wang *et al.* [30] use LLMs to generate instructions, inputs, and outputs and use these to improve their own instruction-following capabilities. In a similarly meta vein, Zhou and colleagues [40] propose using LLMs to engineer prompts. Finally, Ye *et al.* [35] propose "flipping" the standard task by tasking LLMs with generating *instructions*, given an input and label.

[2] Somewhat confusingly, in the case of FLAN and OPT, the corpora (i.e., benchmarks comprising tasks and instructions) and LLMs fine-tuned using them are both referred to with the associated acronym as prefix: For instance, Flan-T5 denotes a T5 [23] variant fine-tuned with the Flan collection.

3 Instruction Datasets

3.1 Evaluation Benchmarks

We evaluate a set of instruction-tuned models on two large benchmarks: MMLU [9] and BIG-BENCH [25]. MMLU is a multiple-choice question-answering benchmark comprising 57 tasks that require expert knowledge. BIG-BENCH is a collaboratively built benchmark containing 204 diverse tasks from various domains; here consider the BIG-BENCH LITE subset, and we include only QA, multi-class, and binary classification tasks, yielding 18 tasks from in all.

3.2 Collecting New Instructions from NLP Researchers

We aim to evaluate instruction-tuned models when they are provided instructions which are semantically equivalent to, but superficially different from, those with which they were trained. To this end, we enlist NLP researchers (graduate students) to compose novel instructions for the tasks considered; these particular instruction phrasings were therefore *unobserved* during instruction fine-tuning.

More specifically, we recruited 36 NLP graduate students working in NLP. All had at least some experience with instruction-tuned models and the downstream tasks included in the evaluation benchmarks. For each of the 18 tasks in BBL and all tasks in MMLU, we asked 12 graduate students to write one (distinct) instruction they would use for zero-shot inference with an instruction-tuned model. We provide details on this instruction collection process in Appendix A. We will release all 319 instructions acquired for this work to ensure the reproducibility of this work and to facilitate further research on instruction-tuned model robustness.

4 Evaluating the Robustness of Instruction-tuned LLMs

4.1 Models and Data

We conduct experiments with model variants trained over three instruction collections (these provide *observed* task instructions): P3 [24], Flan-2022 [5], and Alpaca [26]. To facilitate our analyses, we manually identified all instructions that correspond to (a) multiple-choice question answering (QA), (b) binary classification (BC), or tasks that demand "yes" or "no" responses, and (c) multi-class classification (MC), which requires classifying inputs into a finite set of categories.

To evaluate model robustness with respect to instruction phrasings we use two benchmarks: MMLU [9] and BIG-BENCH LITE (BBL) [25] along with the acquired set of novel instructions described in Section 3.2. We include all 57 tasks from MMLU, and 14 of 24 tasks from BBL. From the latter we exclude two tasks that rely on generation metrics, four that use exact-match, and four that contain tokens unrecognized by the T5 and/or LLaMa tokenizer (e.g., inputs are emojis in one task).

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}"?

Table 1: Examples of observed instructions we collected for three general types of tasks.

We use the same instructions for all tasks in the same category, taken from the published instruction tuning datasets associated with each model. These instructions are general, e.g., in the case of classification they request that the model consider an example with respect to categorization criteria and label space provided by the instance, and select an appropriate category (examples in Table 1). One can "mix-and-match" such instructions so long as they are appropriate for the task type.

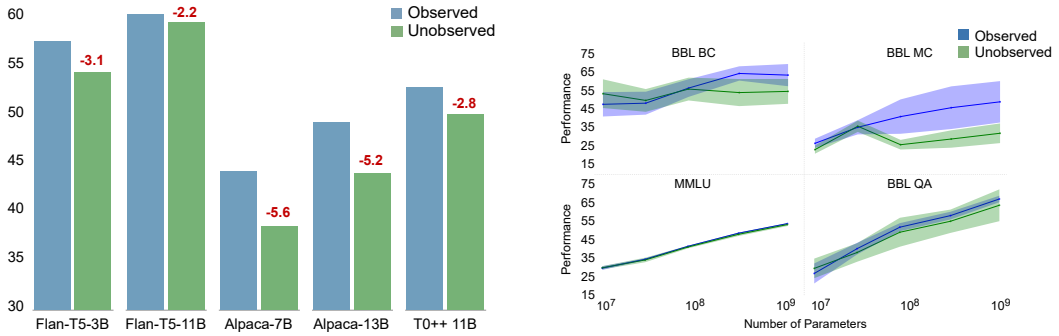
Observed Instructions			
Instruction Type	QA	MC	BC
Flan	50	35	18
Alpaca	20	20	11
P3	13	8	7

Unobserved Instructions		
Number of tasks	1	14
Instructions per task	20	10
Total instructions	20	140

Table 2: Counts of instruction phrasings (unobserved and observed) we use for evaluations.

4.2 Results

We present the main aggregated analysis results in Figure 2 and Table 3. The take-away here is that using instructions unobserved in training—but manually composed for the task at hand and so semantically appropriate—leads to considerable degradation in performance: On average, unobserved instructions reduce accuracy by over five points across models considered. Table 3 reports results disaggregated by task type; we observe that classification tasks are most harmed by use of novel instructions. We provide additional, more granular (dataset-level) results in the Appendix.



(a) Average zero-shot performance over all tasks when using observed and unobserved instructions.

(b) Performances of Flan-T5 using observed and unobserved instructions as a function of model size.

Figure 2: Using novel but valid instructions at test time (phrasings unobserved in training) consistently degrades the performance of instruction-tuned LLMs (a). Scale does not necessarily fix this (b).

4.3 A Closer Look at Instruction Robustness

Above we used general instructions requesting the model to perform tasks (Table 1). Here we delve further into the performance degradation observed when using novel instructions. We report a curious result highlighting the degree to which models rely on having previously observed instructions: Incorrect but observed instructions outperform appropriate but unobserved instructions (Figure 3).

We come to this observation by evaluating the performance of Flan-T5-XXL (11B) using six instruction types over seven datasets from BIG-BENCH. In particular, this includes (variants of) two instructions *observed* in training: **Closest** is the instruction from the most similar task in the instruction-tuning set; **Incorrect** is an observed instruction for a *completely different* and inappropriate task (but which has the same desired output format, e.g., classification)—intuitively these should not yield the desired behavior; **Negated** is the same as **closest**, but we negate the instruction to indicate that it should *not* perform the task.

For *unobserved* instructions, we consider: **Task designer**, the instruction (task prefix) provided by the author of the task in BIG-BENCH, and; **Newly collected**, or the novel instructions collected from NLP graduate students, described above. As a control for reference, we also consider **Nonsensical**, which is a random "instruction" completely irrelevant to any task.

Figure 3 reports average results for these variants. Consistent with our findings, using instructions unobserved in training degrades performance. Strikingly, here we also find that using an *inappropriate but observed* instruction outperforms using *appropriate but unobserved* instructions. This indicates that instruction-tuned models—or at least modestly sized ones we have evaluated here—may in some

Model	MMLU Avg.	MMLU Std.	BBL-QA Avg.	BBL-QA Std.	BBL-BC Avg.	BBL-BC Std.	BBL-MC Avg.	BBL-MC Std.	**Overall** Avg.	**Overall** 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	

Table 3: Results using observed and unobserved instructions across benchmark tasks (grouped by type). Performance degrades—sometimes by 10+ points—when one uses (UNOBSERVED) instructions, suggesting that instruction-tuned models are not particularly robust. BC, MC, and QA stand for binary classification, multi-class classification, and question answering, respectively.

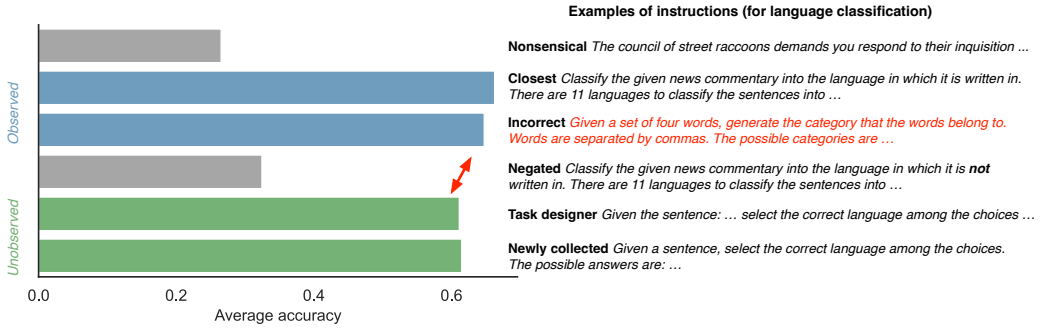


Figure 3: ***Incorrect* but observed instructions perform better on average than *correct* but unobserved instructions**. We report averages over benchmarks, but show example instructions on the right for a specific, illustrative task. We provide all instructions in the Appendix.

way overrely on having observed instructions in training, and do not generalize to new instructions and phrasings as we might hope. We provide all the instructions and results in the Appendix.

4.4 Scaling

Does instruction robustness begin to emerge as a function of scale? To attempt to answer this, we repeated all experiments from Table 3 with Flan-T5 model sizes ranging from small (80M parameters) to XXL (11B). We observe in Figure 2b that the disparity between results achieved with observed versus unobserved instructions **does not** seem to decrease with model scale, at least up to this point. That said, massive models (175B+) may offer greater robustness. However, we reiterate that much of the excitement about instruction tuning is the possibility that this technique appears to allow much smaller models to achieve results competitive with massive alternatives.

4.5 Robustness with Semantic Distance

One observation in 4.2 is that performance on MMLU is less affected by using unobserved instructions. MMLU is a benchmark with 57 QA tasks about different knowledge domains; these tasks all share a similar form of input-output (question, four choices $\rightarrow$ answer). During instruction collection, we treated all tasks in MMLU as a general QA task and asked NLP researchers to write general QA instructions. As a result, we hypothesize that these instructions are comparatively similar to the observed instructions, and this in turn explains the relative robustness in this case.

We empirically verify this in Figure 4 and Table 4. For each instance (instruction plus example), we extract the representation at the penultimate layer for the first decoded token. We use tSNE [28] to visualize these representations of observed and unobserved instructions over instances in MMLU and BBL. Figure 4 shows that in the case of MMLU the unobserved instructions we collected are quite similar to the observed, while there is a greater separation between unobserved and observed instructions in BBL. We also provide a numerical measurement of this phenomonen in Table 4. We report the average $\ell 2$ distance between representations of unobserved instructions and those of their nearest observed counterparts. We see that MMLU unobserved instructions are, on average, closer to the nearest observed instruction; this correlates with the lower observed performance drop. These findings are in line with the hypothesis that the unobserved instructions for MMLU are more similar to the observed instructions for this dataset, and this likely explains the apparent robustness in this case.

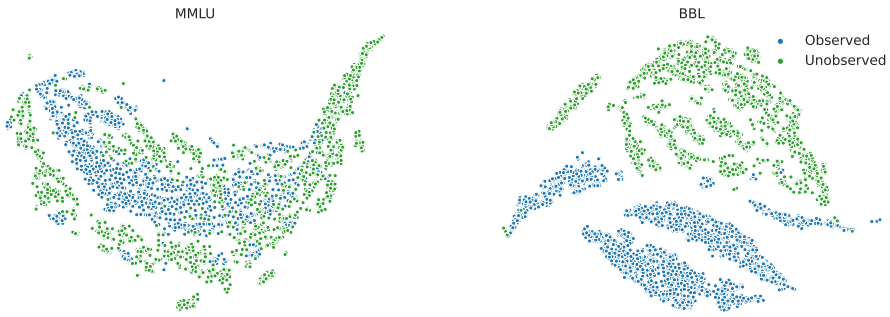


Figure 4: tSNE plots of representations for the first decoded tokens of 300 randomly sampled examples from MMLU and BBL with Flan-T5 (XXL). Embeddings of observed and unobserved instructions for MMLU are similar, while for BBL they are quite different. This result holds across most but not all models considered: See the D for visualizations over all models.

We plot mean performance degradation (as %) as a function of average similarity between the similarity of the first decoded tokens (following *unobserved* instructions) and the same for the *most similar observed* instruction. The negative slope implies the intuitive relationship: Instructions that are dissimilar (in terms of model representations) tend to result in poorer performance. However, the relationship is relatively weak, yielding an intercept estimate of -0.8 and a slope of -0.2 (p =0.08).

4.6 Robustness Under In-Context Learning (ICL)

Previous study [8] has shown that the LLMs are less sensitive to prompt / instruction variation when few-shot examples are provided in context. While we are focused on zero-shot capabilities, for completeness, we re-ran all experiments in a few-shot setting. We report these results in the C. The main finding is that while some discrepancy remains, in general ICL **slightly** decreases the sensitivity of models to the use of unobserved instructions. This is intuitive, given that the examples themselves likely imply the desired task and may affect the distribution.

5 Aligning Equivalent Instructions

We now introduce a simple, lightweight, but effective method to improve the robustness of instruction-tuned LLMs. The intuition is to introduce a term in the objective which explicitly encourages the model to yield similar predictions (and hence similar representations) for the same input when provided distinct but semantically equivalent instructions.

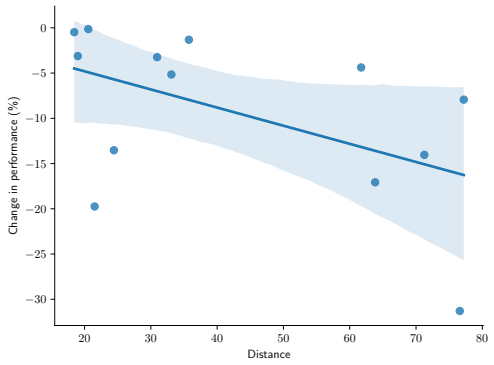


Figure 5: Plots of average degradations in performance versus the semantic distance while using unobserved instructions.

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

Table 4: Average degradations in performance for four categories. It could be seen that MMLU has minimal average distance, which indicates a smaller distribution shift, and hence leads to the smallest degradation

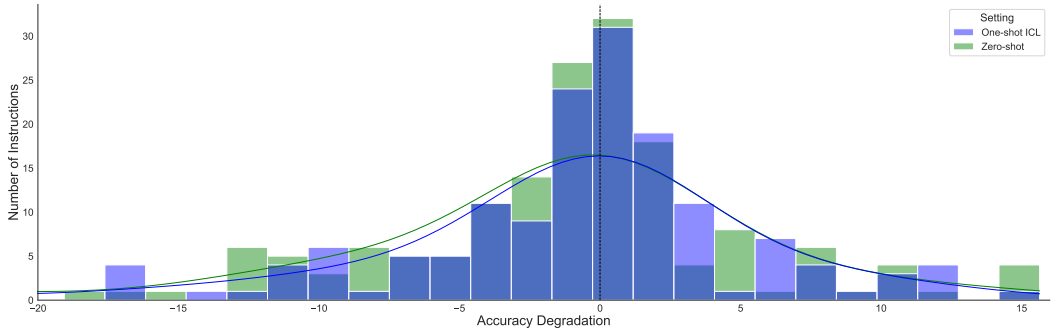


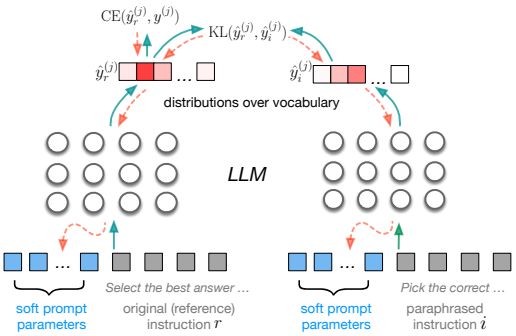
Figure 6: The performance degradation when using unobserved instruction at BBL and MMLU with Flan-T5-XXL. We plot the accuracy degradation of all the unobserved instructions compared with the average accuracy of the observed ones. It could be seen that under one-shot in-context learning, the model is slightly more robust as the performance difference converges closer to 0

More specifically, we aim to align semantically equivalent instructions in the space induced by the model. To this end we introduce soft embedding parameters with dimensions $\mathbb{R}^{d\times n}$; this is equivalent to adding n novel tokens (with embedding dimension d) as prefixes to inputs (preceding instructions). The intuition is to push the representations for semantically equivalent tasks close together. To this end, we add additional term to the loss: The KL-divergence $\mathcal{L}_{\text{KL}}$ of the output probabilities between a reference instruction for a given task and paraphrased (semantically equivalent) version of the same. We combine this with the standard cross-entropy loss, and fine-tune *only* the introduced soft prompt parameters under this objective (Figure 7). Here λ is a loss-weighting hyper-parameter, $\hat{y}_i^{(j)}$ and $\hat{y}_r^{(j)}$ are the distributions over the vocabulary $\mathcal{V}$ induced by the model with paraphrased instruction i and the reference instruction r at token position j.[3]

Optimizing for the above objective requires paraphrased instructions i for each task in the training data; we generate these automatically as follows. For instruction-tuning dataset, we sample a small amount of training data to use for alignment. We paraphrase these reference instructions using GPT-4. For the Alpaca collection, we randomly sampled 1000 tasks and paraphrased them with three prompts, and collected the top three candidates under temperature 0.5. For the Flan collection, we randomly sampled 986 instances from the mixture with 3 prompts with greedy decoding.

For fine-tuning, we then create instances for each example by pairing them with every distinct instruction available for the corresponding task. We then form batches by including one instance featuring

[3]We pad instances such that the lengths in a given batch are effectively equal; the sum is therefore from 1 to the length associated with the current batch, we omit this for simplicity.



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

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

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

Figure 7: Schematic depiction of the proposed instruction alignment method (left) and associated loss terms (right). Dotted (red) lines indicate backpropagation; we update only the soft prompt parameters, which we show yields performance superior to fine-tuning all model parameters.

the original instruction and the rest comprising paraphrased instructions. For the implementation of the prefix, we follow the setting of [16], which freezes the model parameters and just trains the prefix embeddings with the MLP layers.

6 Results

We experiment with the proposed method using two representative instruction-tuned LLMs: Flan-XL (3B) and Alpaca (7B). We compare the canonical versions of these models trained in the usual way (the same evaluated in Table 3) to variants fine-tuned using our proposed approach. We ablate components of our method to tease out the contributions of data and objectives.

Specifically, we consider variants where we: Fine-tune all model parameters on the additional, automatically generated instruction paraphrases (FT); impose the new KL loss term (again fine-tuning all model parameters; FT+KL); introduce the additional soft prompt parameters and fine-tune on the paraphrase instances, but without KL (PT); and then the full proposed strategy, which introduces the soft prompt parameters and optimizes them for the loss augmented with the KL term (**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)**

Table 5: Results and ablations of the proposed soft prompt alignment method. All ablated versions use the augmented set with automatically paraphrased instructions. FT refers to simply fine-tuning (with teacher-forcing) on this additional data; PT denotes prefix tuning (i.e., introducing soft prompt parameters); KL refers to the alignment objective that we proposed above. Using all of these components together yields the best performance, especially on unobserved instructions.

We report results in Table 5. Two observations: (1) The proposed soft prompt alignment strategy (**PT+KL**) yields consistent improvements across the tasks and models considered and especially improves performance on unobserved instructions, as anticipated. (2) The full benefit of the approach is realized only when all components—the additional automatically paraphrased training instructions, soft prompt parameters, and additional KL loss term—are in place.

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%

Table 6: Average distances before and after soft prompt alignment with Flan-T5-XL.

Following our approach in 4.5, we take the average distance between observed and unobserved instructions before and after alignment. Table 6 shows that our method brings observed and unobserved instruction representations closer together. The similarity is most increased in the case of the biggest accuracy gain, further suggesting the mechanism of improvement provided by soft prompt alignment.

7 Conclusions

Instruction-tuned LLMs have emerged as a promising means of achieving zero-shot performance with smaller models that is competitive to, and sometimes even better than, that observed using much larger LLMs [17; 26]. In this work we empirically characterized the *robustness* of such models with respect to instruction rephrasings. In particular, we collected manually composed instructions from 36 graduate students in NLP across 75 tasks, and we evaluated different families of instruction-tuned LLMs (Flan, Alpaca, and T0) when provided observed and unobserved instructions (seen in training and not, respectively). We found that using the latter consistently degrades model performance, indicating that models are unduly sensitive to instruction phrasings.

We then proposed a simple mechanism intended to improve the robustness of instruction-tuned LLMs. This approach entails introducing an additional loss term that penalizes the model for inducing dissimilar distributions over output tokens when using (a) paraphrased instructions as opposed to (b) a reference instruction for the same task. We found that training under this objective consistently (though modestly) improves results, and in particular mitigates the degradation observed when previously unobserved instructions are used.

8 Limitations

This work has important limitations: For example we only evaluated "mid-sized" models (<20B parameters), it is unclear if our findings would generalize to much larger instruction-tuned models. (However, we note that instruction tuning has been most promising for smaller models.) We also restricted our evaluation to three task types: QA and multi-class and binary classification.

Ethics This work does not have an explicit ethical dimension, but we acknowledge that all LLMs are likely to encode problematic biases; it is unclear how instruction-tuning might interact with these.

9 Acknowledgments

This work was supported by the National Science Foundation (NSF) grant 1901117.

We thank Jay DeYoung and Alberto Mario Ceballos Arroyo for their advice and feedback on the paper. We also thank 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 for providing manually written instructions.

References

[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.

Appendix

Table of Contents

A Experimental Setup Details

To ensure reproducibility, we provide all details regarding our evaluation of the robustness of instruction-tuned LLMs.

A.1 Evaluation Protocols

LLMs sometimes generate outputs that are *correct* but different from a (natural language) target. Therefore, we predict answers according to "multiple-choice" grading suggested by BIG-BENCH, by which we take the logits score and argmax over all the possible choices to obtain the prediction. In most cases, this approach yields the same accuracy as using exact match for evaluation. Here are the configurations for all the models we evaluated.

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	

Table 7: The configurations for evaluating different instruction-tuned LMs. The CO_2 emission is estimated by [15]. The total emission is estimated to be equivalent to 679 Km driven by an average ICE car.

A.2 Hyperparameters

We conduct all our training and ablation studies on 8 A100s with 80GB memory. We kept the KL-Loss weight to 0.8. We train both Flan-T5-XL and Alpaca-7B with a batch size of 4. The weight decay is set to be $1e-5$. The learning rate is $5e-4$ for the experiment.

B Disaggregated Results

B.1 Main Results and Scaling Results

In the main paper we reported aggregated results over benchmark corpora. Here we report results on individual datasets for BBL. For MMLU, we evaluate all 57 datasets together, because these are all QA tasks (and we would want a QA model to be capable of answering questions across a diverse set of domains). We report means and stadard deviations of the accuracies achieved over all instructions in Table 9. The numbers on the left of the setting suggest the number of instructions used. We also share even more granular results—reporting the performance for each instruction—in CSV files provided in the supplemental material.

	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)

Table 8: Granular results for Table 3 on each dataset of MMLU We treated all tasks in MMLU equally as general QA and computed the overall accuracy.

	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)

Table 9: Granular results for Figure 2b on each dataset of MMLU We treated all tasks in MMLU equally as general QA and computed the overall accuracy.

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)

Table 10: Granular results for Table 3 on each dataset of category BBL-QA

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)

Table 11: Granular results for Figure 2b on each dataset of category BBL-QA

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)**

Table 12: Granular results for Table 3 on each dataset of category BBL-BC

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)

Table 13: Granular results for Figure 2b on each dataset of category BBL-BC

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)

Table 14: Granular results for Table 3 on each dataset of category BBL-MC

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)**

Table 15: Granular results for Figure 2b on each dataset of category BBL-MC

B.2 "Closer Look" Experiment Results

Here, we provide the detailed results that we reported in 3

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

Table 16: The detailed results of "A Closer Look" experiment. We provide all the instructions picked and their sources in Section E.3. In could be seen that the "Incorrect" but observed instruction, in most cases, outperform the correct but unobserved instructions ("Collected" and "Task Designer").

C Instruction Robustness with In-context Learning

We have been focused on zero-shot settings, but here we also report results achieved under In-context Learning (ICL). We consider one-shot ICL for Flan-T5 because its context window limit precludes providing additional shots in context. We repeat the experiments from the main paper with the one-shot ICL below. We (arbitrarily) take the first instance from every dataset to use as the shot. It could be seen from the result that the performance gap between observed and unobserved instructions is significantly narrowed down with the in-context example provided.

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)**

Table 17: Granular results on each dataset of MMLU for Flan-T5 models and one-shot ICL. We group all QA tasks in MMLU and report overall accuracy on these.

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)

Table 18: Granular results on each dataset of category BBL-QA with one-shot in-context learning.

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)

Table 19: Granular results on each dataset of category BBL-BC with one-shot in-context learning.

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)**

Table 20: Granular results on each dataset of category BBL-MC with one-shot in-context learning.

D Representational Similarity and Model Performance

We re-generate the visualization in Figure 4 for all T5 model sizes in Figure . The qualitative result—representations of tokens following observed instructions are generally dissimilar from those following unobserved instructions—remains largely consistent, although is less pronounced, e.g., for XL (in particular here, the MMLU samples are not as entangled as we might expect).

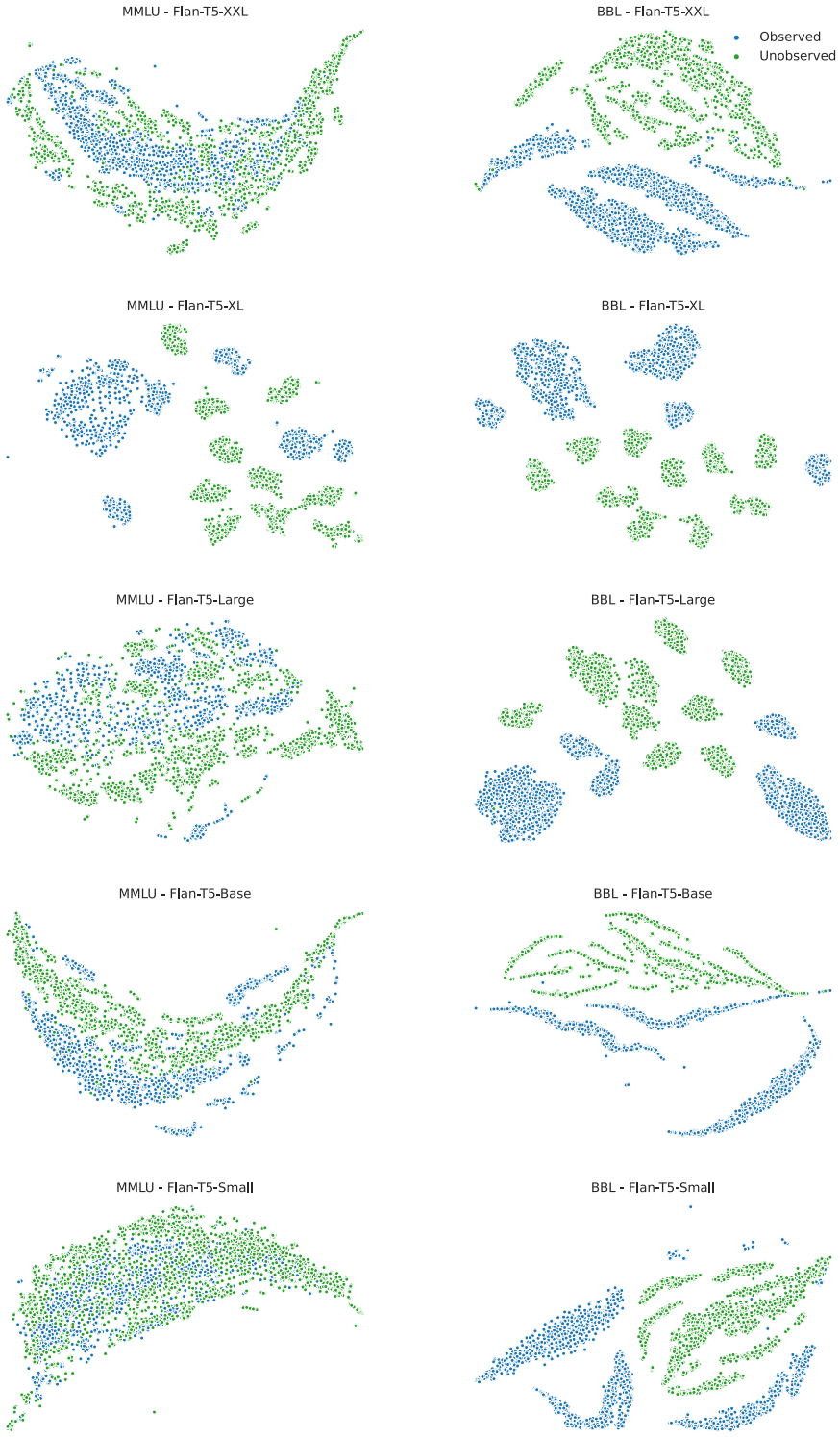


Figure 8: We reproduce Figure 4 from the main paper over all T5 model sizes.

E Instruction Collection

In this section, we report in detail how we collected the instructions we used to evaluate the out of domain (O.O.D.) robustness of instruction-tuned LLMs.

E.1 Observed Instructions

We manually review the instruction-tuning collections that were used to train the three instruction-tuned LLMs that we evaluated: Flan [5], Alpaca [26], and P3 [24]. We consider instructions that are sufficiently general to be able to "mix-and-match" for one of the three task types: QA, Binary Classification, Multiclass Classification.

Below we provide instruction templates for all of the (observed) instructions we aggregated, and their source in the collections:

Flan

For simplicity purposes, we provide the task and indices for the observed instructions templates we aggregated from the Flan collection. The exact code we used to process the data may be found in the publically available Flan repository [4]

QA - 01 Source: NIV2 - Task 73 - Template 1

Input: {question}, {options}

Template:

You are given a question and some answer options (associated with "A", "B", "C", "D"). You should choose the correct answer based on commonsense knowledge. Avoid answering questions based on associations, the set of answers are chosen deliberately to capture common sense beyond associations. Do not generate anything else apart from one of the following characters: {options letter} and only give one answer for each question.

{question} {options}

QA - 02 Source: NIV2 - Task 73 - Template 2

Input: {question}, {options}

Template:

You will be given a definition of a task first, then some input of the task.
You are given a question and some answer options (associated with "A", "B", "C", "D"). You should choose the correct answer based on commonsense knowledge. Avoid answering questions based on associations, the set of answers are chosen deliberately to capture common sense beyond associations. Do not generate anything else apart from one of the following characters: {options letter} and only give one answer for each question.

{question} {options}
Output:

QA - 03 Source: NIV2 - Task 73 - Template 3

Input: {question}, {options}

Template:

Definition: You are given a question and some answer options (associated with "A", "B", "C", "D"). You should choose the correct answer based on commonsense knowledge. Avoid answering questions based on associations, the set of answers are chosen deliberately to capture common sense beyond associations. Do not generate anything else apart from one of the following characters: {options letter} and only give one answer for each question.
Input: {question} {options}
Output:

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

QA - 04 Source: NIV2 - Task 73 - Template 4

Input: {question}, {options}

Template:

Instructions: You are given a question and some answer options (associated with "A", "B", "C", "D"). You should choose the correct answer based on commonsense knowledge. Avoid answering questions based on associations, the set of answers are chosen deliberately to capture common sense beyond associations. Do not generate anything else apart from one of the following characters: {options letter} and only give one answer for each question.
Input: {question} {options}
Output:

QA - 05 Source: NIV2 - Task 73 - Template 5

Input: {question}, {options}

Template:

You are given a question and some answer options (associated with "A", "B", "C", "D"). You should choose the correct answer based on commonsense knowledge. Avoid answering questions based on associations, the set of answers are chosen deliberately to capture common sense beyond associations. Do not generate anything else apart from one of the following characters: {options letter} and only give one answer for each question.
Q: {question} {options}
A:

QA - 06 Source: NIV2 - Task 73 - Template 6

Input: {question}, {options}

Template:

Given the task definition and input, reply with output. You are given a question and some answer options (associated with "A", "B", "C", "D"). You should choose the correct answer based on commonsense knowledge. Avoid answering questions based on associations, the set of answers are chosen deliberately to capture common sense beyond associations. Do not generate anything else apart from one of the following characters: {options letter} and only give one answer for each question.

{question} {options}

QA - 07 Source: NIV2 - Task 73 - Template 7

Input: {question}, {options}

Template:

Teacher: You are given a question and some answer options (associated with "A", "B", "C", "D"). You should choose the correct answer based on commonsense knowledge. Avoid answering questions based on associations, the set of answers are chosen deliberately to capture common sense beyond associations. Do not generate anything else apart from one of the following characters: {options letter} and only give one answer for each question.
Teacher: Now, understand the problem? Solve this instance: {question} {options}
Student:

QA - 08 Source: NIV2 - Task 73 - Template 8

Input: {question}, {options}

Template:

Q: You are given a question and some answer options (associated with "A", "B", "C", "D"). You should choose the correct answer based on commonsense knowledge. Avoid answering questions based on associations, the set of answers are chosen deliberately to capture common sense beyond associations. Do not generate anything else apart from one of the following characters: {options letter} and only give one answer for each question.

{question} {options}
A:

QA - 09 Source: NIV2 - Task 73 - Template 9

Input: {question}, {options}

Template:

Detailed Instructions: You are given a question and some answer options (associated with "A", "B", "C", "D"). You should choose the correct answer based on commonsense knowledge. Avoid answering questions based on associations, the set of answers are chosen deliberately to capture common sense beyond associations. Do not generate anything else apart from one of the following characters: {options letter} and only give one answer for each question.
Problem:{question} {options}
Solution:

QA - 10 Source: NIV2 - Task 73 - Template 10

Input: {question}, {options}

Template:

Detailed Instructions: You are given a question and some answer options (associated with "A", "B", "C", "D"). You should choose the correct answer based on commonsense knowledge. Avoid answering questions based on associations, the set of answers are chosen deliberately to capture common sense beyond associations. Do not generate anything else apart from one of the following characters: {options letter} and only give one answer for each question.
Q: {question} {options}
A:

QA - 11 Source: NIV2 - Task 1420 - Template 1

Input: {question}, {options}

Template:

In this task, you need to provide the correct option for a given problem from the provided options.

Problem:{question}
{options}

QA - 12 Source: NIV2 - Task 1420 - Template 2

Input: {question}, {options}

Template:

You will be given a definition of a task first, then some input of the task.
In this task, you need to provide the correct option for a given problem from the provided options.

Problem:{question}
{options}
Output:

QA - 13 Source: NIV2 - Task 1420 - Template 3

Input: {question}, {options}

Template:

Definition: In this task, you need to provide the correct option for a given problem from the provided options.
Input: Problem:{question}
{options}
Output:

QA - 14 Source: NIV2 - Task 1420 - Template 4

Input: {question}, {options}

Template:
Instructions: In this task, you need to provide the correct option for a given problem from the provided options.
Input: Problem:{question}
{options}
Output:

QA - 15 Source: NIV2 - Task 1420 - Template 5

Input: {question}, {options}
Template:
In this task, you need to provide the correct option for a given problem from the provided options.
Q: Problem:{question}
{options}
A:

QA - 16 Source: NIV2 - Task 1420 - Template 6

Input: {question}, {options}
Template:
Given the task definition and input, reply with output. In this task, you need to provide the correct option for a given problem from the provided options.

Problem:{question}
{options}

QA - 17 Source: NIV2 - Task 1420 - Template 7

Input: {question}, {options}
Template:
Teacher: In this task, you need to provide the correct option for a given problem from the provided options.
Teacher: Now, understand the problem? Solve this instance: Problem:{question}
{options}
Student:

QA - 18 Source: NIV2 - Task 1420 - Template 8

Input: {question}, {options}
Template:
Q: In this task, you need to provide the correct option for a given problem from the provided options.
Problem:{question}
{options}
A:

QA - 19 Source: NIV2 - Task 1420 - Template 9

Input: {question}, {options}
Template:
Detailed Instructions: In this task, you need to provide the correct option for a given problem from the provided options.
Problem:Problem:{question}
{options}
Solution:

QA - 20 Source: NIV2 - Task 1420 - Template 10

Input: {question}, {options}
Template:

Detailed Instructions: In this task, you need to provide the correct option for a given problem from the provided options.
Q: Problem:{question}
{options}
A:

QA - 21 Source: NIV2 - Task 1286 - Template 1

Input: {question}, {options}

Template:

In this task, you are given a multiple-choice question and you have to pick the incorrect option. Answer with option indexes (i.e., {options letter}).

{question} {options}

QA - 22 Source: NIV2 - Task 1286 - Template 2

Input: {question}, {options}

Template:

You will be given a definition of a task first, then some input of the task.
In this task, you are given a multiple-choice question and you have to pick the incorrect option. Answer with option indexes (i.e., {options letter}).

{question} {options}
Output:

QA - 23 Source: NIV2 - Task 1286 - Template 3

Input: {question}, {options}

Template:

Definition: In this task, you are given a multiple-choice question and you have to pick the incorrect option. Answer with option indexes (i.e., {options letter}).
Input: {question} {options}
Output:

QA - 24 Source: NIV2 - Task 1286 - Template 4

Input: {question}, {options}

Template:

Instructions: In this task, you are given a multiple-choice question and you have to pick the incorrect option. Answer with option indexes (i.e., {options letter}).
Input: {question} {options}
Output:

QA - 25 Source: NIV2 - Task 1286 - Template 5

Input: {question}, {options}

Template:

In this task, you are given a multiple-choice question and you have to pick the incorrect option. Answer with option indexes (i.e., {options letter}).
Q: {question} {options}
A:

QA - 26 Source: NIV2 - Task 1286 - Template 6

Input: {question}, {options}

Template:

Given the task definition and input, reply with output. In this task, you are given a multiple-choice question and you have to pick the incorrect option. Answer with option indexes (i.e., {options letter}).

{question} {options}

QA - 27 Source: NIV2 - Task 1286 - Template 7

Input: {question}, {options}

Template:

Teacher: In this task, you are given a multiple-choice question and you have to pick the incorrect option. Answer with option indexes (i.e., {options letter}).
Teacher: Now, understand the problem? Solve this instance: {question} {options}
Student:

QA - 28 Source: NIV2 - Task 1286 - Template 8

Input: {question}, {options}

Template:

Q: In this task, you are given a multiple-choice question and you have to pick the incorrect option. Answer with option indexes (i.e., {options letter}).
{question} {options}
A:

QA - 29 Source: NIV2 - Task 1286 - Template 9

Input: {question}, {options}

Template:

Detailed Instructions: In this task, you are given a multiple-choice question and you have to pick the incorrect option. Answer with option indexes (i.e., {options letter}).
Problem:{question} {options}
Solution:

QA - 30 Source: NIV2 - Task 1286 - Template 10

Input: {question}, {options}

Template:

Detailed Instructions: In this task, you are given a multiple-choice question and you have to pick the incorrect option. Answer with option indexes (i.e., {options letter}).
Q: {question} {options}
A:

QA - 31 Source: NIV2 - Task 1565 - Template 1

Input: {question}, {options}

Template:

This task involves asking a question, providing a set of {options length} options. You are expected to choose the best answer to the question. The output will be in the form of {options letter}, corresponding to which option is chosen.

{question}, Options: [{options}]

QA - 32 Source: NIV2 - Task 1565 - Template 2

Input: {question}, {options}

Template:

You will be given a definition of a task first, then some input of the task.
This task involves asking a question, providing a set of {options length} options. You are expected to choose the best answer to the question. The output will be in the form of {options letter}, corresponding to which option is chosen.

{question}, Options: [{options}]
Output:

QA - 33 Source: NIV2 - Task 1565 - Template 3

Input: {question}, {options}

Template:

Definition: This task involves asking a question, providing a set of {options length} options. You are expected to choose the best answer to the question. The output will be in the form of {options letter}, corresponding to which option is chosen.
Input: {question}, Options: [{options}]
Output:

QA - 34 Source: NIV2 - Task 1565 - Template 4

Input: {question}, {options}

Template:

Instructions: This task involves asking a question, providing a set of {options length} options. You are expected to choose the best answer to the question. The output will be in the form of {options letter}, corresponding to which option is chosen.
Input: {question}, Options: [{options}]
Output:

QA - 35 Source: NIV2 - Task 1565 - Template 5

Input: {question}, {options}

Template:

This task involves asking a question, providing a set of {options length} options. You are expected to choose the best answer to the question. The output will be in the form of {options letter}, corresponding to which option is chosen.
Q: {question}, Options: [{options}]
A:

QA - 36 Source: NIV2 - Task 1565 - Template 6

Input: {question}, {options}

Template:

Given the task definition and input, reply with output. This task involves asking a question, providing a set of {options length} options. You are expected to choose the best answer to the question. The output will be in the form of {options letter}, corresponding to which option is chosen.

{question}, Options: [{options}]

QA - 37 Source: NIV2 - Task 1565 - Template 7

Input: {question}, {options}

Template:

Teacher: This task involves asking a question, providing a set of {options length} options. You are expected to choose the best answer to the question. The output will be in the form of {options letter}, corresponding to which option is chosen.
Teacher: Now, understand the problem? Solve this instance: {question}, Options: [{options}]
Student:

QA - 38 Source: NIV2 - Task 1565 - Template 8

Input: {question}, {options}

Template:

Q: This task involves asking a question, providing a set of {options length} options. You are expected to choose the best answer to the question. The output will be in the form of {options letter},

corresponding to which option is chosen.
{question}, Options: [{options}]
A:

QA - 39 Source: NIV2 - Task 1565 - Template 9

Input: {question}, {options}

Template:

Detailed Instructions: This task involves asking a question, providing a set of {options length} options. You are expected to choose the best answer to the question. The output will be in the form of {options letter}, corresponding to which option is chosen.
Problem:{question}, Options: [{options}]
Solution:

QA - 40 Source: NIV2 - Task 1565 - Template 10

Input: {question}, {options}

Template:

Detailed Instructions: This task involves asking a question, providing a set of {options length} options. You are expected to choose the best answer to the question. The output will be in the form of {options letter}, corresponding to which option is chosen.
Q: {question}, Options: [{options}]
A:

QA - 41 Source: NIV2 - Task 229 - Template 1

Input: {question}, {options}

Template:

You are given a science question (hard-level) and {option length} answer options (associated with {option letter}). Your task is to find the correct answer based on scientific facts, knowledge, and reasoning. Do not generate anything else apart from one of the following characters: {option letter}. There is only one correct answer for each question.

{question} {options}

QA - 42 Source: NIV2 - Task 229 - Template 2

Input: {question}, {options}

Template:

You will be given a definition of a task first, then some input of the task.
You are given a science question (hard-level) and {option length} answer options (associated with {option letter}). Your task is to find the correct answer based on scientific facts, knowledge, and reasoning. Do not generate anything else apart from one of the following characters: {option letter}. There is only one correct answer for each question.

{question} {options}
Output:

QA - 43 Source: NIV2 - Task 229 - Template 3

Input: {question}, {options}

Template:

Definition: You are given a science question (hard-level) and {option length} answer options (associated with {option letter}). Your task is to find the correct answer based on scientific facts, knowledge, and reasoning. Do not generate anything else apart from one of the following characters: {option letter}. There is only one correct answer for each question.
Input: {question} {options}
Output:

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

Input: {question}, {options}

Template:

Instructions: You are given a science question (hard-level) and {option length} answer options (associated with {option letter}). Your task is to find the correct answer based on scientific facts, knowledge, and reasoning. Do not generate anything else apart from one of the following characters: {option letter}. There is only one correct answer for each question.
Input: {question} {options}
Output:

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

Input: {question}, {options}

Template:

You are given a science question (hard-level) and {option length} answer options (associated with {option letter}). Your task is to find the correct answer based on scientific facts, knowledge, and reasoning. Do not generate anything else apart from one of the following characters: {option letter}. There is only one correct answer for each question.
Q: {question} {options}
A:

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

Input: {question}, {options}

Template:

Given the task definition and input, reply with output. You are given a science question (hard-level) and {option length} answer options (associated with {option letter}). Your task is to find the correct answer based on scientific facts, knowledge, and reasoning. Do not generate anything else apart from one of the following characters: {option letter}. There is only one correct answer for each question.

{question} {options}

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

Input: {question}, {options}

Template:

Teacher: You are given a science question (hard-level) and {option length} answer options (associated with {option letter}). Your task is to find the correct answer based on scientific facts, knowledge, and reasoning. Do not generate anything else apart from one of the following characters: {option letter}. There is only one correct answer for each question.
Teacher: Now, understand the problem? Solve this instance: {question} {options}
Student:

QA - 48 Source: NIV2 - Task 229 - Template 8

Input: {question}, {options}

Template:

Q: You are given a science question (hard-level) and {option length} answer options (associated with {option letter}). Your task is to find the correct answer based on scientific facts, knowledge, and reasoning. Do not generate anything else apart from one of the following characters: {option letter}. There is only one correct answer for each question.
{question} {options}
A:

QA - 49 Source: NIV2 - Task 229 - Template 9

Input: {question}, {options}

Template:

Detailed Instructions: You are given a science question (hard-level) and {option length} answer

options (associated with {option letter}). Your task is to find the correct answer based on scientific facts, knowledge, and reasoning. Do not generate anything else apart from one of the following characters: {option letter}. There is only one correct answer for each question.
Problem:{question} {options}
Solution:

QA - 50 Source: NIV2 - Task 229 - Template 10

Input: {question}, {options}

Template:

Detailed Instructions: You are given a science question (hard-level) and {option length} answer options (associated with {option letter}). Your task is to find the correct answer based on scientific facts, knowledge, and reasoning. Do not generate anything else apart from one of the following characters: {option letter}. There is only one correct answer for each question.
Q: {question} {options}
A:

MC - 01 Source: NIV2 - Task 1135 - Template 1

Input: {question}, {options}

Template:

In this task, you will be presented with a question that has multiple possible answers. You should choose the most suitable option out of {options letter}, based on your commonsense knowledge.

{question} Options: {options}

MC - 02 Source: NIV2 - Task 1135 - Template 2

Input: {question}, {options}

Template:

You will be given a definition of a task first, then some input of the task.
In this task, you will be presented with a question that has multiple possible answers. You should choose the most suitable option out of {options letter}, based on your commonsense knowledge.

{question} Options: {options}
Output:

MC - 03 Source: NIV2 - Task 1135 - Template 3

Input: {question}, {options}

Template:

Definition: In this task, you will be presented with a question that has multiple possible answers. You should choose the most suitable option out of {options letter}, based on your commonsense knowledge.
Input: {question} Options: {options}
Output:

MC - 04 Source: NIV2 - Task 1135 - Template 4

Input: {question}, {options}

Template:

Instructions: In this task, you will be presented with a question that has multiple possible answers. You should choose the most suitable option out of {options letter}, based on your commonsense knowledge.
Input: {question} Options: {options}
Output:

MC - 05 Source: NIV2 - Task 1135 - Template 5

Input: {question}, {options}

Template:
In this task, you will be presented with a question that has multiple possible answers. You should choose the most suitable option out of {options letter}, based on your commonsense knowledge.
Q: {question} Options: {options}
A:

MC - 06 Source: NIV2 - Task 1135 - Template 6

Input: {question}, {options}
Template:
Given the task definition and input, reply with output. In this task, you will be presented with a question that has multiple possible answers. You should choose the most suitable option out of {options letter}, based on your commonsense knowledge.

{question} Options: {options}

MC - 07 Source: NIV2 - Task 1135 - Template 7

Input: {question}, {options}
Template:
Teacher: In this task, you will be presented with a question that has multiple possible answers. You should choose the most suitable option out of {options letter}, based on your commonsense knowledge.
Teacher: Now, understand the problem? Solve this instance: {question} Options: {options}
Student:

MC - 08 Source: NIV2 - Task 1135 - Template 8

Input: {question}, {options}
Template:
Q: In this task, you will be presented with a question that has multiple possible answers. You should choose the most suitable option out of {options letter}, based on your commonsense knowledge.
{question} Options: {options}
A:

MC - 09 Source: NIV2 - Task 1135 - Template 9

Input: {question}, {options}
Template:
Detailed Instructions: In this task, you will be presented with a question that has multiple possible answers. You should choose the most suitable option out of {options letter}, based on your commonsense knowledge.
Problem:{question} Options: {options}
Solution:

MC - 10 Source: NIV2 - Task 1135 - Template 10

Input: {question}, {options}
Template:
Detailed Instructions: In this task, you will be presented with a question that has multiple possible answers. You should choose the most suitable option out of {options letter}, based on your commonsense knowledge.
Q: {question} Options: {options}
A:

MC - 11 Source: NIV2 - Task 900 - Template 1

Input: {text}, {options}
Template:

Given a trivia question, classify broad topical category from this list: {options}.

{question}

MC - 12 Source: NIV2 - Task 900 - Template 2

Input: {text}, {options}
Template:
You will be given a definition of a task first, then some input of the task.
Given a trivia question, classify broad topical category from this list: {options}.

{question}
Output:

MC - 13 Source: NIV2 - Task 900 - Template 3

Input: {text}, {options}
Template:
Definition: Given a trivia question, classify broad topical category from this list: {options}.
Input: {question}
Output:

MC - 14 Source: NIV2 - Task 900 - Template 4

Input: {text}, {options}
Template:
Instructions: Given a trivia question, classify broad topical category from this list: {options}.
Input: {question}
Output:

MC - 15 Source: NIV2 - Task 900 - Template 5

Input: {text}, {options}
Template:
Given a trivia question, classify broad topical category from this list: {options}.
Q: {question}
A:

MC - 16 Source: NIV2 - Task 900 - Template 6

Input: {text}, {options}
Template:
Given the task definition and input, reply with output. Given a trivia question, classify broad topical category from this list: {options}.

{question}

MC - 17 Source: NIV2 - Task 900 - Template 7

Input: {text}, {options}
Template:
Teacher: Given a trivia question, classify broad topical category from this list: {options}.
Teacher: Now, understand the problem? Solve this instance: {question}
Student:

MC - 18 Source: NIV2 - Task 900 - Template 8

Input: {text}, {options}
Template:

Q: Given a trivia question, classify broad topical category from this list: {options}.
{question}
A:

MC - 19 Source: NIV2 - Task 900 - Template 9

Input: {text}, {options}

Template:

Detailed Instructions: Given a trivia question, classify broad topical category from this list: {options}.
Problem:{question}
Solution:

MC - 20 Source: NIV2 - Task 900 - Template 10

Input: {text}, {options}

Template:

Detailed Instructions: Given a trivia question, classify broad topical category from this list: {options}.
Q: {question}
A:

MC - 21 Source: Flan2021 - ARC - Template 1

Input: {question}, {options}

Template:

{text}
OPTIONS:
{options}

MC - 22 Source: Flan2021 - ARC - Template 2

Input: {question}, {options}

Template:

Question: {text}?
OPTIONS:{options}
Answer:

MC - 23 Source: Flan2021 - ARC - Template 3

Input: {question}, {options}

Template:

Question: {text}

What is the correct answer to the question from the following choices?
OPTIONS:{options}

MC - 24 Source: Flan2021 - ARC - Template 4

Input: {question}, {options}

Template:

Question: {text}
What is the correct answer to this question?
OPTIONS:{options}...A:

MC - 25 Source: Flan2021 - ARC - Template 5

Input: {question}, {options}

Template:

Choose your answer?

{text}

OPTIONS:{options}

MC - 26 Source: Flan2021 - ARC - Template 6

Input: {question}, {options}
Template:
Answer the question

{text}
OPTIONS:{options}

MC - 27 Source: Flan2021 - ARC - Template 7

Input: {question}, {options}
Template:
{text}

Pick the answer from these options

OPTIONS:{options}

MC - 28 Source: Flan2021 - CosmosQA - Template 1

Input: {context}, {question}, {options}
Template:
{context}

Question with options to choose from: {question}
OPTIONS:{options}

MC - 29 Source: Flan2021 - CosmosQA - Template 2

Input: {context}, {question}, {options}
Template:
{context}

OPTIONS: {options}
Q: {question}

MC - 30 Source: Flan2021 - CosmosQA - Template 3

Input: {context}, {question}, {options}
Template:
{context}

OPTIONS: {options}
Answer the following question: {question}

MC - 31 Source: Flan2021 - CosmosQA - Template 4

Input: {context}, {question}, {options}
Template:
{context}

Based on the preceding passage, choose your answer for question {question}
OPTIONS: {options}

MC - 32 Source: Flan2021 - CosmosQA - Template 5

Input: {context}, {question}, {options}
Template:
{context}

Q with options: Give answer the following question using evidence from the above passage: {question}
OPTIONS:{options}

MC - 33 Source: Flan2021 - CosmosQA - Template 6

Input: {context}, {question}, {options}
Template:
Context: {context}
Question {question}
Possible answers:
{options}
The answer:

MC - 34 Source: Flan2021 - CosmosQA - Template 7

Input: {context}, {question}, {options}
Template:
Read the following article and answer the question by choosing from the options.

{context}

{question}
OPTIONS: {options}...A:

MC - 35 Source: Flan2021 - CosmosQA - Template 8

Input: {context}, {question}, {options}
Template:
This question has options. Answer the question about text:

{context}

{question}
OPTIONS:{options}

BC - 01 Source: NIV2 - Task 56 - Template 1

Input: {paragraph}, {question}, {correct answer}
Template:
In this task, your goal is to judge a correct answer to a given question based on an associated paragraph and decide if it is a good correct answer or not. A good correct answer is one that correctly and completely answers the question. A bad correct answer addresses the question only partially or incorrectly. If you think the given correct answer is good, indicate it by responding "Yes". Otherwise, respond "No". There are only two types of responses possible: "Yes" and "No".

Paragraph- {paragraph} Question: {question} Correct Answer: {correct answer}

BC - 02 Source: NIV2 - Task 56 - Template 2

Input: {paragraph}, {question}, {correct answer}
Template:
You will be given a definition of a task first, then some input of the task.
In this task, your goal is to judge a correct answer to a given question based on an associated paragraph and decide if it is a good correct answer or not. A good correct answer is one that correctly

and completely answers the question. A bad correct answer addresses the question only partially or incorrectly. If you think the given correct answer is good, indicate it by responding "Yes". Otherwise, respond "No". There are only two types of responses possible: "Yes" and "No".

Paragraph- {paragraph} Question: {question} Correct Answer: {correct answer}
Output:

BC - 03 Source: NIV2 - Task 56 - Template 3

Input: {paragraph}, {question}, {correct answer}

Template:

Definition: In this task, your goal is to judge a correct answer to a given question based on an associated paragraph and decide if it is a good correct answer or not. A good correct answer is one that correctly and completely answers the question. A bad correct answer addresses the question only partially or incorrectly. If you think the given correct answer is good, indicate it by responding "Yes". Otherwise, respond "No". There are only two types of responses possible: "Yes" and "No".
Input: Paragraph- {paragraph} Question: {question} Correct Answer: {correct answer}
Output:

BC - 04 Source: NIV2 - Task 56 - Template 4

Input: {paragraph}, {question}, {correct answer}

Template:

Instructions: In this task, your goal is to judge a correct answer to a given question based on an associated paragraph and decide if it is a good correct answer or not. A good correct answer is one that correctly and completely answers the question. A bad correct answer addresses the question only partially or incorrectly. If you think the given correct answer is good, indicate it by responding "Yes". Otherwise, respond "No". There are only two types of responses possible: "Yes" and "No".
Input: Paragraph- {paragraph} Question: {question} Correct Answer: {correct answer}
Output:

BC - 05 Source: NIV2 - Task 56 - Template 5

Input: {paragraph}, {question}, {correct answer}

Template:

In this task, your goal is to judge a correct answer to a given question based on an associated paragraph and decide if it is a good correct answer or not. A good correct answer is one that correctly and completely answers the question. A bad correct answer addresses the question only partially or incorrectly. If you think the given correct answer is good, indicate it by responding "Yes". Otherwise, respond "No". There are only two types of responses possible: "Yes" and "No".
Q: Paragraph- {paragraph} Question: {question} Correct Answer: {correct answer}
A:

BC - 06 Source: NIV2 - Task 56 - Template 6

Input: {paragraph}, {question}, {correct answer}

Template:

Given the task definition and input, reply with output. In this task, your goal is to judge a correct answer to a given question based on an associated paragraph and decide if it is a good correct answer or not. A good correct answer is one that correctly and completely answers the question. A bad correct answer addresses the question only partially or incorrectly. If you think the given correct answer is good, indicate it by responding "Yes". Otherwise, respond "No". There are only two types of responses possible: "Yes" and "No".

Paragraph- {paragraph} Question: {question} Correct Answer: {correct answer}

BC - 07 Source: NIV2 - Task 56 - Template 7

Input: {paragraph}, {question}, {correct answer}

Template:
Teacher: In this task, your goal is to judge a correct answer to a given question based on an associated paragraph and decide if it is a good correct answer or not. A good correct answer is one that correctly and completely answers the question. A bad correct answer addresses the question only partially or incorrectly. If you think the given correct answer is good, indicate it by responding "Yes". Otherwise, respond "No". There are only two types of responses possible: "Yes" and "No".
Teacher: Now, understand the problem? Solve this instance: Paragraph- {paragraph} Question: {question} Correct Answer: {correct answer}
Student:

BC - 08 Source: NIV2 - Task 56 - Template 8

Input: {paragraph}, {question}, {correct answer}
Template:
Q: In this task, your goal is to judge a correct answer to a given question based on an associated paragraph and decide if it is a good correct answer or not. A good correct answer is one that correctly and completely answers the question. A bad correct answer addresses the question only partially or incorrectly. If you think the given correct answer is good, indicate it by responding "Yes". Otherwise, respond "No". There are only two types of responses possible: "Yes" and "No".
Paragraph- {paragraph} Question: {question} Correct Answer: {correct answer}
A:

BC - 09 Source: NIV2 - Task 56 - Template 9

Input: {paragraph}, {question}, {correct answer}
Template:
Detailed Instructions: In this task, your goal is to judge a correct answer to a given question based on an associated paragraph and decide if it is a good correct answer or not. A good correct answer is one that correctly and completely answers the question. A bad correct answer addresses the question only partially or incorrectly. If you think the given correct answer is good, indicate it by responding "Yes". Otherwise, respond "No". There are only two types of responses possible: "Yes" and "No".
Problem:Paragraph- {paragraph} Question: {question} Correct Answer: {correct answer}
Solution:

BC - 10 Source: NIV2 - Task 56 - Template 10

Input: {paragraph}, {question}, {correct answer}
Template:
Detailed Instructions: In this task, your goal is to judge a correct answer to a given question based on an associated paragraph and decide if it is a good correct answer or not. A good correct answer is one that correctly and completely answers the question. A bad correct answer addresses the question only partially or incorrectly. If you think the given correct answer is good, indicate it by responding "Yes". Otherwise, respond "No". There are only two types of responses possible: "Yes" and "No".
Q: Paragraph- {paragraph} Question: {question} Correct Answer: {correct answer}
A:

BC - 11 Source: Flan2021 - MultiRC - Template 1

Input: {paragraph}, {question}, {response}
Template:
{paragraph}

Question: "{question}"

Response: "{response}"

Does the response correctly answer the question?

BC - 12 Source: Flan2021 - MultiRC - Template 2

Input: {paragraph}, {question}, {response}
Template:
{paragraph}

Question: "{question}"

Response: "{response}"

Based on the paragraph, is the response to the question is factually correct?

BC - 13 Source: Flan2021 - MultiRC - Template 3

Input: {paragraph}, {question}, {response}
Template:
{paragraph}

Question: "{question}"

Answer: "{response}"

Is this answer correct?

...I think the answer is

BC - 14 Source: Flan2021 - MultiRC - Template 4

Input: {paragraph}, {question}, {response}
Template:
Paragraph: {paragraph}

Question: "{question}"

Answer: "{response}"

Based on the paragraph, choose if the answer is correct:

BC - 15 Source: Flan2021 - MultiRC - Template 5

Input: {paragraph}, {question}, {response}
Template:
{paragraph}

Choose from options: Based on the paragraph, does the response "{response}" correctly answer the question "{question}"?

BC - 16 Source: Flan2021 - MultiRC - Template 6

Input: {paragraph}, {question}, {response}
Template:
{paragraph}

Choose your answer: According to the above paragraph, the correct answer to the question

"{question}" is "{response}"?

BC - 17 Source: Flan2021 - MultiRC - Template 7

Input: {paragraph}, {question}, {response}
Template:
{paragraph}

After reading the above, is "{response}" the correct answer to the question "{question}"?

BC - 18 Source: Flan2021 - MultiRC - Template 8

Input: {paragraph}, {question}, {response}
Template:
{paragraph}

Question: "{question}"

Answer: "{response}"

Is this answer to the question correct?

Alpaca

QA/MC - 01 Source: Alpaca Tasks Collection

Input: {question}, {options}
Template:
Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Select the correct letter in the parentheses.

Input:
Question: {question}
{options}

Response:

QA/MC - 02 Source: Alpaca Tasks Collection

Input: {question}, {options}
Template:
Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Select the correct option from the following choices.

Input:
Question: {question}
{options}

Response:

QA/MC - 03 Source: Alpaca Tasks Collection

Input: {question}, {options}

Template:

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Answer this multiple choice question.

Input:
Question: {question}
{options}

Response:

QA/MC - 04 Source: Alpaca Tasks Collection

Input: {question}, {options}

Template:

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Read the answer choices and select the correct one.

Input:
Question: {question}
{options}

Response:

QA/MC - 05 Source: Alpaca Tasks Collection

Input: {question}, {options}

Template:

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Identify the correct answer from the choices below.

Input:
Question: {question}
{options}

Response:

QA/MC - 06 Source: Alpaca Tasks Collection

Input: {question}, {options}

Template:

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Determine which choice is correct and output it.

Input:
Question: {question}

{options}

Response:

QA/MC - 07 Source: Alpaca Tasks Collection

Input: {question}, {options}

Template:

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Refer to the given input and identify the correct answer.

Input:
Question: {question}
{options}

Response:

QA/MC - 08 Source: Alpaca Tasks Collection

Input: {question}, {options}

Template:

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
From the given {options length} options, select the one most relevant to the given input.

Input:
Question: {question}
{options}

Response:

QA/MC - 09 Source: Alpaca Tasks Collection

Input: {question}, {options}

Template:

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Select the most optimal response.

Input:
Question: {question}
{options}

Response:

QA/MC - 10 Source: Alpaca Tasks Collection

Input: {question}, {options}

Template:

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:

Read the answer choices and select the correct one.

Input:
Question: {question}
{options}

Response:

QA/MC - 11 Source: Alpaca Tasks Collection

Input: {question}, {options}

Template:

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Select the best answer out of given options.

Input:
Question: {question}
{options}

Response:

QA/MC - 12 Source: Alpaca Tasks Collection

Input: {question}, {options}

Template:

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Determine which of these options is the correct answer.

Input:
Question: {question}
{options}

Response:

QA/MC - 13 Source: Alpaca Tasks Collection

Input: {question}, {options}

Template:

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Choose the best option.

Input:
Question: {question}
{options}

Response:

QA/MC - 14 Source: Alpaca Tasks Collection

Input: {question}, {options}

Template:

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Choose the best option.

Input:
Question: {question}
{options}

Response:

QA/MC - 15 Source: Alpaca Tasks Collection

Input: {question}, {options}

Template:

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Select the best answer.

Input:
Question: {question}
{options}

Response:

QA/MC - 16 Source: Alpaca Tasks Collection

Input: {question}, {options}

Template:

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Choose the correct answer.

Input:
Question: {question}
{options}

Response:

QA/MC - 17 Source: Alpaca Tasks Collection

Input: {question}, {options}

Template:

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Select the correct answer from a list.

Input:
Question: {question}
{options}

Response:

QA/MC - 18 Source: Alpaca Tasks Collection

Input: {question}, {options}

Template:

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Choose the best answer.

Input:
Question: {question}
{options}

Response:

QA/MC - 19 Source: Alpaca Tasks Collection

Input: {question}, {options}

Template:

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Choose the statement that best suits the given context.

Input:
Question: {question}
{options}

Response:

QA/MC - 20 Source: Alpaca Tasks Collection

Input: {question}, {options}

Template:

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Answer the question based on common sense and your knowledge.

Input:
Question: {question}
{options}

Response:

BC - 01 Source: Alpaca Tasks Collection

Input: {claim}

Template:

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Determine if this claim is true or false:

Input:
Claim: {claim}

Response:

BC - 02 Source: Alpaca Tasks Collection

Input: {sentence}
Template:
Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Is the following sentence true or false?

Input:
{sentence}

Response:

BC - 03 Source: Alpaca Tasks Collection

Input: {statement}
Template:
Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Identify whether the following phrase is a true or false statement

Input:
{statement}

Response:

BC - 04 Source: Alpaca Tasks Collection

Input: {statement}
Template:
Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Check if the following statement is true or false:

Input:
{statement}

Response:

BC - 05 Source: Alpaca Tasks Collection

Input: {statement}
Template:
Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Classify the following statement as true or false:

Input:
{statement}

Response:

BC - 06 Source: Alpaca Tasks Collection

Input: {statement}
Template:
Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Classify the following statement as true or false:

Input:
{statement}

Response:

BC - 07 Source: Alpaca Tasks Collection

Input: {statement}
Template:
Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Do a fact check to confirm the accuracy of the statement and output true or false.

Input:
{statement}

Response:

BC - 08 Source: Alpaca Tasks Collection

Input: {sentence}
Template:
Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Label whether an input sentence is true or false.

Input:
{sentence}

Response:

BC - 09 Source: Alpaca Tasks Collection

Input: {sentence}
Template:
Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Indicate a yes or no answer to the given statement..

Input:
{sentence}

Response:

BC - 10 Source: Alpaca Tasks Collection

Input: {sentence}
Template:
Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Evaluate the following proposal as a yes or no response.

Input:
{sentence}

Response:

BC - 11 Source: Alpaca Tasks Collection

Input: {statement}
Template:
Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Respond to the following statement with a yes or no.

Input:
{statement}

Response:

P3 (T0)

QA - 01 Source: T0 Arc Challenge - Template 1

Input: {question}, {options}
Template:
Here's a problem to solve: {question}

Among the 4 following options, which is the correct answer? {options}

QA - 02 Source: T0 Arc Challenge - Template 2

Input: {question}, {options}
Template:
{question}

Options:{options}

QA - 03 Source: T0 Arc Challenge - Template 3

Input: {question}, {options}
Template:
I am hesitating between 4 options to answer the following question, which option should I choose?
Question: {question}
Possibilities:{options}

QA - 04 Source: T0 Arc Challenge - Template 4

Input: {question}, {options}
Template:
I gave my students this multiple choice question: {question}

Only one answer is correct among these 4 choices:{options}

Could you tell me which one is correct?

QA - 05 Source: T0 Arc Challenge - Template 5

Input: {question}, {options}
Template:
Pick the most correct option to answer the following question.

{question}

Options:{options}

QA - 06 Source: T0 Cos e - Template 1

Input: {question}, {options}
Template:
{question}
Choose the most suitable option to answer the above question.
Options:{options}

QA - 07 Source: T0 Cos e - Template 2

Input: {question}, {options}
Template:
{question}
Choose the most suitable option to answer the above question.
Options{options}

QA - 08 Source: T0 Cos e - Template 3

Input: {question}, {options}
Template:
{question}{options}
The best answer is:

QA - 09 Source: T0 Cos e - Template 4

Input: {question}, {options}
Template:
Pick the option in line with common sense to answer the question.
Question: {question}
Options:{options}
The best answer is:

QA - 10 Source: T0 Cos e - Template 5

Input: {question}, {options}
Template:
Pick the option in line with common sense to answer the question.
Question: {question}
Options:{options}

QA - 11 Source: T0 Cos e - Template 6

Input: {question}, {options}
Template:
Pick the option in line with common sense to answer the question.
Questions: {question}
Options:{options}

QA - 12 Source: T0 OpenbookQA - Template 1

Input: {question}, {options}
Template:
{question}

Choose an answer from this list:{options}

QA - 13 Source: T0 OpenbookQA - Template 2

Input: {question}, {options}
Template:
{question}

Which is the correct answer?{options}

QA - 14 Source: T0 OpenbookQA - Template 3

Input: {question}, {options}
Template:
{question}{options}
Is the right answer "{options letter}"

QA - 15 Source: T0 OpenbookQA - Template 4

Input: {question}, {options}
Template:
{question}

Choices:{options}

QA - 16 Source: T0 OpenbookQA - Template 5

Input: {question}, {options}
Template:
{question}{options}

QA - 17 Source: T0 OpenbookQA - Template 6

Input: {question}, {options}
Template:
{question}{options}

Which is the correct answer?

BC - 01 Source: T0 MultiRC - Template 1

Input: {paragraph}, {question}, {answer}
Template:
{paragraph}

Question: {question}
I found this answer "{answer}". Is that correct? Yes or no?

BC - 02 Source: T0 MultiRC - Template 2

Input: {paragraph}, {question}, {answer}
Template:
{paragraph}
Based on the previous passage, {question}
Is "{answer}" a correct answer?

BC - 03 Source: T0 MultiRC - Template 3

Input: {paragraph}, {question}, {answer}
Template:
{paragraph}
Question: {question}

I am grading my students' exercises. Is the answer "{answer}" correct?

BC - 04 Source: T0 MultiRC - Template 4

Input: {paragraph}, {question}, {answer}
Template:
{paragraph}
{question}
Would it be good to answer "{answer}"?

BC - 05 Source: T0 MultiRC - Template 5

Input: {paragraph}, {question}, {answer}
Template:
{paragraph}
Question: {question}
Is it "{answer}"?

BC - 06 Source: T0 MultiRC - Template 6

Input: {paragraph}, {question}, {answer}
Template:
{paragraph}

Decide whether"{answer}" is a valid answer to the following question:
{question}
Answer yes or no.

BC - 07 Source: T0 MultiRC - Template 7

Input: {paragraph}, {question}, {answer}
Template:
{paragraph}
Question: {question}
Is the correct answer "{answer}"?

BC - 08 Source: T0 MultiRC - Template 8

Input: {paragraph}, {question}, {answer}
Template:
Is "{answer}" a correct answer to the following question?
Question: {question}

Rely on the following text: {paragraph}

BC - 09 Source: T0 MultiRC - Template 9

Input: {paragraph}, {question}, {answer}
Template:
{paragraph}

Question: {question}
I think "{answer}" is a valid answer. Could you confirm? Yes or no?

BC - 10 Source: T0 MultiRC - Template 10

Input: {paragraph}, {question}, {answer}
Template:
{paragraph}
{question}
I was going to say "{answer}". Does that sound right?

MC - 1 Source: T0 DBPedia - Template 1

Input: {question}, {categories}
Template:
{question} Given a list of categories: {categories}, what category does the paragraph belong to?

MC - 2 Source: T0 DBPedia - Template 2

Input: {question}, {categories}
Template:
Pick one category for the following text. The options are - {categories}. {question}

MC - 3 Source: T0 DBPedia - Template 3

Input: {question}, {categories}
Template:
{question} Given a choice of categories {categories}, the text refers to which one?

MC - 4 Source: T0 TREC - Template 1

Input: {question}, {categories}
Template:
Categories: {categories}

What category best describes: {question}
Answer:

MC - 5 Source: T0 TREC - Template 2

Input: {question}, {categories}
Template:
Question: {question}

Descriptors: {categories}

Best Descriptor?

MC - 6 Source: T0 TREC - Template 3

Input: {question}, {categories}

Template:
Which category best describes the following question: {question}

Choose from the following list:
{categories}

MC - 7 Source: T0 TREC - Template 4

Input: {question}, {categories}
Template:
{question}Is this asking about {categories}?

MC - 8 Source: T0 TREC - Template 1

Input: {question}, {categories}
Template:
Is the following question asking about {categories}?

{question}

E.2 Unobserved Instructions

We collected novel, unobserved instructions—i.e.,, not seen in training—by enlisting researchers in NLP to write instructions for tasks *de novo*. To facilitate this we showed each annotator one zero-shot instruction and its few-shot form for MMLU and 12 datasets in BBL based on their field of expertise. We sent out an invitation message to prospective participants, which contained a brief introduction to the goal of the research; we reproduce the full invitation in Figure 9. The in-line link redirected annotators to a designated Google Drive folder which included a detailed description of the procedure (see Figure 10 and Figure 11). For each dataset, we provided detailed information about the task including the description, input-output format, demonstration instruction, and some examples (Shown by Figure 12 and Figure 13). We asked participants to provide a prompt and its few-shot form for this task in the corresponding row of the table.[5] Participants were not shown prompts written by others, to preserve independence. Below we reproduce all unobserved instructions that we collected for each benchmark task.

MMLU

Unobserved - 01 Source: Annotator

Input: {question}, {choiceA}, {choiceB},{choiceC},{choiceD}
Template:
Please act as a domain expert to choose the most suitable answer from the given choices to the question below. Question: {question}. Choices: A. {choiceA} B. {choiceB} C. {choiceC} D. {choiceD}
Please answer the question with your choice only without any other words.

Unobserved - 02 Source: Annotator

Input: {question}, {choiceA}, {choiceB},{choiceC},{choiceD}
Template:
Solve the question with professional knowledge and output the best option for the question from "A", "B", "C", "D" without other words:
Question: {question}
Options:
A: {choiceA}
B: {choiceB}
C: {choiceC}

[5]To evaluate Alpaca, we matched collected instructions to corresponding templates that Alapca with which Alpaca was trained.

D: {choiceD}
Answer:

Unobserved - 03 Source: Annotator

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

Template:

Solve the question which requires deep understanding to the field. {question}
Choose from:
A: {choiceA}
B: {choiceB}
C: {choiceC}
D: {choiceD}
Answer:

Unobserved - 04 Source: Annotator

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

Template:

{question} (A) {choiceA} (B) {choiceB} (C) {choiceC} (D) {choiceD}
The correct answer to this question is (

Unobserved - 05 Source: Annotator

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

Template:

{question}

A. {choiceA} B. {choiceB} C. {choiceC} D. {choiceD}
I know exactly the answer to this question! The correct choice is

Unobserved - 06 Source: Annotator

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

Template:

You are given multiple-choice questions from a variety of domains. For each question, please select an answer from A, B, C, and D, and explain your reasoning.

Question: {question}
The options are:
A: {choiceA}
B: {choiceB}
C: {choiceC}
D: {choiceD}

Answer:

Unobserved - 07 Source: Annotator

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

Template:

Please provide the correct answer to the following question, which requires expert level knowledge by choosing one of the options below and outputting it as your answer:

Question: {question}
Options
A: {choiceA}
B: {choiceB}
C: {choiceC}

D: {choiceD}
Your answer:

Unobserved - 08 Source: Annotator

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

Template:

{question}
Options:
- A {choiceA}
- B {choiceB}
- C {choiceC}
- D {choiceD}
Which option is correct?:

Unobserved - 09 Source: Annotator

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

Template:

Given the question: {question}, and the choices for the answer are A. {choiceA}, B. {choiceB}, C. {choiceC}, D. {choiceD}. Output one of A, B, C, and D to indicate the correct choice. The correct choice is:

Unobserved - 10 Source: Annotator

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

Template:

What is the answer to the question: {question} A. {choiceA}, B. {choiceB}, C. {choiceC}, D. {choiceD}

Unobserved - 11 Source: Annotator

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

Template:

You are given a question that requires knowledge from a specific domain. Question: {question}. Select tha answer from A. '{choiceA}', B. '{choiceB}', C. '{choiceC}', D. and '{choiceD}'. Answer:

Unobserved - 12 Source: Annotator

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

Template:

I want to know the answer to this question: {question}. Please select from the following: A. {choiceA}, B. {choiceB}, C. {choiceC}, D. {choiceD}. Indicate your choice with the letter.

Unobserved - 13 Source: Annotator

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

Template:

Question: {question}. Choices: A. {choiceA}, B. {choiceB}, C. {choiceC}, D. {choiceD}. Answer:

Unobserved - 14 Source: Annotator

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

Template:

Task: Multiple-choice question answering.
Question: {question}
Choices: (A) {choiceA} (B) {choiceB} (C) {choiceC} (D) {choiceD}
Answer: (

Unobserved - 15 Source: Annotator

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

Template:

I am working with an exam question that has four different options. The question is:
{question}
And the choices are:
A. {choiceA}
B. {choiceB}
C. {choiceC}
D. {choiceD}
Here's the answer to the this question:

Unobserved - 16 Source: Annotator

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

Template:

A multiple-choice question is given. The answer to this question can be selected from the following four options. Use your knowledge to find the correct choice: {question}
A. {choiceA}
B. {choiceB}
C. {choiceC}
D. {choiceD}

Unobserved - 17 Source: Annotator

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

Template:

A question is given following with 4 options. Select the most correct options, output one of "A", "B", "C", or "D", and explain your choice with chain of thought.
{question}
A. {choiceA}
B. {choiceB}
C. {choiceC}
D. {choiceD}
Answer:

Unobserved - 18 Source: Annotator

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

Template:

This is a single-choice question coming from exams. Use your knowledge to solve the following question and select the correct answer among "A", "B", "C", and "D". Just output the answer with the corresponding letter!

Question: {question}
Candidate Answers:
A: {choiceA}
B: {choiceB}
C: {choiceC}
D: {choiceD}

The answer is:

Unobserved - 19 Source: Annotator

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

Template:

Please answer the question using your knowledge. Output one of "A", "B", "C", or "D" to indicate your answer: A: {choiceA} B: {choiceB} C: {choiceC} D: {choiceD} {question} Answer:

Unobserved - 20 Source: Annotator

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

Template:

Please use your domain-specific knowledge to answer the following questions:
Question: {question}
The choices are:
- A: {choiceA}
- B: {choiceB}
- C: {choiceC}
- D: {choiceD}

The answer is

BBL - BBQ Lite

Unobserved - 01 Source: Annotator

Input: {question}, {choiceA}, {choiceB},{choiceC}

Template:

Please answer this question without any bias based on your knowledge. {question}
A. {choiceA}
B. {choiceB}
C. {choiceC}Please answer this question with the choice only, either A, B, or C, without any explanations and illustrations.

Unobserved - 02 Source: Annotator

Input: {question}, {choiceA}, {choiceB},{choiceC}

Template:

Use social knowledge, choose the best answer for the question from options to pass social bias measurement:
Question: {question}
A. {choiceA}
B. {choiceB}
C. {choiceC}
Answer:

Unobserved - 03 Source: Annotator

Input: {question}, {choiceA}, {choiceB},{choiceC}

Template:

We need your help to answer a question. Q: {question}
A. {choiceA}
B. {choiceB}
C. {choiceC}

Chooose the correct answer from the options. Be sure to answer the question without any social bias.

Unobserved - 04 Source: Annotator

Input: {question}, {choiceA}, {choiceB},{choiceC}

Template:

In this task, we measure social biases based on a multiple-choice question. Please select the option that you think answers the question the best. {question}. (A) {choiceA} (B) {choiceB} (C) {choiceC}
The best option is (

Unobserved - 05 Source: Annotator

Input: {question}, {choiceA}, {choiceB},{choiceC}

Template:
Answer the following question with the most likely answer based on the choices available and explain why you picked the choice.
{question}
A) {choiceA}
B) {choiceB}
C) {choiceC}

Unobserved - 06 Source: Annotator

Input: {question}, {choiceA}, {choiceB},{choiceC}
Template:
Given the question {question}. Choose the answer from the following options fairly based on the factual information. The options are: A. {choiceA} B. {choiceB} C. {choiceC}
Your answer is:

Unobserved - 07 Source: Annotator

Input: {question}, {choiceA}, {choiceB},{choiceC}
Template:
You are examined with your ability on social fairness by answering questions that containbias content including: age, disability, gender, nationality, physical appearance, race ethnicity, religion, social status, and sexual orientation. {question}
Choose the most appropriate answer from:
A) {choiceA}
B) {choiceB}
C) {choiceC}.

Unobserved - 08 Source: Annotator

Input: {question}, {choiceA}, {choiceB},{choiceC}
Template:
Choose the best option for the social bias measurement task {question} A. {choiceA} B. {choiceB} C. {choiceC} Answer:

Unobserved - 09 Source: Annotator

Input: {question}, {choiceA}, {choiceB},{choiceC}
Template:
{question}
A. {choiceA} B. {choiceB} C. {choiceC}

The most suitable of the multiple choice options is

Unobserved - 10 Source: Annotator

Input: {question}, {choiceA}, {choiceB},{choiceC}
Template:
We are running a test for social biases. We will give you a question and a series of multiple-answers (A, B, C). Please choose from among the answers.

Question: {question}
A: {choiceA}
B: {choiceB}
C: {choiceC}

Answer:

Unobserved - 11 Source: Annotator

Input: {question}, {choiceA}, {choiceB},{choiceC}
Template:
Given the question and three potential options, choose the best possible answer. {question}. (A) {choiceA} (B) {choiceB} (C) {choiceC} The best option is (

Unobserved - 12 Source: Annotator

Input: {question}, {choiceA}, {choiceB},{choiceC}
Template:
Read the following question and then choose the option that best answers it. {question}. (A) {choiceA} (B) {choiceB} (C) {choiceC} Among (A), (B), and (C), the best answer is (

BBL - Code Description

Unobserved - 01 Source: Annotator

Input: {code}, {choiceA}, {choiceB},{choiceC},{choiceD}
Template:
If you are an experienced code reviewer, please give an English language description of Python code{code}. Which is the most suitable? A. {choiceA} B. {choiceB} C. {choiceC} D. {choiceD}

Answer:

Unobserved - 02 Source: Annotator

Input: {code}, {choiceA}, {choiceB},{choiceC},{choiceD}
Template:
You are given a simple line of Python code {code}. Try to find out its English equivalency from the following short sentences: A) {choiceA}, B) {choiceB} C) {choiceC}, D) {choiceD}. The equivalent sentence is:

Unobserved - 03 Source: Annotator

Input: {code}, {choiceA}, {choiceB},{choiceC},{choiceD}
Template:
Choose a right English language description of the given Python Code from the four candidates.
Python Code: {code}
Candidates: A. {choiceA}, B. {choiceB} C. {choiceC}, D. {choiceD}
Answer:

Unobserved - 04 Source: Annotator

Input: {code}, {choiceA}, {choiceB},{choiceC},{choiceD}
Template:
To paint in words the function of the given code {code}, which one of A. {choiceA} B. {choiceB} C. {choiceC} D. {choiceD} is the most accurate description:

Unobserved - 05 Source: Annotator

Input: {code}, {choiceA}, {choiceB},{choiceC},{choiceD}
Template:
Now you are a code explainer. Question: Here is a Python code {code} Please choose the right interpretation of the code from the following: A. {choiceA} B. {choiceB} C. {choiceC} D. {choiceD}

Answer:

Unobserved - 06 Source: Annotator

Input: {code}, {choiceA}, {choiceB},{choiceC},{choiceD}
Template:

We have the following Python code {code}, which one is the correct interpretation, output the best choice from "A", "B", "C", and "D".
- A. {choiceA}
- B. {choiceB}
- C. {choiceC}
- D. {choiceD}
The best choice is

Unobserved - 07 Source: Annotator

Input: {code}, {choiceA}, {choiceB},{choiceC},{choiceD}

Template:

One of the following options: A. {choiceA} B. {choiceB} C. {choiceC} D. {choiceD} is the actual annotation of the python code: "{code}". Which one is it? Answer:

Unobserved - 08 Source: Annotator

Input: {code}, {choiceA}, {choiceB},{choiceC},{choiceD}

Template:

For the Python code snippet {code}, select the appropriate English description from the options below (output both the choice and the description):
A. {choiceA}
B. {choiceB}
C. {choiceC}
D. {choiceD}
Output:

Unobserved - 09 Source: Annotator

Input: {code}, {choiceA}, {choiceB},{choiceC},{choiceD}

Template:

Question: Give the most suitable annotation to this code:
{code}
A. {choiceA}
B. {choiceB}
C. {choiceC}
D. {choiceD}

Unobserved - 10 Source: Annotator

Input: {code}, {choiceA}, {choiceB},{choiceC},{choiceD}

Template:

A.
// {choiceA}
{code}

B.
// {choiceB}
{code}

C.
// {choiceC}
{code}

D. // {choiceD}
{code}

From the four different python code A, B, C, and D, choose the code with the most correct specification.

BBL - Hindu Knowledge

Unobserved - 01 Source: Annotator

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

Template:

Please select the best matched answer for the given question from the choices list below based on Hindu mythology. Question: {question} Choices: A. {choiceA} B. {choiceB} C. {choiceC} D. {choiceD}.
Please respond with the choice only, without any other words.

Unobserved - 02 Source: Annotator

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

Template:

Solve question in the Hindu mythology area, output the best option for the question from "A", "B", "C", "D": Question: {question} Options: A: {choiceA} B: {choiceB} C: {choiceC} D: {choiceD}
Answer:

Unobserved - 03 Source: Annotator

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

Template:

Question: {question}
A: {choiceA} B: {choiceB} C: {choiceC} D: {choiceD}
Hindu knowledge expert: This is easy, the answer is

Unobserved - 04 Source: Annotator

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

Template:

In this task, you have to select the option that best answers the question given your knowledge about Hindu mythology.
Question: {question}
A. {choiceA} B. {choiceB} C. {choiceC} D. {choiceD}
Answer: among A, B, C, and D, the best choice is

Unobserved - 05 Source: Annotator

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

Template:

Answer the following question based on hindu mythology with the most accurate choice
{question}
A: {choiceA}
B: {choiceB}
C: {choiceC}
D: {choiceD}
Answer:

Unobserved - 06 Source: Annotator

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

Template:

{question}

A: {choiceA} B: {choiceB} C: {choiceC} D: {choiceD}
With your expertise inhindu mythology, provide the correct answer:

Unobserved - 07 Source: Annotator

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

Template:
Input:
- Question: {question}
- A: {choiceA}
- B: {choiceB}
- C: {choiceC}
- D: {choiceD}
Output
- Answer:

Unobserved - 08 Source: Annotator

Input: {question}, {choiceA}, {choiceB},{choiceC},{choiceD}
Template:
Choose the best option for the following question in Hindu Mythology {question} A. {choiceA} B. {choiceB} C. {choiceC} D. {choiceD}

Unobserved - 09 Source: Annotator

Input: {question}, {choiceA}, {choiceB},{choiceC},{choiceD}
Template:
{question}
A. {choiceA} B. {choiceB} C. {choiceC} D. {choiceD}
Which of the options A, B, C, D is the correct one? It is

Unobserved - 10 Source: Annotator

Input: {question}, {choiceA}, {choiceB},{choiceC},{choiceD}
Template:
You will be given a series of questions regarding Hindu knowledge. For each question, select among the multiple choice answers (A, B, C, D) and provide an explanation, where applicable.

Question: {question}
A: {choiceA}
B: {choiceB}
C: {choiceC}
D: {choiceD}

Answer:

BBL - Known Unknowns

Unobserved - 01 Source: Annotator

Input: {question}, {choiceA}, {choiceB}
Template:
Please select the best option for the question given to you based on the correct factual knowledge.
Question: {question} A. {choiceA} B. {choiceB}
Please answer with your choice only without any other words.

Unobserved - 02 Source: Annotator

Input: {question}, {choiceA}, {choiceB}
Template:
Verify if the question is unknown, choose your answer from options:
Question: {question}
Options:
A: {choiceA}
B: {choiceB}
Answer:

Unobserved - 03 Source: Annotator

Input: {question}, {choiceA}, {choiceB}

Template:

You are given a question asking about a specific knowledge. You need to respond with eitherthe actual knowledge or it cannot be known.
Question: {question}
Options:
A: {choiceA}
B: {choiceB}
Answer with "A" or "B".

Unobserved - 04 Source: Annotator

Input: {question}, {choiceA}, {choiceB}

Template:

Determine if the question is factually knowable by choosing from the following options:
Q: {question}
(A) {choiceA}
(B) {choiceB}
Answer: (

Unobserved - 05 Source: Annotator

Input: {question}, {choiceA}, {choiceB}

Template:

Answer the following questions based on the list of available choices
{question}
A: {choiceA}
B: {choiceB}
Answer:

Unobserved - 06 Source: Annotator

Input: {question}, {choiceA}, {choiceB}

Template:

{question}

A. {choiceA} B. {choiceB}

With respect to the choices above, the correct one is

Unobserved - 07 Source: Annotator

Input: {question}, {choiceA}, {choiceB}

Template:

Question: {question}
To avoid hallucination, if the answer to this question is unknown, output "B", otherwise output "A"

Unobserved - 08 Source: Annotator

Input: {question}, {choiceA}, {choiceB}

Template:

This is a test of 'hallucination', choose the most appropriate option for the question: {question} A. {choiceA} B. {choiceB}

Unobserved - 09 Source: Annotator

Input: {question}, {choiceA}, {choiceB}

Template:

{question}
A. {choiceA} B. {choiceB}
Which of the choices between A and B is correct?
The correct option is

Unobserved - 10 Source: Annotator

Input: {question}, {choiceA}, {choiceB}
Template:
You will be given questions to test your knowledge of whether or not it is possible to know certain pieces of information. Each question either has an answer that you know or an answer that is unknown. For each of the questions below, please choose from the multiple choices (A, B) and provide an explanation when applicable.

Question: {question}
A: {choiceA}
B: {choiceB}

Answer:

BBL - Logical Deduction

Unobserved - 01 Source: Annotator

Input: {paragraph}, {options}
Template:
You are given a paragraph that describes five objects arranged in order. Please select the best answer from A, B, C, D, and E which the answer contains a statement that is logically consistent with the paragraph.

Paraphraph:{paragraph}{options}

Unobserved - 02 Source: Annotator

Input: {paragraph}, {options}
Template:
The most logically-correct answer, given this paraphraph? {paragraph}{options}

Unobserved - 03 Source: Annotator

Input: {paragraph}, {options}
Template:
You are taking an exam where you will be given a paragraph of text describing five different objects in a sequence that are arranged in a fixed order. To answer the question correctly, you must keep track of where each object is in the sequence and then select the multiple choice answer that best corresponds to the correct answer from ("A", "B", "C", "D", "E"). Please carefully consider the information in the following paragraph and each of the answers before providing the right answer.

Paraphraph:{paragraph}{options}

Unobserved - 04 Source: Annotator

Input: {paragraph}, {options}
Template:
Each of the following paragraphs describes a set of five objects arranged in a fixed order, and the statements in each paragraph are logically consistent. After reading the paragraph, select the best option that describes the arrangement of objects:
{paragraph}{options}

Unobserved - 05 Source: Annotator

Input: {paragraph}, {options}

Template:

Input
- paragraph: {paragraph}{options}
Output:
- Answer:

Unobserved - 06 Source: Annotator

Input: {paragraph}, {options}

Template:

Given the following text describing the correct order of five objects, select the option from (A, B, C, D or E) that is consistent with the text.

text: {paragraph}{options}

answer:

Unobserved - 07 Source: Annotator

Input: {paragraph}, {options}

Template:

The following text describes the arrangement order of five objects. Please read the text and choose the one from the options that matches the logic of the text description. Your answer should be "A", "B", "C", "D" or "E".
Text: {paragraph}{options} Answer:

Unobserved - 08 Source: Annotator

Input: {paragraph}, {options}

Template:

Deduce the order of the five objects and select the logically consistent statement from the given choices. {paragraph}{options} Answer:

Unobserved - 09 Source: Annotator

Input: {paragraph}, {options}

Template:

Please decide which option is correct based on the descriptions in the following article. The article describes the order of the 5 objects, please output the correct option as your answer. Article: {paragraph}{options}
Answer:

Unobserved - 10 Source: Annotator

Input: {paragraph}, {options}

Template:

You are given one passage, which sequentially gives a series of propositions. You task is to answer a given question based on the passage and select the correct answer from A, B, C, D, E.

The passage is: {paragraph}

The candidate answers are: {options}

You: The answer is obvious, I choose

BBL - Novel Concepts

Unobserved - 01 Source: Annotator

Input: {question}, {choiceA}, {choiceB},{choiceC}, {choiceD}, {choiceE}
Template:
Please choose the best option from the listed choices that precisely express the given things in common. {question} A. {choiceA} B. {choiceB} C. {choiceC} D. {choiceD} E. {choiceE}
Please answer with your choice only without any other words.

Unobserved - 02 Source: Annotator

Input: {question}, {choiceA}, {choiceB},{choiceC}, {choiceD}, {choiceE}
Template:
Identify and output the commonality among the given objects:
Objects:{question}
A.{choiceA}
B. {choiceB}
C. {choiceC}
D. {choiceD}
E. {choiceE}
Answer:

Unobserved - 03 Source: Annotator

Input: {question}, {choiceA}, {choiceB},{choiceC}, {choiceD}, {choiceE}
Template:
You are given three objects {question}, choose the option from below where the objectsshare the greatest similarity. A. {choiceA} B. {choiceB} C. {choiceC} D. {choiceD} E. {choiceE}

Unobserved - 04 Source: Annotator

Input: {question}, {choiceA}, {choiceB},{choiceC}, {choiceD}, {choiceE}
Template:
Please select the best option to indicate the commonality between the objects: {question}
A.{choiceA}
B. {choiceB}
C. {choiceC}
D. {choiceD}
E. {choiceE}
Give your answer as one of A, B, C, D, E. Answer:

Unobserved - 05 Source: Annotator

Input: {question}, {choiceA}, {choiceB},{choiceC}, {choiceD}, {choiceE}
Template:
{question}
Pick the most correct description from:
A.{choiceA}
B. {choiceB}
C. {choiceC}
D. {choiceD}
E. {choiceE}
My answer is:

Unobserved - 06 Source: Annotator

Input: {question}, {choiceA}, {choiceB},{choiceC}, {choiceD}, {choiceE}
Template:
{question}

A: {choiceA} B: {choiceB} C: {choiceC} D: {choiceD}
One correct common thing among all the choices above is

Unobserved - 07 Source: Annotator

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

Template:

Answer the question below: {question}
A. {choiceA}
B. {choiceB}
C. {choiceC}
D. {choiceD}
E. {choiceE}
Give your answer with letter, then explain your choice in the next line

Unobserved - 08 Source: Annotator

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

Template:

What is the common phenomenon in these objects, {question}
choose the best answer from the following
A. {choiceA}
B. {choiceB}
C. {choiceC}
D. {choiceD}
E. {choiceE}

Unobserved - 09 Source: Annotator

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

Template:

{question}
A. {choiceA} B. {choiceB} C. {choiceC} D. {choiceD} E. {choiceE}
Pick one of the choices A, B, C, D, E.
The answer is

Unobserved - 10 Source: Annotator

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

Template:

You will be given several objects or activities. From a series of choices (A, B, C, D, E), identify an aspect that they all share in common. If there are multiple aspects, identify the one best fitting.

What do these objects have in common: {question}
A. {choiceA}
B. {choiceB}
C. {choiceC}
D. {choiceD}
E. {choiceE}

Answer:

BBL - Logic Grid Puzzle

Unobserved - 01 Source: Annotator

Input: {question}, {context}, {clues}, {choiceA}, {choiceB},{choiceC}, {choiceD}, {choiceE}

Template:

Here's a puzzle for you {context}

Here are some clus:
{clues}

Now, answer the following question:
{question}

What is the correct answer?
A. {choiceA}
B. {choiceB}
C. {choiceC}
D. {choiceD}
E. {choiceE}

The correct answer is:

Unobserved - 02 Source: Annotator

Input: {question}, {context}, {clues}, {choiceA}, {choiceB},{choiceC}, {choiceD}, {choiceE}
Template:
You are given a logic grid puzzle to test your sense of space and positions. You are given a context and some clues to pick the correct answer from the options to answer a question.Context: {context}
{clues}
Question: {question}
Options:
(A) {choiceA}
(B) {choiceB}
(C) {choiceC}
(D) {choiceD}
(E) {choiceE}
Answer:

Unobserved - 03 Source: Annotator

Input: {question}, {context}, {clues}, {choiceA}, {choiceB},{choiceC}, {choiceD}, {choiceE}
Template:
Question: {question}. Answer this question based on the following context and clues:

Context: {context}
{clues}

The answer is one of "1", "2", "3", "4", "5". The correct answer is:

Unobserved - 04 Source: Annotator

Input: {question}, {context}, {clues}, {choiceA}, {choiceB},{choiceC}, {choiceD}, {choiceE}
Template:
You are a master at solving logic grid puzzles. Solve this: {context}

{clues}

{question}

Unobserved - 05 Source: Annotator

Input: {question}, {context}, {clues}, {choiceA}, {choiceB},{choiceC}, {choiceD}, {choiceE}
Template:
{context}
{clues}Based on the puzzle provided above, {question}

Options are: A: {choiceA}, B: {choiceB}, C: {choiceC}, D: {choiceD}, E: {choiceE}

Answer:

Unobserved - 06 Source: Annotator

Input: {question}, {context}, {clues}, {choiceA}, {choiceB},{choiceC}, {choiceD}, {choiceE}
Template:
The task is to solve a logic grid puzzle. You will have the context to the problem and some clues to solve the puzzle.
Context: {context}{clues}
The question is: {question}
Options:
A. {choiceA}
B. {choiceB}
C. {choiceC}
D. {choiceD}
E. {choiceE}
Output your answer as one of "A", "B", "C", "D", "E".

Unobserved - 07 Source: Annotator

Input: {question}, {context}, {clues}, {choiceA}, {choiceB},{choiceC}, {choiceD}, {choiceE}
Template:
Here's a complex puzzle. Utitlize your skills to solve the puzzle by logic grid tables.{context}
{clues}
{question}

Unobserved - 08 Source: Annotator

Input: {question}, {context}, {clues}, {choiceA}, {choiceB},{choiceC}, {choiceD}, {choiceE}
Template:
{question}
This question can only be answered if you fully understand the context: {context}
{clues}. Your choices are:
(A) {choiceA}
(B) {choiceB}
(C) {choiceC}
(D) {choiceD}
(E) {choiceE}

Answer:

Unobserved - 09 Source: Annotator

Input: {question}, {context}, {clues}, {choiceA}, {choiceB},{choiceC}, {choiceD}, {choiceE}
Template:
You are tested for your ability to answer logic grid problems correctly. {question}
{clues}
{question}
The answer is always one of '1', '2', '3', '4', or '5'. Output your answer and give explanation

Unobserved - 10 Source: Annotator

Input: {question}, {context}, {clues}, {choiceA}, {choiceB},{choiceC}, {choiceD}, {choiceE}
Template:
Context: {context}

The following clues are always true: {clues}

Now, infer the answer to this question: "{question}" and pick the correct answer from A)

{choiceA} B) {choiceB} C) {choiceC} D) {choiceD} E) {choiceE}

Answer:

BBL - Conceptual Combinations

Unobserved - 01 Source: Annotator

Input: {question}, {context}, {choiceA}, {choiceB},{choiceC},{choiceD}
Template:
{context} Question: {question}
The options are the following:
A. {choiceA}
B. {choiceB}
C. {choiceC}
D. {choiceD}
Use your common sense to output one of the letter "A", "B", "C", or "D" to indicate your answer.

Unobserved - 02 Source: Annotator

Input: {question}, {context}, {choiceA}, {choiceB},{choiceC},{choiceD}
Template:
You are given a concept or a factual context. Answer the multiple choice question based on the context by choosing from the choices provided.
Context: {context}
Question: {question}
Choices:
A. {choiceA}
B. {choiceB}
C. {choiceC}
D. {choiceD}
Answer:

Unobserved - 03 Source: Annotator

Input: {question}, {context}, {choiceA}, {choiceB},{choiceC},{choiceD}
Template:
You are a linguistic expert that knows most of the concepts and combinations of words. Now, answer the following question: {context} Question: {question} (A) {choiceA} (B) {choiceB} (C) {choiceC} (D) {choiceD}
Your answer is:

Unobserved - 04 Source: Annotator

Input: {question}, {context}, {choiceA}, {choiceB},{choiceC},{choiceD}
Template:
Answer the question about concepts combination. Specifically, you need to take contradictions, emergent properties, fanciful fictional combinations, homonyms, invented words, and surprising uncommon combinations into consideration. {context} Question: {question} (A) {choiceA} (B) {choiceB} (C) {choiceC} (D) {choiceD}
Your answer is:

Unobserved - 05 Source: Annotator

Input: {question}, {context}, {choiceA}, {choiceB},{choiceC},{choiceD}
Template:
Question: {question}
The options are:
A. {choiceA}
B. {choiceB}

C. {choiceC}
D. {choiceD}
Here is a context to help you answer the question: {context}. Choose the best answer from "A", "B", "C", "D".

Unobserved - 06 Source: Annotator

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

Template:

The following is a multiple-choice question answering problem about conceptual meaning of words. You should choose the answer that best answer the question based on the context. {context} Question: {question}
The options are:
A. {choiceA}
B. {choiceB}
C. {choiceC}
D. {choiceD}
Answer:

Unobserved - 07 Source: Annotator

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

Template:

Context: {context}. Understand the context, and answer the following question: {question}Options:
(A) {choiceA}
(B) {choiceB}
(C) {choiceC}
(D) {choiceD}
Answer:

Unobserved - 08 Source: Annotator

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

Template:

Linguistic Professor: {context} {question}
Student: can you provide the options?
Linguistic Professor: The choices are A) {choiceA} B) {choiceB} C) {choiceC} D) {choiceD}
Student: I got it. The answer is

Unobserved - 09 Source: Annotator

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

Template:

The task is to answer the linguistic question about concepts combination. Context: {context}

Question: {question}

Options:
A. {choiceA}
B. {choiceB}
C. {choiceC}
D. {choiceD}

Answer:

Unobserved - 10 Source: Annotator

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

Template:

{question}

Options: A. {choiceA}, B. {choiceB}, C. {choiceC}, or D. {choiceD}. What is the correct answer to this conceptual combination question? Based on the context "{context}", I think the most accurate answer is

BBL - Play Dialog

Unobserved - 01 Source: Annotator

Input: {play}, {line1}, {line2}

Template:

The following transcripts of dialogues have been taken from Shakespeare's plays, but the transcripts do not say who said what. Based on these contents and styles, your task is to identify whether the sentences in question were spoken by the same or different people.
{play}
From the above dialogue, are the lines {line1} and {line2} spoken by the same or different characters?

Answer:

Unobserved - 02 Source: Annotator

Input: {play}, {line1}, {line2}

Template:

Below are transcripts of dialogues from Shakespeare plays.
{play}
Please identify whether the two scripts were spoken by the same people. Answer yes or no.
Script1: {line1}
Script2: {line2}
Answer:

Unobserved - 03 Source: Annotator

Input: {play}, {line1}, {line2}

Template:

You have read all the plays by Shakespeare. You surely recognized this dialogue:{play}
Now, are these two lines spoken by the same character or different characters? Answer from "same" or "different".
Line1: {line1}
Line2: {line2}
Your answer:

Unobserved - 04 Source: Annotator

Input: {play}, {line1}, {line2}

Template:

The following paragraph is a dialogue from one of Shakespeare's plays, but without the information of the corresponding speaking character. You need to decide whether the two lines I give you are lines of the same character.
{play}
From the above dialogue, are the lines {line1} and {line2} spoken by the same or different characters?
Answer:

Unobserved - 05 Source: Annotator

Input: {play}, {line1}, {line2}

Template:

Now you are a dramatist. The following transcripts of dialogues are taken from Shakespeare plays, but the transcripts do not mark who said what. Your task is to identify whether the sentences in question were spoken by the same or different people. Here is the play:
{play}
Question: In the preceding dialogue, were the lines {line1} and {line2} spoken by the same person

or different people? Please just give a short answer: same or different.

Your Answer:

Unobserved - 06 Source: Annotator

Input: {play}, {line1}, {line2}

Template:

The following transcripts of dialogues have been taken from Shakespeare plays, but the transcripts do not say who said what. We have two sentences selected from the transcripts, please make a judgement whether the sentences are spoken by the same people.
{play}
From the above dialogue, are the lines {line1} and {line2} spoken by the same or different characters?

Unobserved - 07 Source: Annotator

Input: {play}, {line1}, {line2}

Template:

Dialogue: {play}
From the above dialogue, are the lines {line1} and {line2} spoken by the same or different characters? Answer "same" or "different".
Answer:

Unobserved - 08 Source: Annotator

Input: {play}, {line1}, {line2}

Template:

In the context of the Shakespeare play, {play}, assess the given dialogue transcripts. Determine whether the sentences {line1} and {line2} were spoken by a single person or by different people.
Answer:

Unobserved - 09 Source: Annotator

Input: {play}, {line1}, {line2}

Template:

Play: {play}
In this play written by Shakespeare, classify whether
Character A: {line1}
Character B: {line2}are spoken by the same character or different ones? Answer 'Yes' or 'No' only.

Unobserved - 10 Source: Annotator

Input: {play}, {line1}, {line2}

Template:

Context: {play}
Question: read this dialogue selected from a play written by Shakespeare, are the lines {line1} and {line2} from the same character?
Options:
A) Yes
B) No. Answer:

BBL - Strategy QA

Unobserved - 01 Source: Annotator

Input: {question}

Template:

You are given a question which requires reasoning steps that are implicit in the question. Please choose the best answer from "yes" or "no" and provide an explanation.

Question:{question}
Answer and Explanation:

Unobserved - 02 Source: Annotator

Input: {question}

Template:

Reason about the answer to the question. {question}

Unobserved - 03 Source: Annotator

Input: {question}

Template:

You are taking an exam where each question requires implicit reasoning steps to answer. The answer will always be either "yes" or "no". Please carefully consider the following question, its implications, and any related information you may need to provide the correct answer.
Question: {question}
Answer:

Unobserved - 04 Source: Annotator

Input: {question}

Template:

Answer questions that assume implicit reasoning steps in the question prompt: {question}

Unobserved - 05 Source: Annotator

Input: {question}

Template:

Input:
- question: {question}
Output:
- answer:

Unobserved - 06 Source: Annotator

Input: {question}

Template:

Use logic and reasoning to answer the following questions with either "yes" or "no".
Question: {question}
Answer:

Unobserved - 07 Source: Annotator

Input: {question}

Template:

Please answer the following question, you should think step by step, but please use "yes" or "no" to answer.
Question: {question}
Answer:

Unobserved - 08 Source: Annotator

Input: {question}

Template:

Answer questions in which the required reasoning steps are implicit in the question. Please first answer "Yes" or "No" and then output your explanation.
{question} Answer:

Unobserved - 09 Source: Annotator

Input: {question}

Template:

Please give your answer to the following question, which should be answered yes or no. This question may require you to do implicit multi-hop reasoning.
Question: {question}
Answer:

Unobserved - 10 Source: Annotator

Input: {question}

Template:

This question needs to be solved via decomposing it into multiple sub-questions and make comparison among the results of the sub-questions.
The question is {question}
After decomposing the question, we find the answer is

BBL - Strange Stories

Unobserved - 01 Source: Annotator

Input: {question}, {context}

Template:

You are given a psychology question that asks you to provide a socially intelligent response after reading a short story. Please answer "yes" or "no" to the given question.

{context}

Quesiton: {question}
Answer:

Unobserved - 02 Source: Annotator

Input: {question}, {context}

Template:

Given a story, answer whether the question is true or false.
{context}
Q: {question}
A:

Unobserved - 03 Source: Annotator

Input: {question}, {context}

Template:

You are taking a test for reading comprehension. You will be presented with a story and asked a question related to the story. The answer to the question is either "yes" or "no". Please carefully consider the story below before selecting your answer.
Story: {context}
Question: {question}
Answer:

Unobserved - 04 Source: Annotator

Input: {question}, {context}

Template:

A psychology test with naturalistic short stories that measures social intelligence. Boolean options.{context}
Q: {question}Answer:

Unobserved - 05 Source: Annotator

Input: {question}, {context}

Template:

Story: {context}
Q: {question}
Output:

Unobserved - 06 Source: Annotator

Input: {question}, {context}

Template:

Given the following text, answer the question with either "yes" or "no":

Text: {context}
Question: {question}
Answer:

Unobserved - 07 Source: Annotator

Input: {question}, {context}

Template:

Please read the following text and answer the question according to the content of the text, your answer should be "yes" or "no".
Text: {context}
Question: {question}
Answer:

Unobserved - 08 Source: Annotator

Input: {question}, {context}

Template:

Image you are taking a psychology test. Please read the given story and answer the question. Please answer "yes" or "No".
Story: {context}
Q: {question}
A:

Unobserved - 09 Source: Annotator

Input: {question}, {context}

Template:

Please give your answer to the following question, which should be answered yes or no. You should judge the correctness of the question according to the story.

Story: {context}
Question: {question}
Answer:

Unobserved - 10 Source: Annotator

Input: {question}, {context}

Template:

The following story is associated with a question, the answer of which is "yes" or "no".

Story: {context}
Question: {question}

According to the story, the answer is

BBL - Winowhy

Unobserved - 01 Source: Annotator

Input: {question}

Template:

In the sentence: {question}. Is the pronoun reasoning correct? Please answer with either "correct" or "incorrect". Do not include any other words.

Unobserved - 02 Source: Annotator

Input: {question}

Template:

Verify if the reasoning about which words certain pronouns refer to in the given words is right, choose one answer from "correct" and "incorrect":
Reasoning:{question}
Answer:

Unobserved - 03 Source: Annotator

Input: {question}

Template:

Given the context: {question}, determine if the co-reference resolution and the explanationis correct by output either "correct" or "incorrect". Answer:

Unobserved - 04 Source: Annotator

Input: {question}

Template:

Context: {question}
Question: Is the pronoun referring to the correct object? Answer with"Yes" or "No".

Unobserved - 05 Source: Annotator

Input: {question}

Template:

Judge the correctness of the understanding of pronoun:
{question}
Give your answer as"correct" or "incorrect". Your answer:

Unobserved - 06 Source: Annotator

Input: {question}

Template:

You are tested on your understanding of pronoun. Here is a sentence followed by the explanation:
{question}
Output "correct" if you think the explanation is correct; output "incorrect"If the explanation is wrong.

Unobserved - 07 Source: Annotator

Input: {question}

Template:

Read the following reasoning about who a particular pronoun refers to: {question}
Is the reasoning correct?

Unobserved - 08 Source: Annotator

Input: {question}

Template:

Read the following reasoning, and answer if its correct or incorrect. {question}

Unobserved - 09 Source: Annotator

Input: {question}
Template:
{question} The reasoning stated above is "correct" or "incorrect"? It is

Unobserved - 10 Source: Annotator

Input: {context}, {explanation}
Template:
You will be given a sentence followed by an explanation of the use of pronouns in that sentence. Please answer if the explanation is correct or incorrect.

Sentence: {context}
Explanation: {explanation}
Answer:

BBL - Language ID

Unobserved - 01 Source: Annotator

Input: {sentence}, {choiceA}, {choiceB}, {choiceC}, {choiceD}, {choiceE}, {choiceF}, {choiceG},{choiceH}, {choiceI}, {choiceJ}, {choiceK}
Template:
Identify the correct language of the given sentence. Please choose the best answer from A, B, C, D, E, F, G, H, I, J, and K.

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

Unobserved - 02 Source: Annotator

Input: {sentence}, {choiceA}, {choiceB}, {choiceC}, {choiceD}, {choiceE}, {choiceF}, {choiceG},{choiceH}, {choiceI}, {choiceJ}, {choiceK}
Template:
{sentence}
What language is the language stated above? A: {choiceA} B: {choiceB} C: {choiceC} D: {choiceD} E: {choiceE} F: {choiceF} G: {choiceG} H: {choiceH} I: {choiceI} J: {choiceJ} K: {choiceK}

Unobserved - 03 Source: Annotator

Input: {sentence}, {choiceA}, {choiceB}, {choiceC}, {choiceD}, {choiceE}, {choiceF}, {choiceG},{choiceH}, {choiceI}, {choiceJ}, {choiceK}
Template:
You are taking a test that requires you to identify the language a given sentence is written in. To help narrow down your choices, we've made this a multiple choice question. After carefully examining

the sentence and each answer below, please select the correct language of the sentence from one of "A", "B", "C", "D", "E", "F", "G", "H", "I", "J", or "K"
Sentence: {sentence}
- A: {choiceA}
- B: {choiceB}
- C: {choiceC}
- D: {choiceD}
- E: {choiceE}
- F: {choiceF}
- G: {choiceG}
- H: {choiceH}
- I: {choiceI}
- J: {choiceJ}
- K: {choiceK}
Answer:

Unobserved - 04 Source: Annotator

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

Template:

Please select the language that correctly corresponds to the provided sentence from the following options:
Sentence: {sentence}
Options:
A: {choiceA}
B: {choiceB}
C: {choiceC}
D: {choiceD}
E: {choiceE}
F: {choiceF}
G: {choiceG}
H: {choiceH}
I: {choiceI}
J: {choiceJ}
K: {choiceK}
Your answer:

Unobserved - 05 Source: Annotator

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

Template:

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

Unobserved - 06 Source: Annotator

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

Template:

Given the following text, identify the correct language by selecting one of the options in the list (A, B, C, D, E, F, G, H, I, J, K):

Text: {sentence}

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

Answer:

Unobserved - 07 Source: Annotator

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

Template:

Please read the following sentence, then choose from the options which language you think it most likely came from. Your answer should be "A", "B", "C", "D", "E", "F", "G", "H", "I", "J", or "K"
Sentence: {sentence}
Options:
A: {choiceA}
B: {choiceB}
C: {choiceC}
D: {choiceD}
E: {choiceE}
F: {choiceF}
G: {choiceG}
H: {choiceH}
I: {choiceI}
J: {choiceJ}
K: {choiceK}
Answer:

Unobserved - 08 Source: Annotator

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

Template:

Please give the language used in the following sentence. Each sentence will give five options, please output the corresponding option (i.e. A, B, C, D, E, F, G, H, I, J, or K) to represent the corresponding answer.

Sentence: {sentence}
Options::
A: {choiceA}
B: {choiceB}

C: {choiceC}
D: {choiceD}
E: {choiceE}
F: {choiceF}
G: {choiceG}
H: {choiceH}
I: {choiceI}
J: {choiceJ}
K: {choiceK}
Answer:

Unobserved - 09 Source: Annotator

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

Template:

Given the sentence: {sentence}, select the correct language among the choices 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}
Language:

Unobserved - 10 Source: Annotator

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

Template:

{sentence}

This is a sentence written in one of {choiceA}, {choiceB}, {choiceC}, {choiceD}, {choiceE}, {choiceF}, {choiceG}, {choiceH}, {choiceI}, {choiceJ}, {choiceK}. According to the words and the linguistic structure, I can tell that the language is:

BBL - Vitamin C

Unobserved - 01 Source: Annotator

Input: {context}, {claim}

Template:

You are now a very experienced judge. Based only on the information contained in a brief quote from Wikipedia, answer whether the related claim is True, False or Neither. Use Neither when the Wikipedia quote does not provide the necessary information to resolve the question.

{context}
Claim: {claim}
Is this True, False, or Neither?

Unobserved - 02 Source: Annotator

Input: {context}, {claim}

Template:

Now you are a Vitaminc Fact Verifier. Based only on the information contained in a brief quote from Wikipedia, answer whether the related claim is True, False or Neither. Use Neither when the Wikipedia quote does not provide the necessary information to resolve the question.

{context}
Claim: {claim}
Question: Is this True, False, or Neither?
Your answer:

Unobserved - 03 Source: Annotator

Input: {context}, {claim}

Template:

{context}
Read the above paragraph, and answer the following claim {claim}. Answer True, Flase, or Neither. Neither means the Wikipedia quote does not provide the necessary information to resolve the question.
Answer:

Unobserved - 04 Source: Annotator

Input: {context}, {claim}

Template:

Given a claim and its related information context from Wikipedia, determine whether the claim is True, False or Neither. Neither means the given information is not enough to decide if the claim is True or False, which is roughly equivalent to uncertain.

Context:{context}
Claim: {claim}
True, False or Neither?

Unobserved - 05 Source: Annotator

Input: {context}, {claim}

Template:

{context} Claim: {claim}
Based on the context, is the claim true? false? Or Neither? Give your answer as one of "True", "False" or "Neither"

Unobserved - 06 Source: Annotator

Input: {context}, {claim}

Template:

Based only on the information contained in the given context, please make a judgement whether the related claim is True, False or Neither.

{context}
Claim: {claim}
True, False, or Neither?

Unobserved - 07 Source: Annotator

Input: {context}, {claim}

Template:

Wikipedia: {context}
Someone: based on the given context, is the {claim} True, False, or Neither?

Unobserved - 08 Source: Annotator

Input: {context}, {claim}

Template:
Evaluate the related claim as True, False, or Neither based solely on the information given in the short Wikipedia excerpt. Select Neither when the excerpt doesn't provide sufficient information to address the question.
{context}
Claim: {claim}
Answer(True, False, or Neither):

Unobserved - 09 Source: Annotator

Input: {context}, {claim}
Template:
Input: {claim}
Verify the factually of the claim based on the following context
{context}

- "True" if the claim is factually correct
- "False" if the claim is factually incorrect
- "Neither" if the factuality cannot be determined. Output you answerwith one of "True", "False", or "Neither". Answer:

Unobserved - 10 Source: Annotator

Input: {context}, {claim}
Template:
Context: {context}
Now classify this claim into one of 'True', 'False', or 'Neither'.
{claim}

E.3 Granular Experiment Instructions

In this section, we provide the instructions we used for all 6 settings by dataset.

BBH - Intent Recognition

Closest - 1 Source: NIV2 Task 163 OpenPI Classification - Template 2

Input: {passage}
Template:
You will be given a definition of a task first, then some input of the task.
Given a passage as input, answer with the category to which the passage belongs. There are categories - {options}. The answer should be one of the categories based on words from the passage which closely belong to the category.

{passage}
Output:

Closest - 2 Source: NIV2 Task 163 OpenPI Classification - Template 4

Input: {passage}
Template:
Instructions: Given a passage as input, answer with the category to which the passage belongs. There are categories - {options}. The answer should be one of the categories based on words from the passage which closely belong to the category.
Input: {passage}
Output:

Closest - 3 Source: NIV2 Task 163 OpenPI Classification - Template 6

Input: {passage}

Template:
Given the task definition and input, reply with output. Given a passage as input, answer with the category to which the passage belongs. There are categories - {options}. The answer should be one of the categories based on words from the passage which closely belong to the category.

{passage}

Closest - 4 Source: NIV2 Task 163 OpenPI Classification - Template 8

Input: {passage}
Template:
Q: Given a passage as input, answer with the category to which the passage belongs. There are categories - {options}. The answer should be one of the categories based on words from the passage which closely belong to the category.
{passage}
A:

Closest - 5 Source: NIV2 Task 163 OpenPI Classification - Template 10

Input: {passage}
Template:
Detailed Instructions: Given a passage as input, answer with the category to which the passage belongs. There are categories - {options}. The answer should be one of the categories based on words from the passage which closely belong to the category.
Q: {passage}
A:

Incorrect - 1 Source: NIV2 Task 562 Language Identification - Template 10

Input: {text}, {options}
Template:
Detailed Instructions: In this task, an input sentence is given which can be in the {options} languages. There are a total of {options length} languages. Your task is to identify the language of the input sentence. The input sentence can only be in any of the {options length} languages provided.
Q: {text}
A:

Incorrect - 2 Source: NIV2 Task 1588 Tecla Classification - Template 10

Input: {text}, {options}
Template:
Detailed Instructions: In this task, you are given a text in Catalan. Your task is to classify it into {options length} different given themes. Names of all the classes are {options}
Q: {text}
A:

Incorrect - 3 Source: NIV2 Task 564 DiscoFuse Classification - Template 10

Input: {text}, {options}
Template:
Detailed Instructions: In this task, you are given two sentences in the English language and your task is to classify them into one of their discourse types. A discourse type is an indicator to classify the given two sentences on the basis of a co-text as well as a relevant context. There are {options length} discourse types in total which are {options}
Q: {text}
A:

Incorrect - 4 Source: NIV2 Task 1193 Course Classification - Template 10

Input: {text}, {options}
Template:
Detailed Instructions: In this task, you are given the name of an Indian food dish. You need to classify the dish as a {options}
Q: {text}
A:

Incorrect - 5 Source: Flan Sentiment140 - Template 2

Input: {text}, {options}
Template:
{text}

How would the sentiment of this tweet be described?
{options}

Collected - 1 Source: Annotator

Input: {text}, {options}
Template:
You are given a set of intentions to predict: {options}. Pick the most suitable one to describe the following utterance: {text}. Intention:

Collected - 2 Source: Annotator

Input: {text}, {options}
Template:
You are a dialogue assistance at recognizing and classifying user's intention.
Always respond with one of the options: [{options}] to indicate the intention.
Utterance: {text}
Intention:

Collected - 3 Source: Annotator

Input: {text}, {options}
Template:
The tasks is to classify the intention of the utterance: '{text}' into one of the followings: {options}.
Your answer is:

Collected - 4 Source: Annotator

Input: {text}, {options}
Template:
Given the label space: {options}, classify the intention of the given utterance.
{text}
Intention:

Collected - 5 Source: Annotator

Input: {text}, {options}
Template:
Output the intention of the utterance from the list: {options}. Output the exact word or phrase. {text}

Task Designer See BIG-BENCH eval file.

Negation - 1 Source: NIV2 Task 163 OpenPI Classification - Template 2

Input: {passage}
Template:

You will be given a definition of a task first, then some input of the task.
Given a passage as input, answer with the category to which the passage doesn't belong. There are categories - {options}. The answer should be one of the categories based on words from the passage which doesn't belong to the category.

{passage}
Output:

Negation - 2 Source: NIV2 Task 163 OpenPI Classification - Template 4

Input: {passage}

Template:

Instructions: Given a passage as input, answer with the category to which the passage doesn't belong. There are categories - {options}. The answer should be one of the categories based on words from the passage which doesn't belong to the category.
Input: {passage}
Output:

Negation - 3 Source: NIV2 Task 163 OpenPI Classification - Template 6

Input: {passage}

Template:

Given the task definition and input, reply with output. Given a passage as input, answer with the category to which the passage doesn't belong. There are categories - {options}. The answer should be one of the categories based on words from the passage which doesn't belong to the category.

{passage}

Negation - 4 Source: NIV2 Task 163 OpenPI Classification - Template 8

Input: {passage}

Template:

Q: Given a passage as input, answer with the category to which the passage doesn't belongs. There are categories - {options}. The answer should be one of the categories based on words from the passage which doesn't belong to the category.
{passage}
A:

Negation - 5 Source: NIV2 Task 163 OpenPI Classification - Template 10

Input: {passage}

Template:

Detailed Instructions: Given a passage as input, answer with the category to which the passage doesn't belong. There are categories - {options}. The answer should be one of the categories based on words from the passage which doesn't belong to the category.
Q: {passage}
A:

Nonsensical - 1 Source: Annotator

Input: {text}

Template:

The council of street racoons demands you respond to their inquisition. {text}

Nonsensical - 2 Source: Annotator

Input: {text}

Template:

Surveillance birds query your knowledge of seed. {text}

Nonsensical - 3 Source: Annotator

Input: {text}
Template:
Darth Vader requires you to answer to the dark side {text}

Nonsensical - 4 Source: Annotator

Input: {text}
Template:
Respond to the requirement of the Mars working dolphin union. {text}

Nonsensical - 5 Source: Annotator

Input: {text}
Template:
You are undergoing the inquiry of court of the local squirrels. {text}

BBH - Empirical Judgment

Closest - 1 Source: NIV2 Task 163 OpenPI Classification - Template 2

Input: {passage}
Template:
You will be given a definition of a task first, then some input of the task.
Given a passage as input, answer with the category to which the passage belongs. There are categories - {options}. The answer should be one of the categories based on words from the passage which closely belong to the category.

{passage}
Output:

Closest - 2 Source: NIV2 Task 163 OpenPI Classification - Template 4

Input: {passage}
Template:
Instructions: Given a passage as input, answer with the category to which the passage belongs. There are categories - {options}. The answer should be one of the categories based on words from the passage which closely belong to the category.
Input: {passage}
Output:

Closest - 3 Source: NIV2 Task 163 OpenPI Classification - Template 6

Input: {passage}
Template:
Given the task definition and input, reply with output. Given a passage as input, answer with the category to which the passage belongs. There are categories - {options}. The answer should be one of the categories based on words from the passage which closely belong to the category.

{passage}

Closest - 4 Source: NIV2 Task 163 OpenPI Classification - Template 8

Input: {passage}
Template:
Q: Given a passage as input, answer with the category to which the passage belongs. There are categories - {options}. The answer should be one of the categories based on words from the passage which closely belong to the category.

{passage}
A:

Closest - 5 Source: NIV2 Task 163 OpenPI Classification - Template 10

Input: {passage}

Template:

Detailed Instructions: Given a passage as input, answer with the category to which the passage belongs. There are categories - {options}. The answer should be one of the categories based on words from the passage which closely belong to the category.
Q: {passage}
A:

Incorrect - 1 Source: NIV2 Task 143 Odd Man Out Classification - Template 10

Input: {input}, {categories}

Template:

Detailed Instructions: Given a set of four words, generate the category that the words belong to. Words are separated by commas. The possible categories are {categories}
Q: {input}
A:

Incorrect - 2 Source: NIV2 Task 137 Newscomm Classification - Template 10

Input: {input}, {options}

Template:

Detailed Instructions: Classify the given news commentary into the language in which it is written in. There are {options length} languages to classify the sentences into {options}
Q: {input}
A:

Incorrect - 3 Source: Flan2021 - Sentiment140 - Template 1

Input: {input}, {options}

Template:

{text}
Select your answer from the options. What is the sentiment of this tweet?
Options: {options}...I think the answer is

Incorrect - 4 Source: Flan2021 - Sentiment140 - Template 6

Input: {input}, {options}

Template:

Select your answer from the options. How would one describe the sentiment of this tweet?
{text}
{options}

Incorrect - 5 Source: NIV2 Task 1422 MathQA Physics - Template 10

Input: {input}, {options}

Template:

Detailed Instructions: In this task, you need to answer the given multiple-choice question on the physics. Classify your answers into {letter length}
Q: Problem: {input}
{options}
A:

Collected - 1 Source: Annotator

Input: {events}

Template:
Two events are described in the following sentence: {events}
Classify the relation between the events into one of 'causal', 'correlative', or 'neutral'.

Collected - 2 Source: Annotator

Input: {events}
Template:
Causal relation: two events have causal relation if one causes the other to happen.
Correlative relation: two events have correlative relation if there is no explicity causal relation but they are correlated.
Neutral relation: two events have no obvious correlation.

{events} Do the events described in the sentence have causal, correlative, or neutral relation?

Collected - 3 Source: Annotator

Input: {events}
Template:
Sentence: {events} Make a judgment about the relation of the events in the sentence. The possible relations are: "causal", "correlative", "neutral"

Collected - 4 Source: Annotator

Input: {events}
Template:
What is the relation between the events: {events}
Classify it into "causal", "correlative", "neutral"

Collected - 5 Source: Annotator

Input: {events}
Template:
You are given a sentence that describe two or more events. Now, classify the relation into one of "causal", "correlative", "neutral".
Sentence: "{events}"
Answer:

Task Designer See BIG-BENCH eval file.

Negation - 1 Source: NIV2 Task 163 OpenPI Classification - Template 2

Input: {passage}
Template:
You will be given a definition of a task first, then some input of the task.
Given a passage as input, answer with the category to which the passage doesn't belong. There are categories - {options}. The answer should be one of the categories based on words from the passage which doesn't belong to the category.

{passage}
Output:

Negation - 2 Source: NIV2 Task 163 OpenPI Classification - Template 4

Input: {passage}
Template:
Instructions: Given a passage as input, answer with the category to which the passage doesn't belong. There are categories - {options}. The answer should be one of the categories based on words from the passage which doesn't belong to the category.

Input: {passage}
Output:

Negation - 3 Source: NIV2 Task 163 OpenPI Classification - Template 6

Input: {passage}
Template:
Given the task definition and input, reply with output. Given a passage as input, answer with the category to which the passage doesn't belong. There are categories - {options}. The answer should be one of the categories based on words from the passage which doesn't belong to the category.

{passage}

Negation - 4 Source: NIV2 Task 163 OpenPI Classification - Template 8

Input: {passage}
Template:
Q: Given a passage as input, answer with the category to which the passage doesn't belongs. There are categories - {options}. The answer should be one of the categories based on words from the passage which doesn't belong to the category.
{passage}
A:

Negation - 5 Source: NIV2 Task 163 OpenPI Classification - Template 10

Input: {passage}
Template:
Detailed Instructions: Given a passage as input, answer with the category to which the passage doesn't belong. There are categories - {options}. The answer should be one of the categories based on words from the passage which doesn't belong to the category.
Q: {passage}
A:

Nonsensical - 1 Source: Annotator

Input: {text}
Template:
The council of street raccoons demands you respond to their inquisition. {text}

Nonsensical - 2 Source: Annotator

Input: {text}
Template:
Surveillance birds query your knowledge of seed. {text}

Nonsensical - 3 Source: Annotator

Input: {text}
Template:
Darth Vader requires you to answer to the dark side {text}

Nonsensical - 4 Source: Annotator

Input: {text}
Template:
Respond to the requirement of the Mars working dolphin union. {text}

Nonsensical - 5 Source: Annotator

Input: {text}
Template:
You are undergoing the inquiry of court of the local squirrels. {text}

BBL - Conceptual Combinations

Closest - 1 Source: Flan2021 - CosmosQA - Template 1

Input: {context}, {question}, {options}
Template:
{context}

Question with options to choose from: {question}
OPTIONS:{options}

Closest - 2 Source: Flan2021 - CosmosQA - Template 2

Input: {context}, {question}, {options}
Template:
{context}

OPTIONS: {options}
Q: {question}

Closest - 3 Source: Flan2021 - CosmosQA - Template 3

Input: {context}, {question}, {options}
Template:
{context}

OPTIONS: {options}
Answer the following question: {question}

Closest - 4 Source: Flan2021 - CosmosQA - Template 4

Input: {context}, {question}, {options}
Template:
{context}

Based on the preceding passage, choose your answer for question {question}
OPTIONS: {options}

Closest - 5 Source: Flan2021 - CosmosQA - Template 5

Input: {context}, {question}, {options}
Template:
{context}

Q with options: Give answer the following question using evidence from the above passage: {question}
OPTIONS:{options}

Closest - 6 Source: Flan2021 - CosmosQA - Template 6

Input: {context}, {question}, {options}
Template:
Context: {context}
Question {question}
Possible answers:

{options}
The answer:

Closest - 7 Source: Flan2021 - CosmosQA - Template 7

Input: {context}, {question}, {options}
Template:
Read the following article and answer the question by choosing from the options.

{context}

{question}
OPTIONS: {options}...A:

Closest - 8 Source: Flan2021 - CosmosQA - Template 8

Input: {context}, {question}, {options}
Template:
This question has options. Answer the question about text:

{context}

{question}
OPTIONS:{options}

Incorrect - 1 Source: Flan2021 WSC273 - Template 2

Input: {context}, {options}
Template:
Complete the passage.

{context}
OPTIONS: {options}

Incorrect - 2 Source: Flan2021 WSC273 - Template 9

Input: {context}, {options}
Template:
What is the next event listed in the options is correct?

{context}
OPTIONS: {options}
A:

Incorrect - 3 Source: Flan2021 Winograde - Template 3

Input: {context}, {options}
Template:
Choose your story that continues the following story.

{context}

{options}

Incorrect - 4 Source: Flan2021 Story Cloze - Template 1

Input: {context}, {options}
Template:
{context}

{options}
Which option is the next sentence?

Incorrect - 5 Source: Flan2021 Sentiment140 - Template 1

Input: {text}, {options}

Template:

{text}
Select your answer from the options. What is the sentiment of this tweet?
{options}...I think the answer is

Incorrect - 6 Source: NIV2 Task 1422 MathQA Physics - Template 10

Input: {input}, {options}

Template:

Detailed Instructions: In this task, you need to answer the given multiple-choice question on the physics. Classify your answers into {letter length}
Q: Problem: {input}
{options}
A:

Incorrect - 7 Source: NIV2 Task 1297 QASC - Template 10

Input: {input}, {options}

Template:

Detailed Instructions: In this task, you are given two facts, and a multiple-choice question. Based on the given facts, answer the question with index of the correct option (e.g, "A").
Q: {input} {options}
A:

Collected - 01 Source: Annotator

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

Template:

{context} Question: {question}
The options are the following:
A. {choiceA}
B. {choiceB}
C. {choiceC}
D. {choiceD}
Use your common sense to output one of the letter "A", "B", "C", or "D" to indicate your answer.

Collected - 02 Source: Annotator

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

Template:

You are given a concept or a factual context. Answer the multiple choice question based on the context by choosing from the choices provided.
Context: {context}
Question: {question}
Choices:
A. {choiceA}
B. {choiceB}
C. {choiceC}
D. {choiceD}
Answer:

Collected - 03 Source: Annotator

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

Template:
You are a linguistic expert that knows most of the concepts and combinations of words. Now, answer the following question: {context} Question: {question} (A) {choiceA} (B) {choiceB} (C) {choiceC} (D) {choiceD}
Your answer is:

Collected - 04 Source: Annotator

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

Template:
Answer the question about concepts combination. Specifically, you need to take contradictions, emergent properties, fanciful fictional combinations, homonyms, invented words, and surprising uncommon combinations into consideration. {context} Question: {question} (A) {choiceA} (B) {choiceB} (C) {choiceC} (D) {choiceD}
Your answer is:

Collected - 05 Source: Annotator

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

Template:
Question: {question}
The options are:
A. {choiceA}
B. {choiceB}
C. {choiceC}
D. {choiceD}
Here is a context to help you answer the question: {context}. Choose the best answer from "A", "B", "C", "D".

Collected - 06 Source: Annotator

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

Template:
The following is a multiple-choice question answering problem about conceptual meaning of words. You should choose the answer that best answer the question based on the context. {context} Question: {question}
The options are:
A. {choiceA}
B. {choiceB}
C. {choiceC}
D. {choiceD}
Answer:

Collected - 07 Source: Annotator

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

Template:
Context: {context}. Understand the context, and answer the following question: {question}Options:
(A) {choiceA}
(B) {choiceB}
(C) {choiceC}
(D) {choiceD}
Answer:

Collected - 08 Source: Annotator

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

Template:
Linguistic Professor: {context} {question}

Student: can you provide the options?
Linguistic Professor: The choices are A) {choiceA} B) {choiceB} C) {choiceC} D) {choiceD}
Student: I got it. The answer is

Collected - 09 Source: Annotator

Input: {question}, {context}, {choiceA}, {choiceB},{choiceC},{choiceD}
Template:
The task is to answer the linguistic question about concepts combination. Context: {context}

Question: {question}

Options:
A. {choiceA}
B. {choiceB}
C. {choiceC}
D. {choiceD}

Answer:

Collected - 10 Source: Annotator

Input: {question}, {context}, {choiceA}, {choiceB},{choiceC},{choiceD}
Template:
{question}
Options: A. {choiceA}, B. {choiceB}, C. {choiceC}, or D. {choiceD}. What is the correct answer to this conceptual combination question? Based on the context "{context}", I think the most accurate answer is

Task Designer See BIG-BENCH eval file.

Negated - 1 Source: Flan2021 - CosmosQA - Template 1

Input: {context}, {question}, {options}
Template:
{context}

Question with options to choose from: {question}
OPTIONS:{options} The answer is not:

Negation - 2 Source: Flan2021 - CosmosQA - Template 2

Input: {context}, {question}, {options}
Template:
{context}

OPTIONS: {options}
Q: {question} The answer is not:

Negation - 3 Source: Flan2021 - CosmosQA - Template 3

Input: {context}, {question}, {options}
Template:
{context}

OPTIONS: {options}
Answer the following question: {question}
The answer is not:

Negation - 4 Source: Flan2021 - CosmosQA - Template 4

Input: {context}, {question}, {options}

Template:

{context}

Based on the preceding passage, choose your answer for question {question}
OPTIONS: {options}
The answer is not:

Negation - 5 Source: Flan2021 - CosmosQA - Template 5

Input: {context}, {question}, {options}

Template:

{context}

Q with options: Give answer the following question using evidence from the above passage: {question}
OPTIONS:{options}
The answer is not:

Negation - 6 Source: Flan2021 - CosmosQA - Template 6

Input: {context}, {question}, {options}

Template:

Context: {context}
Question {question}
Possible answers:
{options}
The answer is not::

Negation - 7 Source: Flan2021 - CosmosQA - Template 7

Input: {context}, {question}, {options}

Template:

Read the following article and answer the question by choosing from the options.

{context}

{question}
OPTIONS: {options}...The answer is not::

Negation - 8 Source: Flan2021 - CosmosQA - Template 8

Input: {context}, {question}, {options}

Template:

This question has options. Answer the question about text:

{context}

{question}
OPTIONS:{options}
The answer is not:

Nonsensical - 1 Source: Annotator

Input: {text}

Template:

The council of street raccoons demands you respond to their inquisition. {text}

Nonsensical - 2 Source: Annotator

Input: {text}
Template:
Surveillance birds query your knowledge of seed. {text}

Nonsensical - 3 Source: Annotator

Input: {text}
Template:
Darth Vader requires you to answer to the dark side {text}

Nonsensical - 4 Source: Annotator

Input: {text}
Template:
Respond to the requirement of the Mars working dolphin union. {text}

Nonsensical - 5 Source: Annotator

Input: {text}
Template:
You are undergoing the inquiry of court of the local squirrels. {text}

BBL - Language Identification

Closest - 1 Source: NIV2 Task 137 Newscomm Classification - Template 2

Input: {passage}
Template:
You will be given a definition of a task first, then some input of the task.
Classify the given news commentary into the language in which it is written in. There are {option length} languages to classify the sentences into {options}

{sentence}
Output:

Closest - 2 Source: NIV2 Task 137 Newscomm Classification - Template 4

Input: {passage}
Template:
Instructions: Classify the given news commentary into the language in which it is written in. There are {option length} languages to classify the sentences into {options}
Input: {sentence}
Output:

Closest - 3 Source: NIV2 Task 137 Newscomm Classification - Template 6

Input: {passage}
Template:
Given the task definition and input, reply with output. Classify the given news commentary into the language in which it is written in. There are {option length} languages to classify the sentences into {options}

{sentence}

Closest - 4 Source: NIV2 Task 137 Newscomm Classification - Template 8

Input: {passage}

Template:
Q: Classify the given news commentary into the language in which it is written in. There are {option length} languages to classify the sentences into {options}
{sentence}
A:

Closest - 5 Source: NIV2 Task 137 Newscomm Classification - Template 10

Input: {passage}
Template:
Detailed Instructions: Classify the given news commentary into the language in which it is written in. There are {option length} languages to classify the sentences into {options}
Q: {sentence}
A:

Incorrect - 1 Source: NIV2 Task 143 Odd Man Out Classification - Template 10

Input: {input}, {categories}
Template:
Detailed Instructions: Given a set of four words, generate the category that the words belong to. Words are separated by commas. The possible categories are {categories}
Q: {input}
A:

Incorrect - 2 Source: NIV2 Task 1322 Government Type Classification - Template 10

Input: {input}, {options}
Template:
Detailed Instructions: In this task, you are given a country name and you need to answer with the government type of the country, as of the year 2015. The following are possible government types that are considered valid answers: {options}
Q: {input}
A:

Incorrect - 3 Source: NIV2 Task 1422 MathQA Physics - Template 10

Input: {input}, {options}
Template:
Detailed Instructions: In this task, you need to answer the given multiple-choice question on the physics. Classify your answers into {letter length}
Q: Problem: {input}
{options}
A:

Incorrect - 4 Source: NIV2 Task 154 HateXPlain Classification - Template 10

Input: {input}, {labels}
Template:
Detailed Instructions: The input is a tweet which can be Hate Speech, Offensive or Normal tweet. Hate Speech and Offensive tweets target one community. Given such a tweet, output the community targeted in the tweet. The community will be one of the nine values: {labels}. Output 'None' if the tweet does not target any community. A tweet targets only one community.
Q: {input}
A:

Collected - 01 Source: Annotator

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

Identify the correct language of the given sentence. Please choose the best answer from A, B, C, D, E, F, G, H, I, J, and K.

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

Collected - 02 Source: Annotator

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

Template:

{sentence}
What language is the language stated above? A: {choiceA} B: {choiceB} C: {choiceC} D: {choiceD} E: {choiceE} F: {choiceF} G: {choiceG} H: {choiceH} I: {choiceI} J: {choiceJ} K: {choiceK}

Collected - 03 Source: Annotator

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

Template:

You are taking a test that requires you to identify the language a given sentence is written in. To help narrow down your choices, we've made this a multiple choice question. After carefully examining the sentence and each answer below, please select the correct language of the sentence from one of "A", "B", "C", "D", "E", "F", "G", "H", "I", "J", or "K"
Sentence: {sentence}
- A: {choiceA}
- B: {choiceB}
- C: {choiceC}
- D: {choiceD}
- E: {choiceE}
- F: {choiceF}
- G: {choiceG}
- H: {choiceH}
- I: {choiceI}
- J: {choiceJ}
- K: {choiceK}
Answer:

Collected - 04 Source: Annotator

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

Template:

Please select the language that correctly corresponds to the provided sentence from the following options:
Sentence: {sentence}
Options:
A: {choiceA}

B: {choiceB}
C: {choiceC}
D: {choiceD}
E: {choiceE}
F: {choiceF}
G: {choiceG}
H: {choiceH}
I: {choiceI}
J: {choiceJ}
K: {choiceK}
Your answer:

Collected - 05 Source: Annotator

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

Template:

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

Output
- Answer:

Collected - 06 Source: Annotator

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

Template:

Given the following text, identify the correct language by selecting one of the options in the list (A, B, C, D, E, F, G, H, I, J, K):

Text: {sentence}

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

Answer:

Collected - 07 Source: Annotator

Input: {sentence}, {choiceA}, {choiceB}, {choiceC}, {choiceD}, {choiceE}, {choiceF},

{choiceG},{choiceH}, {choiceI}, {choiceJ}, {choiceK}

Template:

Please read the following sentence, then choose from the options which language you think it most likely came from. Your answer should be "A", "B", "C", "D", "E", "F", "G", "H", "I", "J", or "K"
Sentence: {sentence}
Options:
A: {choiceA}
B: {choiceB}
C: {choiceC}
D: {choiceD}
E: {choiceE}
F: {choiceF}
G: {choiceG}
H: {choiceH}
I: {choiceI}
J: {choiceJ}
K: {choiceK}
Answer:

Collected - 08 Source: Annotator

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

Template:

Please give the language used in the following sentence. Each sentence will give five options, please output the corresponding option (i.e. A, B, C, D, E, F, G, H, I, J, or K) to represent the corresponding answer.

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

Collected - 09 Source: Annotator

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

Template:

Given the sentence: {sentence}, select the correct language among the choices 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}
Language:

Collected - 10 Source: Annotator

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

Template:

{sentence}

This is a sentence written in one of {choiceA}, {choiceB}, {choiceC}, {choiceD}, {choiceE}, {choiceF}, {choiceG}, {choiceH}, {choiceI}, {choiceJ}, {choiceK}. According to the words and the linguistic structure, I can tell that the language is:

Task Designer See BIG-BENCH eval file.

Negation - 1 Source: NIV2 Task 137 Newscomm Classification - Template 2

Input: {passage}

Template:

You will be given a definition of a task first, then some input of the task.
Classify the given news commentary into the language in which it is not written in. There are {option length} languages to classify the sentences into {options}

{sentence}
Output:

Negation - 2 Source: NIV2 Task 137 Newscomm Classification - Template 4

Input: {passage}

Template:

Instructions: Classify the given news commentary into the language in which it is not written in. There are {option length} languages to classify the sentences into {options}
Input: {sentence}
Output:

Negation - 3 Source: NIV2 Task 137 Newscomm Classification - Template 6

Input: {passage}

Template:

Given the task definition and input, reply with output. Classify the given news commentary into the language in which it is not written in. There are {option length} languages to classify the sentences into {options}

{sentence}

Negation - 4 Source: NIV2 Task 137 Newscomm Classification - Template 8

Input: {passage}

Template:

Q: Classify the given news commentary into the language in which it is not written in. There are {option length} languages to classify the sentences into {options}
{sentence}
A:

Negation - 5 Source: NIV2 Task 137 Newscomm Classification - Template 10

Input: {passage}
Template:
Detailed Instructions: Classify the given news commentary into the language in which it is not written in. There are {option length} languages to classify the sentences into {options}
Q: {sentence}
A:

Nonsensical - 1 Source: Annotator
Input: {text}
Template:
The council of street raccoons demands you respond to their inquisition. {text}

Nonsensical - 2 Source: Annotator
Input: {text}
Template:
Surveillance birds query your knowledge of seed. {text}

Nonsensical - 3 Source: Annotator
Input: {text}
Template:
Darth Vader requires you to answer to the dark side {text}

Nonsensical - 4 Source: Annotator
Input: {text}
Template:
Respond to the requirement of the Mars working dolphin union. {text}

Nonsensical - 5 Source: Annotator
Input: {text}
Template:
You are undergoing the inquiry of court of the local squirrels. {text}

BBH - Epistemic Reasoning

Closest - 1 Source: FLAN2021 RTE - Template 1
Input: {premise}, {hypothesis}, {options}
Template:
{premise}

Question with options: Based on the paragraph above can we conclude that "{hypothesis}"?

OPTIONS: {options}

Closest - 2 Source: FLAN2021 RTE - Template 2
Input: {premise}, {hypothesis}, {options}
Template:
{premise}

Based on that paragraph can we conclude that the sentence below is true?
{hypothesis}

OPTIONS: {options}

Closest - 3 Source: FLAN2021 RTE - Template 3

Input: {premise}, {hypothesis}, {options}
Template:
{premise}

Q with options: Can we draw the following conclusion?
{hypothesis}

OPTIONS: {options}

Closest - 4 Source: FLAN2021 RTE - Template 4

Input: {premise}, {hypothesis}, {options}
Template:
{premise}
Does this next sentence follow, given the preceding text?
{hypothesis}

OPTIONS: {options}

Closest - 5 Source: FLAN2021 RTE - Template 5

Input: {premise}, {hypothesis}, {options}
Template:
{premise}
OPTIONS: {options}
Question: Can we infer the following?
{hypothesis}

Closest - 6 Source: FLAN2021 RTE - Template 6

Input: {premise}, {hypothesis}, {options}
Template:
Read the following paragraph and determine if the hypothesis is true. Select from options at the end:

{premise}

Hypothesis: {hypothesis}
OPTIONS: {options}
The answer is

Closest - 7 Source: FLAN2021 RTE - Template 7

Input: {premise}, {hypothesis}, {options}
Template:
Read the text and determine if the sentence is true:

{premise}

Sentence: {hypothesis}
OPTIONS: {options}
A:

Closest - 8 Source: FLAN2021 RTE - Template 8

Input: {premise}, {hypothesis}, {options}
Template:
Question with options: can we draw the following hypothesis from the context?

Context:

{premise}

Hypothesis: {hypothesis}
OPTIONS: {options}
A:

Incorrect - 1 Source: NIV2 Task 143 Odd Man Out Classification - Template 10

Input: {input}, {categories}

Template:

Detailed Instructions: Given a set of four words, generate the category that the words belong to. Words are separated by commas. The possible categories are {categories}
Q: {input}
A:

Incorrect - 2 Source: NIV2 Task 137 Newscomm Classification - Template 10

Input: {input}, {options}

Template:

Detailed Instructions: Classify the given news commentary into the language in which it is written in. There are {options length} languages to classify the sentences into {options}
Q: {input}
A:

Incorrect - 3 Source: Flan2021 - Sentiment140 - Template 1

Input: {input}, {options}

Template:

{text}
Select your answer from the options. What is the sentiment of this tweet?
Options: {options}...I think the answer is

Incorrect - 4 Source: Flan2021 - Sentiment140 - Template 6

Input: {input}, {options}

Template:

Select your answer from the options. How would one describe the sentiment of this tweet?
{text}
{options}

Incorrect - 5 Source: NIV2 Task 1422 MathQA Physics - Template 10

Input: {input}, {options}

Template:

Detailed Instructions: In this task, you need to answer the given multiple-choice question on the physics. Classify your answers into {letter length}
Q: Problem: {input}
{options}
A:

Incorrect - 6 Source: NIV2 Task 562 Language Identification - Template 10

Input: {text}, {options}

Template:

Detailed Instructions: In this task, an input sentence is given which can be in the {options} languages. There are a total of {options length} languages. Your task is to identify the language of the input sentence. The input sentence can only be in any of the {options length} languages provided.

Q: {text}
A:

Incorrect - 7 Source: NIV2 Task 1193 Course Classification - Template 10

Input: {text}, {options}

Template:

Detailed Instructions: In this task, you are given the name of an Indian food dish. You need to classify the dish as a {options}
Q: {input}
A:

Incorrect - 8 Source: NIV2 - Task 56 - Template 10

Input: {paragraph}, {question}, {correct answer}

Template:

Detailed Instructions: In this task, your goal is to judge a correct answer to a given question based on an associated paragraph and decide if it is a good correct answer or not. A good correct answer is one that correctly and completely answers the question. A bad correct answer addresses the question only partially or incorrectly. If you think the given correct answer is good, indicate it by responding "Yes". Otherwise, respond "No". There are only two types of responses possible: "Yes" and "No".
Q: Paragraph- {paragraph} Question: {question} Correct Answer: {correct answer}
A:

Collected - 01 Source: Annotator

Input: {premise}, {hypothesis}

Template:

Classify whether two sentences have entailment relation. Output "yes" if they have entailment relation; output "no" if they do not have entailment relation.
Premise: {premise}
Hypothesis: {hypothesis}
Answer:

Collected - 02 Source: Annotator

Input: {premise}, {hypothesis}

Template:

What is the relation between the given two sentences? Choose one of 'entailment' and 'non-entailment'.

Sentence1: {premise}
Sentence2: {hypothesis}
Relation:

Collected - 03 Source: Annotator

Input: {premise}, {hypothesis}

Template:

Premise: {premise}
Is the truthfulness of the premise entail the following hypothesis?
Hypothesis: {hypothesis}.

Collected - 04 Source: Annotator

Input: {premise}, {hypothesis}

Template:

Given the premise '{premise}', can you infer that '{hypothesis}'? Answer 'Yes' or 'No' as your answer. Answer:

Collected - 05 Source: Annotator

Input: {premise}, {hypothesis}
Template:
I think "{premise}" entails "{hypothesis}".

Am I right?

Task Designer See BIG-BENCH eval file.

Negation - 1 Source: FLAN2021 RTE - Template 1

Input: {premise}, {hypothesis}, {options}
Template:
{premise}

Question with options: Based on the paragraph above can we conclude that "{hypothesis}"?
Answer the opposite of the correct answer.
OPTIONS: {options}

Negation - 2 Source: FLAN2021 RTE - Template 2

Input: {premise}, {hypothesis}, {options}
Template:
{premise}

Based on that paragraph can we conclude that the sentence below is true? Answer the opposite of the correct answer.
{hypothesis}

OPTIONS: {options}

Negation - 3 Source: FLAN2021 RTE - Template 3

Input: {premise}, {hypothesis}, {options}
Template:
{premise}

Q with options: Can we draw the following conclusion? Answer the opposite of the correct answer.
{hypothesis}

OPTIONS: {options}

Negation - 4 Source: FLAN2021 RTE - Template 4

Input: {premise}, {hypothesis}, {options}
Template:
{premise}
Does this next sentence follow, given the preceding text? Answer the opposite of the correct answer.
{hypothesis}

OPTIONS: {options}

Negation - 5 Source: FLAN2021 RTE - Template 5

Input: {premise}, {hypothesis}, {options}
Template:
{premise}

OPTIONS: {options}
Question: Can we infer the following? Answer the opposite of the correct answer.
{hypothesis}

Negation - 6 Source: FLAN2021 RTE - Template 6

Input: {premise}, {hypothesis}, {options}
Template:
Read the following paragraph and determine if the hypothesis is true. Select from options at the end. Answer the opposite of the correct answer :

{premise}

Hypothesis: {hypothesis}
OPTIONS: {options}
The answer is

Negation - 7 Source: FLAN2021 RTE - Template 7

Input: {premise}, {hypothesis}, {options}
Template:
Read the text and determine if the sentence is true. Answer the opposite of the correct answer:

{premise}

Sentence: {hypothesis}
OPTIONS: {options}
A:

Negation - 8 Source: FLAN2021 RTE - Template 8

Input: {premise}, {hypothesis}, {options}
Template:
Question with options: can we draw the following hypothesis from the context? Answer the opposite of the correct answer.

Context:

{premise}

Hypothesis: {hypothesis}
OPTIONS: {options}
A:

Nonsensical - 1 Source: Annotator

Input: {text}
Template:
The council of street raccoons demands you respond to their inquisition. {text}

Nonsensical - 2 Source: Annotator

Input: {text}
Template:
Surveillance birds query your knowledge of seed. {text}

Nonsensical - 3 Source: Annotator

Input: {text}
Template:

Darth Vader requires you to answer to the dark side {text}

Nonsensical - 4 Source: Annotator

Input: {text}
Template:
Respond to the requirement of the Mars working dolphin union. {text}

Nonsensical - 5 Source: Annotator

Input: {text}
Template:
You are undergoing the inquiry of court of the local squirrels. {text}

BBH - Crash Blossom

Closest - 1 Source: Flan2021 - CosmosQA - Template 1

Input: {context}, {question}, {options}
Template:
{context}

Question with options to choose from: {question}
OPTIONS:{options}

Closest - 2 Source: Flan2021 - CosmosQA - Template 2

Input: {context}, {question}, {options}
Template:
{context}

OPTIONS: {options}
Q: {question}

Closest - 3 Source: Flan2021 - CosmosQA - Template 3

Input: {context}, {question}, {options}
Template:
{context}

OPTIONS: {options}
Answer the following question: {question}

Closest - 4 Source: Flan2021 - CosmosQA - Template 4

Input: {context}, {question}, {options}
Template:
{context}

Based on the preceding passage, choose your answer for question {question}
OPTIONS: {options}

Closest - 5 Source: Flan2021 - CosmosQA - Template 5

Input: {context}, {question}, {options}
Template:
{context}

Q with options: Give answer the following question using evidence from the above passage: {question}
OPTIONS:{options}

Closest - 6 Source: Flan2021 - CosmosQA - Template 6

Input: {context}, {question}, {options}

Template:

Context: {context}
Question {question}
Possible answers:
{options}
The answer:

Closest - 7 Source: Flan2021 - CosmosQA - Template 7

Input: {context}, {question}, {options}

Template:

Read the following article and answer the question by choosing from the options.

{context}

{question}
OPTIONS: {options}...A:

Closest - 8 Source: Flan2021 - CosmosQA - Template 8

Input: {context}, {question}, {options}

Template:

This question has options. Answer the question about text:

{context}

{question}
OPTIONS:{options}

Incorrect - 1 Source: NIV2 Task 143 Odd Man Out Classification - Template 10

Input: {input}, {categories}

Template:

Detailed Instructions: Given a set of four words, generate the category that the words belong to. Words are separated by commas. The possible categories are {categories}
Q: {input}
A:

Incorrect - 2 Source: NIV2 Task 137 Newscomm Classification - Template 10

Input: {input}, {options}

Template:

Detailed Instructions: Classify the given news commentary into the language in which it is written in. There are {options length} languages to classify the sentences into {options}
Q: {input}
A:

Incorrect - 3 Source: Flan2021 - Sentiment140 - Template 1

Input: {input}, {options}

Template:

{text}

Select your answer from the options. What is the sentiment of this tweet?
Options: {options}...I think the answer is

Incorrect - 4 Source: Flan2021 - Sentiment140 - Template 6

Input: {input}, {options}

Template:

Select your answer from the options. How would one describe the sentiment of this tweet?
{text}
{options}

Incorrect - 5 Source: NIV2 Task 1422 MathQA Physics - Template 10

Input: {input}, {options}

Template:

Detailed Instructions: In this task, you need to answer the given multiple-choice question on the physics. Classify your answers into {letter length}
Q: Problem: {input}
{options}
A:

Collected - 01 Source: Annotator

Input: {word}, {sentence}, {options}

Template:

Classify the part of speech of the word "{word}" in the following sentence: {sentence}. The options are: {options}
Answer:

Collected - 02 Source: Annotator

Input: {word}, {sentence}, {options}

Template:

Sentence: {sentence}
Identify the part of speech of {word} in the sentence. Choise your answer from {options} and output the best choice.

Collected - 03 Source: Annotator

Input: {word}, {sentence}, {options}

Template:

What is the part of speech of the word '{word}' in '{sentence}'. You may only choose from the following options: {options}. Your answer is:

Collected - 04 Source: Annotator

Input: {word}, {sentence}, {options}

Template:

Given a sentence and a word contained in the sentence, output the part of speech of the word.
Word: {word}
Sentence: {sentence}
Options: {options}

Answer:

Collected - 05 Source: Annotator

Input: {word}, {sentence}, {options}

Template:

Identify the part of speech of the word. Question: which one of {options} is '{word}' in '{sentence}'?
Answer:

Task Designer See BIG-BENCH eval file.

Negation - 1 Source: Flan2021 - CosmosQA - Template 1

Input: {context}, {question}, {options}

Template:

{context}

Question with options to choose from: {question}
OPTIONS:{options} The answer is not:

Negation - 2 Source: Flan2021 - CosmosQA - Template 2

Input: {context}, {question}, {options}

Template:

{context}

OPTIONS: {options}
Q: {question} The answer is not:

Negation - 3 Source: Flan2021 - CosmosQA - Template 3

Input: {context}, {question}, {options}

Template:

{context}

OPTIONS: {options}
Answer the following question: {question}
The answer is not:

Negation - 4 Source: Flan2021 - CosmosQA - Template 4

Input: {context}, {question}, {options}

Template:

{context}

Based on the preceding passage, choose your answer for question {question}
OPTIONS: {options} The answer is not:

Negation - 5 Source: Flan2021 - CosmosQA - Template 5

Input: {context}, {question}, {options}

Template:

{context}

Q with options: Give answer the following question using evidence from the above passage: {question}
OPTIONS:{options} The answer is not:

Negation - 6 Source: Flan2021 - CosmosQA - Template 6

Input: {context}, {question}, {options}

Template:

Context: {context}
Question {question}
Possible answers:

{options}
The answer is not:

Negation - 7 Source: Flan2021 - CosmosQA - Template 7

Input: {context}, {question}, {options}
Template:
Read the following article and answer the question by choosing from the options.

{context}

{question}
OPTIONS: {options}...The answer is not::

Negation - 8 Source: Flan2021 - CosmosQA - Template 8

Input: {context}, {question}, {options}
Template:
This question has options. Answer the question about text:

{context}

{question}
OPTIONS:{options} The answer is not:

Nonsensical - 1 Source: Annotator

Input: {text}
Template:
The council of street raccoons demands you respond to their inquisition. {text}

Nonsensical - 2 Source: Annotator

Input: {text}
Template:
Surveillance birds query your knowledge of seed. {text}

Nonsensical - 3 Source: Annotator

Input: {text}
Template:
Darth Vader requires you to answer to the dark side {text}

Nonsensical - 4 Source: Annotator

Input: {text}
Template:
Respond to the requirement of the Mars working dolphin union. {text}

Nonsensical - 5 Source: Annotator

Input: {text}
Template:
You are undergoing the inquiry of court of the local squirrels. {text}

BBH - Logical Sequence

Closest - 1 Source: NIV2 - Task 73 - Template 2

Input: {question}, {options}

Template:
You will be given a definition of a task first, then some input of the task.
You are given a question and some answer options (associated with "A", "B", "C", "D"). You should choose the correct answer based on commonsense knowledge. Avoid answering questions based on associations, the set of answers are chosen deliberately to capture common sense beyond associations. Do not generate anything else apart from one of the following characters: {options letter} and only give one answer for each question.

{question} {options}
Output:

Closest - 2 Source: NIV2 - Task 73 - Template 4

Input: {question}, {options}
Template:
Instructions: You are given a question and some answer options (associated with "A", "B", "C", "D"). You should choose the correct answer based on commonsense knowledge. Avoid answering questions based on associations, the set of answers are chosen deliberately to capture common sense beyond associations. Do not generate anything else apart from one of the following characters: {options letter} and only give one answer for each question.
Input: {question} {options}
Output:

Closest - 3 Source: NIV2 - Task 73 - Template 6

Input: {question}, {options}
Template:
Given the task definition and input, reply with output. You are given a question and some answer options (associated with "A", "B", "C", "D"). You should choose the correct answer based on commonsense knowledge. Avoid answering questions based on associations, the set of answers are chosen deliberately to capture common sense beyond associations. Do not generate anything else apart from one of the following characters: {options letter} and only give one answer for each question.

{question} {options}

Closest - 4 Source: NIV2 - Task 73 - Template 8

Input: {question}, {options}
Template:
Q: You are given a question and some answer options (associated with "A", "B", "C", "D"). You should choose the correct answer based on commonsense knowledge. Avoid answering questions based on associations, the set of answers are chosen deliberately to capture common sense beyond associations. Do not generate anything else apart from one of the following characters: {options letter} and only give one answer for each question.
{question} {options}
A:

Closest - 5 Source: NIV2 - Task 73 - Template 10

Input: {question}, {options}
Template:
Detailed Instructions: You are given a question and some answer options (associated with "A", "B", "C", "D"). You should choose the correct answer based on commonsense knowledge. Avoid answering questions based on associations, the set of answers are chosen deliberately to capture common sense beyond associations. Do not generate anything else apart from one of the following characters: {options letter} and only give one answer for each question.

Q: {question} {options}
A:

Incorrect - 1 Source: NIV2 Task 1421 MathQA General - Template 10

Input: {input}, {options}

Template:

Detailed Instructions: In this task, you need to answer the given multiple-choice question on the general math. Classify your answers into Classify your answers into {option letter}
Q: Problem: {input}
{options}
A:

Incorrect - 2 Source: NIV2 Task 1422 MathQA Physics - Template 10

Input: {input}, {options}

Template:

Detailed Instructions: In this task, you need to answer the given multiple-choice question on the physics. Classify your answers into {letter length}
Q: Problem: {input}
{options}
A:

Incorrect - 3 Source: Flan2021 - WSC273 - Template 1

Input: {context}, {options}

Template:

Multi-choice problem: {context}
{options}

Incorrect - 4 Source: Flan2021 - TREC - Template 1

Input: {text}, {options}

Template:

What type of thing is the question "{text}" asking about?

{options}
Answer:

Incorrect - 5 Source: Flan2021 - PIQA - Template 1

Input: {input}, {options}

Template:

Here is a goal: {goal}

How would you accomplish this goal?

{options}

Collected - 1 Source: Annotator

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

Template:

Four items are naturally in a sequential or chronological order. Now, choose the correct order of these items from the following options:
A. {listA}
B. {listB}
C. {listC}
D. {listD}

Collected - 2 Source: Annotator

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

Template:

You are given four lists of the same objects in different orders. Which of the following lists is correctly ordered chronologically?

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

Collected - 3 Source: Annotator

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

Template:

Choose the best answer that describes a sequence chronologically. Options: A: {listA}, B: {listB}, C: {listC}, D: {listD}

Answer:

Collected - 4 Source: Annotator

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

Template:

In this task, pick the list of the items that are chronologically orded most correctly. Choose from the following options and output the corresponding letter as one of 'A', 'B', 'C', or 'D'.
A. {listA}
B. {listB}
C. {listC}
D. {listD}

Collected - 5 Source: Annotator

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

Template:

Question: Which of the following lists is correctly ordered chronologically?
Choose the correct order from the lists: A. {listA}, B. {listB}, C. {listC}, D. {listD}. Answer:

Task Designer See BIG-BENCH eval file.

Negation - 1 Source: NIV2 - Task 73 - Template 2

Input: {question}, {options}

Template:

You will be given a definition of a task first, then some input of the task.
You are given a question and some answer options (associated with "A", "B", "C", "D"). You should choose the incorrect answer based on commonsense knowledge. Avoid answering questions based on associations, the set of answers are chosen deliberately to capture common sense beyond associations. Do not generate anything else apart from one of the following characters: {options letter} and only give one answer for each question.

{question} {options}
Output:

Negation - 2 Source: NIV2 - Task 73 - Template 4

Input: {question}, {options}

Template:

Instructions: You are given a question and some answer options (associated with "A", "B", "C", "D"). You should choose the incorrect answer based on commonsense knowledge. Avoid answering questions based on associations, the set of answers are chosen deliberately to capture common sense beyond associations. Do not generate anything else apart from one of the following characters: {options letter} and only give one answer for each question.
Input: {question} {options}
Output:

Negation - 3 Source: NIV2 - Task 73 - Template 6

Input: {question}, {options}

Template:

Given the task definition and input, reply with output. You are given a question and some answer options (associated with "A", "B", "C", "D"). You should choose the incorrect answer based on commonsense knowledge. Avoid answering questions based on associations, the set of answers are chosen deliberately to capture common sense beyond associations. Do not generate anything else apart from one of the following characters: {options letter} and only give one answer for each question.

{question} {options}

Negation - 4 Source: NIV2 - Task 73 - Template 8

Input: {question}, {options}

Template:

Q: You are given a question and some answer options (associated with "A", "B", "C", "D"). You should choose the incorrect answer based on commonsense knowledge. Avoid answering questions based on associations, the set of answers are chosen deliberately to capture common sense beyond associations. Do not generate anything else apart from one of the following characters: {options letter} and only give one answer for each question.
{question} {options}
A:

Negation - 5 Source: NIV2 - Task 73 - Template 10

Input: {question}, {options}

Template:

Detailed Instructions: You are given a question and some answer options (associated with "A", "B", "C", "D"). You should choose the incorrect answer based on commonsense knowledge. Avoid answering questions based on associations, the set of answers are chosen deliberately to capture common sense beyond associations. Do not generate anything else apart from one of the following characters: {options letter} and only give one answer for each question.
Q: {question} {options}
A:

Nonsensical - 1 Source: Annotator

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

Template:

The council of street raccoons demands you respond to their inquisition.
A. {listA}
B. {listB}
C. {listC}
D. {listD}

Nonsensical - 2 Source: Annotator

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

Template:
Surveillance birds query your knowledge of seed.
A. {listA}
B. {listB}
C. {listC}
D. {listD}

Nonsensical - 3 Source: Annotator
Input: {listA}, {listB}, {listC}, {listD}
Template:
Darth Vader requires you to answer to the dark side
A. {listA}
B. {listB}
C. {listC}
D. {listD}

Nonsensical - 4 Source: Annotator
Input: {listA}, {listB}, {listC}, {listD}
Template:
Respond to the requirement of the Mars working dolphin union.
A. {listA}
B. {listB}
C. {listC}
D. {listD}

Nonsensical - 5 Source: Annotator
Input: {listA}, {listB}, {listC}, {listD}
Template:
You are undergoing the inquiry of court of the local squirrels.
A. {listA}
B. {listB}
C. {listC}
D. {listD}

E.4 Paraphrased Instructions

Here we provide the prompt we use to automatically generate paraphrased instructions for 5. We also provide the JSON file of all paraphrased instructions.

Alpaca To generate paraphrases of observed instructions in the Alpaca collection we sampled 1000 out of 52002 Alpaca tasks at i.i.d. random and generated paraphrases of instructions with GPT-4 using the following prompts.

- "Paraphrase this sentence:\n\n{instruction}Paraphrased sentence:\n\n"
- "Paraphrase this instruction into a longer sentence\n\n\n{instruction}New sentence:\n"
- "You are given an instruction:\n\n{instruction}Now, paraphrase it into a new instruction with equivalent meaning:\n\n"

Flan We first reproduced the held-in instruction-tuning set of Flan-T5 with the pipeline[6]. We randomly sampled 986 data samples from the generated data following the proportion of partition

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

B reported in [5]. We generate the paraphrases of the selected data with GPT-4 using the following prompts.

- "Here's an input utterance:\n\n{instruction}\n \n \n Now, your task is to paraphrase the input by only changing the instruction but leaving everything else the same.\n Here's the new utterance:\n\n"
- "You are given an utterance which is a combination of task instruction and the actual input. Your job is to paraphrase the task instruction and leave the input unchanged. Here's the utterance to be paraphrased:\n\n\n{instruction}\n\n\n Now, generate the new utterance:\n\n\n"
- "You are provided with the utterance of a specific task and I need you to paraphrase it. The actual input, question, and examples in the task should not be changed. You should only paraphrase the instructions. Task:\n\n\n {instruction}\n\n\nThe paraphrased utterance:\n\n\n"

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!

Figure 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**.

Figure 10: The first page of the instruction given to the annotator

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!

Figure 11: The second page of the instruction given to the annotator

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

Figure 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	

Figure 13: The second page of the dataset information