

# Dancing in Chains: Reconciling Instruction Following and Faithfulness in Language Models

Zhengxuan Wu<sup>1◊</sup>, Yuhao Zhang<sup>2◊\*</sup>, Peng Qi<sup>3◊\*</sup>, Yumo Xu<sup>4\*</sup>,  
Rujun Han<sup>5</sup>, Yian Zhang<sup>6◊</sup>, Jifan Chen<sup>4</sup>, Bonan Min<sup>4</sup>, Zhiheng Huang<sup>7◊</sup>

<sup>1</sup>Stanford University <sup>2</sup>Samaya AI <sup>3</sup>Orby AI <sup>4</sup>AWS AI Labs

<sup>5</sup>Google <sup>6</sup>NVIDIA <sup>7</sup>Denser.ai

wuzhengx@stanford.edu yuhao@samaya.ai peng@orby.ai yumomxu@amazon.com

## Abstract

Modern language models (LMs) need to follow human instructions while being faithful; yet, they often fail to achieve both. Here, we provide concrete evidence of a trade-off between instruction following (i.e., follow open-ended instructions) and faithfulness (i.e., ground responses in given context) when training LMs with these objectives. For instance, fine-tuning LLaMA-7B on instruction following datasets renders it less faithful. Conversely, instruction-tuned Vicuna-7B shows degraded performance at following instructions when further optimized on tasks that require contextual grounding. One common remedy is multi-task learning (MTL) with data mixing, yet it remains far from achieving a synergic outcome. We propose a simple yet effective method that relies on **Rejection Sampling for Continued Self-instruction Tuning (RESET)**, which significantly outperforms vanilla MTL. Surprisingly, we find that less is more, as training RESET with high-quality, yet substantially smaller data (three-fold less) yields superior results. Our findings offer a better understanding of objective discrepancies in alignment training of LMs.

## 1 Introduction

Aligning language models (LMs) with human preferences becomes increasingly important. One main objective is to train LMs to follow human instructions (e.g., answering open-ended questions) while being faithful (e.g., grounding responses in the given context). However, LMs often suffer from failing to follow human instructions (Kadavath et al., 2023; Chen et al., 2023; Ji et al., 2023) or making up facts that are not grounded in context (Zhang et al., 2023; Wang et al., 2023a; Huang et al., 2023; Ghosh et al., 2024).

We trace this problem back to commonly used alignment training datasets, often collected from naturalistic conversations covering a wide range of

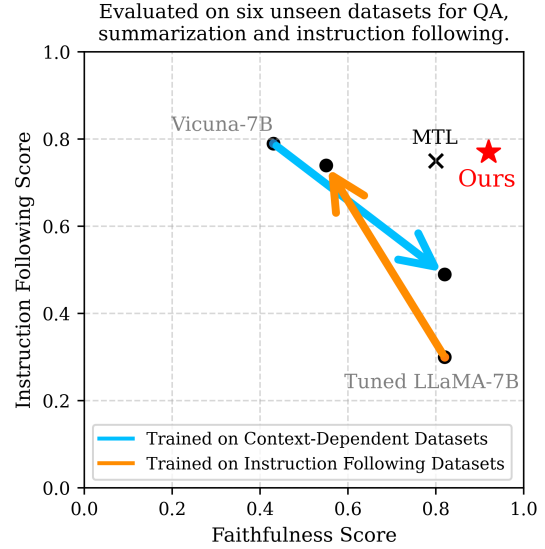


Figure 1: Faithfulness scores on context-dependent tasks (QA and summarization) decrease when we fine-tune grounded LLaMA-7B checkpoint with instruction following datasets (orange), and instruction following scores (assessed by GPT-4) decrease when we fine-tune Vicuna-7B with context-dependent tasks (blue). Our method, RESET surpasses the vanilla MTL with data mixing, approaching the North Star (upper right corner).

domains (Taori et al., 2023; Geng et al., 2023; Chiang et al., 2023; Christiano et al., 2017; Ouyang et al., 2022). For instance, Alpaca (Taori et al., 2023) and Dolly-15K (Conover et al., 2023) cover tasks from creative writing to context-dependent QA. Particularly, these tasks may have distinct objectives, and, when mixed, may induce potential conflicts of interest during alignment.

In this context, we examine the interaction between instruction following (i.e., how well does the LM follow open-ended instructions) and faithfulness (i.e., is the LM’s response grounded in the context) during alignment training. Specifically, we study how instruction following and faithfulness scores change when adapting LMs to two types of datasets: 1) *instruction-tuning datasets* that are commonly used to train chat-models such as Chat-

\*Equal contribution. ◊Work done at AWS AI Labs.

GPT (OpenAI, 2022) and Llama-2-Chat (Touvron et al., 2023); and 2) *context-dependent datasets* that require grounding to a provided *context*, and are commonly used to train Retrieval-Augmented Generation (RAG) models (Lewis et al., 2020) such as Atlas (Izacard et al., 2023) and DSPy (Khat-tab et al., 2024). We observe a clear trade-off between the two scores, as shown in Figure 1: a fine-tuned LM with a competitive faithfulness score becomes much less faithful when separately fine-tuned on instruction following datasets. Conversely, an instruction-tuned LM becomes worse at following instructions when fine-tuned on context-dependent datasets. Our findings suggest fine-tuning an LM with either instruction following or context-dependent datasets exclusively may impair its original ability in the other aspect.

One natural mitigation strategy is to use multi-task learning (MTL) by mixing datasets, which we find as a strong yet sub-optimal baseline. To achieve a more synergic outcome, we propose a simple yet effective method **Rejection Sampling for Continued Self-instruction Tuning (RESET)**. Inspired by recent works in self-instruct (Zelikman et al., 2022; Wang et al., 2023b), RESET leverages the LM to sample generations for instruction following and task-specific datasets, which is different from vanilla MTL. Generations are then rated by external judges for instruction following and faithfulness scores, where top-rated generations are collected and used to further fine-tune the LM. Our experiments show that when trained with RESET using only a single iteration and 8,000 additional fine-tuning examples, LMs see substantial gains in faithfulness scores (up to +18.8%) compared to the MTL baseline, while maintaining their instruction-following scores. Furthermore, we find that less is more: training with RESET on higher quality yet three-fold less data yields up to 31.3% improvements on faithfulness among datasets compared to MTL. Our analyses shed lights on finding and mitigating objective discrepancies in alignment training where datasets encompass different or even conflicting goals.<sup>1</sup>

## 2 Background

To recap, our goal is to study the interplay of training on instruction following datasets and on datasets that require grounding to a specific context.

<sup>1</sup>We will release our code and evaluation data at <https://github.com/frankaging/dancing-in-chains>.

Dataset	Train	Dev	Test	Avg. Length
Instruction Datasets	257,307	2,500	-	305
<i>Alpaca-15K</i>	-	-	31,323	79
<i>Vicuna-eval</i>	-	-	80	367
<i>Koala-eval</i>	-	-	180	444
NQ	69,639	8,757	13,368	6
CNN/Daily Mail	287,113	13,368	11,490	76
MS MARCO	153,725	2,500	12,466	23
<i>BioASQ</i>	-	-	1,956	9
<i>SearchQA</i>	-	-	31,760	4
<i>WikiSum</i>	-	-	2,000	140

Table 1: Data statistics of two-stage fine-tuning experiments. *Italicized* datasets are held-out unseen evaluation datasets, while the rest are used for training. “Instruction Datasets” here consist of publicly available datasets such as Dolly-15K (Conover et al., 2023), ShareGPT, Self-Instruct (Wang et al., 2023b) and OASST-1 (Köpf et al., 2023). Average length is the averaged response token length, which is calculated with training or testing sets for each dataset with the LLaMA-7B tokenizer.

To do this, we first construct two separate groups of datasets, each resembling a setting above. We then adopt a two-stage fine-tuning paradigm that allows us to closely study the effect of each type of training. We introduce the setup in this section.

### 2.1 Datasets

We outline different datasets used for instruction following and context-dependent fine-tuning. Table 1 shows data statistics. For each dataset we carefully design the instructions to well-align the input and output (see Section 2.3).

#### 2.1.1 Instruction Following Datasets

We curated an instruction following training dataset by compiling unique examples from publicly available datasets such as Dolly-15K (Conover et al., 2023), ShareGPT<sup>2</sup>, Self-Instruct (Wang et al., 2023b), and OASST-1 (Köpf et al., 2023). For evaluation, we gather unique examples from Alpaca-15K (Taori et al., 2023), Vicuna-eval (Chiang et al., 2023), and Koala-eval (Geng et al., 2023).

**Data Pre-processing** We exclude examples that originally come with context (e.g., examples labeled as summarization type in Dolly-15K are filtered out) to prevent overlaps with our context-dependent datasets. We retain only unseen examples in our evaluation sets. For OASST-1, we only include examples with an average human rating higher than 0.5 (i.e., the higher the rating is, the

<sup>2</sup><https://huggingface.co/datasets/LLMs/Alpaca-ShareGPT>

```

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:
[Task-specific Instruction Abbreviated]

### Input:
[Passage: Abbreviated]
[... ...]

###Response:
Model Generated Answer Goes Here

```

Table 2: The prompt template we used for training and evaluation for LLaMA-7B. The instruction field contains task-specific instruction, the input field contains contexts if applicable, and the response field is followed by model’s generation. We use a different template for Vicuna-7B since it uses a different template during instruct-tuning phase. Actual input examples for each dataset and model are included in Appendix A.6.

higher the quality is) and that are rated by at least two annotators. To refine ShareGPT, we include only examples with responses that are longer than 10 words, split by whitespace. For other instruction following training datasets, we exclude examples with empty instructions or responses.

### 2.1.2 Context-Dependent Datasets

We aim to evaluate the model’s generation for faithfulness to the given input context. We select a range of context-dependent datasets from three task domains: (1) extractive QA including NQ (Kwiatkowski et al., 2019), BioASQ and SearchQA, taken from the RobustQA benchmark (Han et al., 2023); (2) abstractive QA with MS MARCO (Bajaj et al., 2016) where the answers are well-formed sentences grounded in context; and (3) abstractive summarization including CNN DailyMail (Hermann et al., 2015; Nallapati et al., 2016) and WikiSum (Liu et al., 2018). BioASQ, SearchQA and WikiSum are hold out for evaluation and the rest are for training. One crucial advantage of using these context-dependent datasets is that they provide us with a reliable way of measuring faithfulness, in terms of how well the response is grounded in the given context.

**Data Pre-processing** For QA datasets, we include five retrieved passages maximally as the context,<sup>3</sup> where the gold answer is at least mentioned

<sup>3</sup>Details about the retrieving process can be found in the original papers of RobustQA (Han et al., 2023) and MS

in one of the passages. For MS MARCO, we only include examples where there exist at least one well-formed answers. In cases involving multiple retrieved passages, we concatenate all passages with line breaks inserted between them.

## 2.2 Evaluation Metrics

In the context of our datasets, we evaluate our models with three metrics: instruction following score, faithfulness score and task performance score. The standard methods for measuring instruction following and faithfulness of language models are subject to ongoing debate. In this work, we employ widely adopted approaches for these measurements and compile a set of metrics for more stable evaluations.<sup>4</sup> To present our findings, we report macro-averaged results across all test datasets.

**Instruction Following Score** We adopt the commonly used evaluation paradigm proposed by LLM-as-a-Judge (Zheng et al., 2023), and zero-shot prompt GPT-4 to provide a rating followed by a short explanation (i.e., named as LLM-as-a-Judge (R) in the paper).<sup>5</sup> For the GPT-4 evaluator, we set the temperature to 0 for stability with a maximum generation length of 512. We check instruction following scores only for instruction following evaluation datasets. See Appendix A.1 for our actual evaluation prompt.

**Faithfulness Score** For extractive QA datasets, we utilize the span coverage score as our metric (i.e., whether the predicted answer is a span within the context). We apply standard normalization to both the predicted answer and the context (see Appendix A.2 for details). A score of 1.0 is assigned if the span is covered, and 0.0 otherwise. Additionally, we include unigram and bigram coverage for selected datasets to further refine our faithfulness evaluation in the Appendix (see Figure 9 and Figure 10). For abstractive QA and summarization datasets, we employ SummaC-ZeroShot (SummaC-ZS; Laban et al. (2022)) to assess whether the provided context (with the question concatenated as a prefix for QA datasets) entails the model-generated answer. Specifically, we segment both the context

MARCO (Bajaj et al., 2016).

<sup>4</sup>Throughout the paper, we sample a subset of the full evaluation data which include 6,000 examples (1,000 examples from each context-dependent evaluation set), and sample 300 examples (100 examples from each instruction following evaluation set) due to limited compute resources.

<sup>5</sup>We use gpt-4-0613.

Dataset Type	Instruction
<b>Extract QA</b>	Answer to the question by <b>extracting a specific text span</b> from the given passages. Do not include new information beyond the given passages.
<b>Abstractive QA</b>	Answer the question with well-formed sentences. Paraphrase the context in the passages if necessary. <b>Do not include new information</b> beyond the given passages.
<b>Summarization</b>	Summarize the text in a few sentences. Using original phrases or paraphrasing them if necessary. <b>Do not include new information</b> beyond the given passages.

Table 3: Task specific instructions. We bold texts that indicate that our prompts are designed to be objective-aligned with our instruction following training data (i.e., fine-tuning our model on instruction following datasets should with keeping it to be faithful as well).

and the answer into sentences and then use a pre-trained NLI model to compute an entailment score between pairs. These NLI scores are aggregated across pairs using max-pooling.

**Task Performance** Task performance complements the aforementioned metrics, as task performance is generally expected to decline when models are fine-tuned across domains. We utilize Exact Match (EM) for extractive QA datasets and the ROUGE-L score (Lin, 2004) for abstractive QA and summarization datasets.

### 2.3 Instruction Template

Our datasets follow the same instruction-tuning template format as used in the Alpaca setting (Taori et al., 2023). This template includes a header that outlines general guidelines, followed by task-specific instructions, as illustrated in Table 2. For QA tasks, the question is presented after the retrieved passages.

**Objective-Aligned Instructions** We design our task-specific instructions to minimize *objective-conflict* among datasets. For example, in prompting our model for a context-dependent extractive QA task, we explicitly instruct the model to “extract a specific text span from the given passages”. This template helps models to reduce hallucination when fine-tuned with instruction following datasets: as models getting better at understanding human instructions, they also get better at understanding how to *extracting a span* which ensures the answer to be faithful. The instructions for all datasets are depicted in Table 3. Our task-specific templates ensure there is no objective conflict when we fine-tune our models with mixed of datasets. We use the same template for training and evaluation for each task.

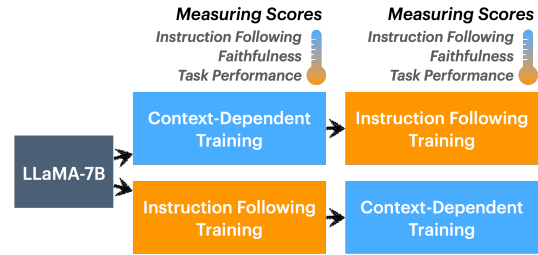


Figure 2: Two-stage fine-tuning with LLaMA-7B.

### 2.4 Two-stage Fine-tuning Paradigm

To understand the trade-off between instruction following and faithfulness when training LMs with both objectives, we formulate a two-stage fine-tuning paradigm to answer our research questions (as illustrated in Figure 2).<sup>6</sup> For our first pipeline, we initially fine-tune our LM with context-dependent datasets that require grounding. We then take the best checkpoint from the initial stage to further fine-tune it on instruction following datasets (CD→IF). Conversely, in our second pipeline, we fine-tune instruct-tuned LM (e.g., Vicuna-7B) with context-dependent datasets (IF→CD). For both pipelines, we measure instruction following and faithfulness scores before and after training, to gauge the impact of the second-stage training on both capabilities. We follow this paradigm to find evidence of the trade-off in Section 3 and Section 4.

**Models** We use two models in our two-stage fine-tuning paradigm. We use a base LM LLaMA-7B (Touvron et al., 2023), one of the most widely used open-source LM, for the CD→IF pipeline. For our IF→CD pipeline, we use Vicuna-7B off-the-shelf as our instruct-tuned LLaMA-7B without re-training one from scratch. Vicuna-7B is one of the most competitive open-source chat-model, and is a

<sup>6</sup>Our approach is akin to the data inoculation paradigm proposed by Liu et al. (2019a), albeit with significantly larger models and datasets.



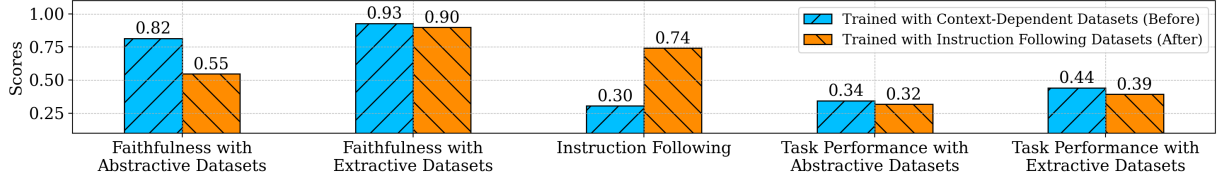


Figure 3: Macro-averaged faithfulness, instruction following, and task performance scores on corresponding evaluation datasets before and after fine-tuning with instruction following datasets.

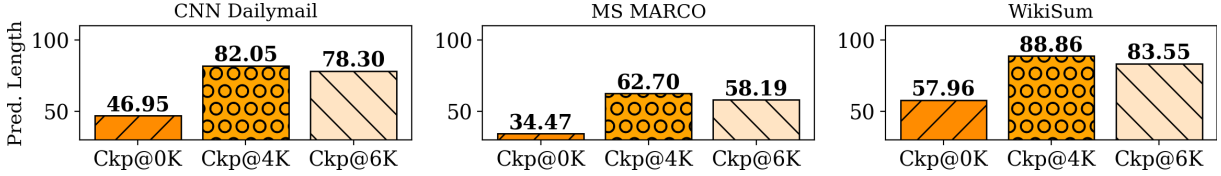


Figure 4: Average generation token length throughout the instruction following training stage. The first checkpoint is the best checkpoint from the context-dependent training stage. The middle checkpoint is with the lowest evaluation loss during the second stage.

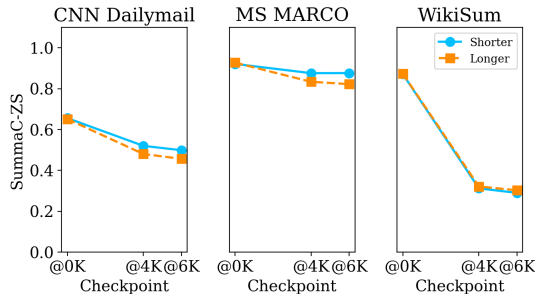


Figure 5: Faithfulness scores for abstraction QA and summarization datasets categorized by whether the generation length is strictly shorter or much longer ( $> 100$  tokens) than the golden answer.

fine-tuned LLaMA-7B on conversational data from ShareGPT (Zheng et al., 2023). Our hypothesis and paradigm are transferable to other base LMs at different scales, although we pick these two models as they are among the first open-sourced LMs during the time frame of this project. Other experimental setup details are included in Appendix A.3.

### 3 Does Fine-tuning with Instruction Following Data Hurt Faithfulness?

To answer this question, we follow the first pipeline outlined in Figure 2, where we take our LLaMA-7B that is fine-tuned on context-dependent datasets and further fine-tune it with instruction following datasets. Our results are shown in Figure 3. First of all, instruction following scores increase drastically as expected, from 0.30 to 0.74. Meanwhile, faithfulness scores on abstractive datasets drop 33.0% from 0.82 to 0.55, while the task performance is largely maintained (from 0.34 to 0.32). For ex-

tractive datasets, both faithfulness scores and task performance see a relatively small drop. To provide a fine-grained view of these datasets, we provide additional analysis of per-task and per-metric changes of faithfulness score as well as task performance in Appendix A.4 and Appendix A.5.

One potential confounding factor for faithfulness scores dropping on abstractive tasks is the length of the generated response. The model could simply have generated longer responses (Figure 4) as a result of training on instruction tuning data with longer responses (Table 1). To rule this out, we re-evaluate our models separately with two contrasting groups: evaluated with only those examples with shorter generation length (less than or equal to), and those examples with much longer generation length (100 more tokens) compared with the golden answer. As shown in Figure 5, both short and long generations see very substantial drops in faithfulness, while longer generations indeed see larger drops. This nevertheless supports our conclusion that instruction following training hurts faithfulness.

### 4 Does Context-Dependent Fine-tuning Hurt Instruction Following?

To answer this question, we fine-tune Vicuna-7B (LLaMA-7B fine-tuned on ShareGPT) on context-dependent datasets. Figure 6 shows our results. As expected, our model becomes more faithful, with improved task performance. This improvement is partly because Vicuna-7B does not follow our instruction template out-of-the-box. On the other hand, instruction following scores drop by

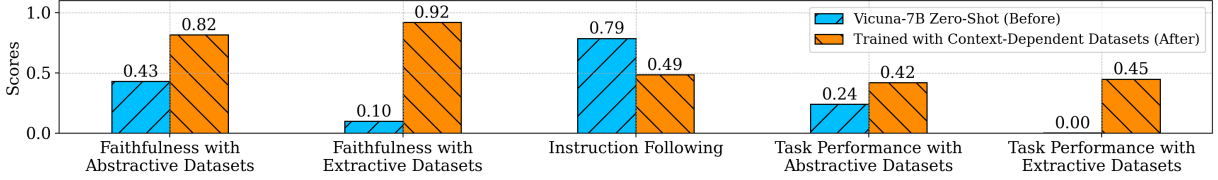


Figure 6: Macro-averaged faithfulness, instruction following, and task performance scores on corresponding evaluation datasets before and after fine-tuning with context-dependent datasets.

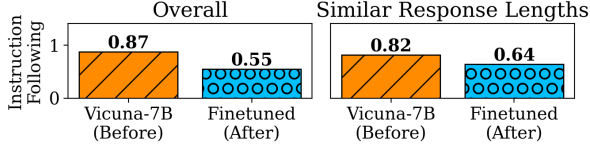


Figure 7: Instruction following scores for 1,000 randomly selected examples from Alpaca-15K (left), and for filtered examples with similar generation lengths (maximally 10 tokens longer) before and after fine-tuning with context-dependent datasets (right).

37.9%, suggesting our model becomes much less human-aligned. Similar to previous findings, one natural concern is that context-dependent training makes our model generate much shorter responses, which leads to lower instruction following scores. To rule out this concern, we randomly sample 200 examples from the Alpaca-15K dataset and only keep those with a minimal sequence length change (a maximum of 10 new tokens) compared to the corresponding Vicuna-7B zero-shot generation. As shown in Figure 7, instruction following scores consistently drop considerably for the filtered setting, regardless of length (36.8% vs. 22.0%). Overall, context-dependent fine-tuning negatively affects the instruction following score, as we expected.

## 5 RESET: Reconciling Instruction Following and Faithfulness

As shown in the previous two sections, there are clear objective discrepancies between diverse instruction following datasets and context-grounded ones. In this section, we explore two different mitigation strategies, starting from the MTL baseline to our novel method based on **Rejection Sampling for Continued Self-instruction Tuning (RESET)**, which significantly outperforms MTL.

**Our MTL Baseline** Our MTL simply mixes up the instruction following as well as context-dependent datasets. We up-sample smaller datasets to have an equal number of examples per dataset.

### 5.1 Our Method

RESET starts with a model that is fine-tuned on a mixture of instruction following and context-dependent datasets and reconciles the two objectives with the following steps (Figure 8):

**Sample Generations** For a random subset of the training datasets, we sample generations from the checkpoint with different decoding settings. Specifically, we focus on two hyperparameters by randomly changing one of them at a time by enumerating all possible values. For the decoding temperature, it takes on a value from  $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$ . For top-k, we set it to take on a value from  $\{5, 10, 20, 50, 70, 90, 100\}$ . When we vary the temperature, we fix top-k to be 0. Similarly, when we vary the top-k, we fix the temperature at 0. For each example, if we run the sampling procedure once, it will sample 7 examples in total.

**External Judges** We then use a set of external judges to rate our collected generations as in Section 2.2. These judges are offline evaluators. We evaluate generation based on instruction following score, faithfulness score and task performance. One potential limitation is that our model may overfit to existing judges on seen datasets. We thus include unseen datasets to test generalizability.

**Top-1 Weighted Score** For each generation, we collect a set of scores from our external judges. Then, we take a weighted sum of these scores before ranking the generations and picking the top rated one. The score is weighted score of task performance  $s_{\text{task}}$ , instruction following  $s_{\text{instr}}$  and faithfulness scores  $s_{\text{faith}}$ ,

$$\text{score} = s_{\text{task}} + 2.0 * (\mathbb{I}_{\text{instr}} * s_{\text{instr}} + \mathbb{I}_{\text{faith}} * s_{\text{faith}})$$

where we use  $\mathbb{I}_{\text{instr}}$  and  $\mathbb{I}_{\text{faith}}$  to indicate whether the example is from our instruction following datasets or context-dependent datasets. We pick the top rated sample per example and combine them into a continued fine-tuning dataset as described next.

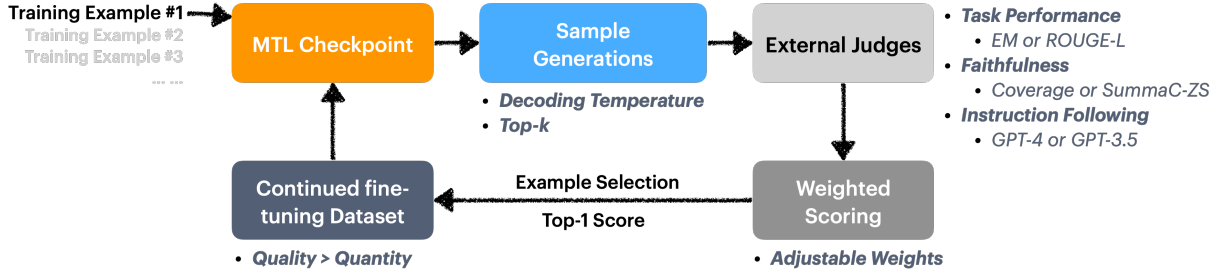


Figure 8: The illustration of our proposed method RESET. It samples generations from the initial vanilla multi-task learning (MTL) checkpoint with seen examples from instruction following and context-dependent datasets. For each example, it generates a set of possible responses with different decoding strategies. Generations are rated by external judges with a weighted scores of task performance, faithfulness and instruction following scores. Then, the top rated generations will be collected to further fine-tune the initial model.

**Continued Fine-tuning** For RESET, we randomly sample training data from each of our four dataset in Table 1 and collect 2,000 additional fine-tuning examples per dataset. In total, our new fine-tuning dataset has 8,000 examples. With this small collected dataset, we further fine-tune our starting checkpoint model for a single epoch with a smaller learning rate to avoid overfitting. This continued fine-tuning step is very lightweight as the training data is usually less than 1% of the MTL training step. Overall, RESET resembles a reject sampling based preference learning paradigm, which has been proven effective (Touvron et al., 2023) while drastically saving training costs as well as increasing stability. Other experimental setup details are included in Appendix A.3.

**Supercharged RESET (RESET-S)** In addition, we evaluate RESET with a different setting to test the impact of the quality of our continued fine-tuning dataset. Specifically, we *supercharge* the quality of our additional fine-tuning dataset by sampling  $1\times$  more generations. We swap our instruction following judge from GPT-4 to the weaker ChatGPT<sup>7</sup>. One potential benefits is that ChatGPT is a weaker judge with a lower recall for good generations. As a result, examples rated high by ChatGPT may have higher quality. While sampling for generation, we decrease the curated fine-tuning dataset by 3-fold (2,000 examples in total). We fine-tune our checkpoint model with the same setup.

## 5.2 Results

Table 4 shows results on models fine-tuned with other baselines: (1) **Vicuna-7B**: Evaluated with Vicuna-7B in a zero-shot manner. (2) **w/o MTL**:

We fine-tune Vicuna-7B with context-dependent datasets without any MTL objective. The purpose is to establish a potential upper bound of faithfulness score without any further training for instruction following. (3) **w/ MTL**: We fine-tune Vicuna-7B with baseline MTL by mixing instruction following and context-dependent datasets together. (4) **w/ RESET**: We fine-tune Vicuna-7B directly with our collected continued fine-tuning dataset (8,000 examples) from our RESET pipeline without MTL training first. Note that the curated dataset comprises of sampled generations from our MTL checkpoint. The purpose is to evaluate whether MTL is necessary for model improvements. (5) **w/ MTL+RESET**: We follow Figure 8 to fine-tune our MTL checkpoint with curated dataset. (6) **w/ MTL+RESET-S**: We follow Figure 8 to fine-tune our MTL checkpoint with the high-quality version of our curated dataset (2,000 examples).

We evaluate our models on both seen and unseen testing datasets. As shown in Table 4, fine-tuning Vicuna-7B with a mixture of datasets close the gap on instruction following score substantially (from 0.49 to 0.75) while leaving headroom across the board. Next, models fine-tuned with RESET significantly outperform the MTL baseline on both seen and unseen testing datasets. Our results also suggest directly fine-tuning Vicuna-7B performs worse compared to fine-tuning a model checkpoint after MTL training. Last but not least, our results with RESET-S provide strong evidence that data quality is more important than data quantity and that using multiple iterations of sampling helps, as RESET-S achieves similar or better performance with 3-fold less training data. Due to space limit, we present qualitative model generations in Appendix A.7.

<sup>7</sup>We use gpt-4-0613.

	Faithfulness								Instruction Following Datasets
	Abstractive Datasets				Extractive Datasets				
	CNN DailyMail	MS MARCO	WikiSum	Overall	NQ	SearchQA	BioASQ	Overall	
Vicuna-7B	0.36	0.59	0.34	0.43	0.06	0.21	0.02	0.10	0.79
w/o MTL	0.67	<b>0.93</b>	0.85	0.82	0.89	0.94	0.93	0.92	0.49
w/ MTL	0.67	0.92	0.80	0.80	0.90	0.94	0.93	0.92	0.75
w/ ReSET	0.80	0.90	0.88	0.86	<b>0.96</b>	<b>0.99</b>	0.92	0.96	0.73
w/ MTL+ReSET	0.77	<b>0.93</b>	<b>0.95</b>	0.85	<b>0.96</b>	0.97	<b>0.98</b>	<b>0.97</b>	0.76
w/ MTL+ReSET-S	<b>0.88</b>	<b>0.93</b>	<b>0.95</b>	<b>0.92</b>	0.94	0.97	0.97	0.96	<b>0.77</b>

Table 4: Faithfulness and alignment scores on testing datasets, and unseen datasets are *italicized*. Scores are averaged across three distinct runs. Higher scores are better. Overall scores are macro-average across datasets.

## 6 Related Work

**Instruction Following of LMs** There are a variety of instruction following training datasets (Taori et al., 2023; Geng et al., 2023; Chiang et al., 2023; Wang et al., 2023b) covering tasks from wildly different domains such as poetry creation to context-dependent summarization. In this paper, we focus on a particular trade-off between instruction following and model faithfulness, which are fundamental objectives for modern LMs. In addition to training datasets, LM generations are often assessed by human experts (Ouyang et al., 2022) or model scorers (Zheng et al., 2023; Ethayarajh et al., 2022). We use GPT-4 to assess instruction following scores as in Chiang et al. (2023).

**Faithfulness and Groundedness of LMs** Being faithful or grounded is crucial for tasks like context-dependent QA or summarizations, reducing hallucinations for LMs (Zhang et al., 2023; Wang et al., 2023a; Huang et al., 2023; Kadavath et al., 2023; Chen et al., 2023; Ji et al., 2023). Unlike task performance, faithfulness measures whether generated answers are based on the given context (Rashkin et al., 2021; Dziri et al., 2022; Paranjape et al., 2022). For context-dependent QA and summarization benchmarks, common metrics include subsequence-based lexical matching (Papineni et al., 2002; Lin, 2004; Banerjee and Lavie, 2005), natural language inference (NLI; Laban et al., 2022; Fabbri et al., 2022), and more recently, LLMs as factuality scorers (Chiang and Lee, 2023; Liu et al., 2023; Kamalloo et al., 2023). We use lexical matching and NLI to measure faithfulness by checking whether the answer is grounded.

**Instruction Following Training with LMs** Base LMs are often tuned to follow human instructions (Wei et al., 2022; Mishra et al., 2022; Wang et al., 2022; Chung et al., 2022). Various post-hoc fine-tuning techniques have been proposed

to align base LMs with human preferences, such as supervised fine-tuning on instruction-following datasets (Taori et al., 2023; Chiang et al., 2023) or variances of RLHF and RLAI (Schulman et al., 2017; Ouyang et al., 2022; Lee et al., 2023; Rafailov et al., 2023; Touvron et al., 2023). These techniques can require training policy and reward models which is not cost-efficient. Inspired by alignment training with reject-sampling (Touvron et al., 2023), we propose a simple yet effective method that uses reject-sampling by self-instruct. On the other hand, there exist various advance MTL techniques (Liu et al., 2019b; Crawshaw, 2020) prior to the development of LMs. However, RE-SET is the only variant that leverages LM’s self-instruct generations to further fine-tune the LM. More importantly, RESET is complimentary to different instruct-tuning techniques by leveraging the LM and external evaluators to generate high-quality continued fine-tuning datasets. Compared with recent works showing how instruct-tuning may cause hallucination (Ghosh et al., 2024), we provide a more rigorous analysis with our two-stage fine-tuning paradigm and focus on the interaction between instruction following and faithfulness.

## 7 Conclusion

In this paper, we demonstrate a clear trade-off between instruction following and faithfulness when fine-tuning LMs on datasets with differing objectives. To alleviate this trade-off, we proposed **RESET**, a simple yet effective iterative method that significantly outperforms the MTL baseline in both instruction following and faithfulness scores. The lightweight and iterative nature of RESET makes it extensible for future refinement at minimum cost and integrable with recent instruct-tuning techniques. We contribute to the broader goal of creating more reliable, accurate, and user-aligned language technologies.



## Limitations

The limitations of our work are as follows:

- Our study primarily focuses on the LLaMA-7B and Vicuna-7B models, that are among the best open-source models at the time of this work. While we posit that our findings and the proposed RESET method could generalize across other language models, our findings could remain speculative without evaluating on more current model types (e.g., LLaMA-2, Mistral or Mixtral at various scales).
- The datasets chosen for fine-tuning and evaluation, though comprehensive, are not exhaustive. There are other interesting datasets that are not covered in this study. For instance, long-form QA where the answers are much longer than 1-2 phrases or sentences. Our instruction following datasets can also be further categorized into creativity-driven, world-knowledge driven and others to help us to disentangle discrepancies in objectives better across datasets.
- Our evaluation relies heavily on automated metrics and external judges like external LMs for assessing instruction following and faithfulness. While these methods are standard, they cannot fully encapsulate the nuanced understanding and preferences of human evaluators. For future research, evaluating responses with human annotators would provide additional validations.
- Although the purpose of our study is to study the objective discrepancies in the datasets and come up with mitigation strategies without another novel training paradigm, it would strengthen our results if we can compare our method with more recent alignment training methods.
- Although we evaluate RESET on unseen datasets, our method still has the potential to overfit to certain evaluators. Future work may use a different set of evaluators for a more robust evaluation. Human evaluation is challenging for our unseen testing datasets, such as WikiSum, because the input and response are extremely long, which could also make human ratings unstable.

## Ethics Statement

In this paper, we delve into the nuances of instruct-tuning LMs, a process that involves balancing a trade-off between instruction following and faithfulness. We feel that the immediate ethical and societal consequences of our research may be limited, and none which we feel must be specifically highlighted here.

## References

- Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2016. [Ms marco: A human generated machine reading comprehension dataset](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.
- Shiqi Chen, Yiran Zhao, Jinghan Zhang, I-Chun Chern, Siyang Gao, Pengfei Liu, and Junxian He. 2023. [Felm: Benchmarking factuality evaluation of large language models](#). In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.
- Cheng-Han Chiang and Hung-yi Lee. 2023. [Can large language models be an alternative to human evaluations?](#) In *Association for Computational Linguistics (ACL)*.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. [Vicuna: An open-source chatbot impressing gpt-4 with 90%\\* chatgpt quality](#).
- Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. [Deep reinforcement learning from human preferences](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny

- Zhou, Quoc V. Le, and Jason Wei. 2022. [Scaling instruction-finetuned language models](#). In *arXiv preprint arXiv:2210.11416*.
- Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. [Free dolly: Introducing the world’s first truly open instruction-tuned llm](#).
- Michael Crawshaw. 2020. [Multi-task learning with deep neural networks: A survey](#). In *arXiv preprint arXiv:2009.09796*.
- Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. [On the origin of hallucinations in conversational models: Is it the datasets or the models?](#) In *North American Chapter of the Association for Computational Linguistics (NAACL)*.
- Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. [Understanding dataset difficulty with  \$\mathcal{V}\$ -usable information](#). In *International Conference on Machine Learning (ICML)*.
- Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. [QAFactEval: Improved QA-based factual consistency evaluation for summarization](#). In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.
- Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. [Koala: A dialogue model for academic research](#). Blog post.
- Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, Ramaneswaran S, Deepali Aneja, Zeyu Jin, Ramani Duraiswami, and Dinesh Manocha. 2024. [A closer look at the limitations of instruction tuning](#). In *arXiv preprint arXiv:2402.05119*.
- Rujun Han, Peng Qi, Yuhao Zhang, Lan Liu, Juliette Burger, William Yang Wang, Zhiheng Huang, Bing Xiang, and Dan Roth. 2023. [RobustQA: Benchmarking the robustness of domain adaptation for open-domain question answering](#). In *Findings of Association for Computational Linguistics (ACL)*.
- Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. [Teaching machines to read and comprehend](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. [A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions](#). *arXiv preprint arXiv:2311.05232*.
- Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. [Atlas: Few-shot learning with retrieval augmented language models](#). In *The Journal of Machine Learning Research (JMLR)*.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. [Survey of hallucination in natural language generation](#). In *ACM Computing Surveys*.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2023. [Language models \(mostly\) know what they know](#). In *Findings of Association for Computational Linguistics (ACL)*.
- Ehsan Kamalloo, Nouha Dziri, Charles L. A. Clarke, and Davood Rafiei. 2023. [Evaluating open-domain question answering in the era of large language models](#). In *Association for Computational Linguistics (ACL)*.
- Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Mateia Zaharia, and Christopher Potts. 2024. [DSPy: Compiling declarative language model calls into self-improving pipelines](#). In *International Conference on Learning Representations (ICLR)*.
- Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. 2023. [Openassistant conversations—democratizing large language model alignment](#). In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. [Natural questions: A benchmark for question answering research](#). In *Transactions of the Association of Computational Linguistics (TACL)*.
- Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. [Summac: Re-visiting nli-based models for inconsistency detection in summarization](#). In *Transactions of the Association of Computational Linguistics (TACL)*.
- Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2023. [Rlaif: Scaling reinforcement learning from human feedback with ai feedback](#). In *arXiv preprint arXiv:2309.00267*.

- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out Workshop at Association for Computational Linguistics (ACL)*.
- Nelson F. Liu, Roy Schwartz, and Noah A. Smith. 2019a. [Inoculation by fine-tuning: A method for analyzing challenge datasets](#). In *North American Chapter of the Association for Computational Linguistics (NAACL)*.
- Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. [Generating wikipedia by summarizing long sequences](#). In *International Conference on Learning Representations (ICLR)*.
- Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019b. [Multi-task deep neural networks for natural language understanding](#). In *Association for Computational Linguistics (ACL)*.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-eval: NLG evaluation using gpt-4 with better human alignment](#). In *Empirical Methods in Natural Language Processing (EMNLP)*.
- Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. [Cross-task generalization via natural language crowdsourcing instructions](#). In *Association for Computational Linguistics (ACL)*.
- Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. [Abstractive text summarization using sequence-to-sequence RNNs and beyond](#). In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.
- OpenAI. 2022. [Introducing chatgpt](#). URL <https://openai.com/blog/chatgpt>.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Association for Computational Linguistics (ACL)*.
- Ashwin Paranjape, Omar Khattab, Christopher Potts, Matei Zaharia, and Christopher D Manning. 2022. [Hindsight: Posterior-guided training of retrievers for improved open-ended generation](#). In *International Conference on Learning Representations (ICLR)*.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. [Increasing faithfulness in knowledge-grounded dialogue with controllable features](#). In *Association for Computational Linguistics and International Joint Conference on Natural Language Processing (ACL-IJCNLP)*.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. [Proximal policy optimization algorithms](#). In *arXiv preprint arXiv:1707.06347*.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. [Stanford alpaca: An instruction-following llama model](#). [https://github.com/tatsu-lab/stanford\\_alpaca](https://github.com/tatsu-lab/stanford_alpaca).
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Auralien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). In *arXiv preprint arXiv:2307.09288*.
- Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xian-gru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. 2023a. [Survey on factuality in large language models: Knowledge, retrieval and domain-specificity](#). In *arXiv preprint arXiv:2310.07521*.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. [Self-instruct: Aligning language](#)



model with self generated instructions. In *Association for Computational Linguistics (ACL)*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krma Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. [Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks](#). In *Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. [Finetuned language models are zero-shot learners](#). In *International Conference on Learning Representations (ICLR)*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. [Star: Bootstrapping reasoning with reasoning](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. [Siren’s song in the ai ocean: A survey on hallucination in large language models](#). In *arXiv preprint arXiv:2309.01219*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.

## A Appendix

### A.1 Evaluation Metric: Instruction Following Score

Table 5 shows our actual prompt template for evaluation. We follow the prompt template provided in the LLM-as-a-Judge (Zheng et al., 2023) paper. We normalize the return rating to [0, 1].

### A.2 Evaluation Metric: Faithfulness Score

**Text Normalization** We use regular expressions to replace spaces around hyphens, slashes, and before “s”. We then remove all the articles (e.g., “a” and “the”) and punctuations. We lowercase all letters for simplicity.

**Common Failure Mode** The authors also human label model’s generation to check whether model’s generations are paraphrased version of the golden answers. The most common failure mode is the model extracting a non-existent span or a wrong span. This supports our findings of models being unfaithful.

### A.3 Experimental Setup

We train our model for a maximum of a single epoch<sup>8</sup> across all training jobs. We up-sample the smaller datasets to match the number of examples in the larger ones when combining datasets for training. The learning rate is set at  $1 \times 10^{-5}$  with a batch size of 16 for our faithfulness-driven training. For training jobs involving instruction following datasets, we increase the batch size to 32. Training is conducted using bfloat16 precision with a maximum sequence length of 2048. The weight decay is set to 0.05, with a cosine learning rate scheduler and a warm-up ratio of 0.03. We save checkpoints every 100 training steps and evaluate them based on perplexity scores on the evaluation set. The best-performing checkpoint is then selected for the next stage of training. Our models are trained using the stage-3 deepspeed library. We train each model with three random seeds and average the results for consistency. For each training job, our models are trained on 8×A100 GPUs within a single-node setup, with the total training time not exceeding 24 hours. For model generation, we employ greedy decoding with a maximum generation length of 480, which aligns with the maximum response length across the training datasets.

For the continued fine-tuning step in RESET, we use a smaller learning rate of  $8 \times 10^{-6}$  and keep other settings the same.

### A.4 Additional Analysis on Instruction Following Fine-tuning

Figure 11 shows how instruction following scores vary during fine-tuning LLaMA-7B with context-dependent tasks. Figure 12 shows how instruction following scores vary during the second-stage fine-tuning with instruction following datasets.

<sup>8</sup>Experiments with up to three epochs showed minimal changes in results.



[System] Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question displayed below. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of the response. Begin your evaluation by providing a short explanation (strictly 1-2 short sentences). Be as objective as possible. After providing your explanation, please rate the response on a scale of 1 to 10 by strictly following this format: "[[rating]]", for example: "Rating: [[5]]".
[Question] [The Question Abbreviated]
[The Start of Assistant's Answer] Answer: <b>Model Generated Answer Goes Here</b> [The End of Assistant's Answer]

Table 5: The template of the prompt we used for evaluating instruction following scores using GPT-4. This template is adopted from LLM-as-a-Judge (Zheng et al., 2023) paper.

A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions.
### USER: [Task-specific Instruction Abbreviated] [Passage: Abbreviated] [... ...] ###ASSISTANT: <b>Model Generated Answer Goes Here</b>

Table 6: The prompt template we used for training and evaluation for Vicuna-7B. The instruction field contains task-specific instruction, the input field contains contexts if applicable, and the response field is followed by the model generation.

## A.5 Quandary of Mixed Training on Abstractive and Extractive QA and Summarization Datasets

Figure 9 shows faithfulness score and task performance across related datasets when we fine-tune LLaMA-7B on context-dependent datasets. One suprising finding is that throughout the fine-tuning process, the faithfulness scores on extractive QA datasets gradually decrease, while task performance scores gradually increase. On the other hand, this trend is not salient for abstractive tasks as shown in Figure 13.

## A.6 Examples of Actual Instructions

From Table 9 to Table 15, we show actual model inputs from each testing datasets.

## A.7 Qualitative Examples

Table 7 and Table 8 show two qualitative examples of actual model generations from our experiments.

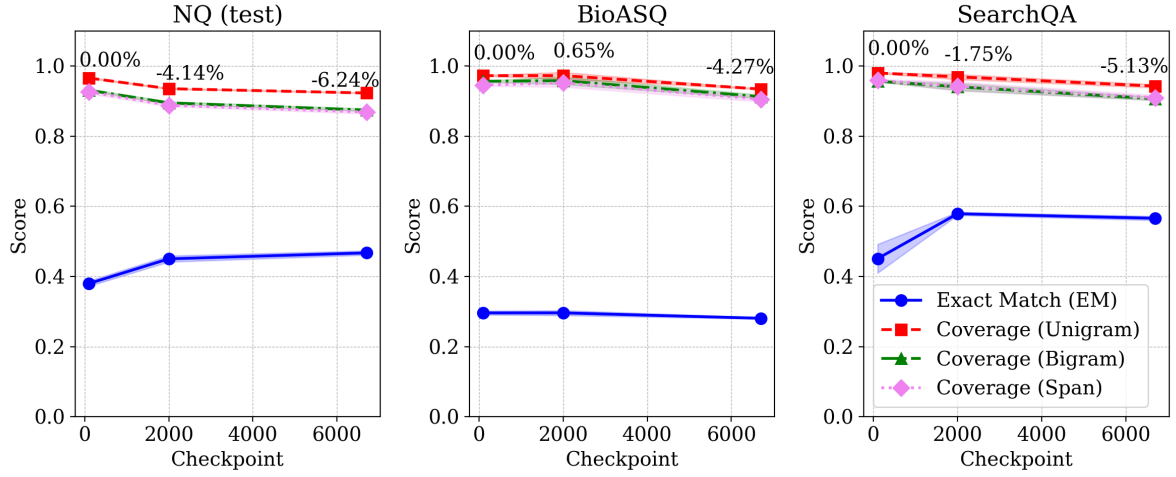


Figure 9: Individual faithfulness score and task performance with extractive QA datasets evaluated with three distinct model checkpoints of LLaMA-7B fine-tuned on context-dependent datasets. The middle checkpoint is the one with lowest in-training evaluation loss.

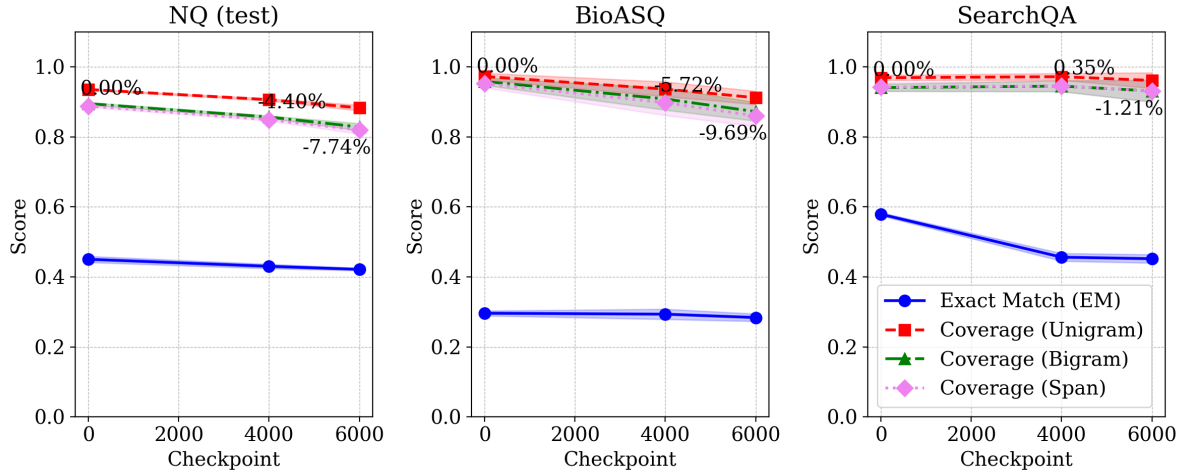


Figure 10: Individual faithfulness score and task performance with extractive QA datasets across evaluated with three distinct model checkpoints through the instruction following fine-tuning. The middle checkpoint is the one with lowest in-training evaluation loss.

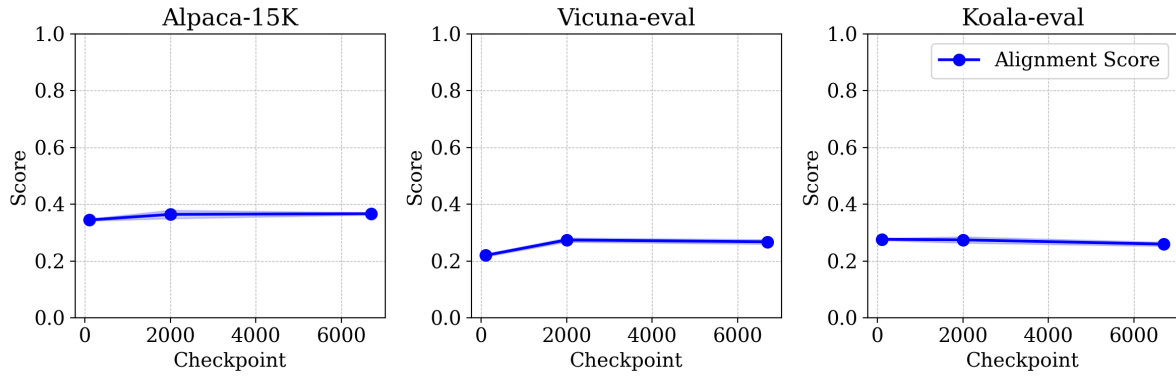


Figure 11: Individual instruction following score on alignment datasets evaluated with three distinct model checkpoints of LLaMA-7B fine-tuned on context-dependent datasets. The middle checkpoint is the one with lowest in-training evaluation loss.

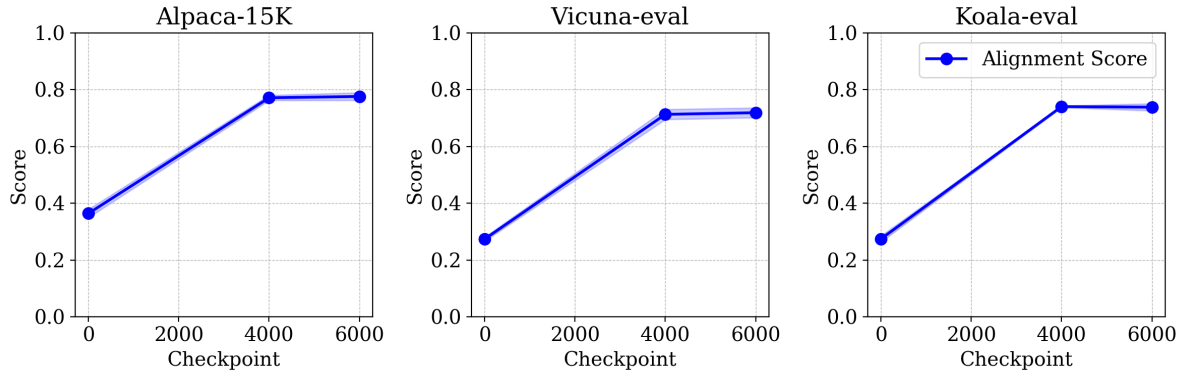


Figure 12: Individual instruction following score on alignment datasets evaluated with three distinct model checkpoints through the instruction following fine-tuning. The middle checkpoint is the one with lowest in-training evaluation loss.

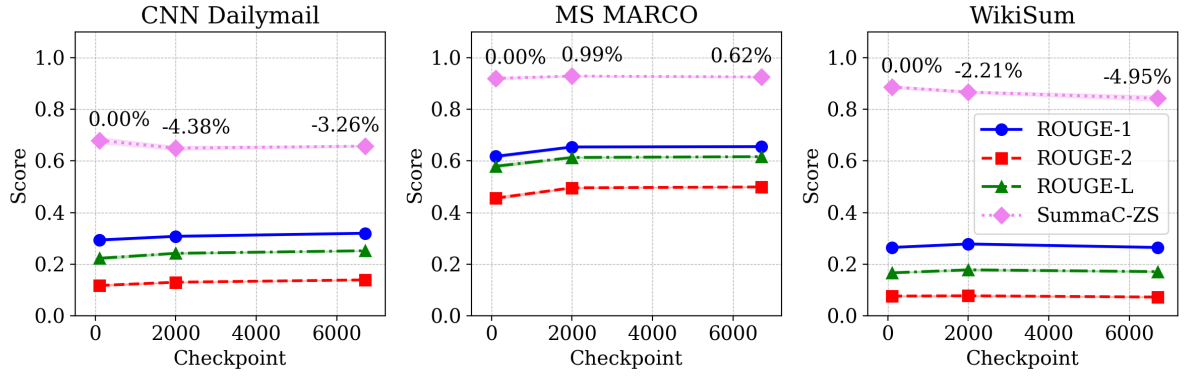


Figure 13: Individual faithfulness score and task performance with abstractive QA and summarization datasets across evaluated with three distinct model checkpoints of LLaMA-7B fine-tuned on context-dependent datasets. The middle checkpoint is the one with lowest in-training evaluation loss.

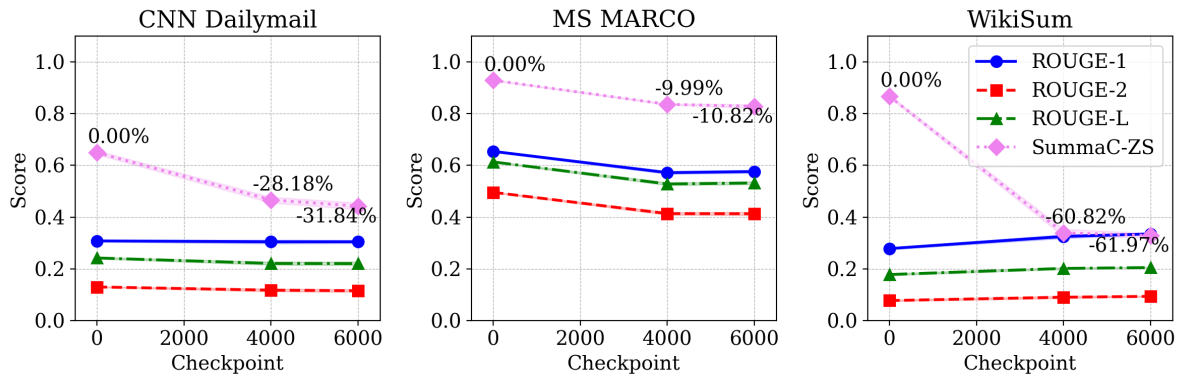


Figure 14: Individual faithfulness score and task performance with abstractive QA and summarization datasets across evaluated with three distinct model checkpoints through the instruction following fine-tuning. The middle checkpoint is the one with lowest in-training evaluation loss.

Testing Example from WikiSum Dataset	
Passage	Take your pencil and on the top of the page, about two inches from the right start drawing a line to 2 inches (5.1cm) below the right top corner. The line should loop down and have points. (see the picture) Draw straight lines from the points in your first line to the corner. Make lines parallel to your fist line going all the way up. You should have 5 or 6 lines. Get a paper and make a cross on it, try to make both lines a similar length (using a ruler will help) Draw diagonal lines through the centre, dividing the paper up from 4 to 8 sections. Make sure they are smaller than the cross you made before. Start connecting the lines with inverted arcs, this is an arc ), from the inside out. Once you have reached the end of the web, elongate the diagonal lines, (this will make it look like it has supports). Draw a spider by making a fuzzy ball, then drawing legs (eight of them) on your web. Or see the spider drawing tip. Finished. Draw a circle and draw a cross section which also extends outside the circle. Draw two diagonal lines at the midpoint of the cross-sections which form an X-mark. Draw squares which descend in size as it nears the centre point. Draw the corners or vertices of the square along the diagonal lines. Draw diamond shapes descending in size as it nears the centre point. Draw the vertices along the lines of the cross-section. Draw curves to connect the lines 2013 from the squares to the diamonds, much like forming bridges. Trace with a pen and erase unnecessary lines. You may add drawings for spiders. Color to your liking!
Golden Answer	
Response	To draw a spider web, start by drawing a cross with 2 lines of equal length. Next, draw diagonal lines through the center to divide your web into 4 to 8 equal sections. Then, connect the lines with inverted arcs, or curved lines, from the inside out. When you reach the end of the web, elongate the diagonal lines to make the web look like it's hanging from something, like a tree or shrub. Finally, draw a spider in your web by making a fuzzy ball with 8 legs.
Response from LLaMA-7B Fine-tuned on Context-Dependent Datasets	
Response	Draw a cross on a piece of paper. Draw a spider by making a fuzzy ball, then drawing legs (eight of them) on your web.
Response from LLaMA-7B Further Fine-tuned on Instruction Following Datasets	
Response	This is a simple spider web drawing tutorial. It is a simple and easy to follow tutorial. It is a great way to practice your drawing skills.

Table 7: WikiSum testing example from the **first** pipeline of our two-stage fine-tuning paradigm where we take LLaMA-7B and fine-tune it on context-dependent datasets first, and then further fine-tune it with human alignment datasets.



Testing Example from the Alpaca-15K Dataset	
Instruction	Construct a sentence that adequately describes a baby duck.
Vicuna-7B Zero-Shot	
Response	A baby duck is a fluffy, downy creature with a round body and webbed feet, often seen quacking and waddling around with its mother.
Examples from Vicuna-7B w/ Context Dependent Fine-tuning	
Response	A baby duck is a young duck that has not yet developed its full adult plumage.
Examples from Vicuna-7B w/ MTL Fine-tuning	
Response	A baby duck is a small, fluffy, and adorable bird that is often seen waddling around with its parents.
Examples from Vicuna-7B w/ RESET-S	
Response	A baby duck is a small, fluffy, and adorable bird that is often seen swimming in ponds or lakes.

Table 8: Qualitative examples from the **second** pipeline of our two-stage fine-tuning paradigm where we take Vicuna-7B and fine-tune it on context-dependent datasets with various fine-tuning methods.

<p>Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.</p> <p>### Instruction:  Answer to the question by extracting a specific text span from the given passages. Do not include new information beyond the given passages.</p> <p>### Input:  Question: a bond that the issuer has the right to pay off before its maturity date</p> <p>Passage: does well. Companies also reserve the right to call their bonds, which mean they can call it sooner than the maturity date. Often there is a clause in the contract that allows this; for example, if a bond issuer wishes to rebuy a 30-year bond at the 25th year, they must pay a premium. If a bond is called, it means that less interest is paid out. Failure to pay a bond effectively means bankruptcy. Bondholders who have not received their interest can throw an offending company into bankruptcy, or seize its assets if that is stipulated in the contract.</p> <p>Passage: Callable bond A callable bond (also called redeemable bond) is a type of bond (debt security) that allows the issuer of the bond to retain the privilege of redeeming the bond at some point before the bond reaches its date of maturity. In other words, on the call date(s), the issuer has the right, but not the obligation, to buy back the bonds from the bond holders at a defined call price. Technically speaking, the bonds are not really bought and held by the issuer but are instead cancelled immediately. The call price will usually exceed the par or issue</p> <p>Passage: options embedded. Callable bond A callable bond (also called redeemable bond) is a type of bond (debt security) that allows the issuer of the bond to retain the privilege of redeeming the bond at some point before the bond reaches its date of maturity. In other words, on the call date(s), the issuer has the right, but not the obligation, to buy back the bonds from the bond holders at a defined call price. Technically speaking, the bonds are not really bought and held by the issuer but are instead cancelled immediately. The call price will usually exceed the par</p> <p>Passage: the amount on which the issuer pays interest, and which, most commonly, has to be repaid at the end of the term. Some structured bonds can have a redemption amount which is different from the face amount and can be linked to the performance of particular assets. The issuer has to repay the nominal amount on the maturity date. As long as all due payments have been made, the issuer has no further obligations to the bond holders after the maturity date. The length of time until the maturity date is often referred to as the term or tenure or</p> <p>Passage: Bond (finance) In finance, a bond is an instrument of indebtedness of the bond issuer to the holders. The most common types of bonds include municipal bonds and corporate bonds. The bond is a debt security, under which the issuer owes the holders a debt and (depending on the terms of the bond) is obliged to pay them interest (the coupon) or to repay the principal at a later date, termed the maturity date. Interest is usually payable at fixed intervals (semiannual, annual, sometimes monthly). Very often the bond is negotiable, that is, the ownership of the instrument can be</p> <p>### Response:  <b>Model Generated Answer Goes Here</b></p>
---

Table 9: An example of NQ dataset with our instruction template for LLaMA-7B. Vicuna-7B’s template is slightly different as shown in Table 6.

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:

Summarize the text in a few sentences. Using original phrases or paraphrasing them if necessary. Do not include new information beyond the given passages.

### Input:

(Billboard)Considering the Academy of Country Music Awards celebrated its 50th anniversary on Sunday night at the Dallas Cowboys stadium, it was bound to be bigger than any previous year's ACMs. Plus, as hosts Blake Shelton and Luke Bryan were quick to point out, everything is bigger in Texas. Billboard: 2015 ACM Awards: See All the Photos . But bigger isn't always better. Here's our breakdown of the 10 best and 5 worst moments at the 2015 ACMs. The Best . Eric Church & Keith Urban provide an opening wallop . With a full stadium and millions of home viewers watching, two of country's leading men successfully lit the fuse for the 50th annual ACM Awards with a slick joint kickoff performance. Eric Church's Pledge Allegiance To The Hagwas a fine throwback, but Keith Urban's powerhouse Raise 'Em Up lived up to its title – and then some. Setting a Guinness World Record . Not only was this the ACMs biggest audience ever, but the 2015 ACM Awards brought in the biggest audience for a live TV awards show ever. Now that's how you celebrate half a century. Reba McEntire demonstrates how it's done . During a night that found a couple performers sounding a little weak in the vocal department, Reba McEntire showed the entire Cowboys stadium – and many younger singers who don't have half her energy – how it's done. Reba is eternal. Taylor Swift singing in the audience . Whether it was during Eric Church or Martina McBride, T-Swift was heating up the audience by singing and swaying along to country staples. Taylor might have gone pop for 1989, but she was deep in the heart of Texas for this annual country extravaganza. Billboard: 2015 ACM Awards: And the Winners Are... Garth Brooks' All-American salute . During Brooks' performance of All-American Kid, the country giant welcomed a slew of U.S. military representatives into AT&T Stadium, in an emotional moment that put the red, white and blue front and center. Brooks ended the performance by saluting the U.S. armed forces who are here and are all around the world for protecting our freedom, and also shouted out his home state of Oklahoma, who are grieving 20 years after the Oklahoma City bombing on Sunday. A truly classy moment from one of the best ever. Taylor Swift's mom makes an appearance . Sure, Andrea Swift's presentation of a special award for her daughter came with a fun origin tale about the creation of Love Story and some lovely words about the 25-year-old superstar. But the sight of the elder Swift – just days after Taylor confirmed that her mother had been diagnosed with cancer – walking to the podium was more than enough to yield one of the night's most poignant moments. Christina Aguilera joins Rascal Flatts . Aguilera is far from country, but with a voice as malleable as hers, she can pass for a song or two. After singing a bit of the tune she did while guesting on Nashville, XtiNa joined longtime hitmakers Rascal Flatts for Riot from their recent album Rewind. Aguilera and Gary LeVox trading vocals was the rare unexpected artist pairing that actually worked. Miranda Lambert domination . In addition to kicking ass during her Mama's Broken Heart/Little Red Wagon medley, Lambert justly owned the night when it came to awards. If there's one thing the country community loves more than Miranda Lambert, it's giving Miranda Lambert awards. Billboard: Watch Little Big Town Bring Provocative 'Girl Crush' to ACMs . Little Big Town don't back down . Despite some mild controversy over their song Girl Crush, Little Big Town brought the poignant ballad to the awards show, giving the ACMs one of its more melancholy moments. Jason Aldean comes on strong . For a night featuring some shaky vocals, Aldean brought his silky yet powerful country croon to the ACMs during a massive medley. It's hard to see an audience get weak in the knees when you're watching at home, but it's fair to assume that's what happened during his performance. The Worst . Tony Romo . The Dallas Cowboys QB was understandably a little stiff on the mic (athletes usually aren't the most charismatic public speakers) but the whole gag with Shelton asking Romo to toss Bryan a pass went on waaaaay too long. On the plus side, Bryan caught the pass. On the other hand, there was a tired play on words about balls. The length . Three hours for the 50th ACMs? Sure, why not. Three and a half hours? That's pushing it. They could have shaved off the last half hour by cutting a couple of the performers who only sang half a song, and shortened a few of the massive commercial breaks. All of the Milestone Awards . Taylor Swift was given an extended honor at this year's ACM Awards, but some of the Milestone Awards – especially those given to Reba McEntire, Kenny Chesney and George Strait – seemed rushed for the country giants they were saluting. It's understandable since 2015 is the 50th anniversary of the ACMs, but sometimes, less (recipients) is more. Steven Tyler's facial hair . See link. Twitter calling out Taylor Swift . Plenty of country fans went after T-Swizzle on Twitter, berating her for attending the ACMs after abandoning country music for pop. The truth is, Swift has just as many country classics under her belt as any other artist in her age range. She might have moved to pop, but don't underplay her importance to the genre that birthed her. Billboard. All Rights Reserved.

### Response:

**Model Generated Answer Goes Here**

Table 10: An example of CNN DailyMail dataset with our instruction template for LLaMA-7B. Vicuna-7B's template is slightly different as shown in Table 6.

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 with well-formed sentences. Paraphrase the context in the passages if necessary. Do not include new information beyond the given passages.

### Input:

Question: 2015 college half term holiday dates

Passage: UK School Half Term Dates 2015, 2016 and 2017. Research the latest school holidays and term dates for England, Scotland, Wales and Ireland. Welcome to Half Term Dates. The place to view School Holidays for the UK, Ireland, France and Australia. We list the latest published half term times on one handy website.

Passage: UK School Half Term Dates 2015, 2016 and 2017. Research the latest school holidays and term dates for England, Scotland, Wales and Ireland. Welcome to Half Term Dates. The place to view School Holidays for the UK, Ireland, France and Australia.

Passage: UK School Half Term Dates 2015, 2016 and 2017. Research the latest school holidays and term dates for England, Scotland, Wales and Ireland.

Passage: Find your child's school term, half term and holiday dates on your local council's website.

Passage: School term dates-guide. These are the school term and holiday dates for schools in Lambeth. The dates shown do not take account of the five professional development days when schools are closed to pupils, or any other changes.

Passage: Find your child's school term, half term and holiday dates on your local council's website. School term and holiday dates vary across the UK.

Passage: Irish School Holidays, Mid Term Dates. Research the official half term breaks and school holidays in Ireland. The Irish authorities do not enforce the same holidays as the UK. Noticeable differences being the lack of early June half term and a 9 week long summer holiday. Commencing end of June for most Irish schools.

Passage: School term dates-guide. These are the school term and holiday dates for schools in Lambeth. The dates shown do not take account of the five professional development days when schools are closed to pupils, or any other changes. Please check with your child's school for more detailed information.

Passage: School term dates and holidays 2014/15. Here are the school term dates and holidays for Sandwell's primary, secondary and special schools in 2014/15. Some academies and voluntary aided church schools may not follow this schedule. You are advised to check with these schools directly.

Passage: Holiday Dates for college students\*: 1 Autumn Half Term: 26 October 2015 to 30 October 2015 (teaching re-starts from 3/4 November 2015 due to staff P&D days 2013 check with your tutor). 2 Winter Break: 18 December 2015 to 5 January 2016 (teaching re-starts from 5/6 January 2016 due to staff P&D days 2013 check with your tutor).

### Response:

**Model Generated Answer Goes Here**

Table 11: An example of MS MARCO dataset with our instruction template for LLaMA-7B. Vicuna-7B's template is slightly different as shown in Table 6.



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 to the question by extracting a specific text span from the given passages. Do not include new information beyond the given passages.

### Input:

Question: abnormalities in which chromosomes were linked to the moyamoya disease?

Passage: moyamoya disease. We postulate that a protein encoded on chromosome 21 may be related to the pathogenesis of moyamoya disease. Although the neuronal substrate is abnormal in Down syndrome patients, recovery from hemiplegic stroke in patients with MM-DS is comparable to recovery in patients with primary moyamoya.

Passage: Moyamoya disease (MIM 252350) is characterized by stenosis or occlusion of the terminal portions of the bilateral internal carotid arteries and by abnormal vascular networks at the base of the brain. There is a high incidence of moyamoya disease in Asia, especially in Japan. Multifactorial inheritance is estimated with  $\lambda(s) > 40$ . Previous linkage studies have indicated that susceptibility loci for the disease are located on chromosomes 3p, 6q, and 17q. In the present study, we searched for loci linked to the disease in 12 Japanese families using 428 microsatellite markers and found significant evidence for linkage to 8q23 [maximum LOD score

Passage: We reported an autopsy case of Down's syndrome with moyamoya syndrome. A 30-year-old male with Down's syndrome suffered from a cerebral infarction and died of brain herniation. Cerebral angiography showed vascular abnormalities that were the same as moyamoya disease. Pathological findings revealed multiple stenosis of main trunk of the cerebral arteries. Pathologically, the stenosed vessels showed eccentric intimal thickness with cholesterol deposit, unlike moyamoya disease. There are only two previous reports of autopsied cases of Down's syndrome with moyamoya syndrome. We postulate that a protein encoded on chromosome 21 may be related to the pathogenesis of Down's syndrome with moyamoya

Passage: other). The karyotype was normal. No mutation in the RFN213 gene was found, and none of the HLA types linked to moyamoya disease or described in similar familial cases were identified. By describing these multisystemic associations, polycystic kidney disease for the second time, and intestinal malformation for the first time in the literature, our report expands the phenotypic variability of moyamoya syndrome. The coexistence of disparate malformations among close relatives suggests an underlying common genetic background predisposing to structural or physiological abnormalities in different tissues and organs.

Passage: OBJECTIVE: We report a detailed description of a family affected by a hereditary multisystem disorder associated with moyamoya syndrome. METHODS: In this family case report, we evaluated 9 members of the same family originating from Algeria. Investigations included neuroimaging, cardiology and ophthalmologic evaluation, hormonal testing, hemoglobin electrophoresis, chromosomal karyotyping, muscle biopsy for morphology, immunohistochemistry and enzyme assays, mtDNA mutation screening, and haplotype analysis of 2 loci previously linked to moyamoya, on chromosomes 10 (ACTA2) and 17. RESULTS: Five males related through a maternal lineage were affected, suggesting an X-linked inheritance. Four of them had symptomatic moyamoya syndrome with an onset of acute

### Response:

**Model Generated Answer Goes Here**

Table 12: An example of BioASQ dataset with our instruction template for LLaMA-7B. Vicuna-7B's template is slightly different as shown in Table 6.

<p>Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.</p> <p>### Instruction:</p> <p>Answer to the question by extracting a specific text span from the given passages. Do not include new information beyond the given passages.</p> <p>### Input:</p> <p>Question: 30 days in the hole if you can't name this old peter frampton band whose name refers to a forced apology</p> <p>Passage: CLASSIC YUMMY ROCKERS   30 Days in the Hole if you can't name this old Peter Frampton band whose name refers to a forced apology   Humble Pie. right:.</p> <p>Passage: 30 Days in the Hole by Humble Pie song meaning, lyric interpretation, video and ... Steve Marriot had said that Peter Frampton had heard the early stages of this ... of mine with the 'Dead End Kids' from the movie of the same name, Dead End; ... A Dirty room: one full of evidence that can get you busted; especially if you...</p> <p>Passage: 30 Days in the Hole is the seventh single by English rock group Humble Pie, from the band's ... The song refers to Borstal - Some seeds and dust, and you got Borstal= referring to Borstal Prison and its borstal ilk ... Humble Pie's Greatest Hits [Featuring Peter Frampton &amp; Steve Marriott]; Best of Humble Pie; Classics Volume...</p> <p>Passage: Apr 1, 2012 ... Perhaps it was my new loose and mellow attitude, perhaps I had ... As I wrote earlier this month, it is now illegal to sell an old piano with ... Needless to say this would devastate the antique industry and force a ..... 30 Days In The Hole ... This song of Humble Pie was recorded after Frampton left the band...</p> <p>Passage: The Irish town of Kerry lends its name to this colorful breed of pooch   A Kerry blue .... Peter Frampton band whose name refers to a forced apology   Humble Pie.</p> <p>### Response:</p> <p><b><u>Model Generated Answer Goes Here</u></b></p>
---

Table 13: An example of SearchQA dataset with our instruction template for LLaMA-7B. Vicuna-7B's template is slightly different as shown in Table 6.

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:

Summarize the text in a few sentences. Using original phrases or paraphrasing them if necessary. Do not include new information beyond the given passages.

### Input:

. Condition score the horse Condition scoring is a process in which you assess the amount of fat the horse has on it. Condition scoring requires that you look at and feel the horse's body and assess the amount of fat it is carrying in specific areas. With condition scoring, you can evaluate whether the horse is in ideal condition or not. Once you look at and feel an area, you will write down your assessment in a chart made specifically for condition scoring. Condition scoring can take some instruction and practice, so you may want to consult with your veterinarian for some guidance on the procedure. A horse needs some fat to get it through the winter but it shouldn't have so much fat on it that it becomes obese. Use a weight tape. A weight tape is a tool used to approximate a horse's weight. It is a measuring tape that is wrapped around a horse's back and chest and the measurement markings are in pounds or kilograms. Using a weight tape will not give you a completely accurate measurement. It is only an estimate. It is best used for assessing change over time. Have the horse weighed. If you are bringing your horse to a veterinary clinic or a center that has a horse scale, then you can have the horse actually weighed. However, this is not usually available to horse owners on a regular basis. Using a scale is the most accurate way to weigh a horse. Measure the horse's weight regularly. In order to get an accurate understanding of changes to the horse's weight over time, you will need to measure it on a regular basis. If you are very concerned about a horse's weight, this can be every week. If you think the horse is maintaining weight fine, then every couple of weeks should be fine. Use the same type of weight measurement every time you measure the horse. This is the only way to really assess changes in weight over time. Take the horse's winter coat into consideration. When a horse is in the cold during the winter, it can grow a thick winter coat. This can interfere with weight tape measurements and condition scoring, which is why it is important to put your hands on your horse as you take these measurements. This will enable you to feel for fat pockets and ribs beneath the horse's wooly coat. With this coat change in mind, try to be consistent with the pressure and placement of weight tape and how you feel the horse's body when condition scoring. If you focus on consistency, any changes you document will usually signal a real trend in weight change. Be sure to remove a horse's blanket daily to assess weight gain or loss. Record your horse's weight over time. To track the horse's weight, you will need to know what its weight was before the winter. Then, you will need to keep a record of its weight throughout the winter. Make sure to write down each weight along with the date in a journal or notebook. This will be helpful to share with your horse's veterinarian later on. Be sure that you measure the horse's weight the same way every time. For instance, if you use a horse tape to measure the weight in the Fall, then use it to make subsequent measurements. Get professional help. If you are unsure if your horse's weight loss is of concern or you are unsure how to help your horse gain weight, you should get some professional advice. Talk to your veterinarian about what and how much you should be feeding your horse to fatten it up. Your veterinarian will also be able to recommend further bloodwork and diagnostics if they suspect that something else, such as parasites or disease, may be causing your horse's weight loss. Weight loss, if at all, should be gradual. Contact your veterinarian immediately if you notice extreme changes in your horse's weight over a short period of time. Check your horse's manure. Your horse's manure can provide some helpful clues about their eating and drinking habits. If you notice anything different, then call your veterinarian. Some instances where you would want to call your horse's vet include stool that is: Runny or wetter than usual, such as diarrhea. Drier than usual and/or less frequent, which may indicate constipation. This could be caused by not being able to access water because it is frozen. Check your horse's water often in cold weather to make sure it is not frozen. A different texture or color than is normal for your horse. Identify the signs of an emergency. If your horse has lost too much weight it can be an emergency that requires immediate veterinary care. Assess your horse's vital signs. Make sure that it is breathing regularly and that it has a strong pulse. A horse with irregular breathing or a weak pulse, or that is weak and unable to move, is in desperate need of emergency veterinary care. If any of its vital signs are not strong, you should call a veterinarian immediately. Help your horse regain weight. Horses should eat 2 - 2.5% of their body weight in good quality bulky foods, such as grass or hay, during winter. For the average 1,000 pound horse, this would mean eating 25 pounds. Make sure that your horse is getting enough. Getting your horse to gain weight requires that you feed it more calories every day than you have been. Discuss making one of the following changes to your horse's diet with your veterinarian: Give the horse unrestricted access to hay around the clock. Change to a higher grade of hay. Add in a complete feed. Add 4 to 8 ounces of corn or cocosoya oil to a small amount of grain. Add beet pulp, sweet feed, or concentrates to your horse's diet.

### Response:

**Model Generated Answer Goes Here**

Table 14: An example of WikiSum dataset with our instruction template for LLaMA-7B. Vicuna-7B's template is slightly different as shown in Table 6.

<p>Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.</p> <p>### Instruction: Compile a list of ethical considerations for the development of AI systems.</p> <p>### Input:</p> <p>### Response: <b><u>Model Generated Answer Goes Here</u></b></p>
--

Table 15: An example of Alpaca-15K with our instruction template for LLaMA-7B. Vicuna-7B’s template is slightly different as shown in Table 6.