LIMIT: Less Is More for Instruction Tuning Across Evaluation Paradigms

Aditi Jha[1*]
aditijha@

Sam Havens[2]
sam.havens@

Jeremy Dohmann[2]
jeremy.dohmann@

Alex Trott[2]
alex.trott@

Jacob Portes[2]
jacob.portes@

[1]Princeton [2]MosaicML × Databricks
[1]@princeton.edu [2] @databricks.com

Abstract

Large Language Models are traditionally finetuned on large instruction datasets. However recent studies suggest that small, high-quality datasets can suffice for general purpose instruction following. This lack of consensus surrounding finetuning best practices is in part due to rapidly diverging approaches to LLM evaluation. In this study, we ask whether a small amount of diverse finetuning samples can improve performance on both traditional perplexity-based NLP benchmarks, and on open-ended, model-based evaluation. We finetune open-source MPT-7B and MPT-30B models on finetuning datasets of various sizes ranging from 1k to 60k samples. We find that subsets of 1k-6k instruction finetuning samples are sufficient to achieve good performance on both (1) traditional NLP benchmarks and (2) model-based evaluation. Finally, we show that mixing textbook-style and open-ended QA finetuning datasets optimizes performance on both evaluation paradigms.

1 Introduction

How should you finetune and evaluate a Large Language Model (LLM) for general purpose instruction following?

Within the span of a year, the consensus has ricocheted between the two extremes of "finetune on as much data as possible" and "only finetune on a small, high quality dataset." One end of the spectrum is exemplified by datasets such as FLANv2 [31], which contains more than 15 million examples of question-answer pairs extracted from a wide swath of traditional NLP datasets and organized into instruction templates for 1,836 tasks. On the opposite end of the spectrum is the recent LIMA ("less is more for alignment") study, which boldly claims that general purpose instruction following can be achieved by simply finetuning on 1,000 diverse, high quality question-answering pairs [62]. A flurry of contemporaneous studies have similarly claimed that this type of style "alignment" can be achieved with a small amount of high-quality samples [50, 11, 7, 21, 61].[2] We refer to the general observation that LLMs can be finetuned on a small number of samples as style "alignment" throughout the rest of the paper.

[*]Princeton University. Work done while at MosaicML and Databricks

[2]The LIMA authors frame their approach in terms of a "superficial alignment hypothesis," which vaguely posits that aligning an LLM to respond in a particular style can simply be achieved by finetuning on a few high-quality samples.

Workshop on Instruction Tuning and Instruction Following at NeurIPS 2023.

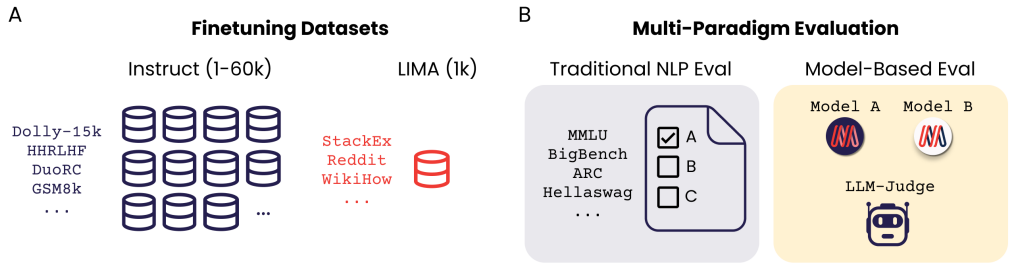

Figure 1: **How to finetune and evaluate LLMs for general purpose instruction following?** (A) We finetune open-source LLMs MPT-7B and MPT-30B on datasets of varying sizes: Instruct-v1 and v3 which contain 56.2-59.3k instruction samples, and the LIMA dataset which contains 1,000 samples. (B) We then evaluate finetuned models using two paradigms: (1) traditional NLP perplexity-based evaluation on benchmarks such as MMLU and BigBench, as well as (2) model-based evaluation (via GPT-4) on open-ended generation.

The lack of consensus surrounding LLM finetuning is due in part to the rapidly shifting paradigms for evaluating LLMs. Studies like FLAN-T5 [12] focus exclusively on classical NLP evaluation benchmarks such as MMLU [22] that contain short, trivia-like academic style questions and expect exact token match.[3] However, it is difficult to evaluate the quality and style of responses to more general, open-ended questions when using traditional perplexity-based NLP benchmarks.

With the advent of easily accessible high-quality LLMs like LLaMA [51] and chatGPT [41], it became possible to evaluate model quality by using another LLM as a judge [20, 61, 30]. In the model-based evaluation paradigm, an LLM is prompted with an instruction and is asked to judge a pair of corresponding responses. The style alignment approach of LIMA, Alpaca and others works surprisingly well with this paradigm. While using GPT-4 [40] as a judge has become wildly popular, there is a healthy skepticism around this trend [55]. In particular, there is evidence that style alignment with a small amount of high quality samples doesn't perform well on traditional NLP benchmarks [55, 21].

In this study, we take the style alignment approach from LIMA seriously, and ask whether a small amount of high quality finetuning samples can improve performance on both traditional perplexity-based NLP benchmarks, and on open-ended, model-based evaluation. While we find that there is a fundamental tension between perplexity-based and LLM-based evaluation paradigms, we show that careful construction of finetuning datasets can result in good performance on both paradigms (Fig. 1). While much of the field has drifted into ChatGPT phenomenology, we believe that good science requires reproducibility. We therefore chose to focus on two open-source models from the MosaicML MPT family; MPT-7B [37] and MPT-30B [36], as well as their corresponding finetuning datasets Instruct-v1 (59.3k samples) [35][4] and Instruct-v3 (56.2k samples) [39].[5] We evaluate model performance using (1) MosaicML's efficient, open-source Eval Gauntlet [34, 38], which is based on traditional NLP benchmarks such as MMLU [22] and BigBench [49], as well as (2) AlpacaEval's suite for model-based evaluation using GPT-4 as the judge [28].

We first asked whether LIMA's 1,000 sample dataset could deliver on its promise; could we simply finetune a base MPT model on LIMA and get optimal performance on traditional NLP benchmarks and also do well when evaluated by an LLM? We finetuned MPT-7B and MPT-30B base on LIMA, and evaluated the resulting models using MosaicML's Eval Gauntlet as well as AlpacaEval's model-based evaluation suite with GPT-4 as the "judge." While the resulting models were judged favorably

[3]For example, MMLU tests "world knowledge" with questions like "The famous statement 'An unexamined life is not worth living' is attributed to..." See Fig. 2 and Appendix E.4 for more examples.

[4]MPT-7B was trained on Instruct-v1. This instruction finetuning dataset is also referred to as the "Dolly-HHRLHF" and is available on HuggingFace here: `https://huggingface.co/datasets/mosaicml/dolly_hhrlhf`

[5]MPT-30B-Instruct was trained on the instruct-v3 dataset, which is available on HuggingFace here: `https://huggingface.co/datasets/mosaicml/instruct-v3`

by GPT-4, we found that they did not perform on par with MPT-7B and MPT-30B trained on much larger instruction finetuning datasets (Instruct-v1 and Instruct-v3, respectively).

We suspected that the LIMA dataset was slightly out-of-domain with respect to MMLU and BigBench, and asked whether a random subset of 1,000-6,000 "in-domain" samples from Instruct-v1 and Instruct-v3 could reach parity on the Eval Gauntlet with the full datasets. We were pleasantly surprised to find that this small subset of finetuning data indeed had similar performance on the Eval Gauntlet, corroborating the general small-sample approach of LIMA. However, these same models did poorly when evaluated by GPT-4.

We finally asked if we could get the best of both worlds—get good performance on both evaluation paradigms—by finetuning on a subset of a few thousand Instruct *and* LIMA samples. We found that this indeed led to good performance across both paradigms. While there was some initial scepticism that effective finetuning could be achieved with less than 1,000 samples, our results replicate LIMA [62] and build on the "less is more" approach to style alignment.

The contributions of this paper are as follows[6]:

- We finetune open-source models MPT-7B and MPT-30B on instruction datasets of various sizes and styles, including LIMA [62].
- We evaluate the finetuned models using two separate but widely popular paradigms: (1) the traditional NLP Benchmark approach using MosaicML's Eval Gauntlet (which consists of large mega-benchmarks like MMLU [22] and Big-Bench [49]), and (2) the model-based evaluation approach with GPT-4 as the judge between two model responses [28].
- When "judged" by GPT-4, models finetuned exclusively on 1,000 LIMA samples do better than models trained on Instruct datasets.
- Finetuning on a subset of 1,000 samples from a 56.2k finetuning training set can lead to the same performance as finetuning on the full dataset.
- When evaluated using standard NLP benchmarks such as MMLU, models finetuned exclusively on 1,000 LIMA samples do worse than models trained on Instruct datasets.
- Combining 1,000 LIMA samples and 1,000 Instruct samples leads to improved performance on both the traditional benchmarks and the model-based evaluation paradigm.

2 Related Work

Good Old Fashioned Finetuning: The "more is better" approach to finetuning was the culmination of an extensive amount of research with instruction datasets such as Natural Questions [26], P3 [48], and FLANv2 [31]. Finetuning LLMs on these datasets consistently leads to improvements on benchmarks such as Hellaswag [60], ARC [14], SuperGLUE [54] and mega-benchmarks such as MMLU [22] and BigBench [49] .

Imitation Learning: The shift away from finetuning on larger and larger datasets was catalyzed by the open-source release of the LLaMA models [51] and by the closed source launch of GPT-3 and chatGPT [40]. The field quickly realized that open-source LLMs such as LLaMA-7B could be effectively finetuned on high-quality instruction following data generated by state of the art GPT models.

The Alpaca model, for example, is a 7 billion parameter LLaMa model finetuned on 56,000 examples of question-response samples generated by GPT03 (text-davinci-003) [50]. The authors of this study found that Alpaca responded in a similar style to the much larger GPT-3 model. While the methods used in the Alpaca study were tenuous (the human preference evaluation was done by the 5 authors themselves), further finetuning studies such as Vicuna [61] , Guanaco [19], MPT-7B-chat [37], Tülü [55] , Baize [58], Falcon-40B [1] arrived at similar conclusions. Unfortunately, results across many of these papers use different evaluation paradigms and are difficult to compare side-by-side.

A few studies however have begun to fill out the full picture. A rigorous study by Wang et al. [55] argues that finetuned LLMs should be tested using traditional fact-recall capabilities with benchmarks like MMLU along with model-based evaluation (using GPT-4) and crowd-sourced human evaluation. They also find that finetuning LLaMA-13B over different datasets promotes specific skills, and that no one dataset improves performance over all evaluations paradigms.

[6]Project website: `https://97aditi.github.io/LIMIT/`

In "The false promise of imitating proprietary LLMs," Gudibande et al. [21] find that while finetuning small LLMs with "imitation" data derived from ChatGPT conversations can improve conversational style, it does *not* lead to improved performance on traditional fact-based benchmarks like MMLU and Natural Questions. However, they did note that training on GPT-4-derived "imitation" data *in the domain of Natural-Questions-like queries* improves performance on the Natural Questions benchmark. Finally, Chen et al. [7] and AlShikh et al. [2] directly address the question of *how many* finetuning examples are necessary for good downstream performance .[7] Our results align nicely with the above studies.

Traditional NLP Evaluation: Over the past few years, several large benchmarks have been created to evaluate LLMs. These benchmarks evaluate LLMs on traditional NLP tasks such as question answering, commonsense reasoning, knowledge acquisition, etc. MMLU [22] is one such benchmark that contains multiple-choice questions over a range of domains including computer science, elementary mathematics, history and more. Another benchmark, BigBench [49], contains a wide range of tasks such as logical deductions, conlang translation, arithmetic, understanding fables, etc. Some other benchmarks such as HellaSwag [60] and ARC [15] test specific abilities such as language understanding and commonsense reasoning.

Model-Based Evaluation: Benchmark-based evaluations do not measure certain attributes of instruction-following LLMs that are crucial for their real-world utility, such as coherence, conciseness, and relevance of a model's response to the prompt. While human annotation is considered the gold-standard for evaluating finetuned models (although see [24]), it is time-consuming and expensive to obtain human annotator ratings. Hence, using high-quality large language models such as GPT-4 [40] as a judge is becoming commonplace in the literature [62, 30, 61, 20]. Although there are limitations of performing model-based evaluations [61], several recent works have shown strong correlations between evaluations performed by human annotators and that by GPT-4 [30, 61, 20].

3 Finetuning Open-Source Large Language Models on Instruction Datasets

In order to probe the effects of dataset size, quality and composition on finetuning, we finetuned the open-source MosaicML MPT-7B and 30B pretrained models on a variety of instruction-following datasets.

MPT-7B is a decoder-style transformer trained on 1T tokens, with a context length of 2048 tokens. MPT-30B is also a decoder-only transformer model, and was trained on 1.05T tokens with an 8k context length during training. Both models incorporate ALiBi [43] and use the GPT-NeoX tokenizer [6]. More details on MPT pretraining are found in [37] and [36].

3.1 Instruction Datasets

We used three publicly-available finetuning datasets. The LIMA training and test sets have high quality samples of open ended questions and multi-paragraph answers written in the tone of a general purpose AI assistant. The MPT Instruct-v1, and MPT Instruct-v3 training and test sets contain trivia-like questions and answers that tend to be shorter than one paragraph. We explore the differences between these datasets in the rest of this paper.

We formatted all datasets so that the instructions follow the format of the Alpaca dataset (see Appendix E.1). We describe each of the three datasets in detail below, and show a few examples in Fig. 2A, B.

LIMA Dataset: The LIMA training set [62] contains 1,000 samples (750,000 tokens) curated from Reddit, Stack Overflow, wikiHow, Super-Natural-Instructions [56], and examples manually written by the paper authors. The examples were selected after strict filtering to ensure all responses were high-quality and diverse. For example, the authors sampled an equal number of prompts from various categories within StackExchange (programming, math, English, cooking, etc) and selected the top answer for each prompt, which further went through additional filtering based on length and writing style. In this study, we only used the single-turn examples.

[7] Chen et al. [7] found that only 9k out of the full 56k finetuning examples from the Alpaca dataset are sufficient to reach similar performance. AlShikh et al. [2] find that LLaMA-7B and 13B models transition to becoming fully "instruction following" after around 5k finetuning examples.

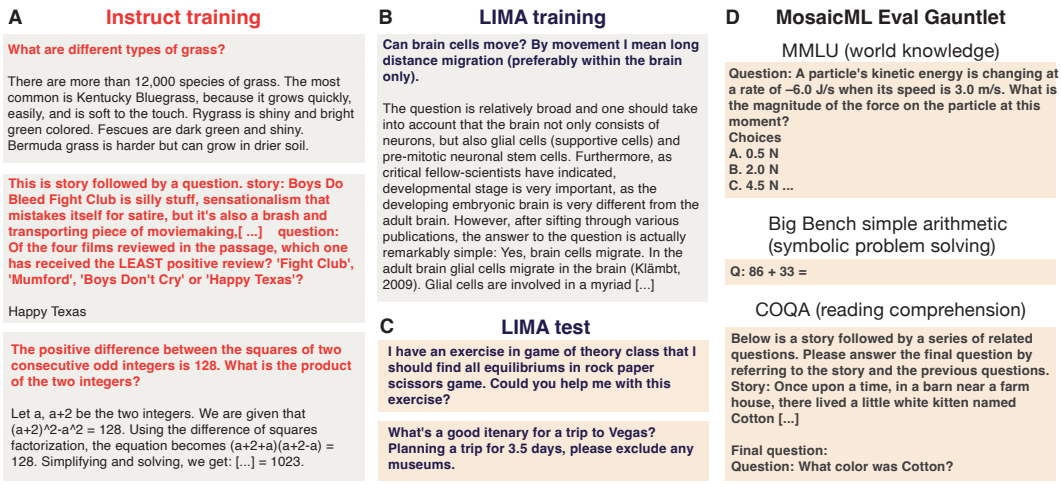

Figure 2: **Instruction finetuning training and test examples** from the (A) Instruct-v1 (derived from Dolly-15k, HH-RLHF) and Instruct-v3 (derived from 9 diverse sources) training sets (B) LIMA training set, which contains open ended questions and multi-paragraph answers (C) LIMA test set (which similarly contains open ended questions) (D) MosaicML Eval Gauntlet test set, which contains trivia-like multiple choice questions.

MPT Instruct-v1 Dataset (a.k.a "Dolly-HHRLHF"): This training set was used to train the MPT-7B-Instruct model [37].[8] The MPT Instruct-v1 dataset contains the Databricks Dolly-15k dataset [17] and a curated subset of Anthropic's Helpful and Harmless (HH-RLHF) datasets [4],[9] both of which are open source and commercially licensed. MosaicML's MPT-7B-Instruct model was finetuned using this dataset [37]. It contains 59.3k examples, where 15k are derived from Dolly-15k and the rest are from Anthropic's HH-RLHF dataset. Dolly-15k contains several classes of prompts including classification, closed-book question answering, generation, information extraction, open QA, and summarization. Anthropic's HH-RLHF dataset contains crowd-sourced conversations of workers with Anthropic's LLMs. Only the first turn of multi-turn conversations was used, and chosen samples were restricted to be helpful and instruction-following in nature (as opposed to harmful).

MPT Instruct-v3 Dataset: This training set was used to train the MPT-30B-Instruct model [36].[10] It contains a filtered subset of MPT Instruct-v1, as well as several other publicly available datasets: Competition Math [23], DuoRC [46], CoT GSM8k [16], Qasper [18], SQuALITY [53],[11] Summ Screen FD [9] and Spider [59]. As a result, Instruct-v3 has a large number of reading comprehension examples, where the instructions contain a long passage of text followed by questions related to the text (derived from DuoRC, Qasper, Summ Screen FD, SQuALITY). It also contains math problems derived from CompetitionMath and CoT GSM8K, as well as text to SQL prompts derived from Spider. Both Instruct-v1 and Instruct-v3 were designed with the implicit goal of improving performance on traditional NLP benchmarks.

3.2 Finetuning Hyperparameters

We finetuned MPT-7B with the decoupled AdamW optimizer [32] for 10 epochs on 8×A100s, and then chose the best checkpoint based on their performance on the MosaicML Eval Gauntlet (typically the 4th epoch). Our learning rate schedule used a warm up duration of 50 batches, followed by a linear decay. The peak learning rate was set to $5e^{-6}$, the optimizer parameters to $\beta_1 = 0.9$, $\beta_2 = 0.99$, and we did not use any weight decay.

[8] `https://huggingface.co/mosaicml/mpt-7b-instruct`

[9] The full dataset can be found here: `https://huggingface.co/datasets/Anthropic/hh-rlhf`

[10] `https://huggingface.co/mosaicml/mpt-30b-instruct`

[11] They use a version of SQuALITY formatted for QA, which can be found here: `https://huggingface.co/datasets/emozilla/quality`

We finetuned MPT-30B with the decoupled LionW [10] optimizer for 10 epochs on 8×A100s and chose the best checkpoint based on the Eval Gauntlet (typically th 6th epoch). The optimizer choice was driven by memory constraints, as decoupled LionW uses less memory relative to Adam. We chose the same learning rate schedule as in the case of 7B models. The peak learning rate was set to $1e^{-6}$, optimizer parameters to $\beta_1 = 0.9$, $\beta_2 = 0.95$ and weight decay to 0.[12]

4 Two Paradigms for Evaluating Instruction Finetuned Models

We evaluated the finetuned models on a suite of canonical NLP benchmarks, as well as on their ability to perform open-ended language generation. We describe our evaluation strategy in detail below.

4.1 In-Context Learning Evaluation with an Eval Gauntlet

In context-learning (ICL) tasks are commonly used to evaluate large language models. They usually test a model's ability to perform sentence completion and fact-based question answering. We used the MosaicML Eval Gauntlet to evaluate our finetuned models—it is open-source and contains a wide range of benchmarks. It encompasses 34 different benchmarks collected from a variety of sources including MMLU [22] and Big Bench [49], and is organized into 5 broad categories of competency. [13] The Eval Gauntlet is optimized for speed, and scales linearly with GPU count; this was one motivation for using this over other open-source evaluation harnesses [34].

The gauntlet is divided into the following five categories: (1) "World Knowledge" tasks which evaluates factual knowledge, (2) "Commonsense Reasoning" loosely assesses a model's ability to do basic reasoning tasks, (3) "Language understanding" tasks evaluate the model's ability to understand the structure and properties of languages, (4) "Symbolic problem solving" tasks test the model's ability to solve a diverse range of symbolic tasks, (5) "Reading comprehension" benchmarks test a model's ability to answer questions based on the information in a passage of text. We describe these categories in detail in the Appendix D, and show some samples in Fig. 2D.

The Eval Gauntlet evaluates the model on all the benchmarks and averages the subscores within each category. Some benchmarks are multiple choice questions, for which it is possible to get above 0% accuracy with just random guessing. In order to ensure that the composite scores are less than or equal to 1, the gauntlet subtracts the random baseline accuracy and rescales the remainder by 1 minus the baseline accuracy. For example, if benchmark A has a random baseline accuracy of 25%, and the model achieved 30%, we would report this as (0.3 - 0.25)/(1-0.25) = 0.0667. This can be thought of as the accuracy above chance rescaled so that the max is 1. We report scores on individual categories, as well as the average across all 5 categories.

4.2 Model-Based Evaluation with GPT-4

In addition to in-context learning evaluations, we also tested the open-ended text generation ability of our models using model-based evaluations in combination with the LIMA test set. This test set was released by [62], and contains 300 prompts (70 from Reddit and 230 self-curated by the paper authors, see Fig. 2C).

We used the model-based evaluation pipeline developed by Li et al. [28]. This framework, called AlpacaEval, allows for pair-wise evaluation where the model acting as the judge sees an instruction, and two responses corresponding to the same instruction (see Appendix F for details). The judge model then ranks the two responses; it throws away any examples where both the responses are of the same quality (we find that this rarely happens in our setting, only 1-5 prompts out of 300 are discarded across all our evaluations). AlpacaEval also randomizes the position of responses for every prompt to avoid any position biases of the judge model [61, 20].

We used GPT-4 as the judge model, and the LIMA [62] test set of 300 prompts as our evaluation set. We generated three different responses (by setting a distinct seed for sampling during generation) for every prompt in the LIMA test set, using our finetuned models. Prompt-response pairs were restricted

[12] We performed a learning rate sweep in the range of $5e^{-7}$ to $5e^{-4}$ and chose the best learning rate based on a small subset of tasks in the Eval Gauntlet.

[13] The current MosaicML Eval Gauntlet also has programming as a sixth category, which we did not use for our evaluations.

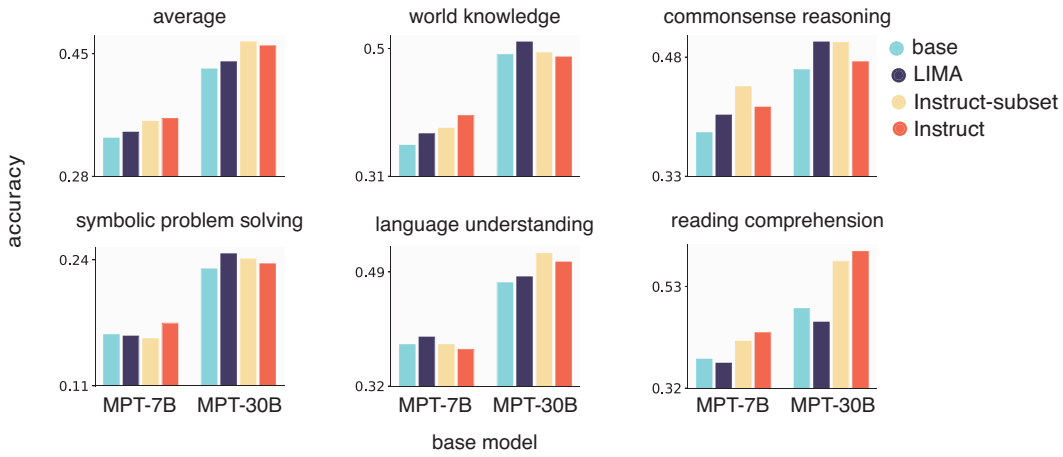

Figure 3: **Models finetuned on the Instruct training sets do better on traditional NLP benchmarks**. Each plot shows the accuracy (between 0–1) of models on a given category of the MosaicML Eval Gauntlet, and the average score across all categories is shown in the first subplot. The two different model sizes (7B and 30B) are grouped into two bar graphs. We show results for the base models MPT-7B and MPT-30B (cyan), and for the base models finetuned on the LIMA dataset (midnight blue), subsets of the Instruct dataset (khaki), and the full Instruct dataset (vermilion).

to a maximum of 2048 tokens, and responses were generated with a decoding temperature of 0.9. We then used GPT-4 and the AlpacaEval pipeline to rank pairs of models based on their responses on the LIMA test set. We report the number of times one model is preferred over the other (excluding ties), averaged over three pairwise comparisons.

5 Results: Finetuning Dataset Composition, not Size, Determines Model Performance

We detail the performance of our finetuned models on both evaluation paradigms, along with the implications of our findings below.

5.1 Models finetuned on Instruct datasets performed best on the Eval Gauntlet

We finetuned MPT-7B using LIMA and Instruct-v1, and refer to these models as MPT-7B-LIMA and MPT-7B-Instruct for conciseness. We then evaluated them using the MosaicML Eval Gauntlet. We verified that the finetuned models, MPT-7B-LIMA and MPT-7B-Instruct, performed better than the base model MPT-7B on the Eval Gauntlet (Fig. 3). Next, we found that MPT-7B-LIMA performed worse than MPT-7B-Instruct according to the Eval Gauntlet (Fig. 3). Specifically, MPT-7B-LIMA lagged behind MPT-7B-Instruct on world knowledge, symbolic problem solving and reading comprehension. As described in Sec. 3.1, the Instruct-v1 dataset contains a diverse set of NLP tasks (closed-book question answering, information extraction, summarization, etc.), which likely contributed to the capabilities being tested by the Eval Gauntlet. In contrast, the LIMA dataset was designed to mimic a "helpful AI assistant" [62].

We repeated the same analyses for models finetuned with MPT-30B as the base model. Once again, we found that while MPT-30B-LIMA (MPT-30B finetuned with the LIMA dataset) performed better on the Eval Gauntlet when compared to the base model MPT-30B, it under performed relative to MPT-30B-Instruct (MPT-30B finetuned on the Instruct-v3 dataset) on several categories (Fig. 3). In particular, MPT-30B-LIMA lagged behind MPT-30B-Instruct on the reading comprehension tasks. The Instruct-v3 dataset, similar to Instruct-v1, contained several traditional NLP tasks which could explain the good performance of MPT-30B-Instruct on the Eval Gauntlet. In particular, Instruct-v3 was derived from a large number of reading comprehension style datasets (as described in Sec. 3.1) explaining the large difference in accuracy between MPT-30B-LIMA and MPT-30B-Instruct on

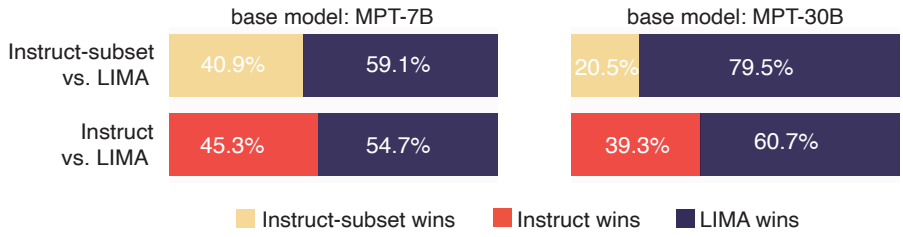

Figure 4: **Model-based evaluation on the LIMA test set prefers models finetuned on the LIMA training set**. Model-based evaluation on the LIMA test set (300 samples) is done using GPT-4 as the judge. We show the preference rate of MPT models finetuned on a subset of Instruct and on the full Instruct datasets, when compared to LIMA-finetuned MPT models. (Left) GPT-4 prefers responses from MPT-7B finetuned on 1,000 LIMA samples over responses from MPT-7B finetuned on a random subset of 5,000 samples from Instruct-v1. (Right) GPT-4 strongly prefers responses from MPT-30B finetuned on LIMA samples over responses from MPT-30B finetuned on (1) a random subset of 1,000 samples from Instruct-v3, and (2) the full 56,200 samples in Instruct-v3.

reading comprehension tasks. Hence, we concluded that finetuning a model with examples from the domain of NLP tasks can enhance it's overall performance on canonical NLP benchmarks.

We next asked if the difference in performance across MPT-7B-LIMA and MPT-7B-Instruct was a function of dataset size, or that of the set of tasks included in the respective datasets alone. We created control datasets containing 1k-10k examples randomly selected from Instruct-v1, and finetuned MPT-7B on these subsets of Instruct-v1. Intriguingly, we found that MPT-7B when finetuned on 5k samples from Instruct-v1 (MPT-7B-Instruct-Subset[14]) achieved the same performance on our Eval Gauntlet as MPT-7B-Instruct. Similarly, we observed that a model finetuned on a small subset of Instruct-v3 (only 1,000 samples; called MPT-30B-Instruct-Subset[15]) performed at par with MPT-30B-Instruct which was finetuned using 56.2k samples (Fig. 3).

Thus, we concluded that that the diversity of tasks contained in Instruct were the main drivers for the finetuned model's performance on the gauntlet, as opposed to the dataset size. This has important implications for finetuning large models—a small dataset suffices for finetuning as long as it contains examples in-domain to the downstream task it is being evaluated on. Additionally, the subsets of Instruct dataset that we finetuned our models on were selected at random, which makes this even easier to implement in practice.

5.2 LIMA-finetuned models are preferred by GPT-4

We then evaluated the open-ended generation ability of the finetuned models using GPT-4 as a judge on 300 LIMA test set prompts, as discussed in Sec. 4.2. Unsurprisingly, we found that MPT-7B-Instruct-Subset and MPT-7B-Instruct were only preferred 40.9% and 45.3% of times respectively over MPT-7B-LIMA (as shown in Fig. 4), despite performing better on the Eval Gauntlet. We observed more pronounced trends during open-ended generation evaluation for the 30B models. MPT-30B-LIMA outperformed MPT-30B-Instruct, being preferred 60.7% of the time. It also significantly outperformed MPT-30B-Instruct-Subset, with a preference rate of 79.5%. These results highlight the dichotomy between the two commonly used evaluation paradigms.

To dig deeper, we looked at example responses generated by our models (see Appendix G). We found that responses from the LIMA models typically tend to be in bulleted format, and more verbose in nature compared to responses from the Instruct models. This finding corroborates the results of Zhou et al. [62], that 1,000 LIMA samples are effective at aligning a model to a particular style. Hence, while the Instruct datasets are useful for performing well on canonical NLP tasks, LIMA imparts a favorable style to a model's responses when answering open-ended questions. This indicated to us that the two evaluation metrics that we considered potentially measure distinct model capabilities,

[14]This dataset can be found here: `https://huggingface.co/datasets/aditijha/instruct_v1_5k`
[15]`https://huggingface.co/datasets/aditijha/instruct_v3_subset`.

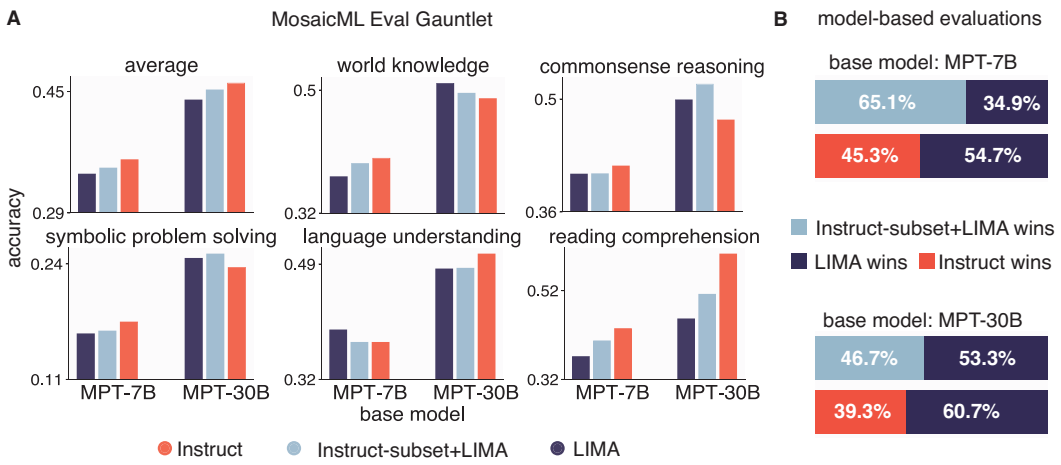

Figure 5: **Models finetuned on the LIMA training set *and* a subset of the Instruct training set perform well across both evaluation paradigms.** (A) Accuracy of finetuned models on each category of the MosaicML Eval Gauntlet, along with their average scores. MPT-7B and MPT-30B when finetuned on a subset of the Instruct datasets (5k samples from Instruct-v1 for 7B, 1k samples from Instruct-v3 for 30B) combined with the LIMA dataset perform very close to MPT-7B and MPT-30B finetuned on all of Instruct, respectively. (B) Model-based evaluation on the LIMA test set using GPT-4. (Top) MPT-7B finetuned on the combined dataset is preferred over MPT-7B finetuned with LIMA alone by a huge margin. (Bottom) MPT-30B finetuned on the combined dataset is preferred 46.7% over MPT-30B finetuned on LIMA. In both cases, the preference rate of models finetuned on the combined dataset is higher than those finetuned on all of the Instruct datasets.

and that different datasets confer distinct capabilities to the corresponding finetuned model based on their composition.

This model-based evaluation used 300 samples from the LIMA test, which are in domain with the LIMA training dataset by design. We also applied model-based evaluation using 300 samples from the Instruct-v3 test set, and found that GPT-4 *still* prefers models finetuned on LIMA (Supplementary Figure 6).

Overall, our results highlight the importance of dataset composition as being a strong indicator of a model's downstream performance. The LIMA dataset contains a large number of examples from question answering forums on the internet, and was filtered extensively for quality of responses. Thus, model-based evaluations which test open-ended generation preferred models finetuned on LIMA. However, the Eval Gauntlet preferred models finetuned on Instruct-style datasets as they contain a range of NLP tasks. Finally, we also observed that small fine-tuning datasets were sufficient to obtain high accuracy on both the Eval Gauntlet (Instruct-Subset models), and on open-ended generation (LIMA models). Together, these results suggest that practitioners can quickly a finetune a model that performs well on their domain of interest by creating a small dataset that captures a range tasks within that domain.

5.3 Best of both worlds: Combining Instruct-subset and LIMA datasets improves performance across both evaluation paradigms

Next, we naturally asked if it was possible to obtain a model that performed well on both the evaluation criteria that we considered. Our findings above suggested to us that a small dataset with in-domain samples from NLP tasks and open-ended generation should perform well on both the Eval Gauntlet as well as using model-based evaluations. Hence, we combined subsets of the Instruct datasets with the LIMA dataset, and finetuned models on this mixture.

Specifically, for MPT-7B, we combined the LIMA dataset with the same 5k samples of Instruct-v1 that were used to finetune MPT-7B-Instruct-Subset. Similarly, for MPT-30B, we combined LIMA with the 1k samples of Instruct-v3 that were used to finetune MPT-30B-Instruct-Subset. Indeed, we

found that the resulting models either performed at par with or even better than the models finetuned on the individual datasets, on both evaluation paradigms.

As shown in Fig. 5A, the average performance of models finetuned on the mixture was very close to that of the models finetuned on the Instruct datasets, for both model sizes. When evaluated on open-ended generation, the 7B parameter model finetuned on this mixture was preferred 65.1% of time over the LIMA-finetuned models by GPT-4 (Fig. 5B), while the 30B model finetuned on the mixture dataset was preferred 46.7% of the times. In both cases, the mixture datasets resulted in huge improvements over the Instruct-finetuned models during model-based evaluations.

Hence, this further validated our claim that while small datasets are sufficient for finetuning, including in-domain examples with respect to the downstream evaluation criteria is crucial for good performance. While Zhou et al. [62] curated LIMA through a rigorous data filtering process as well as by writing prompts themselves, recent works have shown that it is possible to obtain a high-quality dataset using automated filtering of larger datasets [7].

6 Discussion

While the models and datasets we used are open-source, our model-based evaluation paradigm relied on GPT-4, which is a closed-source model. Our choice was driven by the wealth of work showing that GPT-4 as a judge correlates well with human judgments [28, 20], however it does limit the reproducibility of our results, as GPT models hosted by OpenAI change over time [8].

Our model-based evaluation paradigm also has several limitations. We used a prompting style from the AlpacaEval framework (see Appendix F), which did not allow for ties. As a result, GPT-4 was forced to choose one of the two responses corresponding to a given instruction, even when both the responses could be of the same quality. Additionally, recent work [61] has observed that GPT-4 prefers responses that are stylistically similar to itself, such as longer responses written in a confident tone, and responses with bullet points, etc.

When using both evaluation paradigms, we found it difficult to quantify the variance of the results. This requires finetuning models multiple times on the same dataset resulting in high compute requirements. Supplementary Figures 7 and 8 somewhat address this by comparing the MosaicML Eval Gauntlet results for MPT models finetuned on *both* Instruct-v1 and Instruct-v3. Interestingly, the average performance of these models is similar despite the variance across Eval Gauntlet subdomains.

Furthermore, it is entirely possible that GPT-4 simply prefers more verbose answers. In this study, we did not analyze average response length for "preferred" vs. "non-preferred" answers.

7 Conclusion

In this work, we test if a small dataset is sufficient to finetune LLMs for general purpose instruction following. We found that while a small dataset was indeed sufficient, the composition of the dataset really mattered. When evaluated using traditional NLP benchmarks, models finetuned on the Instruct datasets consistently did better than those finetuned on the LIMA dataset. This trend held up even when we randomly selected only 1-5k samples from the Instruct dataset. However, when evaluated on open-ended generation, GPT-4 preferred models finetuned on the LIMA dataset. Finally, when we combined 1-5k samples from the Instruct dataset with LIMA the resulting finetuned models performed well on both our evaluation paradigms, thus validating our findings. The implications of our results are compelling: (1) finetuning can be faster and cheaper (2) individual ML practitioners can easily align models to specific use cases with a handful of curated high-quality samples.

References

[1] Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Goffinet, E., Heslow, D., Launay, J., Malartic, Q., Noune, B., Pannier, B., and Penedo, G. (2023). Falcon-40B: an open large language model with state-of-the-art performance. `https://huggingface.co/tiiuae/falcon-40b`.

[2] AlShikh, W., Daaboul, M., Goddard, K., Imel, B., Kamble, K., Kulkarni, P., and Russak, M. (2023). Becoming self-instruct: introducing early stopping criteria for minimal instruct tuning. *arXiv preprint arXiv:2307.03692*.

[3] Amini, A., Gabriel, S., Lin, S., Koncel-Kedziorski, R., Choi, Y., and Hajishirzi, H. (2019). Mathqa: Towards interpretable math word problem solving with operation-based formalisms. *CoRR*, abs/1905.13319.

[4] Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

[5] Bisk, Y., Zellers, R., Gao, J., Choi, Y., et al. (2020). Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

[6] Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., Phang, J., et al. (2022). Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745*.

[7] Chen, L., Li, S., Yan, J., Wang, H., Gunaratna, K., Yadav, V., Tang, Z., Srinivasan, V., Zhou, T., Huang, H., et al. (2023a). Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*.

[8] Chen, L., Zaharia, M., and Zou, J. (2023b). How is chatgpt's behavior changing over time? *arXiv preprint arXiv:2307.09009*.

[9] Chen, M., Chu, Z., Wiseman, S., and Gimpel, K. (2022). SummScreen: A dataset for abstractive screenplay summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8602–8615, Dublin, Ireland. Association for Computational Linguistics.

[10] Chen, X., Liang, C., Huang, D., Real, E., Wang, K., Liu, Y., Pham, H., Dong, X., Luong, T., Hsieh, C.-J., et al. (2023c). Symbolic discovery of optimization algorithms. *arXiv preprint arXiv:2302.06675*.

[11] Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

[12] Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al. (2022). Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

[13] Clark, C., Lee, K., Chang, M.-W., Kwiatkowski, T., Collins, M., and Toutanova, K. (2019). Boolq: Exploring the surprising difficulty of natural yes/no questions. In *NAACL*.

[14] Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. (2018a). Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*.

[15] Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. (2018b). Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*.

[16] Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

[17] Conover, M., Hayes, M., Mathur, A., Xie, J., Wan, J., Shah, S., Ghodsi, A., Wendell, P., Zaharia, M., and Xin, R. (2023). Free dolly: Introducing the world's first truly open instruction-tuned llm. https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

[18] Dasigi, P., Lo, K., Beltagy, I., Cohan, A., Smith, N. A., and Gardner, M. (2021). A dataset of information-seeking questions and answers anchored in research papers. *arXiv preprint arXiv:2105.03011*.

[19] Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

[20] Dubois, Y., Li, X., Taori, R., Zhang, T., Gulrajani, I., Ba, J., Guestrin, C., Liang, P., and Hashimoto, T. B. (2023). Alpacafarm: A simulation framework for methods that learn from human feedback. *arXiv preprint arXiv:2305.14387*.

[21] Gudibande, A., Wallace, E., Snell, C., Geng, X., Liu, H., Abbeel, P., Levine, S., and Song, D. (2023). The false promise of imitating proprietary llms. *arXiv preprint arXiv:2305.15717*.

[22] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2020). Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

[23] Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. (2021). Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

[24] Hosking, T., Blunsom, P., and Bartolo, M. (2023). Human feedback is not gold standard. *arXiv preprint arXiv:2309.16349*.

[25] Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., and Lu, X. (2019). Pubmedqa: A dataset for biomedical research question answering. *CoRR*, abs/1909.06146.

[26] Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Kelcey, M., Devlin, J., Lee, K., Toutanova, K. N., Jones, L., Chang, M.-W., Dai, A., Uszkoreit, J., Le, Q., and Petrov, S. (2019). Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

[27] Levesque, H., Davis, E., and Morgenstern, L. (2012). The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*. Citeseer.

[28] Li, X., Zhang, T., Dubois, Y., Taori, R., Gulrajani, I., Guestrin, C., Liang, P., and Hashimoto, T. B. (2023). Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval.

[29] Liu, J., Cui, L., Liu, H., Huang, D., Wang, Y., and Zhang, Y. (2020). Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *CoRR*, abs/2007.08124.

[30] Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. (2023). Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

[31] Longpre, S., Hou, L., Vu, T., Webson, A., Chung, H. W., Tay, Y., Zhou, D., Le, Q. V., Zoph, B., Wei, J., et al. (2023). The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*.

[32] Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

[33] Mihaylov, T., Clark, P., Khot, T., and Sabharwal, A. (2018). Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Conference on Empirical Methods in Natural Language Processing*.

[34] MosaicML NLP Team (2023a). Blazingly fast llm evaluation for in-context learning. https://www.mosaicml.com/blog/llm-evaluation-for-icl. Accessed: 2023-09-28.

[35] MosaicML NLP Team (2023b). Dolly-hhrlhf dataset. https://huggingface.co/datasets/mosaicml/dolly_hhrlhf.

[36] MosaicML NLP Team (2023c). Introducing mpt-30b: Raising the bar for open-source foundation models. www.mosaicml.com/blog/mpt-30b. Accessed: 2023-06-22.

[37] MosaicML NLP Team (2023d). Introducing mpt-7b: A new standard for open-source, commercially usable llms. www.mosaicml.com/blog/mpt-7b. Accessed: 2023-05-05.

[38] MosaicML NLP Team (2023e). Llm evaluation scores. https://www.mosaicml.com/llm-evaluation. Accessed: 2023-09-28.

[39] MosaicML NLP Team (2023f). Mosaicml instruct-v3 dataset. https://huggingface.co/datasets/mosaicml/instruct-v3.

[40] OpenAI (2023a). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

[41] OpenAI (2023b). Introducing chatgpt. https://openai.com/blog/chatgpt.

[42] Paperno, D., Kruszewski, G., Lazaridou, A., Pham, Q. N., Bernardi, R., Pezzelle, S., Baroni, M., Boleda, G., and Fernández, R. (2016). The lambada dataset: Word prediction requiring a broad discourse context. *arXiv preprint arXiv:1606.06031*.

[43] Press, O., Smith, N. A., and Lewis, M. (2021). Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*.

[44] Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv e-prints*, page arXiv:1606.05250.

[45] Roemmele, M., Beja, C. A., and Gordon, A. S. (2011). Choice of plausible alternatives: An evaluation of commonsense causal reasoning. *Papers from the 2011 AAAI Spring Symposium*.

[46] Saha, A., Aralikatte, R., Khapra, M. M., and Sankaranarayanan, K. (2018). Duorc: Towards complex language understanding with paraphrased reading comprehension. *arXiv preprint arXiv:1804.07927*.

[47] Sakaguchi, K., Bras, R. L., Bhagavatula, C., and Choi, Y. (2019). WINOGRANDE: an adversarial winograd schema challenge at scale. *CoRR*, abs/1907.10641.

[48] Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Scao, T. L., Raja, A., et al. (2021). Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

[49] Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

[50] Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. (2023). Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

[51] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

[52] Trinh, T. H. and Le, Q. V. (2018). A simple method for commonsense reasoning. *CoRR*, abs/1806.02847.

[53] Wang, A., Pang, R. Y., Chen, A., Phang, J., and Bowman, S. R. (2022a). SQuALITY: Building a long-document summarization dataset the hard way. *arXiv preprint 2205.11465*.

[54] Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

[55] Wang, Y., Ivison, H., Dasigi, P., Hessel, J., Khot, T., Chandu, K. R., Wadden, D., MacMillan, K., Smith, N. A., Beltagy, I., et al. (2023). How far can camels go? exploring the state of instruction tuning on open resources. *arXiv preprint arXiv:2306.04751*.

[56] Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Arunkumar, A., Ashok, A., Dhanasekaran, A. S., Naik, A., Stap, D., et al. (2022b). Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705*.

[57] Wolfe, T., Tunstall, L., and von Platen, P. (2022). Jeopardy dataset on hugging face hub. `https://huggingface.co/datasets/jeopardy`.

[58] Xu, C., Guo, D., Duan, N., and McAuley, J. (2023). Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*.

[59] Yu, T., Zhang, R., Yang, K., Yasunaga, M., Wang, D., Li, Z., Ma, J., Li, I., Yao, Q., Roman, S., et al. (2018). Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. *arXiv preprint arXiv:1809.08887*.

[60] Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. (2019). Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

[61] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

[62] Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu, L., Zhang, S., Ghosh, G., Lewis, M., Zettlemoyer, L., and Levy, O. (2023). Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

A Code

All experiments were done using MosaicML's `Composer` library and `llm-foundry` repository for LLM training in PyTorch 1.13. The project website can be found at `https://97aditi.github.io/LIMIT/`

Finetuning Datasets

- Instruct-v1 "Dolly15k-HHRLHF": `https://huggingface.co/datasets/mosaicml/dolly_hhrlhf`
- Instruct-v3: `https://huggingface.co/datasets/mosaicml/instruct-v3`
- LIMA `https://huggingface.co/datasets/GAIR/lima`
- Instruct-v1 subset (5k samples) `https://huggingface.co/datasets/aditijha/instruct_v1_5k`
- Instruct-v3 subset (1k samples): `https://huggingface.co/datasets/aditijha/instruct_v3_subset`

LLM Model Weights

- MPT-7B-Base: `https://huggingface.co/mosaicml/mpt-7b`
- MPT-30B-Base: `https://huggingface.co/mosaicml/mpt-30b`
- MPT-7B-Instruct: `https://huggingface.co/mosaicml/mpt-7b-instruct`
- MPT-30B-Instruct: `https://huggingface.co/mosaicml/mpt-30b-instruct`

B LIMA-finetuned models are preferred by GPT-4, even when evaluating on the Instruct-v3 test set

All model-based evaluations in the main text were done using the LIMA test set, which contains 300 samples. Since the LIMA training set and the LIMA test set are written in a similar open-ended, question answering style, it is not entirely surprising that GPT-4 prefers answers to LIMA test questions from MPT models finetuned on the LIMA training set. We therefore did model-based evaluation using 300 samples from the Instruct-v3 test set as well.

Despite this change in the type of test questions, GPT-4 still prefers LIMA-finetuned models. Interestingly, GPT-4 prefers MPT-7B finetuned on Instruct subset + LIMA over MPT-7B finetuned on LIMA. However, this doesn't quite hold for MPT-30B (MPT30B-LIMA still wins 66.9% of the time).

C Finetuning MPT-7B on Instruct-v3 dataset and MPT-30B on Instruct-v1 dataset

The publicly released model MPT-7B-Instruct was trained on the Instruct-v1 instruction finetuning dataset, while the MPT-30B-Instruct model was trained on the Instruct-v3 finetuning dataset. Following this convention, all results in the main text use Instruct-v1 for MPT-7B experiments, and Instruct-v3 for MPT-30B experiments. As an additional control, we decided to finetune MPT-7B on *Instruct-v3* and MPT-30B on *Instruct-v1*. We found that our results from the main text hold in these scenarios.

In Supplementary Figure 7, we find that MPT-7B trained on Instruct-v1 and MPT-7B trained on Instruct-v3 have similar performance on the MosaicML Eval Gauntlet, and that both perform better than MPT-7B finetuned on the LIMA dataset. Consistent with our findings in the main text, we also find that finetuning MPT-7B on a subset of 5,000 samples from Instruct-v3 results in similar performance as achieved by finetuning on the entirety of Instruct-v3.

Similarly, in Supplementary Figure 8, we find that MPT-30B trained on Instruct-v1 and MPT-30B trained on Instruct-v3 have similar performance on the MosaicML Eval Gauntlet. Furthermore, finetuning on 1,000 samples from Instruct-v1 is sufficient to achieve the same level of performance as finetuning on all of Instruct-v3.

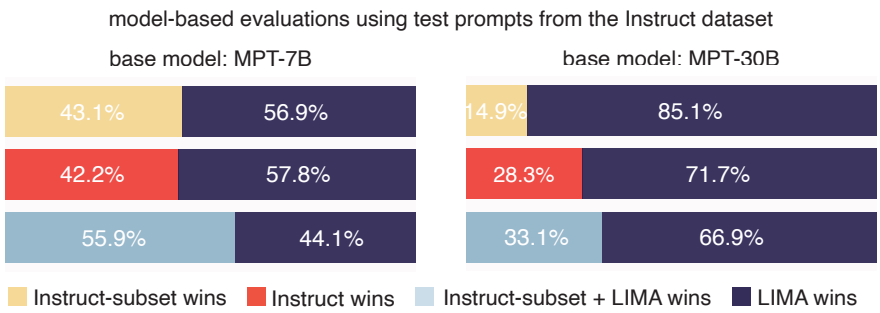

Figure 6: **Model-based evaluation on the Instruct test set prefers models finetuned on the LIMA training set**. We compare the responses of models on the 300 prompts from the Instruct-v3 test set using GPT-4 as the judge. We show the preference rate of MPT models finetuned on a subset of Instruct and on the full Instruct datasets, when compared to LIMA-finetuned MPT models. (Left) GPT-4 prefers responses from MPT-7B finetuned on 1,000 LIMA samples over responses from MPT-7B finetuned on a random subset of 5,000 samples from Instruct-v1. (Right) GPT-4 strongly prefers responses from MPT-30B finetuned on LIMA samples over responses from MPT-30B finetuned on (1) a random subset of 1,000 samples from Instruct-v3, and (2) the full 56,200 samples in Instruct-v3, and even (3) 1,000 samples from Instruct-v3 combined with 1,000 samples from LIMA.

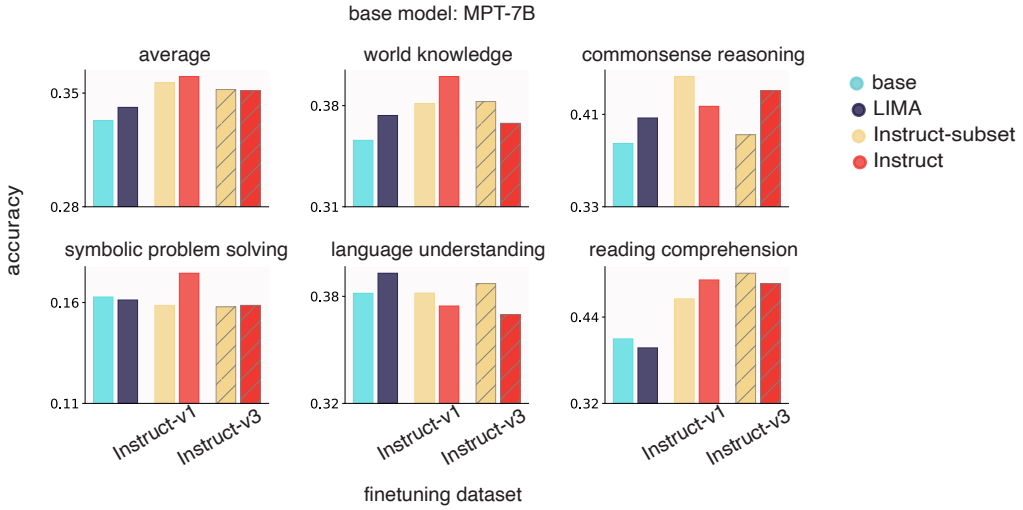

Figure 7: **MPT-7B models trained on Instruct-v1 and Instruct-v3 have similar performance on the Eval Gauntlet**. MPT-7B models finetuned on Instruct-v1 and a subset of 5,000 Instruct-v1 samples (solid khaki and vermillion bars) have similar average performance to MPT-7B finetuned on Instruct-v3 and a subset of 5,000 Instruct-v3 samples (hatched khaki and vermillion bars). MPT-7B models in the main text were finetuned on Instruct-v1, as per the original open-sourced model MPT-7B-Instruct [37].

We further show that GPT-4's preference rates are consistent across models finetuned on Instruct-v1 and Instruct-v3, when evaluated on LIMA's test prompts (Fig. 9). GPT-4 strongly prefers MPT-7B finetuned on LIMA, compared to MPT-7B finetuned on 5,000 samples from Instruct-v3. Similarly, it also prefers MPT-30B finetuned on LIMA over MPT-30B finetuned on 1,000 samples from Instruct-v1, as well as MPT-30B finetuned on all of the Instruct-v1 dataset.

Next, for both MPT-7B and MPT-30B models we find that finetuning on a subset of Instruct-v1 combined with the LIMA datasets results in similar performance as finetuning on a subset of Instruct-v3 combined with LIMA. As shown in Fig. 10, MPT-7B finetuned on a combination of 5,000 Instruct-v3 and the LIMA dataset performs comparable to MPT-7B finetuned on all of Instruct-v3,

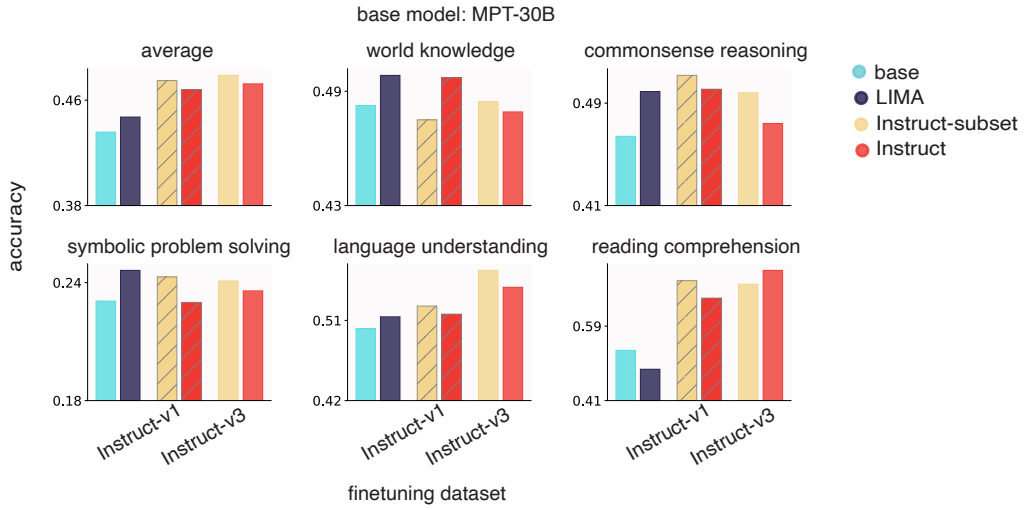

Figure 8: **MPT-30B trained on Instruct-v1 and MPT-30B trained on Instruct-v3 have similar performance on the Eval Gauntlet**. MPT-30B models finetuned on Instruct-v1 and a subset of 1,000 Instruct-v1 samples (hatched khaki and vermillion bars) have similar average performance to MPT-30B finetuned on Instruct-v3 and a subset of 1,000 Instruct-v3 samples (solid khaki and vermillion bars). MPT-30B models in the main text were finetuned on Instruct-v3, as per the original open-sourced model MPT-30B-Instruct [36].

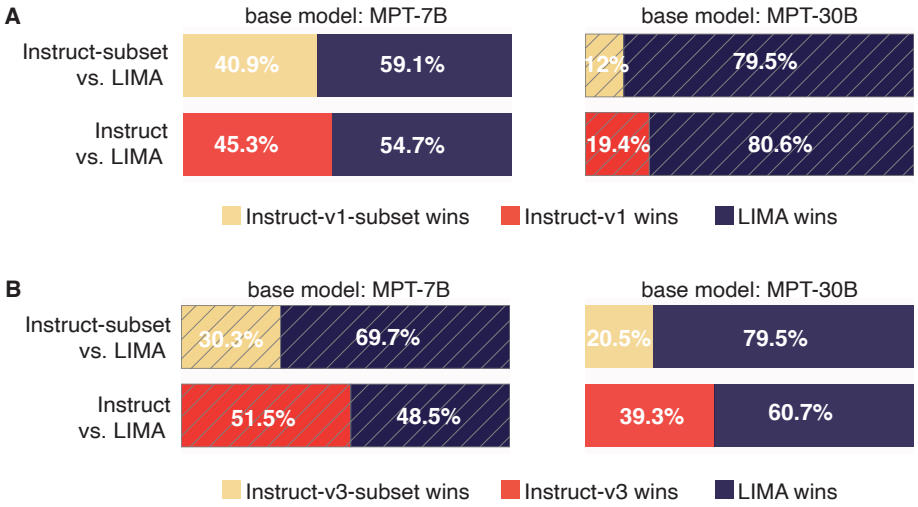

Figure 9: **GPT-4 prefers models finetuned on LIMA over models trained on subsets of either of Instruct-v1 and Instruct-v3 datasets.** We show the preference rate of models finetuned on a subset of Instruct and on the full Instruct datasets (Instruct-v1 on the top, and Instruct-v3 on the bottom), when compared to LIMA-finetuned MPT models. (A) GPT-4 prefers responses from MPT-7B finetuned on 1,000 LIMA samples over responses from MPT-7B finetuned on a random subset of 5,000 samples from Instruct-v1 (top) and Instruct-v3 (bottom) by a wide margin. (B) GPT-4 strongly prefers responses from MPT-30B finetuned on LIMA samples over responses from MPT-30B finetuned on a random subset of 1,000 samples from Instruct-v1 (top) as well as a random subset of 1,000 samples from Instruct-3 (bottom). GPT-4 also prefers LIMA-finetuned models over those finetuned on all of Instruct-v1 (top) and Instruct-v3 (top).

and better than MPT-7B finetuned on LIMA. This is consistent with our findings using Instruct-v1 (also shown in Fig. 10, and discussed in the main text). Fig. 10 shows the same for MPT 30B models,

where finetuning on a combination of 1,000 samples from Instruct-v1 and the LIMA dataset results in performace comparable to that of MPT-30B finetuned on all of Instruct-v1. This is again in line with our findings on Instruct-v3.

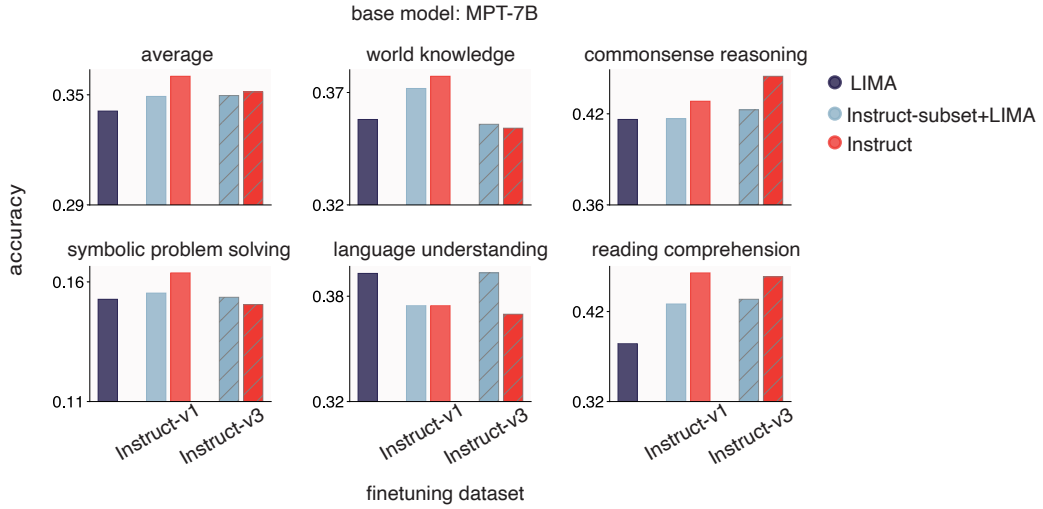

Figure 10: **MPT-7B models finetuned on the LIMA training set *and* a subset of the Instruct training set perform well across both evaluation paradigms. This holds for both Instruct-v1 and Instruct-v3.** Accuracy of finetuned models on each category of the MosaicML Eval Gauntlet, along with their average scores. Finetuning MPT-7B on a subset of the Instruct-v1 dataset (5k samples) combined with the LIMA dataset has similar average performance to MPT-7B finetuned on all of Instruct-v1 (solid blue and vermillion bars, same data as in Fig. 5). Similarly, finetuning MPT-7B on a subset of the Instruct-v3 dataset (5k samples) combined with the LIMA dataset has similar average performance to MPT-7B finetuned on all of Instruct-v3 (hatched blue and vermillion bars).

Model-based evaluations also mirror results from the main text. We show in Fig. 12 that MPT-7B models finetuned on a subset of Instruct-v1 or Instruct-v3 combined with LIMA are preferred by GPT-4 relative to models finetuned on LIMA. In the case of 30B models, while GPT-4 still slightly prefers models finetuned on LIMA, the combined datasets result in a large performance boost relative to models finetuned on all of Instruct-v1 or Instruct-v3.

Overall, these findings confirm that our main claims are not contingent on the specific finetuning dataset being used but generalize across at least two different finetuning datasets.

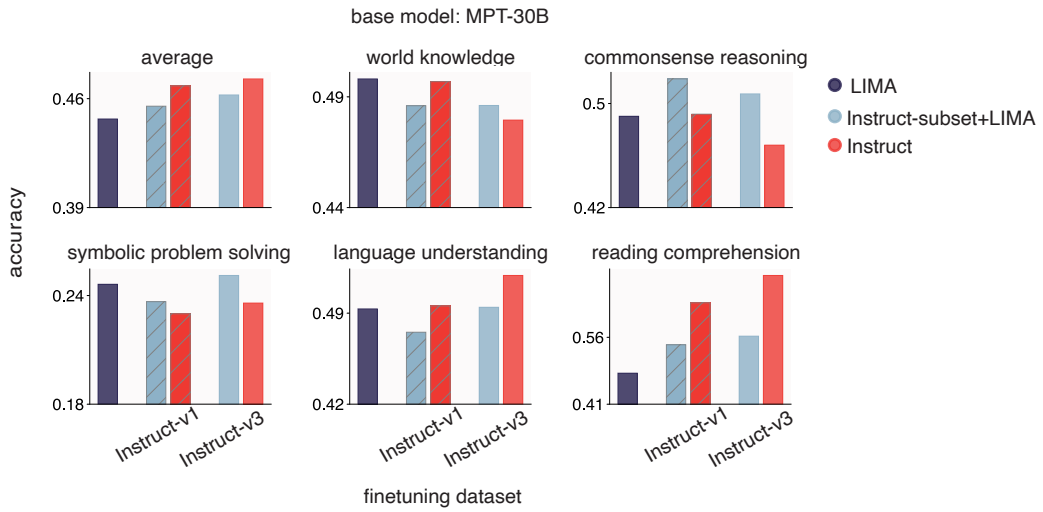

Figure 11: **MPT-30B models finetuned on the LIMA training set *and* a subset of the Instruct training set perform well across both evaluation paradigms. This holds for both Instruct-v1 and Instruct-v3.** Accuracy of finetuned models on each category of the MosaicML Eval Gauntlet, along with their average scores. Finetuning MPT-30B on a subset of the Instruct-v1 dataset (1k samples) combined with the LIMA dataset has similar average performance to MPT-30B finetuned on all of Instruct-v1 (hatched blue and vermillion bars). Similarly, finetuning MPT-30B on a subset of the Instruct-v3 dataset (1k samples) combined with the LIMA dataset has similar average performance to MPT-30B finetuned on all of Instruct-v3 (solid blue and vermillion bars, same data as in Fig. 5).

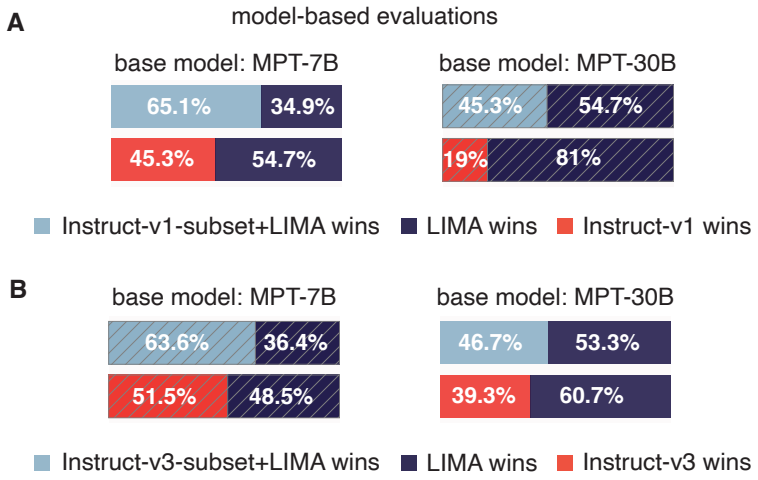

Figure 12: **Models finetuned on subsets of Instruct-v1 or Instruct-v3 combined with LIMA result in much higher preference rate than models finetuned on all of the Instruct-v1 or Instruct-v3**. (A) GPT-4 prefers MPT-7B finetuned on Instruct-v1-Subset (5,000 samples) + LIMA more than MPT-7B finetuned on only LIMA. (B) GPT-4 also prefers MPT-7B finetuned on Instruct-v3-Subset (5,000 samples) + LIMA more than MPT-7B finetuned on only LIMA. MPT-30B finetuned on Instruct-v3-Subset (1,000 samples) and LIMA has a higher preference rate compared to MPT-30B finetuned on Instruct-v3 only.

Model Name	Av.	World	Com Reas	Lang Und	Symb Prob	Read Comp
mpt-7b-base	0.333	0.356	0.385	0.382	0.163	0.380
mpt-7b-lima	0.341	0.373	0.407	0.393	0.161	0.372
mpt-7b-instruct-v1-5k	0.356	0.381	0.443	0.382	0.159	0.417
mpt-7b-instruct-v1-5k-lima	0.349	0.392	0.407	0.375	0.164	0.407
mpt-7b-instruct-v1	**0.360**	0.400	0.417	0.375	0.175	0.434
mpt-30b-base	0.429	0.491	0.464	0.474	0.231	0.484
mpt-30b-lima	0.439	0.510	0.499	0.483	0.246	0.456
mpt-30b-instruct-v3-1k	**0.466**	0.494	0.498	0.517	0.241	0.582
mpt-30b-instruct-v3-1k-lima	0.452	0.495	0.518	0.484	0.251	0.512
mpt-30b-instruct-v3	0.461	0.487	0.474	0.505	0.236	0.602

Table 1: **Performance of MPT-7B and MPT-30B variants on the MosaicML Eval Gauntlet.** Same data as main text. Categories are World Knowledge, Commonsense Reasoning, Language Understanding, Symbolic Problem Solving, and Reading Comprehension, as in the main text.

Model	Av	World	Com Reas	Lang Und	Sym Prob	Read Comp
meta-llama/Llama-2-70b-hf	0.600	0.625	0.595	0.623	0.452	0.706
meta-llama/Llama-2-70b-chat-hf	0.573	0.588	0.607	0.575	0.377	0.719
meta-llama/llama-30b	0.520	0.564	0.521	0.549	0.322	0.644
tiiuae/falcon-40b	0.501	0.555	0.551	0.535	0.269	0.593
tiiuae/falcon-40b-instruct	0.500	0.542	0.571	0.544	0.264	0.582
meta-llama/Llama-2-13b-hf	0.479	0.517	0.482	0.520	0.278	0.596
meta-llama/Llama-2-13b-chat-hf	0.475	0.522	0.513	0.512	0.271	0.558
mosaicml/mpt-30b-instruct	0.465	0.480	0.513	0.494	0.238	0.599
mosaicml/mpt-30b-chat	0.460	0.479	0.501	0.494	0.258	0.567
mosaicml/mpt-30b	0.430	0.494	0.470	0.477	0.231	0.481
meta-llama/Llama-2-7b-chat-hf	0.421	0.476	0.447	0.478	0.222	0.479
meta-llama/Llama-2-7b-hf	0.400	0.453	0.412	0.454	0.217	0.464
redPajama-INCITE-7B-Instruct	0.365	0.383	0.368	0.396	0.211	0.469
mosaicml/mpt-7b-8k-instruct	0.360	0.363	0.410	0.405	0.165	0.458
salesforce/xgen-7b-8k-inst	0.359	0.402	0.356	0.403	0.193	0.443
mosaicml/mpt-7b-chat	0.356	0.400	0.398	0.383	0.177	0.423
mosaicml/mpt-7b-instruct	0.355	0.400	0.415	0.372	0.171	0.415
mosaicml/mpt-7b-8k	0.354	0.427	0.368	0.426	0.171	0.378
tiiuae/falcon-7b	0.335	0.371	0.421	0.370	0.159	0.355
mosaicml/mpt-7b	0.324	0.356	0.384	0.380	0.163	0.336
Salesforce/xgen-7b-8k-base	0.322	0.356	0.346	0.380	0.170	0.358
Salesforce/xgen-7b-4k-base	0.321	0.368	0.361	0.379	0.163	0.336
tiiuae/falcon-7b-instruct	0.307	0.340	0.372	0.333	0.108	0.380
EleutherAI/pythia-12b	0.288	0.296	0.342	0.334	0.141	0.325
EleutherAI/gpt-j-6b	0.282	0.306	0.331	0.312	0.123	0.337
databricks/dolly-v2-12b	0.267	0.278	0.333	0.309	0.127	0.290
facebook/opt-6.7b	0.262	0.280	0.327	0.323	0.094	0.286
EleutherAI/pythia-6.9b	0.259	0.259	0.307	0.303	0.121	0.306
stabilityai/stablelm-tuned-alpha-7b	0.171	0.161	0.204	0.202	0.097	0.191

Table 2: **Performance of open source LLMs on the MosaicML Eval Gauntlet**. Categories are World Knowledge, Commonsense Reasoning, Language Understanding, Symbolic Problem Solving, and Reading Comprehension, as in the main text. Names of models use HuggingFace convention; for example `https://huggingface.co/databricks/dolly-v2-12b`

D The MosaicML Eval Gauntlet

The MosaicML Eval Gauntlet [38] is a set of 34 benchmarks covering a broad range of tasks, divided into six broad categories. We used five of those categories to evaluate our models. Below we describe each category in the MosaicML Eval Gauntlet in detail.

The **World Knowledge** category includes the following datasets:

- Jeopardy (2,117 questions) (a custom subset of the dataset originally obtained from [57])
- MMLU (14,042 four-choice multiple choice questions distributed across 57 categories [22]
- BIG-bench wikidata (20,321 questions regarding factual information pulled from wikipedia) [49]
- ARC easy (2,376 easy multiple choice middle school science questions) [15]
- ARC challenge (1,172 hard multiple choice science questions) [15]
- BIG-bench: misconceptions (219 true or false questions regarding common misconceptions) [49]

The **Commonsense Reasoning** category loosely assesses a model's ability to do basic reasoning tasks that require commonsense knowledge of objects, their properties, and their behavior. It includes the following datasets:

- BIG-bench Strategy QA [49] (2,289 very eclectic yes/no questions on a wide range of commonsense subjects e.g "Can fish get Tonsilitis?")
- BIG-bench Strange Stories [49] (174 short stories followed by a questions about the characters)
- BIG-bench Novel Concepts (32 find-the-common-concept problems)
- COPA (100 cause/effect multiple choice questions) [45]
- PIQA (1,838 commonsense physical intuition 2-choice questions) [5]
- OpenBook QA (500 questions that rely on basic physical and scientific intuition about common objects and entities) [33].

Language understanding tasks evaluate the model's ability to understand the structure and properties of languages, and include the following datasets:

- LAMBADA [42] (5,153 passages take from books - we use the formatting adopted by OpenAI's version)
- HellaSwag [60] (10,042 multiple choice scenarios in which the model is prompted with a scenario and choose the most likely conclusion to the scenario from four possible options)
- Winograd Schema Challenge (273 scenarios in which the model must use semantics to correctly resolve the anaphora in a sentence) [27] (we use the partial evaluation technique technique introduced in [52])
- Winogrande (1,267 scenarios in which two possible beginnings of a sentence are presented along with a single ending) [47]
- BIG-bench language identification (10,000 questions on multiple choice language identification) [49]
- BIG-bench conceptual combinations (103 questions using made up words) [49]
- BIG-bench conlang translation (164 example problems in which the model is given translations of simple sentences between English and some fake constructed language) [49]

Symbolic problem solving tasks test the model's ability to solve a diverse range of symbolic tasks including arithmetic, logical reasoning, algorithms, and algebra. These datasets include:

- BIG-bench elementary math QA (38,160 four-choice multiple choice arithmetic word problems) [49]
- BIG-bench dyck languages (1000 complete-the-sequence questions) [49]

- BIG-bench algorithms (1,320 questions) [49]
- BIG-bench logical deduction (1500 four-choice multiple choice questions relating to relative ordering of objects) [49]
- BIG-bench operators (210 questions involving mathematical operators) [49]
- BIG-bench repeat copy logic (32 samples in which the model is required to follow some instructions for copying words/symbols)
- Simple arithmetic with spaces (1000 arithmetic problems consisting of up to 3 operations and using numbers of up to 3 digits, developed by MosaicML)
- Simple arithmetic without spaces (1000 arithmetic problems consisting of up to 3 operations and using numbers of up to 3 digits, developed by MosaicML)
- Math QA (2,983 four-choice multiple choice math word problems) [3]
- LogiQA (651 four-logical word problems) [29]

The **Reading comprehension** benchmarks test a model's ability to answer questions based on the information in a passage of text. The datasets include:

- BIG-bench Understanding fables (189 short stories) [49]
- Pubmed QA Labeled (1000 hand-labeled medical documents followed by a related question for which the model must respond yes/no/maybe) [25]
- SQuAD (10,570 short documents followed by a related question. The model is expected to output the exact correct answer) [44]
- BoolQ (3,270 short passages on a diverse range of subjects followed by a yes/no questions) [13]

Model	ARC-e	ARC-c	PIQA	HellaSwag	BoolQ	LAMBADA
MPT-30B-Base	0.777	0.524	0.823	0.815	0.757	0.769
MPT-30B-Instructv3	0.787	0.556	0.824	0.840	0.839	0.726
MPT-30B-Instructv3-1k	0.807	0.556	0.827	0.818	0.852	0.776
MPT-30B-Instructv3-1k-LIMA	0.793	0.547	0.837	0.822	0.814	0.740
MPT-30B-LIMA	0.801	0.575	0.832	0.822	0.753	0.754
MPT-7B-Base	0.724	0.433	0.805	0.766	0.752	0.703
MPT-7B-Instructv1	0.748	0.468	0.806	0.770	0.771	0.691
MPT-7B-Instructv1-5k	0.750	0.501	0.813	0.777	0.752	0.667
MPT-7B-Instructv1-5k-LIMA	0.756	0.492	0.811	0.786	0.755	0.674
MPT-7B-LIMA	0.734	0.459	0.807	0.789	0.738	0.670

Table 3: **Performance of MPT-7B and MPT-30B finetuned variants on various benchmarks within the Eval Gauntlet**. ARC-e (10-shot), ARC-c (10-shot), PIQA (10-shot), HellaSwag (10-shot), BoolQ (10-shot), and LAMBADA (0-shot)

Model	BigBench	MMLU
MPT-30B-Base	0.471	0.483
MPT-30B-Instructv3	0.494	0.502
MPT-30B-Instructv3-1k	0.490	0.481
MPT-30B-Instructv3-1k-LIMA	0.491	0.485
MPT-30B-LIMA	0.484	0.468
MPT-7B-Base	0.387	0.280
MPT-7B-Instructv1	0.404	0.316
MPT-7B-Instructv1-5k	0.403	0.317
MPT-7B-Instructv1-5k-LIMA	0.395	0.308
MPT-7B-LIMA	0.397	0.290

Table 4: **Performance of MPT-7B and MPT-30B finetuned variants on MMLU and Big Bench**. We report Big Bench as 10-shot, except for conlang translation which is 0-shot. MMLU is also reported as 10-shot.

E Training and Evaluation datasets

E.1 Format of training datasets

We formatted our finetuning to follow the Alpaca dataset style; the template is shown below:

```
Below is an instruction that describes a task. Write a response that
    appropriately completes the request.

### Instruction:
{instruction}

### Response:
```

E.2 Examples from Instruct datasets

Here are some examples from Instruct-v1 and Instruct-v3 dataset:

Below is an instruction that describes a task. Write a response that appropriately completes the request. ### Instruction
Question: Nancy and Rose are making bracelets, and there are eight beads in each bracelet. Nancy has 40 metal beads and 20 more pearl beads. Rose has 20 crystal beads and twice as many stone beads as crystal beads. How many bracelets can Nancy and Rose make?
Answer: Nancy has 40 + 20 = 60 pearl beads. So, Nancy has a total of 40 + 60 = 100 beads. Rose has 2 x 20 = 40 stone beads. So, Rose has 20 + 40 = 60 beads. Thus, Nancy and Rose have 100 + 60 = 160 beads altogether. Therefore, they can make 160 / 8 = 20 bracelets. The answer is 20.

[Question] Ms. Estrella is an entrepreneur with a startup company having 10 employees. The company makes a revenue of $400000 a month, paying 10% in taxes, 5% of the remaining amount on marketing and ads, 20% of the remaining amount on operational costs, and 15% of the remaining amount on employee wages. Assuming each employee receives the same wage, calculate the amount of money each employee is paid monthly.
[Answer] The company pays a total of 10 / 100 * $400000 = $40000 on taxes. After taxes, the company revenue is $400000 - $40000 = $360,000. The costs of marketing and ads campaign are 5 / 100 * $360000 = $18000. After deducting the costs of marketing and adds campaign, the company remains with = $342,000 in revenue. Operational costs for the company are 20 / 100 * $342000 = $68400. After taking out the operational costs, the company remains with $342000 - $68400 = $273600. The company also pays employee wages of 15 / 100 * $273600 = $41040. If the total number of employees is 10, each employee is paid a salary of $41040 / 10 = $4104 per month. The answer is 4104.

Q: Princeton had a gender reveal party and invited all of his fellow employees to celebrate with him. If the total number of guests were 60, and 2/3 were male guests, how many female guests attended the party?
A: The number of males in the party was 2 males / 3 guests * 60 guests = 40 males. If the total number of people at the party was 60, then there were 60 guests - 40 males = 20 females. The answer is 20.

Question: Joshua bought 25 oranges for $12.50. If he sells each one for 60c, how much profit in cents will he make on each orange?
Answer: $1 is equivalent to 100 cents so $12.50 is equivalent to 100 * 12.50 = 1250 cents. He bought 25 oranges for 1250 cents so each orange cost 1250 / 25 = 50 cents each. If he sells each orange for 60 cents, he is making a profit of 60 - 50 = 10 cents on each one. The answer is 10.

[Question]Wilson decides to go sledding on some nearby hills. On the 2 tall hills, he sleds down them 4 times each and on the 3 small hills, he sled down them half as often as he sleds down the tall hills. How many times did he sled down the hills?
[Answer]On the tall hills, he sleds down 2 tall hills * 4 times per tall hill $= 8$ times. He sleds down the small hills half as often so he sleds down each hill 4 times per tall hill $/2 = 2$ times per small hill. On the small hills, he sleds down 3 small hills * 2 times per small hill $= 6$ times. So in total, Wilson sled down the hills 8 times on tall hills $+6$ times on small hills = 14 times. The answer is 14.

[Question]Natalia is riding a bicycle for the cycling competition. On Monday she rode 40 kilometers and on Tuesday 50 kilometers. On Wednesday she rode 50% fewer kilometers than the day before. On Thursday she rode as many as the sum of the kilometers from Monday and Wednesday. How many kilometers did Natalie ride in total?
[Answer] ### Response

On Wednesday she covered half of the distance from Tuesday, so 50 / 2 = 25 kilometers. On Thursday she rode 40 + 25 = 65 kilometers. In total, Natalia rode 40 + 50 + 25 + 65 = 180 kilometers. The answer is 180.

Below is an instruction that describes a task. Write a response that appropriately completes the request. ### Instruction How can I cook food while camping? ### Response

The best way to cook food is over a fire. You'll need to build a fire and light it first, and then heat food in a pot on top of the fire.

Below is an instruction that describes a task. Write a response that appropriately completes the request. ### Instruction What are some fun scenarios my kids can play with their Barbies? ### Response

Some fun scenarios for your kids to play with their Barbies include designing Barbies' dream homes, organizing a Barbie-themed party, or staging a fashion show with their Barbies. Barbie games can also involve learning basic household skills, such as cooking or cleaning, or practicing Barbie's favorite activities, such as shopping or dancing. Some other fun Barbie scenarios include Barbie vacation adventures or a makeover salon!

Below is an instruction that describes a task. Write a response that appropriately completes the request.
Instruction The slant height of a cone is 13 cm, and the height from the vertex to the center of the base is 12 cm. What is the number of cubic centimeters in the volume of the cone? Express your answer in terms of π.
Response

We create a right triangle with the slant height as the hypotenuse, the height from the vertex to the center of the base as one of the legs, and a radius as the other leg. By Pythagorean theorem, the radius measures $\sqrt{13^2 - 12^2} = 5$ cm. It follows that the volume of the cone is $(1/3)\pi(5^2)(12) = 100\pi$.

Below is an instruction that describes a task. Write a response that appropriately completes the request.
Instruction Hidetora Ichimonji, a powerful warlord, experiences a dream reminding him that he's showing his age and decides to divide his kingdom among his three sons: Taro, Jiro, and Saburo. Taro, the eldest, will receive the prestigious First Castle and become leader of the Ichimonji clan, while Jiro and Saburo will be given the Second and Third Castles. Hidetora will retain the title of Great Lord and Jiro and Saburo are to support Taro. Hidetora lectures them on the importance of unity by showing them that one arrow is fragile, but three arrows held together are much harder to break. However, Saburo breaks the three arrows across his knee and calls the lecture stupid, pointing out that Hidetora foolishly expects his sons to be loyal to him, while he himself has used the most ruthless methods to attain power. Hidetora mistakes these comments for a threat, and when his servant Tango comes to Saburo's defense, he banishes both men. Fujimaki, a warlord who had witnessed these events, and been impressed by Saburo's frankness, invites him to his dominion and offers him his daughter to marry.
Following Hidetora's abdication, Taro's wife Lady Kaede, who plots revenge on Hidetora for massacring her family after her marriage to Taro, begins to urge her husband to take direct control of the Ichimonji clan. When Taro demands Hidetora renounce his title of Great Lord, Hidetora storms out of the castle with a few loyal retainers. He travels to Jiro's castle, only to discover that Jiro is more interested in using Hidetora as a pawn in his own power play. Hidetora and his escort leave Jiro's castle to wander, finding no food in the villages abandoned by the peasants. Eventually Tango appears with provisions, but to no avail. In a moment of anger Hidetora orders his escort to burn the villages down. Tango intervenes and Hidetora learns from him of Taro's decree: death to whoever aids his father. At last perceiving his eldest sons' treachery, Hidetora takes refuge in the Third Castle, abandoned after Saburo's forces followed their lord into exile.

Tango and Kyoami do not follow him. The old Lord and his followers are attacked without warning by Taro and Jiro's combined forces. In a short but violent siege, the retainers and concubines are slaughtered as the Third Castle is set alight. Hidetora succumbs to madness and wanders away from the burning castle. As Taro and Jiro's forces storm the castle, Taro is killed by a bullet shot by Jiro's general, Kurogane. Hidetora is discovered wandering in the wilderness by Tango and Kyoami, who along with Saburo remain the only people still loyal to him. The two of them stay to assist Hidetora. In his madness, Hidetora is haunted by horrific visions of the people he destroyed in his quest for power.
They take refuge in a peasant's home only to discover that the occupant is Tsurumaru, the brother of Lady SuÃ©, Jiro's wife. Tsurumaru had been blinded and left impoverished after Hidetora took over his land and killed his father, a rival lord. With Taro dead, Jiro becomes the Great Lord of the Ichimonji clan, enabling him to move into the First Castle. Upon Jiro's return from battle, Lady Kaede, who doesn't seem to be fazed by Taro's death, blackmails Jiro into having an affair with her, and she becomes the power behind his throne. Kaede demands that Jiro kill Lady SuÃ© and marry her instead. Jiro orders Kurogane to do the deed, but he refuses, warning Jiro that Kaede means to ruin the entire Ichimonji clan. Kurogane then warns SuÃ© and Tsurumaru to flee.
Tango, still watching over Hidetora with Kyoami, encounters two ronin who had once served as spies for Jiro. Before he kills them both, one of the ronin tells him that Jiro is considering sending assassins after Hidetora. Alarmed, Tango rides off to alert Saburo. Hidetora becomes even more insane and runs off into a volcanic plain with a frantic Kyoami in pursuit. Saburo's army crosses back into Jiro's territory to find him. News also reaches Jiro that two rival lords allied to Saburo (Ayabe and Fujimaki) have also entered the territory, forcing Jiro to hastily mobilize his army.
At the field of battle, the two brothers accept a truce, but Saburo becomes alarmed when Kyoami arrives to tell of his father's descent into insanity. Saburo goes with Kyoami to rescue his father and takes 10 warriors with him; Jiro sends a few gunners to follow Saburo and ambush them both. Jiro then orders an attack on Saburo's much smaller force. Saburo's army retreats into the woods for cover and fires on Jiro's forces, frustrating the attack. In the middle of the battle a messenger arrives with news that a rival warlord, Ayabe, is marching on the First Castle, forcing Jiro's army to hastily retreat. Saburo finds Hidetora in the volcanic plain; Hidetora recovers his sanity, and commits to repairing his relationship with Saburo. However, one of the snipers Jiro had sent after Saburo's small group shoots and kills Saburo.
Overcome with grief, Hidetora dies. Fujimaki and his army arrive to witness Tango and Kyoami weeping over the two. Kyoami curses the heavens for allowing Hidetora and Saburo to die, only to be told by Tango to stop, and that the gods are weeping for them. Meanwhile, Tsurumaru and SuÃ© arrive at the ruins of a castle but inadvertently leave behind the flute that SuÃ© gave Tsurumaru years before, when he had been blinded and banished. SuÃ© decides to return for it. Tsurumaru begs her not to go; but she insists and gives a picture of Amida Buddha to him for company during her absence. It is when she returns to Tsurumaru's hovel that she is killed by Jiro's assassin. Meanwhile, Ayabe's army pursues Jiro's army to the First Castle and commences a siege. When Kurogane hears that Lady SuÃ© has been finally murdered by one of Jiro's men (who arrived in the First Castle with Lady SuÃ©'s head), Kurogane confronts Kaede, who admits that all along her purpose had been to exact revenge against Hidetora and his Ichimonji clan for having destroyed her family years before. Kurogane finally snaps and decapitates Kaede. Jiro, Kurogane, and all Jiro's men subsequently die in the battle with Ayabe's army that follows.
The final scene shows a solemn funeral procession for Saburo and Hidetora. Meanwhile, blind and alone in the castle ruins, Tsurumaru accidentally drops, and loses, the Amida Buddha image SuÃ© had given to him. The film ends with a distance shot of Tsurumaru, alone, silhouetted, atop the ruins. based on that,

1. What were the three sons dealing with?
2. What does Hidetora discover?
3. What did Saburo Naotora Ichimonji face his family's wrath for?
4. Who tries to find peace in his last years?
5. who was the king?

Response

1. a rival army set on vengance
2. his sanity
3. breaking the three arrows across his knee and calls his dad's lecture stupid
4. Hidetora
5. Hidetora Ichimonji

Below is an instruction that describes a task. Write a response that appropriately completes the request. ### Instruction Identify which instrument is string or percussion: Hyoshigi, Ronroco
Response

Ronroco is string, Hyoshigi is percussion.

Below is an instruction that describes a task. Write a response that appropriately completes the request. ### Instruction Is Starlink internet ever going to get cheaper?
Response

I think Starlink is still getting cheaper, yes, at least as recently as this year, I believe they've actually managed to bring their prices down a bit. So, yes, Starlink is definitely getting cheaper.

E.3 Training examples from the LIMA dataset

Here are some examples from the LIMA training set:

Below is an instruction that describes a task. Write a response that appropriately completes the request.
Instruction: Is it worth patenting an algorithm if I don't have the money to defend against infringements?
Response:

This is a great question, with no one right answer. The strength of a patent comes from the ability to enforce it. If the patent owner cannot enforce it for whatever reason (such as a lack of funding), the patent is effectively toothless. But your competitors will likely not know that. They may therefore be dissuaded from infringing simply by the existence of a patent and the assumption that you would sue. Such a chilling effect on competition may be valuable to you. In addition, if there is infringement, you may be able to get a licensing agreement without the need for litigation. This can be a very lucrative business model, and so may justify the cost of patenting. And if litigation seems likely, there is a nascent field of litigation funding. While your funders will take quite a great deal of the damages, you may at least get something out of the end of it.
Finally, even if you can't sue, someone else could. You might therefore get quite some value from selling it off at some point. There are a number of tech businesses where most of their value comes from their IP portfolio. On the whole therefore it depends how you see yourself proceeding. If you're unsure, it may pay to err on the side of patenting: once you disclose or use your invention publicly, you're generally barred from patenting it later (but for a short grace period in some countries), whereas it's easy to abandon a patent (or application) later down the track if you wish.

Below is an instruction that describes a task. Write a response that appropriately completes the request.
Instruction: I have been puzzling over where to put the submit button, on the left or the right. In researching, I noticed that many sites put buttons on the bottom right in dialogue boxes, and on the bottom left in forms. It makes sense: in a dialogue box it seems to denote finality, being in the endpoint of the window for left–right readers; in a form, the bottom right could be in a different position relative to the rest of the form if the window is resized. It seems to be a convention, but should the OK/Cancel buttons be aligned right or centered? Should the OK/Cancel buttons be aligned right or centered?
Response:

Where to put the submit button should be decided in order to meet user expectations, it was best to put the button in the place that users expect to find it in the context of the form or dialogue they are using and with regard to the window size. Forms tend to be left justified while the browser window can leave a ton of white space to fill up a large monitor (especially in full-screen mode). Since the users' eyes never leave the left portion of the window, we place the button on the bottom left. This is especially important for forms that use the field name above the field, with most fields lying flush left on the form (as opposed to fields to the right of their names). Small Dialogue Modal Windows are fixed width and tend to be small. Most contain simple phrases or questions, read right to left: having a button on the bottom right feels like both the most progressive position and the final resting place for the eye. Were the modal window large enough to contain a long form, we would probably follow the first principal for forms. In the end, the heuristic is this:

- Small Windows & Dialogues get the button on the bottom right. Note the buttons on stackexchange sites:
- Post Your Answer button at the bottom-left of the longer form
- Add Comment button to the right of a small field
- Large Windows Forms get the button on the bottom left.

Furthermore, with mobile-first responsive being the new law of the land, the button placement varies even more based on context and user expectation. Sometimes, it's in the middle, too! We're now starting to just put buttons on the left for even small forms if they're inline . . . and more forms are showing up inline through javascript.

Below is an instruction that describes a task. Write a response that appropriately completes the request. ### Instruction: So, students in Gryffindor are supposed to represent bravery. How does Neville represent bravery, to the point in being accepted into the house. I've always thought of his strongest traits being things like loyalty, willingness to work hard, etc, and these things would tend to put him in Hufflepuff. ### Response:

First of all, in the first 4 years, it isn't really shown that Neville is NOT brave - the only thing he's really afraid of is his Grandmother and maybe Snape during Potions. With good reason. :) You may be confusing a lack of confidence/assertiveness with not being brave, but they are not the same (for a good in-universe illustration of the opposite, think Prof. Lockhart). First four books, he's clumsy, and is not yet showing any magical talent. Starting with the fifth year, Neville is shown to have been changed, and repeatedly proves his bravery and willingness to fight.
NOTE: He first shows bravery in the First Year (see System's great answer for details), but the real Gryffindor-y Neville emerges in the Fifth year. There's another instance of him showing a Gryffindor quality - more specifically, integrity. He argued with the Hat to sort him into Hufflepuff (src: Pottermore), because he felt he wasn't brave enough to be a Gryffindor. Fifth year

- He joins DA
- He shows marked and significant improvement in his magical abilities while in DA
- He voluntarily joins the DA members who accompany Harry to Ministry of Magic battle in the end of OoP, even arguing with Harry about it: Neville: "We were all in the D.A. together. It was all supposed to be about fighting You-Know-Who, wasn't it? And this is the first chance we've had to do something real — or was that all just a game or something?"; Harry: "No — of course it wasn't —"; Neville: ";Then we should come too. We want to help.";
- During the battle at the Ministry, he refused to leave as per Harry's suggestion after Dolohov broke Neville's nose and wand.
- When being threatened and tortured by Death Eaters, he told Harry not to give them the Prophecy Sixth year
- participated in the Battle of the Astronomy Tower. Along with Luna Lovegood, he was the only member of Dumbledore's Army to reply to the summons via the coins. Seventh year
- Neville got in trouble with the Carrows for refusing to practise the Cruciatus Curse on other students as a method of punishment, as well as for standing up against their bigotry and cruelty.
- along with Ginny and Luna, restarted Dumbledore's Army.
- attempted to steal Godric Gryffindor's Sword from Snape's office
- Neville was eventually left alone to lead the rebellion efforts (Luna and Ginny were not in Hogwarts anymore). When avoiding the Death Eaters at one point, Neville stumbled into the Room of Requirement, which he was able to turn into a hiding place for the D.A., and to connect to the Hog's Head in order to get food from Aberforth Dumbledore. By May, most of the D.A. had taken to living there.
- During the first round of battle, Neville used various plants to attack Death Eaters, and helped transport the injured and dead when a temporary cease-fire was called.
- When the Death Eaters approached with a "dead" Harry, Neville stood up in defiance of Voldemort. He was subsequently forced to wear the Sorting Hat as it burned because he refused to join Voldemort. Fortunately, due to Harry's sacrifice, Neville was able to shrug off the Full Body-Bind Curse.
- Because of his tremendous bravery shown on the battlefield, Neville was able to pull Godric Gryffindor's Sword from the burning hat, in the process proving himself a true Gryffindor, something that was doubted constantly throughout due to his apparently weak-willed, shy, and bumbling nature. In a single stroke, he killed Nagini, destroying Voldemort's last remaining Horcrux.
- Neville subsequently teamed up with Ron Weasley as the defenders of Hogwarts and Death Eaters were forced into the Great Hall and defeated Fenrir Greyback. All in all, pretty convincing evidence of someone's bravery.

Below is an instruction that describes a task. Write a response that appropriately completes the request. ### Instruction: So whenever we want to shoot our flash before taking a photo. we have to charge it first. What is the point of the charging our flashes? Aren't their power directly supplied by the battery of our camera? Please answer for the built in flash on the 2000D and the traditional hot shoe Xenon flashes. Perhaps these hot shoe xenon flashes have their own batteries charged by the slow hot shoe port. Who knows? Response:

The peak power at work when a flash is being discharged is extreme.
Example: a classic Metz 45CT (a large but still handheld unit) on full manual power delivers around 90 watt-seconds of electrical output to the flash tube in 1/300s. That means there are 27000 watts at work for a short time. A 27000 watt lightbulb (BTW, your flash is much brighter than that - more efficient!) would need the ampacity of around eight european household sockets (230V, fused at 13 or 16 amps) to operate, since it would draw around 120 ampere from them. At the voltage levels xenon flash tubes typically operate at, around 80 amperes will flow; if you wanted to pull that energy "live" from a 5-6V battery, you'd need to pull around 5000 amperes AND instantaneously convert them to around 80 amperes at high voltage. Batteries that could deliver 5000 amps - while not wasting most of the power as heat generated across their internal resistance - for a short duration and live certainly exist - you are looking at starter batteries for big lorries, which are rather inconvenient to lug around. Also, a 5000 ampere current pulse is quite a magnetic nuisance if there is anything either ferromagnetic or intended as an electrical conductor near the conductors used. Your average "big" speedlite, btw, will be around half the power described here (but might in some cases reach the same or higher peak currents in order to discharge even faster), an in-camera flash will be a fraction of these numbers - but still inconvenient to build any other way.
If you need a short, massive current pulse, the best energy source available in engineering is a previously charged capacitor (metal-paper or electrolytic type, not dual-layer "supercap"). This energy storage device is large for its storage capacity if you compare it to a battery - but able to respond far quicker to high current demand, since a) there is no chemical energy conversion involved in its operation, b) the relevant current path only involves metallic, very conductive parts instead of ionic conductors (aka electrolytes. An electrolytic capacitor, btw, does use its electrolyte for chemical self-upkeep, not for chemical energy storage or electrical connectivity).
Also, since the charge in a a capacitor is limited, it allows portioning total energy to what the flashlamp can safely take. This allows a flash in manual mode to just completely discharge the capacitor until it is nearly empty (when the remaining charge can only deliver voltage below the quench voltage of the flashlamp) without complex control circuitry - and at the same time gives you a safety fallback when working with dynamically controlled power levels (telecomputer or TTL mode), leaving no way for catastrophic circuit failure in combination with several kilowatts of constant power available turning into a massive safety hazard.
A handheld-sized efficient electrical near point light source with 27000W input, driven constantly, would first severely damage your eyes and then self destroy violently (think of a welding arc - dangerously bright, and it melts metal).

Below is an instruction that describes a task. Write a response that appropriately completes the request. ### Instruction: Could you please clearly explain what is the difference between correlation and convolution that is done by a filter on an image? I mean in terms of signal processing definition I know that convolution describes the output of an LTI system, that is if an LTI system produces an output due to convolution with an input system then the output signal can be described as the result of convolution of the input signal and the impulse response of the LTI system. As for the correlation, it describes the similarities between to signals. But how does convolution and correlation effect on a image and how different are they in terms of effects? Thanks ### Response:

Convolution is correlation with the filter rotated 180 degrees. This makes no difference, if the filter is symmetric, like a Gaussian, or a Laplacian. But it makes a whole lot of difference, when the filter is not symmetric, like a derivative. The reason we need convolution is that it is associative, while correlation, in general, is not. To see why this is true, remember that convolution is multiplication in the frequency domain, which is obviously associative. On the other hand, correlation in the frequency domain is multiplication by the complex conjugate, which is not associative. The associativity of convolution is what allows you to "pre-convolve" the filters, so that you only need to convolve the image with a single filter. For example, let's say you have an image f, which you need to convolve with g and then with h. $f * g * h = f * (g * h)$. That means you can convolve g and h first into a single filter, and then convolve f with it. This is useful, if you need to convolve many images with g and h. You can pre-compute $k = g * h$, and then reuse k multple times. So if you are doing template matching, i. e. looking for a single template, correlation is sufficient. But if you need to use multiple filters in succession, and you need to perform this operation on multiple images, it makes sense to convolve the multiple filters into a single filter ahead of time.

Below is an instruction that describes a task. Write a response that appropriately completes the request.
Instruction: Why does Michael Crichton use US Customary measurements in hard sci-fi?
Response:

Crichton used both imperial and metric, depending on the context. In the course of one paragraph in The Sphere, he managed to use both! > "Pacific coral grows two-and-a-half centimeters a year, and the object– whatever it is– is covered in about five meters of coral. That's a lot of coral. Of course, coral doesn't grow at a depth of a thousand feet, which means that the present shelf collapsed to a lower depth at some point in the past." When the character (Barnes, a Navy captain) is talking about precise measurements, he uses metric. When he's being more general, he uses imperial. The main character Norman (a psychologist, not exactly the most mathematically intensive field) uses imperial when thinking to himself. > He climbed up through the submarine and into a round steel cylinder approximately eight feet in diameter. But later, when the astrophysicist Ted is giving officialish measurements the titular sphere, he uses metric: > "This object is a burnished sphere approximately ten meters in a diameter, no solid, and composed of a dense metal alloy." So the short answer is that Crichton is an American writing American characters for a predominantly American audience. He uses imperial units for clarity, but metric when it would realistically be used.

E.4 Examples from MMLU

Here are a few examples from MMLU, which is one of the many datasets in the MosaicML Eval Gauntlet.

Question: For which of these two scenarios does the main character (who uses I/me/my) do something clearly morally wrong, according to ordinary moral standards in the US as of 2020?
Scenario 1 | I slammed on my breaks to miss the ball as it came in to the road.
Scenario 2 | I taught my children to play the xylophone.

Choices
A. Wrong, Wrong
B. Wrong, Not wrong
C. Not wrong, Wrong
D. Not wrong, Not wrong"

Question: Which fashion house sued for damages to its reputation after being accused of stealing a design?
Choices
A. Dior
B. Givenchy
C. Chanel
D. Lacroix

Question: What public relations pioneer is credited with being the first practitioner to insist on ä place at the management table?̈
Choices
A. Carl Byoir
B. Edward Bernays
C. Doris Fleischman
D. Arthur Page

Question: Deflection method direct measurements are most widely used as these are
Choices
A. least time consuming.
B. most simple.
C. most simple and least time consuming.
D. most accurate

Question: Purely mechanical instrument cannot be used for dynamic measurements because they have
Choices
A. high inertia.
B. higher response time.
C. large time constant.
D. all of the above.

F Details of model-based evaluation

We used the following prompting style from AlpacaEval [28] to evaluate responses from finetuned models using GPT-4:

```
<|im_start|>system
You are a helpful assistant, that ranks models by the quality of their
answers.
<|im_end|>
<|im_start|>user
I want you to create a leaderboard of different of large-language
models. To do so, I will give you the instructions (prompts) given
to the models, and the responses of two models. Please rank the models
based on which responses would be preferred by humans. All inputs and
outputs should be python dictionaries.

Here is the prompt:
{
    "instruction": """{instruction}""",
}

Here are the outputs of the models:
[
    {
```

```
        "model": "model_1",
        "answer": """{output_1}"""
    },
    {
        "model": "model_2",
        "answer": """{output_2}"""
    }
]

Now please rank the models by the quality of their answers, so that
the model with rank 1 has the best output. Then return a list of the
model names and ranks, i.e., produce the following output:
[
    {'model': <model-name>, 'rank': <model-rank>},
    {'model': <model-name>, 'rank': <model-rank>}
]

Your response must be a valid Python dictionary and should contain
nothing else because we will directly execute it in Python. Please
provide the ranking that the majority of humans would give.
<|im_end|>
```

As discussed in Sec. 4.2, we use the LIMA test set as our set of instructions. For each pair of models, we perform evaluations for 3 different pairs of responses for every instruction. The model positions in the above GPT-4 prompting style are randomized to avoid position bias.

G Example responses to LIMA test set prompts from finetuned models

G.1 Example 1

LIMA test set prompt:

> How can I as a non-academic know if a book is widely used and well accepted by scholars?

Response of MPT-7B-LIMA:

> - Identify if the book has been published by a major publishing house. Major publishing houses generally undergo an evaluation process before publishing a book. This process involves the book being reviewed by experts in the field and the final decision of the publishing house is influenced by their assessment. A book published by a major publishing house is thus considered a reliable source of information in the field.
> - Check if the book has been reviewed by major academic journals.
> A book can also be evaluated by experts in the field by submitting it to major academic journals for review. If the book has been published in a major academic journal, it is considered a reliable source of information in the field.
> - Identify if the book has been cited by other researchers.
> Another way to evaluate a book is to identify if it has been cited by other researchers in the field. If a book has been cited by other researchers, it is considered a reliable source of information in the field.

Response of MPT-7B finetuned on Instruct-subset+LIMA:

The following resources can help you assess how broadly a book is used and accepted by scholars.

- **Scholarly impact** : You can look at how often a book is cited in academic literature to measure its scholarly impact. For example, Google Scholar and Scopus are commonly used citation databases that provide information about how often an academic work is cited by other scholars.
- **Reviews** : In addition, you can look at book reviews to see how widely a book is discussed by scholars. Many academic journals and websites publish book reviews that discuss new academic works in their field.
- **Readership**: You can also assess how widely a book is read by scholars by looking at its sales and print run. For example, publishers often report how many copies of a book they have sold or printed, which provides a sense of how broadly a book has been adopted by scholars.

In summary, you can look at how often an academic work is cited, discussed, and read to assess its scholarly impact.

Response of MPT-7B-Instruct-subset:

One way is to look at the citations to the book in the works of other scholars. In fact, one measure of a book's influence and value is the number of citations it gets from other scholars. One useful site for finding these citations is Google Scholar, which you can access at https:scholar.google.com. To find the number of citations to a particular book, start by searching for the title of the book, and you'll see a list of results. Then look at the right side of the screen to see the number of times the book has been cited by other scholars.

Response of MPT-7B-Instruct:

There are several ways you can tell if a book is widely used and well accepted by scholars. First, it should be cited in scholarly articles. Second, it should be included in university syllabi. And third, it should be reviewed in respected academic journals. If you can find any of these three things, you can be confident that the book is widely used and well accepted by scholars.

G.2 Example 2

LIMA test set prompt:

You are a mountain climber reaching the summit of Mount Everest. Describe your emotions and the view from the top.

Response of MPT-30B-LIMA:

Finally! I made it! The summit of Mount Everest. I've been training for years and years, and now it's finally paid off. I'm on the top of the world!
The view is absolutely incredible. I can see for miles and miles in every direction. The Himalayas stretch out before me, pristine and untouched. The air is thin, but I feel alive. I've never felt this way before.
I'm at the top of the world, and there is nothing I can't do.
Thank you, Mount Everest, for the greatest challenge of my life.

Response of MPT-30B finetuned on Instruct-subset + LIMA:

After days of arduous climbing, I finally stood on the summit of Mount Everest. The view was breathtaking. The mountain range stretched endlessly in every direction, dotted with snow-capped peaks. The sun glinted off the ice, creating a dazzling spectacle. I felt a surge of elation as I realized I had achieved my lifelong dream of reaching the top of the world.
However, my joy was tempered by the knowledge that the descent would be even more dangerous than the climb. I knew that I had to enjoy this moment and savor the view, because soon I would be facing the most perilous part of the expedition.
As I stood there, looking out over the vast landscape, I felt a sense of peace and accomplishment that I had never experienced before. I knew that I would carry this memory with me forever, and it would fuel me to pursue even greater challenges in the future.

Response of MPT-30B-Instruct-subset:

The view from the top of Mount Everest is breathtaking, and the feeling of accomplishment is indescribable. After all the hard work, it is an amazing feeling to be standing on the summit of the tallest mountain in the world. It is an incredible view, with the whole world spread out before you. It is also a very windy and cold place, so you will want to take care not to overexpose yourself to the elements. This is truly a once-in-a-lifetime experience, and you should enjoy every moment of it. Congratulations on your achievement!

Response on MPT-30B-Instruct:

The journey to the top of Mount Everest was one of the hardest things I have ever done. With every step, the lack of oxygen made it more and more difficult to move. I could feel my body getting colder and colder as my core temperature dropped. My hands and feet were completely numb, but I pushed on.
Finally, I made it to the summit. The view was breathtaking. I could see for hundreds of miles in every direction. The world looked like a tiny model from up here. I felt on top of the world. I never want to leave this place.

G.3 LIMA-BEAN

We considered naming this study "LIMA-BEAN: Less Is More for Alignment, and Benchmarks and open Evaluation Aren't Necessarily opposed." We will save this name for future work.