

# INSTRUCT-SKILLMIX: A Powerful Pipeline for LLM Instruction Tuning

Simran Kaur<sup>1\*</sup>, Simon Park<sup>1\*</sup>, Anirudh Goyal<sup>2</sup>, Sanjeev Arora<sup>1</sup>

<sup>1</sup> Princeton Language and Intelligence (PLI), Princeton University

<sup>2</sup> Meta

## Abstract

We introduce INSTRUCT-SKILLMIX, an automated approach for creating diverse, high quality SFT data for instruction-following. The pipeline involves two stages, each leveraging an existing powerful LLM: (1) *Skill extraction*: uses the LLM to extract core “skills” for instruction-following by directly prompting the model. This is inspired by “LLM metacognition” of Didolkar et al. [2024]; (2) *Data generation*: uses the powerful LLM to generate (instruction, response) data that exhibit a randomly chosen pair of these skills. Here, the use of random skill combinations promotes diversity and difficulty. The estimated cost of creating the dataset is under \$600.

Vanilla SFT (i.e., no PPO, DPO, or RL methods) on data generated from INSTRUCT-SKILLMIX leads to strong gains on instruction following benchmarks such as AlpacaEval 2.0, MT-Bench, and WildBench. With just 4K examples, LLaMA-3-8B-Base achieves 42.76% length-controlled win rate on AlpacaEval 2.0, a level similar to frontier models like Claude 3 Opus and LLaMA-3.1-405B-Instruct.

Ablation studies also suggest plausible reasons for why creating open instruction-tuning datasets via naive crowd-sourcing has proved difficult. In our dataset, adding 20% low quality answers (“shirkers”) causes a noticeable degradation in performance.

The INSTRUCT-SKILLMIX pipeline seems flexible and adaptable to other settings.

## 1 Introduction

*Instruction tuning* (sometimes also called *imitation learning*) is the first step in converting a base LLM trained on next-word prediction into a helpful and interactive agent. Whereas early versions of instruction tuning involved supervised fine-tuning (SFT) on traditional NLP question-answer datasets [Wei et al., 2022], nowadays, the SFT data is collected at high cost from skilled human annotators. We will use the term “instruction tuning” to refer solely to supervised fine-tuning (SFT) on such Q&A pairs — and not to reinforcement-learning methods such as PPO/DPO/RLHF [Schulman et al., 2017, Rafailov et al., 2023] etc., which usually follow SFT in the pipeline.

Human-generated data is expensive (e.g., even the tiny model Instruct-GPT was estimated to require 20K human hours OpenAI [2022]), which has motivated the creation of open-domain alternatives. ShareGPT [Chiang et al., 2023] contains conversations collected from a model-hosting website, whereas OpenAssistant [Köpf et al., 2023] and Dolly [Conover et al., 2023] contain crowd-sourced human data. Another intriguing method, popularized by Self-Instruct [Wang et al., 2023b] (and its variants, e.g., Alpaca [Taori et al., 2023]) leverages synthetic datasets. Here, a strong LLM is

\*Equal contribution.

<sup>1</sup>Datasets and models can be found at <https://huggingface.co/collections/PrincetonPLI/instruct-skillmix-66de0821238e34213b4968ec>

prompted using a small set of human-created examples to generate a large number of (query, answer) examples on a variety of topics.

Open evaluations of instruction-following ability have also sprung up. The popular AlpacaEval 2.0 [Dubois et al., 2023, 2024] is based upon curated queries from various sources. In such evaluations, a model’s response to a provided query is compared against a strong reference model’s response, and the model is ranked based upon its *win rate* — the percentage of queries for which the model produces a better answer than the reference model, as judged by a powerful LLM. Rankings on AlpacaEval and related benchmarks like WildBench broadly align with the human rankings of a model’s performance [Dubois et al., 2024, Lin et al., 2024].

## 1.1 Surprising difficulty of instruction tuning

A persistent puzzle in this field is that SFT on the above public datasets does *not* yield good performance on the evaluations. It was initially suspected this is due to a lack of diversity in the training data. But, efforts to produce more diverse synthetic data — e.g., UltraChat [Ding et al., 2023], a synthetic dataset of 1.5M multi-turn conversations created via meticulously tracking lexical and topical diversity as well as coherence — did not significantly improve performance.

Another hypothesis places the blame on the uneven quality of open datasets — which are usually a hodge-podge of collected queries (e.g., Dolly [Conover et al., 2023]) — whereas proprietary datasets are produced to careful specifications using strict quality-control. One finding that supports this hypothesis is that SFT on the 1K Q&A pairs in Alpaca-52K with the longest responses, outperforms SFT on all 52K pairs [Zhao et al., 2024]. In other words, the 51K other data-points are redundant, or even interfere with the “signal” present in the best 1K examples. This finding has inspired “less is more” approaches — including an extreme one based upon just a judicious set of in-context examples [Lin et al., 2023] to provide a surprisingly reasonable level of instruction tuning and alignment — but they did not significantly improve the performance either.

Some have cautioned against hopes for a miracle out of instruction tuning. Gudibande et al. [2023] suggest, based upon careful experiments, that basic capabilities of the LLM arise from pre-training and its massive training corpus. Most deficiencies left after pre-training will not be fixable by, say, a million SFT examples. While this perspective feels broadly correct, it does not quite explain why open efforts to instruction tune Mistral-7B-Base-v0.2 and LLaMA-3-8B-Base fail to match the performance of their proprietary *Instruct* counterparts.

The above difficulties have lately lowered interest level in instruction tuning, with many researchers now turning to RL-based methods (eg, PPO, DPO), which have been used in recent open-source projects to greatly improve proprietary chat models [Meng et al., 2024], which had already trained on expensive human data.

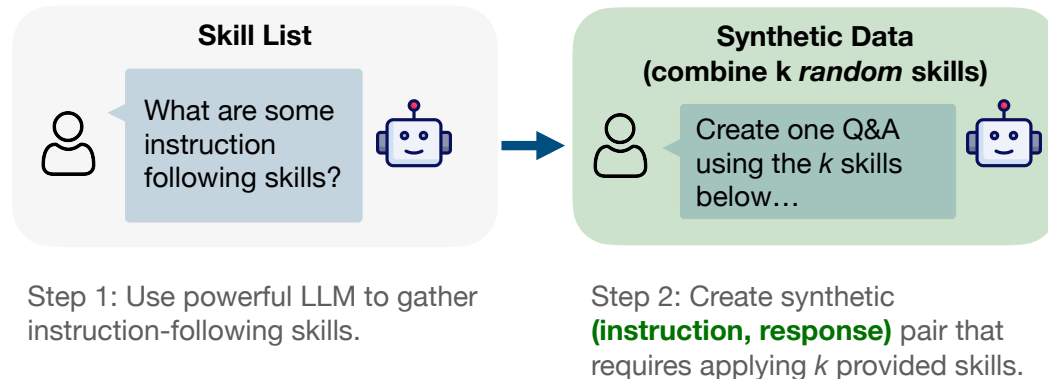


Figure 1: **Sketch of INSTRUCT-SKILLMIX pipeline.** See Figures 2a and 2b for more details on two different implementations of INSTRUCT-SKILLMIX.

## 1.2 Our contributions

We describe a more efficient and effective approach for creating synthetic instruction tuning datasets. Past open efforts invested significant human effort in ensuring *high coverage* of topics and scenarios to sufficiently equip the LLM for scenarios it might encounter at deployment time. We take a subtly different tack. Accepting that pre-training is the dominant source of the LLM’s “inner knowledge,” we focus our instruction tuning on merely teaching the LLM skills to drawing upon that inner knowledge and present it nicely during conversations.

The key idea is to use a strong LLM as a teacher. The recent discovery of *LLM Metacognition* [Didolkar et al., 2024] suggests that frontier models have significant capability to “think about thinking,” which in humans is referred to as *metacognition* [Flavell, 1979]. Specifically, it was shown that given a task dataset, frontier LLMs can help assemble a list of named “skills” needed to solve that task. This requires no human involvement apart from an automated interaction with an LLM<sup>1</sup>.

The first phase (“*Skill Extraction*”) of our pipeline INSTRUCT-SKILLMIX uses this idea and a frontier LLM to identify a list of “basic skills” needed for instruction-following. Unlike Didolkar et al. [2024], which extracts skills from existing SFT datasets, we instead identify skills by directly prompting a strong LLM. (We also tried extracting skills using examples from Alpaca and Ultrachat, and it works quite well, but noticeably worse than our main method.) See Section 2.1.

The second phase of our pipeline, *Data Generation*, uses the list of extracted skills to produce synthetic query-response examples. Here, we repeatedly draw a random pair of skills from the list and prompt the powerful LLM to produce a suitable query that tests those two skills, and to also produce a good response to the query. This generation is inspired by the SKILLMIX evaluation [Yu et al., 2024] for LLMs’ compositional generalization, which also uses a predetermined list of skills. Hence we call our method INSTRUCT-SKILLMIX. See Section 2.2

Using merely 2K to 4K such Q&A examples, vanilla SFT allows popular small base models (Mistral-7B-Base-v0.2, LLaMA-3-8B-Base, and Gemma-2-9B-Base) to match or surpass some apex models on AlpacaEval 2.0, such as the original GPT-4, LLaMA-3.1-405B-Instruct and Claude 3 Opus (Table 1). The estimated cost of creating this 4K dataset using the GPT-4 API is under 600 US dollars.

We stress that although reminiscent of prior efforts using synthetic data such as UltraChat, our pipeline is fully automated with no human design elements (e.g., choice of topics, lexicon etc.). The only human involvement involves the short prompts used for skill extraction and question generation, which we adapted from the math setting of Didolkar et al. [2024]. While our pipeline currently focuses on simple instruction-following, the method seems extensible in future to safety/alignment, as well as domain-specific Q&A.

## 2 INSTRUCT-SKILLMIX

This section describes our methodology for extracting skills from powerful LLMs<sup>1</sup> and how to use these extracted skills to create a diverse, high quality dataset. A simplified version of our pipeline and prompts are depicted in Figures 1 and 2. Section 3 reports the evaluation results when finetuning on this dataset.

### 2.1 Skill Extraction Procedure

The method involves an automated interaction with a frontier LLM (GPT-4-Turbo). We ask the frontier LLM to first generate a list of topics that arise in instruction-following. For each topic returned by the LLM, we further prompt it to generate a list of skills that are needed to answer typical queries on that topic. Additionally, we ask the LLM to create a list of query types (e.g., “Information Seeking”) that might arise in that topic. See Appendix K.4 for details about the prompts used, and Appendix J.2 for the list of all extracted skills. Since this method relies solely upon the LLM’s inner meta-knowledge, this method should extend easily to other types of instruction-following.

<sup>1</sup>Skill lists generated by different frontier models are related but not isomorphic. Skills generated by one model are comprehensible to other models. See Didolkar et al. [2024] for such experiments.

<sup>1</sup>We use GPT-4-Turbo, but any powerful LLM can be used in the pipeline. GPT-4-Turbo refers to the 2024-04-09 checkpoint throughout the paper unless specified otherwise.

Table 1: **Evaluation results of *base* models supervised-finetuned on INSTRUCT-SKILLMIX versus the proprietary *instruct* versions and other proprietary models.** For our models, we report the results for best checkpoint selected using held-out queries. For other models(\*), we report the published numbers available on publicly available leaderboards. “# Data” refers to the number of (instruction, response) pairs in the training data. See Table 7 for a more detailed view, including comparisons to past open datasets.

Model	# Data	AlpacaEval2.0 LC WR(%)	WildBench WB-Reward <sub>gpt4t</sub> <sub>∞</sub>
<b>LLaMA-3-8B</b>			
Ours	4K	<b>42.76</b>	<b>-36.91</b>
*LLaMA-3-8B-Instruct	-	22.90	-46.30
<b>Mistral-7B-v0.2</b>			
Ours	4K	<b>36.70</b>	<b>-29.25</b>
SFT on Alpaca-52K	52K	8.64	-80.47
*Mistral-7B-Instruct-v0.2	-	17.10	-54.70
<b>Gemma-2-9B</b>			
Ours	2K	36.18	-37.83
Gemma-2-9B-Instruct	-	<b>37.21</b>	<b>-28.78</b>
<b>*Other Proprietary Models</b>			
LLaMA-3.1-405B-Instruct	-	39.30	-
Mistral Large	-	32.70	-46.40
Claude 3 Opus	-	40.50	-21.20
Claude 3 Sonnet	-	34.90	-30.30
GPT-4-Omni (2024-05-13)	-	<b>57.50</b>	<b>+1.70</b>
GPT-4 (2023-03-14)	-	35.30	-

**An Earlier Attempt:** Our initial attempt to extract skills leveraged existing instruction tuning datasets, which is a more direct analog of the method in Didolkar et al. [2024]. However, we suspected this to be sub-optimal due to known limitations of past instruction tuning datasets. We therefore designed the method described above, and found it superior. It also has scientific benefit of being independent of existing datasets like Alpaca and Ultrachat. However, the dataset from the initial method, called INSTRUCT-SKILLMIX-SEED-DATASET-DEPENDENT (INSTRUCT-SKILLMIX-D; see Appendix A) is still very useful for ablations that pinpoint ways in which our skill-based pipeline improves on past synthetic datasets for instruction-following (see Tables 2, 3, 4, and 5).

## 2.2 Data Generation

Inspired by the recent SKILLMIX evaluation [Yu et al., 2024], we generate instruction-following examples by randomly picking  $k$  skills as well as a random query type. The frontier LLM is prompted to create Q&A pairs that illustrate these  $k$  skills and the query type. We refer to the resulting dataset as INSTRUCT-SKILLMIX. For example, INSTRUCT-SKILLMIX( $k=2$ )-1K refers to 1,000 examples of data created from random combinations of  $k = 2$  skills. See Appendix K.3 and K.5 for the details about the prompts used for data generation.

See Appendix A for more details, and an estimate of the low cost of INSTRUCT-SKILLMIX pipeline.

**Where does diversity come from?** The first source of diversity is the skill labels. A skill label represents some part of the frontier LLM’s meta-knowledge of human behavior and needs, which it observed in its vast training set or during instruction tuning. Replacing a concrete Q&A example with a skill label converts it into a pointer to a region in the frontier LLM’s meta-knowledge, which the model can then freely draw upon to create new examples. The second source of diversity is the use of random  $k$ -tuples of skills when generating new examples. The motivation here is that, in most cases, distinct tuples will lead to very distinct flavor of examples.

For instance, the skill pair (critical thinking and communication, literature and language skills) leads to the following instruction

I'm a high school English teacher aiming to develop a curriculum unit for my 11th-grade class, focusing on American literature. I want this unit to go beyond just reading and understanding the texts. Specifically, I'm looking to enhance my students' critical thinking and communication skills through engaging activities related to the literature. Can you suggest detailed ways to incorporate these skills, ideally with concrete examples and expected learning outcomes?

whereas the skill pair (critical thinking and communication, skill in virtual and system design) leads to

As an IT manager, I am overseeing the development of a virtual workspace to enhance communication and efficiency among remote teams. This workspace must support multimedia content, including video conferencing and live document editing. What are the critical steps I should take in its design and implementation, balancing technical robustness with ease of use? Could you provide specific technologies to consider and any potential obstacles?

Even though the two skill pairs share a common skill, they lead to rather distinct Q&A pairs, involving creative and nuanced situations with subtle moving parts. Since the number of  $k$ -tuples scales as  $\binom{N}{k}$ , where  $N$  is the number of skills, using pairs of skills foster a lot of diversity — e.g., 125,000 possibilities with  $N = 500, k = 2$ . The pipeline in our experiments mainly uses  $k = 2$ , but generating answers to these queries will certainly end up using many other unnamed skills as well, and thus serve as a rich source for learning how to follow instructions.

### 3 Experiments

#### 3.1 Experimental Setup

**SFT on INSTRUCT-SKILLMIX(k).** We finetune LLaMA-3-8B-Base, Mistral-7B-Base-v0.2, Gemma-2-9B-Base, LLaMA-2-7B-Base, and LLaMA-2-13B-Base on a varying number of examples from INSTRUCT-SKILLMIX-D(k) and INSTRUCT-SKILLMIX(k). We train for multiple epochs and select the best checkpoint by performance on a 100 held-out questions. Similar to Ouyang et al. [2022], Zhou et al. [2023], we observe that using cross entropy-loss on a validation set does not lead to the best checkpoint. See Appendix D.2 for a more detailed discussion of the checkpoint selection procedure. As a baseline, we also finetune on different subsets of Alpaca-52K, including the 1K or 5K examples with the longest completions. For further training details (e.g., hyperparameters), see Appendix D.1.

**Evaluation.** We evaluate our models on popular instruction following benchmarks: AlpacaEval 2.0 [Dubois et al., 2024], MT-Bench [Zheng et al., 2023], and WildBench [Lin et al., 2024]. For AlpacaEval, we report the length-controlled win rate (LC WR) of the responses of our model against a reference response, which corrects for the length bias of the judge model. For MT-Bench, we report the average score of the responses of our model graded by a judge model. For WildBench, we report the WB-Reward (weighted win-rate) of the response of our model against one reference response when graded by a judge model. For further evaluation details, see Appendix C. See Table 8 in Appendix B for evaluations on additional benchmarks.

#### 3.2 Main Results

For the main results of the paper, we report the evaluation results when models are finetuned on INSTRUCT-SKILLMIX in Table 1, and summarize our findings below. For a more detailed version of Table 1, see Table 7. For additional ablations, see Appendix E. For evaluations on other LLM benchmarks, see Table 8.

**INSTRUCT-SKILLMIX achieves SOTA performance amongst SFT models.** LLaMA-3-8B-Base finetuned on 4K examples from INSTRUCT-SKILLMIX(k=2) yields LC win rate of 42.76% on AlpacaEval 2.0. This score is higher than Claude 3 Opus, LLaMA-3.1-405B-Instruct, and GPT-4 (2023-03-14). Mistral-7B-Base-v0.2 finetuned on the same data achieves -29.25 on WildBench, which outperforms Claude 3 Sonnet and Mistral Large. Gemma-2-9B-Base finetuned on 2K examples from INSTRUCT-SKILLMIX(k=2) gets a score of 8.12 on MT-Bench, which is better than GPT-3.5-Turbo

(2023-03-01). To best of our knowledge, these scores are much higher than any base model that has *only* undergone supervised instruction finetuning (i.e., no RLHF, DPO, PPO, or variants).

**Early saturation.** Performance from our method rises rapidly, reaching unprecedented levels with 1K examples. Unfortunately, improvements stop already with 4K examples. This turns out to be a consequence of its high efficiency at inducing good instruction-following. Specifically, with 4K examples, the win-rate against GPT-4 approaches 50% on *heldout* queries from our pipeline, and thus overfitting sets in.

**Observed limitations.** The open benchmarks used in this study have known limitations, related to the insufficient number of under-specified or ambiguous queries, and no testing of long-form generations such as multi-page essays. Our current pipeline shares some of these limitations. Fixing this seems very doable via suitable modification to our INSTRUCT-SKILLMIX pipeline, but this is left for future work. This aligns with the observation in Bai et al. [2024] that a model’s effective generation length seems to be limited by the typical length of examples seen during SFT, and is exacerbated by the relative scarcity of long-form samples in the SFT data. This underscores the critical influence of training data composition on a model’s post-fine-tuning capabilities, and would be interesting to investigate in future work.

## 4 Ablation study

Whereas pretraining is the source of an LLM’s basic capabilities [Gudibande et al., 2023], the sole goal of instruction tuning is to impart skills, such as answer-structuring, empathy, helpfulness, etc. Vanilla SFT on Q&A data generated by a teacher LLM is akin to *imitation learning*. Our ablation studies below help understand the contribution of different elements to the effectiveness of imitation learning method using INSTRUCT-SKILLMIX Q&A. The main finding is that the source of largest improvement is the skill extraction step.

### 4.1 Benefits of Skill Extraction (with mixing turned off)

To highlight the benefits of our skill-based method versus current synthetic approaches, we use the pioneering Alpaca dataset, whose responses are rewritten by GPT-4 (2023-03-14) [Peng et al., 2023]. The fairest comparison here would be with our INSTRUCT-SKILLMIX-D(k=1) data, where the underlying skills were derived from a random sample of *Alpaca-52K*, and each of our datapoints uses one of those extracted skills. All results below involve finetuning Mistral-7B-Base-v0.2 on different subsets of the Alpaca-52K dataset : (1) *Alpaca-1K Longest*: 1,000 examples with the longest responses [Zhao et al., 2024]; (2) *Alpaca-5K Longest*: 5,000 examples with the longest responses; (3) *Alpaca-5K Random*: 5,200 randomly sampled examples from which we extracted our skills; and (4) *Alpaca-52K*: the full 52,002 examples.

Table 2: **Evaluation results of Mistral-7B-Base-v0.2 finetuned on INSTRUCT-SKILLMIX-D vs. on Alpaca-52K.** Note that skills extracted from Alpaca-5K Random were used to create the INSTRUCT-SKILLMIX-D datasets.

SFT Dataset	# Data	AlpacaEval 2.0 LC WR(%)	MT-Bench	WildBench WB-Reward $_{\infty}^{\text{gpt4t}}$
INSTRUCT-SKILLMIX-D(k=2)	4K	<b>29.77</b>	7.17	<b>-39.06</b>
INSTRUCT-SKILLMIX-D(k=1)	1K	27.04	<b>7.22</b>	-46.83
Alpaca-1K Longest	1K	10.09	6.88	-63.38
Alpaca-5K Longest	5K	8.92	6.90	-62.55
Alpaca-5K Random	5K	11.10	6.86	-74.41
Alpaca-52K Full	52K	8.64	6.45	-80.47

As shown in Table 2, finetuning on 1,000 examples with the longest completions from Alpaca-52K yields 10.09% LC win rate on AlpacaEval 2.0. On the other hand, finetuning on only 1K examples of INSTRUCT-SKILLMIX-D(k=1) yields 27.04% LC win rate. Note that since the skills in INSTRUCT-SKILLMIX-D are mostly derived from Alpaca-52K, the observed improvements in the win rate are indicative of the improved quality of INSTRUCT-SKILLMIX-D queries.

## 4.2 Mixing skills helps, but not as much as skill extraction

In Table 3, models finetuned on INSTRUCT-SKILLMIX-D( $k=2$ ) data marginally outperform models SFT on INSTRUCT-SKILLMIX-D( $k=1$ ) on AlpacaEval and WildBench, whereas performance on MT-bench is about the same. The marginal improvements from increasing  $k$  are less noticeable for INSTRUCT-SKILLMIX.

Table 3: **Evaluation results of Mistral-7B-Base-v0.2 SFT on INSTRUCT-SKILLMIX where  $k=1$  vs.  $k=2$ .** In each entry, we report **INSTRUCT-SKILLMIX-D/INSTRUCT-SKILLMIX**

Model	# Data	AlpacaEval 2.0		MT-Bench	WildBench
		WR(%)	LC WR(%)		WB-Reward <sub>∞</sub> <sup>gpt4t</sup>
SFT on INSTRUCT-SKILLMIX(k=2)					
Mistral-7B-Base-v0.2	1K	33.87/42.48	27.48/38.34	6.92/7.33	-41.46/-30.65
	2K	37.05/40.83	31.57/36.18	7.04/7.20	-43.46/-31.92
	4K	35.08/40.74	29.77/36.70	7.17/7.16	-39.06/-29.25
SFT on INSTRUCT-SKILLMIX(k=1)					
Mistral-7B-Base-v0.2	1K	30.06/41.75	27.04/38.34	7.22/7.49	-46.83/-30.95
	2K	35.07/-	31.66/-	7.39/-	-46.97/-
	4K	33.57/-	28.85/-	7.13/-	-44.43/-

## 4.3 Quality of Queries (and Skills) Matters

The effectiveness of this approach depends on the quality of the queries used in the fine-tuning process, where high-quality queries enable the frontier LLM teacher to provide richer instruction to the student model undergoing instruction tuning. This relationship between the quality of queries and the skills being imparted is supported by two key observations. First, the frontier LLM proves to be a more effective teacher when the skill list being used was also entirely generated using its help (as opposed to giving it skills derived from existing datasets). Across all model types, dataset size, and the evaluation benchmark, we generally see an improvement when finetuning on INSTRUCT-SKILLMIX compared to INSTRUCT-SKILLMIX-D (see Table 7 for more details). Second, incorporating these sub-optimal skills from existing datasets as a part of “teaching” (e.g., with INSTRUCT-SKILLMIX-D) is still more effective than using an equal number of random (or even the longest examples) from Alpaca-52K when responses are also by the same frontier LLM.

These findings suggest that the quality of the queries (and the skills used to create those queries) drives how well data generated by the frontier LLM is able to impart its skills on the model undergoing instruction tuning.

## 4.4 Effect of Teacher and Grader

SFT performance derives from the model used to generate Q&A data, which plays the *teacher* role in imitation learning. The student’s performance is evaluated by the grader model. The main results reported in this paper used GPT-4-Turbo as the teacher, and some checkpoint of GPT-4 or GPT-4-Turbo as the grader.

**Effect of the teacher** Many SFT efforts in 2023 used earlier versions of GPT-4 or GPT-3.5, which were weaker than GPT-4-Turbo. To pin-point the effect of this change, we try doing a head-to-head comparison once we fix the teacher. The responses in Alpaca-1K Longest are written by GPT-4 (2023-03-14), whereas INSTRUCT-SKILLMIX data is generated by GPT-4-Turbo. Thus, we use GPT-4-Turbo to regenerate answers to Alpaca-1K Longest [Zhao et al., 2024], and we also use GPT-4 (2023-03-14) to regenerate INSTRUCT-SKILLMIX-D.

Table 4 compares the performance of Mistral-7B-Base-v0.2 when finetuned on the two datasets using the two versions of GPT-4. For each fixed data generator model, the INSTRUCT-SKILLMIX dataset leads to a better performance. Furthermore, replacing GPT-4 with the stronger GPT-4-Turbo in data generation makes INSTRUCT-SKILLMIX pull even further ahead of Alpaca-1K Longest, which highlights that our pipeline is better positioned than Alpaca dataset to elicit better supervision from a more powerful LLM teacher.

Table 4: **Evaluation results of Mistral-7B-Base-v0.2 finetuned on INSTRUCT-SKILLMIX-D vs. Alpaca-1K Longest generated from two different versions of GPT-4.** For a fixed data generator model, SFT Mistral-7B-Base-v0.2 on INSTRUCT-SKILLMIX-D outperforms SFT on Alpaca-1K Longest.

Model for Data Generation	Dataset	AlpacaEval 2.0		MT-Bench
		WR(%)	LC WR(%)	
GPT-4 (2023-03-14)	Alpaca-1K Longest	12.75	10.09	6.83
	INSTRUCT-SKILLMIX-D-1K	13.29	15.01	7.10
GPT-4-Turbo (2024-04-09)	Alpaca-1K Longest	35.23	19.62	6.99
	INSTRUCT-SKILLMIX-D-1K	33.87	27.48	6.92

**Effect of choice of grader** We use GPT-4-Turbo to generate data and AlpacaEval 2.0 uses GPT-4 for grading, creating a scenario where both the teacher model and grader model are from the same family. This raises the question of whether model family overlap leads to a potential grading bias and inflated scores. To quantify this effect, we used Claude 3 Opus as the grader for AlpacaEval 2.0. Table 5 shows that although Claude is a more generous grader across the board, it generally preserves the relative rankings among the models. Importantly, it exhibits even stronger preference for our student models’ generations than does GPT-4.

Table 5: **Evaluation results when using two different graders for AlpacaEval 2.0.** Relative ranking of evaluated models are generally preserved when using different graders. Here, ISM-D refers to INSTRUCT-SKILLMIX-D.

Model	Grader: GPT-4 (2023-11-06)		Grader: Claude 3 Opus	
	WR(%)	LC WR(%)	WR(%)	LC WR(%)
Mistral-7B-Base-v0.2 SFT on ISM-D-1K	33.87	27.48	50.56	38.50
Mistral-7B-Base-v0.2 SFT on ISM-D-2K	37.05	31.57	48.94	38.29
Mistral-7B-Base-v0.2 SFT on ISM-D-4K	35.08	29.77	52.55	44.16
(Reference Model) LLaMA-3-70B-Instruct	33.20	34.40	39.68	42.33
(Reference Model) Mistral-7B-Instruct-v0.2	14.70	17.10	15.16	18.89
(Reference Model) LLaMA-2-70B-Chat	13.90	14.70	16.67	17.85

## 5 Effect of Low Quality Data

Our fully synthetic pipeline produces a large number of high-quality questions and answers that look impressive but also (for want of a better word) “robotic.” Data sourced from human workers shows greater variation, and one begins to wonder if that additional diversity could be beneficial. We tried interventions such as generating 20% using a different prompt — e.g., require a shorter answer, or a poor quality answer. In a human pipeline, this variation would be expected. We can think of this as “data from shirkers,” and one would expect a fair bit of it in naive crowdsourcing. (In corporate settings it would be mitigated via quality control measures.) See Appendix H for an example of a poor quality response.

We replace 20% of the responses in INSTRUCT-SKILLMIX(k=2)-2K with short responses (“respond in one paragraph”) to create BREV-INSTRUCT-SKILLMIX(k=2)-2K. Finetuning Mistral-7B-Base-v0.2 on BREV-INSTRUCT-SKILLMIX-D was surprising: brevity constraint on just 20% of data almost halved the average response length on AlpacaEval, from 2817 to 1746 characters. LC win rate dropped from 31.57% to 23.93%.

We alternatively replace 20% of the responses in the same datasets with responses that are still long but have poor quality (i.e., deliberately sloppy and unhelpful) to create JUNK-INSTRUCT-SKILLMIX(k=2)-2K. Mistral-7B-Base-v0.2 finetuned on the JUNK-INSTRUCT-SKILLMIX-D yields less than 1% win rate on AlpacaEval and 5.01 on MT-Bench.

**Lower-quality data harms performance.** As shown in Table 6, replacing just 20% of the data with poor quality responses harms performance. For INSTRUCT-SKILLMIX-D, the harm is super-proportionate. These observation may help explain why creating open-domain instruction tuning data has proved so difficult via naive crowd-sourcing.



Table 6: **Evaluation results of models finetuned on low quality INSTRUCT-SKILLMIX.** Replacing just 20% of the dataset with low quality data has a super-proportionate harm on the model performance. Amount of harm greatly differs between the two versions of the pipeline.

Model	# Data	AlpacaEval 2.0		MT-Bench	WildBench
		LC WR(%)	Avg Len		WB-Reward <sub>gpt4t</sub> <sup>∞</sup>
SFT on INSTRUCT-SKILLMIX-D(k=2)					
Mistral-7B-Base-v0.2	2K	31.57	2817	7.04	-43.46
	2K (Brevity 20%)	23.93	1746	6.69	-49.85
	2K (Junk 20%)	0.77	1104	5.01	-47.50
SFT on INSTRUCT-SKILLMIX(k=2)					
Mistral-7B-Base-v0.2	2K	36.18	2936	7.20	-31.92
	2K (Brevity 20%)	31.61	2336	7.32	-32.27
	2K (Junk 20%)	24.60	2435	6.90	-47.50

**High-quality data’s protective effect.** While adding some low-quality data to INSTRUCT-SKILLMIX already causes a noticeable performance drop, doing the same to INSTRUCT-SKILLMIX-D is catastrophic. This suggests that INSTRUCT-SKILLMIX is more robust to “shirkers,” corroborating our previous observations in Table 7 of the superior performance of INSTRUCT-SKILLMIX over INSTRUCT-SKILLMIX-D. This finding suggests that higher quality data can somewhat protect against negative effects of “shirkers,” which needs further study.

## 6 Related Work

Prior works observe improvements from instruction finetuning on *fewer*, but *higher quality* data generated by humans [Zhou et al., 2023, Touvron et al., 2023]. However, efforts to curate high quality data from humans are quite expensive, and licensing can become complicated. This has led to an increase in the popularity of semi-automated and less expensive approaches.

**Selecting high quality data.** Synthetic data creation has become a predominant approach for curating instruction tuning datasets, especially in the academic realm [Wang et al., 2023b, Dubois et al., 2023, Xu et al., 2024, Gunasekar et al., 2023]. These synthetic datasets are generally created by providing in-context examples to a powerful LLM to produce the synthetic data, followed by some post-filtering [Wang et al., 2023b]. Recent efforts have also focused on data selection strategies for high quality subsets of the original dataset, which lead to performance gains [Tunstall et al., 2023, Chen et al., 2024, Liu et al., 2024, Zhao et al., 2024]. Notably, Zhao et al. [2024] show that finetuning on just the 1K longest completions from Alpaca-52K outperforms finetuning on the entire Alpaca-52K dataset. Whereas the data selection methods just described focus on *general-purpose* instruction tuning, Xia et al. [2024] explore an optimizer-aware data selection strategy for *targeted* instruction tuning.

**Encouraging data diversity.** Common approaches to elicit diversity in datasets include mixing multiple datasets [Wang et al., 2022, Longpre et al., 2023, Wang et al., 2023a], as well as rewriting the data in multiple ways and changing formatting [Allen-Zhu and Li, 2024, Honovich et al., 2023]. The Self-Instruct framework [Wang et al., 2023b] and variants such as Alpaca-52K [Dubois et al., 2023] encourage diversity by identifying similar pairs using ROUGE-L similarity. Other approaches to ensure diversity impose constraints on the topic in order to enhance wide coverage [Ding et al., 2023, Xu et al., 2024], or require synthetic data to use a random subset of words or concepts chosen from some vocabulary [Eldan and Li, 2023, Gunasekar et al., 2023, Li et al., 2024]. The latter approach is also suggested by recent work that provides a mathematical model for emergence via LLM scaling [Arora and Goyal, 2023] and used in the evaluation setting in Yu et al. [2024].

**AlpacaEval.** AlpacaEval [Li et al., 2023, Dubois et al., 2024] is a popular evaluation for assessing instruction-following capabilities of LLMs. The tested model provides answers 805 carefully curated instructions, and its answers are compared against reference outputs of a designated baseline model. For each instruction, another evaluator LLM outputs a preference between the two responses (output of the model being evaluated vs. reference output by the baseline mode). The primary evaluation

metric is the *win rate*, which represents the expected probability that the grader model favors the response generated by the evaluated model over the response produced by the baseline model. Given that a raw win rate shows bias towards longer responses, AlpacaEval 2.0 [Dubois et al., 2024] introduces the *length-corrected (LC) win rate* as a proxy for what the raw win rate would be if the evaluated model’s response lengths and baseline model’s response lengths matched.

**WildBench.** WildBench [Lin et al., 2024] is another benchmark for assessing the instruction following capabilities of LLMs. Unlike the AlpacaEval instructions, 50% of which are only “information seeking” type questions, the instructions for WildBench cover a more diverse distribution of task categories, including coding and creative writing. Whereas the grading in AlpacaEval is more liberal (since there is no penalty for poor responses), the grading in WildBench is more finegrained: a model answer is compared against a reference answer, but is graded on a scale of (1) win by a big margin, (2) win by a small margin, (3) tie, (4) lose by a small margin, and (5) lose by a big margin. This ensures that models that output bad answers to certain types of questions are penalized.

**RL-inspired approaches.** Since we do not use RL, we defer discussion of these approaches to Appendix F.

## 7 Conclusion

While one would have certainly expected the cost factor as well as scaling ability to ultimately favor synthetic data, the surprising finding in this paper is that, when done well, synthetic data can be much more *effective* than human data for instruction tuning. Our INSTRUCT-SKILLMIX pipeline, uses the recent discovery of LLM Metacognition [Didolkar et al., 2024] to extract skills using a powerful LLM and then leverages an LLM to create quality instruction data using random pairs of those skills.

Vanilla SFT of base models on just 1K to 4K examples from our pipeline outperforms the proprietary *instruct* versions of the same model, as well as older and larger instruction tuning efforts like Vicuna and Ultrachat that used orders of magnitude more datapoints. The performance also approaches those of frontier models, which trained on expensive human data as well as with RL techniques. Unfortunately, our method saturates at 4K examples, when win-rate on heldout queries approaches 50%.

Ablation studies in Section 4.4 rule out potential confounding factors, such as the use of a strong teacher, or bias due to teacher and grader belonging to the same family. These ablations reinforce that the improvement is primarily due to the uniformly high quality of examples produced by our skill-based pipeline. Each example contains a query with nontrivial scenarios and lots of moving parts, which improve imitation learning.

Section 5 offers a preliminary exploration of pitfalls of naive collection of instruction tuning data. In particular, the presence of some lower quality data noticeably harms the model’s performance. This insight should be more rigorously investigated, including via new theory. The experiment also suggests that our less preferred INSTRUCT-SKILLMIX-D method (which involves extracting skills from an existing dataset) is more susceptible to such bad data than our preferred INSTRUCT-SKILLMIX.

One potential benefit of INSTRUCT-SKILLMIX-D may be that it gives some insights into an efficient method for dataset distillation [Wang et al., 2020] for text datasets, which has not yet proved possible.

Finally, it should be noted that our results look stronger on paper than they actually are. Open evaluations such as AlpacaEval 2.0 have blind spots, especially the fact that win rate of even 50% against a frontier model still allows unacceptably high frequency of unsuitable responses in a deployment setting. The new WildBench evaluation does test for more corner cases. We hope that INSTRUCT-SKILLMIX ideas can also leverage LLM metacognition to create a better evaluation.

Although our SFT data does not address safety and alignment, our skill-based ideas may be useful there. A related next step would be to leverage our ideas of skill extraction to improve RL-based methods (whether for instruction-following or alignment). We hope to address these in future work.

## References

- Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws, 2024. URL <https://arxiv.org/abs/2404.05405>.
- Sanjeev Arora and Anirudh Goyal. A theory for emergence of complex skills in language models, 2023. URL <https://arxiv.org/abs/2307.15936>.
- Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. Longwriter: Unleashing 10,000+ word generation from long context llms, 2024. URL <https://arxiv.org/abs/2408.07055>.
- Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. Alpapasus: Training a better alpaca with fewer data. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=FdVXgSJhvz>.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality, March 2023. URL <https://lmsys.org/blog/2023-03-30-vicuna/>.
- Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world’s first truly open instruction-tuned llm, 2023. URL <https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm>.
- Aniket Didolkar, Anirudh Goyal, Nan Rosemary Ke, Siyuan Guo, Michal Valko, Timothy Lillicrap, Danilo Rezende, Yoshua Bengio, Michael Mozer, and Sanjeev Arora. Metacognitive capabilities of llms: An exploration in mathematical problem solving, 2024. URL <https://arxiv.org/abs/2405.12205>.
- Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 3029–3051. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.183. URL <https://doi.org/10.18653/v1/2023.emnlp-main.183>.
- Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. AlpacaFarm: A simulation framework for methods that learn from human feedback. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL [http://papers.nips.cc/paper\\_files/paper/2023/hash/5fc47800ee5b30b8777fdd30abcaaf3b-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2023/hash/5fc47800ee5b30b8777fdd30abcaaf3b-Abstract-Conference.html).
- Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaEval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024. URL <https://arxiv.org/abs/2404.04475>.
- Ronen Eldan and Yuanzhi Li. Tinstories: How small can language models be and still speak coherent english?, 2023. URL <https://arxiv.org/abs/2305.07759>.
- John H. Flavell. Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34:906–911, 1979. URL <https://psycnet.apa.org/record/1980-09388-001>.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika,

- Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 12 2023. URL <https://zenodo.org/records/10256836>.
- Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. The false promise of imitating proprietary llms, 2023. URL <https://arxiv.org/abs/2305.15717>.
- Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. Textbooks are all you need, 2023. URL <https://arxiv.org/abs/2306.11644>.
- Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. Unnatural instructions: Tuning language models with (almost) no human labor. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2023, Toronto, Canada, July 9-14, 2023, pages 14409–14428. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.806. URL <https://doi.org/10.18653/v1/2023.acl-long.806>.
- Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. Openassistant conversations - democratizing large language model alignment. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL [http://papers.nips.cc/paper\\_files/paper/2023/hash/949f0f8f32267d297c2d4e3ee10a2e7e-Abstract-Datasets\\_and\\_Benchmarks.html](http://papers.nips.cc/paper_files/paper/2023/hash/949f0f8f32267d297c2d4e3ee10a2e7e-Abstract-Datasets_and_Benchmarks.html).
- Haoran Li, Qingxiu Dong, Zhengyang Tang, Chaojun Wang, Xingxing Zhang, Haoyang Huang, Shaohan Huang, Xiaolong Huang, Zeqiang Huang, Dongdong Zhang, Yuxian Gu, Xin Cheng, Xun Wang, Si-Qing Chen, Li Dong, Wei Lu, Zhifang Sui, Benyou Wang, Wai Lam, and Furu Wei. Synthetic data (almost) from scratch: Generalized instruction tuning for language models, 2024. URL <https://arxiv.org/abs/2402.13064>.
- Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. AlpacaEval: An automatic evaluator of instruction-following models. [https://github.com/tatsu-lab/alpaca\\_eval](https://github.com/tatsu-lab/alpaca_eval), 2023.
- Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. The unlocking spell on base llms: Rethinking alignment via in-context learning, 2023. URL <https://arxiv.org/abs/2312.01552>.
- Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. Wildbench: Benchmarking llms with challenging tasks from real users in the wild, 2024. URL <https://arxiv.org/abs/2406.04770>.
- Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=BTKAeLqLMw>.
- Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. The flan collection: Designing data and methods for effective instruction tuning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, volume 202 of Proceedings of Machine Learning Research*, pages 22631–22648. PMLR, 2023. URL <https://proceedings.mlr.press/v202/longpre23a.html>.

- Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward, 2024. URL <https://arxiv.org/abs/2405.14734>.
- OpenAI. Our approach to alignment, 2022. URL <https://openai.com/index/our-approach-to-alignment-research/>.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL [http://papers.nips.cc/paper\\_files/paper/2022/hash/blefde53be364a73914f58805a001731-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2022/hash/blefde53be364a73914f58805a001731-Abstract-Conference.html).
- Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4, 2023. URL <https://arxiv.org/abs/2304.03277>.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL [http://papers.nips.cc/paper\\_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html).
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL <http://arxiv.org/abs/1707.06347>.
- Gokul Swamy, Christoph Dann, Rahul Kidambi, Zhiwei Steven Wu, and Alekh Agarwal. A minimaximalist approach to reinforcement learning from human feedback, 2024. URL <https://arxiv.org/abs/2401.04056>.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. [https://github.com/tatsu-lab/stanford\\_alpaca](https://github.com/tatsu-lab/stanford_alpaca), 2023.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL <https://arxiv.org/abs/2307.09288>.
- Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Cl  mentine Fourier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. Zephyr: Direct distillation of lm alignment, 2023. URL <https://arxiv.org/abs/2310.16944>.
- Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A. Efros. Dataset distillation, 2020. URL <https://arxiv.org/abs/1811.10959>.

- Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hannaneh Hajishirzi, and Daniel Khashabi. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.
- Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. How far can camels go? exploring the state of instruction tuning on open resources. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023a. URL [http://papers.nips.cc/paper\\_files/paper/2023/hash/ec6413875e4ab08d7bc4d8e225263398-Abstract-Datasets\\_and\\_Benchmarks.html](http://papers.nips.cc/paper_files/paper/2023/hash/ec6413875e4ab08d7bc4d8e225263398-Abstract-Datasets_and_Benchmarks.html).
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023b.
- Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. In *International Conference on Representation Learning (ICLR)*, 2022.
- Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. Self-play preference optimization for language model alignment, 2024. URL <https://arxiv.org/abs/2405.00675>.
- Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. Less: Selecting influential data for targeted instruction tuning. In *International Conference on Machine Learning (ICML)*, 2024.
- Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions. In *International Conference on Representation Learning (ICLR)*, 2024.
- Dingli Yu, Simran Kaur, Arushi Gupta, Jonah Brown-Cohen, Anirudh Goyal, and Sanjeev Arora. Skill-mix: a flexible and expandable family of evaluations for ai models. In *International Conference on Representation Learning (ICLR)*, 2024.
- Hao Zhao, Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Long is more for alignment: A simple but tough-to-beat baseline for instruction fine-tuning. In *International Conference on Machine Learning (ICML)*, 2024.
- Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. Pytorch fsdp: Experiences on scaling fully sharded data parallel, 2023. URL <https://arxiv.org/abs/2304.11277>.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *CoRR*, abs/2306.05685, 2023. doi: 10.48550/ARXIV.2306.05685. URL <https://doi.org/10.48550/arXiv.2306.05685>.
- Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke

Zettlemoyer, and Omer Levy. LIMA: less is more for alignment. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL [http://papers.nips.cc/paper\\_files/paper/2023/hash/ac662d74829e4407ce1d126477f4a03a-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2023/hash/ac662d74829e4407ce1d126477f4a03a-Abstract-Conference.html).

# List of Appendices

A	INSTRUCT-SKILLMIX Pipeline (More Details)	17
B	Full Evaluation Results (More Detailed)	18
C	Evaluation Details	20
D	Training Details	21
E	Ablations	23
F	INSTRUCT-SKILLMIX is Competitive with RL-Inspired Methods.	26
G	Robustness of INSTRUCT-SKILLMIX Across Random Skill Combinations for SFT	27
H	Examples of BREV-INSTRUCT-SKILLMIX and JUNK-INSTRUCT-SKILLMIX	28
I	Stats on Different Datasets	29
J	List of Skills	30
K	Skill Extraction Prompts	68
L	Comparison of Responses	76



## A INSTRUCT-SKILLMIX Pipeline (More Details)

### A.1 INSTRUCT-SKILLMIX-D and INSTRUCT-SKILLMIX Pipelines

**Method 1: Leveraging existing instruction datasets.** Even though existing instruction-following datasets may not induce good chat capability via vanilla SFT, these datasets still exhibit (possibly in an uneven fashion) some “skills” needed by the model. Thus, we adapt the methodology presented in Didolkar et al. [2024] and use GPT-4-Turbo to extract instruction-following skills from random samples of existing instruction and alignment datasets (5,200 samples from Alpaca-52K and 1,000 samples from UltraChat). We then use GPT-4-Turbo to cluster similar skills into broader categories, forming our final list of instruction-following skills. See Appendix J.1 for the list of all extracted skills and Appendix K.1 and K.2 for details about the prompts used for skill extraction.

**Method 2: Directly prompting a powerful LLM.** While Method 1 works surprisingly well, it generated unease about possibly relying on existing seed datasets of uneven quality, and thus potentially inheriting their limitations and biases. Therefore we also tried an alternative pipeline that solely relies on the powerful LLM’s ideas about list of skills it leverages for instruction-tuning.

We will refer to the datasets generated from the seed-dataset dependent and the seed-dataset agnostic versions as INSTRUCT-SKILLMIX-SEED-DATASET-DEPENDENT and INSTRUCT-SKILLMIX-SEED-DATASET-AGNOSTIC, respectively. Unless stated otherwise, INSTRUCT-SKILLMIX refers to the INSTRUCT-SKILLMIX-SEED-DATASET-AGNOSTIC data.

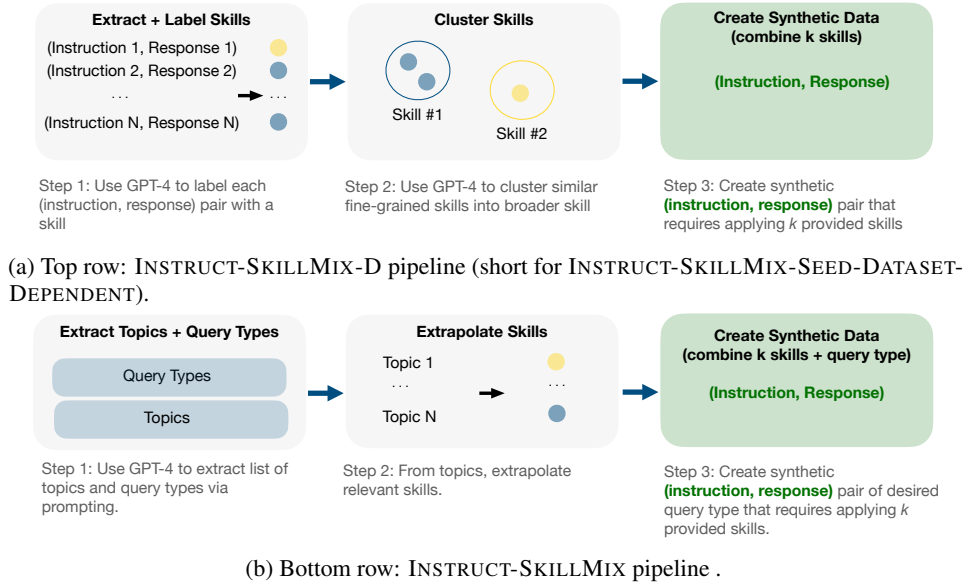


Figure 2: **Two variants of the INSTRUCT-SKILLMIX pipeline.** INSTRUCT-SKILLMIX(k) involves two steps: (1) skill extraction using similar ideas as Didolkar et al. [2024]; (2) data generation from random  $k$ -tuples of skills.

### A.2 Dataset Curation Costs

Generating synthetic data using the INSTRUCT-SKILLMIX pipeline is more cost effective compared to using human annotators. To extract the skill clusters for INSTRUCT-SKILLMIX-D, it costs less than \$120 to extract and cluster skills from 6,200 examples from various existing datasets. For INSTRUCT-SKILLMIX, extracting skills via direct prompting costs under \$5. Additionally, producing 4,000 examples of INSTRUCT-SKILLMIX(k=2) data costs under \$570.

## B Full Evaluation Results (More Detailed)

Table 7 contains the full evaluation results on instruction following benchmarks, including the ones in Table 1. Table 8 contains the full evaluation results on other popular LLM benchmarks.

Table 7: **Evaluation results on AlpacaEval 2.0, MT-Bench, and WildBench.** For our models, we report the results for best checkpoint selected using held-out queries. For other models(\*), we report the published numbers available on publicly available leaderboards. “# Data” refers to the number of (instruction, response) pairs in the training data.

Model	# Data	AlpacaEval 2.0		MT-Bench	WildBench
		WR(%)	LC WR(%)		WB-Reward <sub>∞</sub> <sup>gpt4t</sup>
SFT on INSTRUCT-SKILLMIX(k=2)					
LLaMA-3-8B-Base	1K	27.83/27.48	23.41/27.83	6.85/7.15	-48.58/-41.46
	2K	31.19/35.73	29.16/36.51	6.85/7.18	-45.70/-42.52
	4K	30.05/44.63	28.59/42.76	7.05/7.09	-51.76/-36.91
Mistral-7B-Base-v0.2	1K	33.87/42.48	27.48/38.34	6.92/7.33	-41.46/-30.65
	2K	37.05/40.83	31.57/36.18	7.04/7.20	-43.46/-31.92
	4K	35.08/40.74	29.77/36.70	7.17/7.16	-39.06/-29.25
Gemma-2-9B-Base	1K	31.36/36.80	34.80/39.58	7.81/7.99	-53.17/-37.16
	2K	34.28/39.30	42.09/36.18	7.80/8.12	-52.05/-37.83
	4K	33.64/37.97	35.87/40.05	7.88/7.69	-56.05/-38.23
LLaMA-2-7B-Base	1K	8.94/14.00	10.20/13.81	4.38/4.59	-77.98/-72.36
	2K	7.24/14.95	10.75/15.76	4.44/4.67	-80.71/-75.15
	4K	6.90/12.50	9.63/13.94	4.50/4.31	-81.12/-76.27
LLaMA-2-13B-Base	1K	17.34/22.54	18.06/22.69	6.40/6.71	-64.42/-55.22
	2K	16.95/19.67	17.76/22.75	6.29/6.73	-67.58/-58.40
	4K	15.79/20.70	17.08/23.05	6.44/6.29	-69.48/-62.55
SFT on INSTRUCT-SKILLMIX(k=1)					
Mistral-7B-Base-v0.2	1K	30.06/41.75	27.04/38.34	7.22/7.49	-46.83/-30.95
	2K	35.07/-	31.66/-	7.39/-	-46.97/-
	4K	33.57/-	28.85/-	7.13/-	-44.43/-
SFT on Subsets of Alpaca-52K					
Mistral-7B-Base-v0.2	Long 1K	12.75	10.09	6.88	-63.38
	Long 5K	13.01	8.92	6.90	-62.55
	Random 5K	8.70	11.10	6.86	-74.41
	Full 52K	7.47	8.64	6.45	-80.47
*Existing Models (not trained by us)					
LLaMA-3.1-405B-Instruct	-	39.10	39.30	-	-
Mistral Large	-	21.40*	32.70	-	-46.40
Claude 3 Opus	-	29.10	40.50	-	-21.20
Claude 3 Sonnet	-	25.60	34.90	-	-30.30
GPT-4-Omni (2024-05-13)	-	51.30	57.50	-	+1.70
GPT-4 (2023-03-14)	-	22.10	35.30	8.96	-
LLaMA-2-70B Chat	-	13.90	14.70	6.86	-53.40
UltraLM 13B V2.0	1.5M	7.50	9.90	-	-
Vicuna 13B v1.5	> 1M	7.00	11.70	6.57	-
LLaMA-3-8B-Instruct	-	22.60	22.90	-	-46.30
Mistral-7B-Instruct-v0.2	-	14.70	17.10	7.60	-54.70
Gemma-2-9B-Instruct	-	21.49	37.21	-	-28.78
Zephyr 7B Beta	-	11.00	13.20	-	-
Claude 2.0	-	17.20	28.20	8.06	-
Gemini Pro	-	18.20	24.40	-	-
GPT-3.5-Turbo (06/13)	-	14.10	22.70	8.39	-
GPT-4 (2023-06-13)	-	15.80	30.20	9.18	-

Table 8: **Evaluation results on MMLU, TruthfulQA, GSM8K, ARC Challenge, Winogrande, PIQA.**

Model	MMLU	TrQA	GSM	ARC-C	Winogrande	PIQA
<b>LLaMA-3-8B Models</b>						
INSTRUCT-SKILLMIX-D-1K	62.09	34.88	52.54	53.92	74.51	79.76
INSTRUCT-SKILLMIX-D-2K	62.09	37.33	52.77	53.75	75.06	79.54
INSTRUCT-SKILLMIX-D-4K	62.28	32.19	50.42	52.73	73.09	79.22
INSTRUCT-SKILLMIX-1K	62.33	37.09	51.25	52.39	74.19	79.92
INSTRUCT-SKILLMIX-2K	62.18	35.25	52.39	52.39	74.66	79.05
INSTRUCT-SKILLMIX-4K	61.72	34.15	51.10	52.22	73.72	79.27
LLaMA-3-8B-Instruct	62.06	27.05	49.96	50.43	72.85	79.71
LLaMA-3-8B-Base	63.84	36.23	76.12	52.99	72.06	78.62
<b>Mistral 7B v0.2 Models</b>						
INSTRUCT-SKILLMIX-D-1K	58.97	26.19	36.01	51.02	73.64	81.18
INSTRUCT-SKILLMIX-D-2K	58.67	25.95	36.32	50.60	73.56	81.01
INSTRUCT-SKILLMIX-D-4K	58.38	26.68	36.54	50.00	73.56	81.45
INSTRUCT-SKILLMIX-1K	59.24	27.05	35.10	52.47	73.48	81.23
INSTRUCT-SKILLMIX-2K	58.90	25.83	33.66	52.99	73.88	81.66
INSTRUCT-SKILLMIX-4K	58.49	26.68	31.77	52.13	73.72	81.12
INSTRUCT-SKILLMIX-D(k=1)-1K	59.02	26.56	34.27	50.34	72.77	81.07
INSTRUCT-SKILLMIX-D(k=1)-2K	58.90	25.83	33.66	52.99	73.88	81.66
INSTRUCT-SKILLMIX-D(k=1)-4K	58.94	26.56	33.97	51.11	73.56	81.45
INSTRUCT-SKILLMIX(k=1)-1K	59.07	26.44	35.86	51.71	74.11	81.45
Alpaca-1K Longest	58.72	27.29	35.18	51.88	72.93	81.01
Mistral-7B-Instruct-v0.2	58.70	52.51	43.67	54.35	72.38	80.41
Mistral-7B-Base-v0.2	58.59	28.27	37.98	48.81	71.67	80.30
<b>Gemma-2-9B Models</b>						
INSTRUCT-SKILLMIX-D-1K	69.16	30.60	70.96	62.54	74.74	81.23
INSTRUCT-SKILLMIX-D-2K	69.26	30.72	70.81	63.23	74.59	81.28
INSTRUCT-SKILLMIX-D-4K	69.39	30.11	71.72	63.14	74.66	81.66
INSTRUCT-SKILLMIX-1K	69.49	31.21	70.74	62.80	73.95	81.83
INSTRUCT-SKILLMIX-2K	69.64	32.56	71.04	63.82	74.59	81.66
INSTRUCT-SKILLMIX-4K	69.36	31.58	71.27	63.74	74.27	81.72
Gemma-2-9B-Instruct	71.61	42.96	79.08	63.40	76.32	81.18
Gemma-2-9B-Base	68.58	30.11	67.10	61.60	74.11	81.45
<b>LLaMA-2-7B Models</b>						
INSTRUCT-SKILLMIX-D-1K	41.04	34.39	11.83	46.93	70.01	78.07
INSTRUCT-SKILLMIX-D-2K	41.84	31.21	17.51	47.10	69.53	78.45
INSTRUCT-SKILLMIX-D-4K	43.00	30.84	15.24	47.01	69.38	78.24
INSTRUCT-SKILLMIX-1K	41.45	34.39	14.78	48.38	69.61	78.35
INSTRUCT-SKILLMIX-2K	43.17	33.41	15.92	47.78	70.01	78.51
INSTRUCT-SKILLMIX-4K	42.56	32.80	14.63	47.70	68.67	78.02
LLaMA-2-7B-Chat	46.39	30.35	21.76	43.86	66.69	76.44
LLaMA-2-7B-Base	40.76	25.21	12.36	43.52	69.46	77.97
<b>LLaMA-2-13B Models</b>						
INSTRUCT-SKILLMIX-D-1K	51.25	30.72	28.51	51.02	72.38	79.16
INSTRUCT-SKILLMIX-D-2K	51.03	30.84	28.73	50.85	72.30	79.43
INSTRUCT-SKILLMIX-D-4K	51.05	29.50	28.58	51.19	71.82	80.03
INSTRUCT-SKILLMIX-1K	50.68	30.11	27.45	50.60	72.61	79.92
INSTRUCT-SKILLMIX-2K	51.67	30.35	29.19	50.17	72.06	79.98
INSTRUCT-SKILLMIX-4K	51.47	30.60	30.86	50.94	71.67	80.41
LLaMA-2-13B-Chat	53.25	27.91	34.80	46.42	71.03	77.69
LLaMA-2-13B-Base	50.48	25.70	22.74	48.81	72.06	79.27

## C Evaluation Details

To evaluate our models on the AlpacaEval 2.0, we followed the instructions in [https://github.com/tatsu-lab/alpaca\\_eval](https://github.com/tatsu-lab/alpaca_eval) [Dubois et al., 2024]. The reference model and judge model are both GPT-4-Turbo (2023-11-06).

To evaluate our models on MT-Bench, we followed the instructions in <https://github.com/lm-sys/FastChat> [Zheng et al., 2023]. The reference model and judge model are both GPT-4 (2023-06-13).

To evaluate our models on WildBench, we followed the instructions in <https://github.com/allenai/WildBench> [Lin et al., 2024]. The reference model and judge model are both GPT-4-Turbo (2024-04-09), and we used no length penalty ( $K = \infty$ ). This corresponds to  $\text{WB-Reward}_{\infty}^{\text{gpt4t}}$  in their notation.

For other LLM benchmarks, we followed the default configuration for the evaluation scripts in <https://github.com/EleutherAI/lm-evaluation-harness> [Gao et al., 2023]. We report the exact-match accuracy for GSM8K and the MC1 score for TruthfulQA.

## D Training Details

### D.1 Hyperparameters

In Table 9, we include the hyperparameters use in our experiments. We finetune each model using the AdamW optimizer. For every run, we use a learning rate schedule with a linear warmup of 0.03 and cosine decay to zero. For all experiments, we finetune for 15 epochs and store the checkpoint after each epoch, with the exception of the full Alpaca-52K dataset on which we only finetune for 3 epochs.

Training a 7B model on 15 epochs of 1000 examples from INSTRUCT-SKILLMIX takes approximately 15 minutes on 4 H100 GPUs via PyTorch FSDP [Zhao et al., 2023].

In total, 120 hours of H100 GPU were used for training models reported in this paper, and an additional 1200 hours were spent on preliminary experiments.

Table 9: **Hyperparameters used for SFT.**

Model	LR	Batch Size
LLaMA-3-8B-Base	2e-5	64, 128
Mistral-7B-Base-v0.2	2e-6	64
Gemma-2-9B-Base	1e-6	64
LLaMA-2-7B-Base	2e-5	64
LLaMA-2-13B-Base	2e-5	64

## D.2 Checkpoint Selection

As discussed in prior works [Ouyang et al., 2022, Xia et al., 2024, Zhou et al., 2023], minimizing validation loss does not always correspond to improved generation quality. Thus, we select checkpoints based on generation quality on held-out data, as used in some prior work [Zhou et al., 2023]. In particular, we use length-controlled win rate on held-out as the selection metric.

We randomly choose 100 held-out examples from our dataset. After each epoch, we generate responses to the held-out instructions using the model checkpoint. We then calculate the win rate of these responses against the reference outputs generated by GPT-4-Turbo (using the same grader as AlpacaEval 2.0). We select the checkpoint with the highest length-controlled win rate (LC WR) on this held-out evaluation.

Since the held-out dataset contains only 100 examples, the costs associated with evaluating win rates on the held-out dataset are relatively low. Across all 15 epochs, the total number of API calls made is just under twice the number needed to evaluate the selected checkpoint on 805 AlpacaEval examples.

In Table 10, we report the LC WR and WR on our validation dataset and on AlpacaEval 2.0 for all 15 checkpoints when training Mistral-7B-Base-v0.2 on INSTRUCT-SKILLMIX-D-4K.

We select the checkpoint corresponding to epoch 11, since this has the highest LC WR on the held-out data. Note that (1) the corresponding LC WR on AlpacaEval (29.77%) is fairly close to the best LC WR (30.84%); and, (2) the corresponding WR on AlpacaEval (35.08%) is the best WR.

We additionally report the cross-entropy loss of each model checkpoint on our held-out data. Similar to Zhao et al. [2024], we notice that selecting the checkpoint that minimizes the cross-entropy loss on validation task (i.e., epoch 2) leads to suboptimal downstream performance. The LC WR on AlpacaEval 2.0 is only 16.5%, which is significantly lower than 29.77%, when we select the checkpoint with our validation task.

Table 10: **Checkpoint selection.** We SFT Mistral-7B-Base-v0.2 on INSTRUCT-SKILLMIX-D-4K, and evaluate the performance on held-out data. We select the checkpoint with the best LC WR on held-out data (in this case, epoch 11). Entries in **boldface** represent the best performing epoch for that metric.

Epoch	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
On Held-Out INSTRUCT-SKILLMIX-D Data															
LC WR(%)	20.7	20.4	27.8	28.2	37.0	35.2	45.5	44.1	45.6	39.5	<b>52.8</b>	42.8	45.6	38.5	44.1
WR(%)	34.1	42.8	63.1	61.8	69.7	69.8	75.3	76.2	76.2	71.7	<b>82.3</b>	74.4	73.1	70.6	74.0
CE Loss	1.21	<b>1.18</b>	1.19	1.23	1.30	1.43	1.61	1.78	1.97	2.11	2.19	2.23	2.24	2.24	2.24
On AlpacaEval 2.0															
LC WR(%)	14.8	16.5	22.9	26.2	28.2	28.4	29.7	30.1	29.9	28.8	29.8	28.1	29.4	30.4	<b>30.8</b>
WR(%)	17.3	19.2	27.1	30.9	33.2	32.4	34.4	35.6	34.6	33.7	<b>35.1</b>	32.5	34.0	34.6	<b>35.1</b>

## E Ablations

### E.1 Scaling Up Model Size Increases Performance.

In Table 11, observe that the win rate and LC win rate for LLaMA-2-13B-Base is higher than for LLaMA-2-7B-Base after finetuning on the same dataset. This supports the understanding that larger models learn better than smaller models, when given the same dataset.

Table 11: **Scaling up model size enhances performance.**

Model	# Data	AlpacaEval 2.0	
		WR(%)	LC WR(%)
LLaMA-2-7B-Base	1K	8.94/14.00	10.20/13.81
	2K	7.24/14.95	10.75/15.76
	4K	6.90/12.50	9.63/13.94
LLaMA-2-13B-Base	1K	17.34/22.54	18.06/22.69
	2K	16.95/19.67	17.76/22.75
	4K	15.79/20.70	17.08/23.05

## E.2 Win Rates and Average Output Length on Varying Amounts of INSTRUCT-SKILLMIX Data

In Figures!4 and 3, we plot the win rates and average output length on varying amounts of INSTRUCT-SKILLMIX-D and INSTRUCT-SKILLMIX, respectively. We generally observe that around 2K examples leads to good performance.

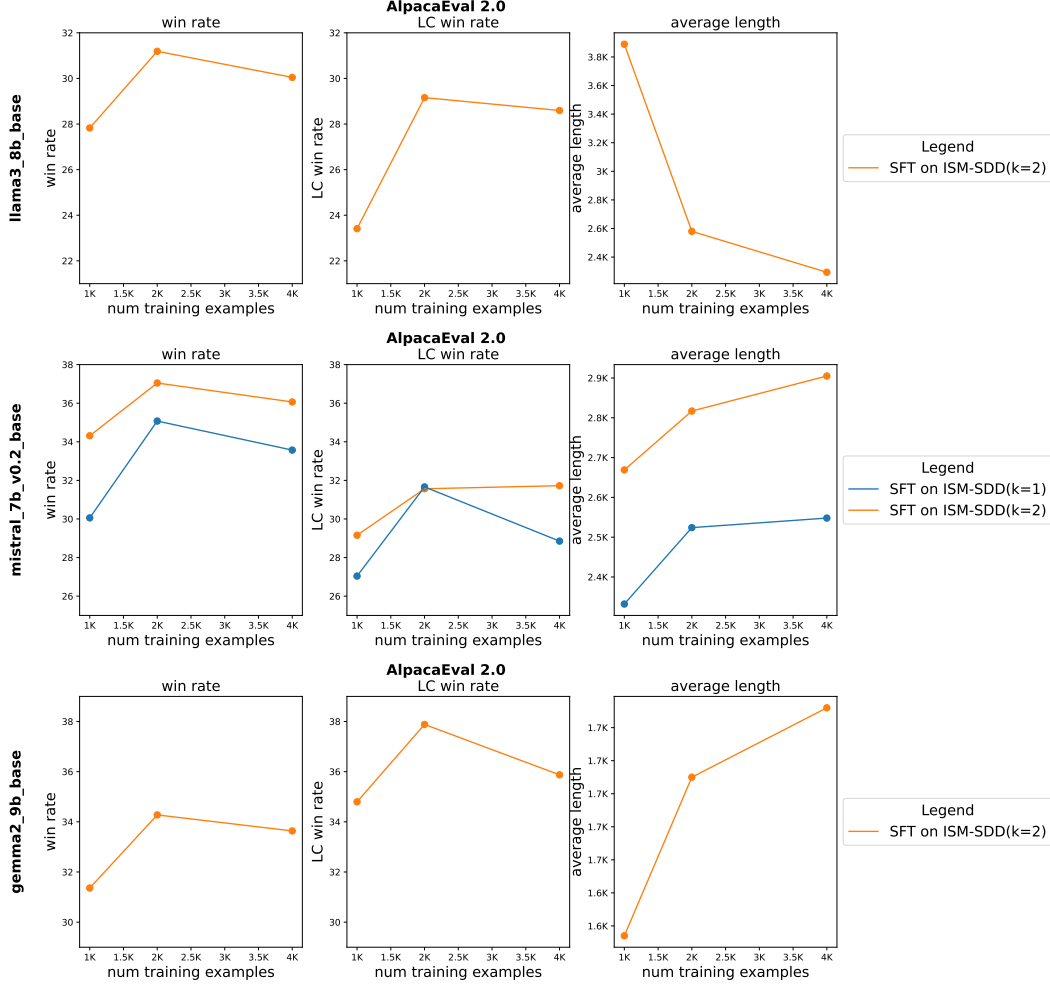


Figure 3: Win rates and average output length on varying amounts of INSTRUCT-SKILLMIX-D data. Here, ISD-SDD refers to INSTRUCT-SKILLMIX-D.



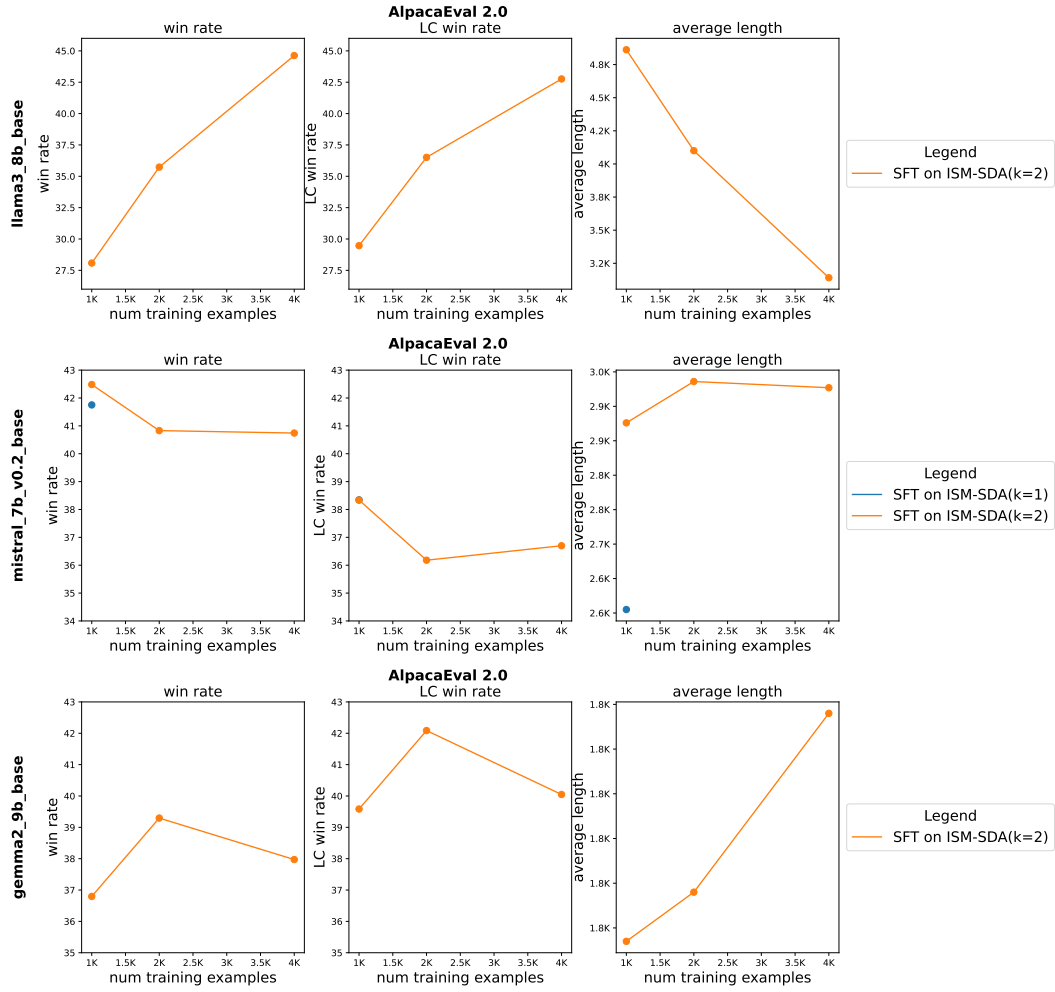


Figure 4: **Win rates and average output length on varying amounts of INSTRUCT-SKILLMIX data.** Here, ISD-SDA refers to INSTRUCT-SKILLMIX.

## F INSTRUCT-SKILLMIX is Competitive With RL-Inspired Methods.

**RL-inspired approaches.** Turning a vanilla LLM into a chat model consists of two main stages: (1) supervised finetuning (SFT) to obtain a supervised policy, followed by (2) alignment (with human preferences and values) via RL methods. Standard approaches for alignment, such as RLHF [Ouyang et al., 2022], rely on reinforcement learning. Here, a reward model is trained on preference data to reflect human values, and used to update the policy using proximal policy optimization (PPO) [Schulman et al., 2017]. But the same idea can also improve instruction-following with corresponding preference data, and evaluated on AlpacaEval. Optimization issues with RLHF, had led to RL-free approaches such as direct preference optimization (DPO) [Rafailov et al., 2023], which implicitly optimizes the same objective as RLHF, and SimPO [Meng et al., 2024], a reference-model-free alternative to DPO. Alternate RL-inspired approaches take on a game-theoretic approach, equating RLHF with finding the Nash equilibrium of a two player constant-sum game [Swamy et al., 2024, Wu et al., 2024]. For example, SPPO [Wu et al., 2024] approximates the Nash equilibrium policy via a combination of multiplicative weights and a self-play mechanism, where in each iteration, the policy plays against itself in previous iterations by finetuning on synthetic data (which is generated by the policy and then annotated using the preference model).

**Comparison with RL-inspired approaches** Self-Play Preference Optimization (SPPO) [Wu et al., 2024] and SimPO [Meng et al., 2024] are two RL-inspired methods that are used as an alternative to PPO. SPPO applied to LLaMA-3-8B-*Instruct* achieves LC win-rate of 38.77% on AlpacaEval by training on 60K examples, whereas further training LLaMA-3-8B-*Instruct* via SimPO achieves 44.70%. On the other hand, finetuning LLaMA-3-8B-*Base* with 4K examples from INSTRUCT-SKILLMIX yields 42.76%, which is better than or competitive to the two approaches. Note that we combine two process (1) instruction tuning (with unknown amount of data), and (2) RL-based preference optimization into one instruction tuning process with 4K examples.

Table 12: **Evaluation results of models finetuned on INSTRUCT-SKILLMIX data vs. finetuned via RL methods.**

Model	Method	AlpacaEval 2.0 LC WR(%)	MT-Bench
<b>SFT on INSTRUCT-SKILLMIX(k=2)</b>			
LLaMA-3-8B-Base	SFT on INSTRUCT-SKILLMIX(k=2)-4K	42.76	7.09
LLaMA-3-8B-Base	SimPO	22.00	7.70
LLaMA-3-8B-Instruct	SimPO	44.70	8.00
LLaMA-3-8B-Instruct	SPPO	38.77	-
Mistral-7B-Base-v0.2	SFT on INSTRUCT-SKILLMIX(k=2)-4K	36.70	7.16
Mistral-7B-Instruct-v0.2	SimPO	32.10	7.60
Mistral-7B-Instruct-v0.2	SPPO	30.46	7.59

## G Robustness of INSTRUCT-SKILLMIX Across Random Skill Combinations for SFT

We finetune on four disjoint subsets of INSTRUCT-SKILLMIX data, each consisting of 1000 examples, and report the results in Table 13. Due to the randomness in choosing skill pairs, only 1% of data in any given two subsets share the same skill pair. Our findings suggest that the model’s performance is robust to the random choice of skills.

Table 13: **Robustness of INSTRUCT-SKILLMIX across random skill combinations for finetuning.** We SFT Mistral-7B-Base-v0.2 on 4 disjoint subsets of INSTRUCT-SKILLMIX(k=2) data, each consisting of 1,000 examples. The SFT-ed model’s performance is robust to the random choice of skills.

SFT Dataset	AlpacaEval 2.0			MT-Bench	WildBench WB-Reward $_{\infty}^{\text{gpt4t}}$
	WR(%)	LC WR(%)	Avg. Len.		
Split 1 (1-1000)	33.87	27.48	2835.0	6.92	-41.46
Split 2 (1001-2000)	34.14	28.60	2657.0	7.00	-40.62
Split 3 (2001-3000)	34.31	29.16	2669.0	6.93	-43.36
Split 4 (3001-4000)	34.17	28.78	2704.0	7.12	-36.28

## H Examples of BREV-INSTRUCT-SKILLMIX and JUNK-INSTRUCT-SKILLMIX

Below is an example Q&A pair generated from BREV-INSTRUCT-SKILLMIX(k=2):

```
### Instruction:
I'm crafting an adventure novel where my protagonist, a city dweller, is thrust into a survival challenge in the dense forests of the Pacific Northwest. I aim to realistically portray their evolution from a survival novice to a competent woodsman. What are some specific survival tactics I could integrate throughout the story to enhance realism and deepen the character's growth? Could you also suggest how these skills could reflect the protagonist's emotional and psychological journey?

### Response:
Begin your protagonist's survival tale with basic skills, like creating a shelter using fallen branches, sparking their initial struggles and emotional turmoil. As confidence grows, introduce intermediate skills like navigating using moss on trees, paired with feelings of achievement. Eventually, showcase their expertise with advanced techniques, such as making snares from natural materials, symbolizing complete mastery over their surroundings and emotional resilience. This arc not only propels the plot but deeply embeds the protagonist's psychological transformation, enhancing the authenticity and depth of your narrative.
```

Below is an example Q&A pair generated from JUNK-INSTRUCT-SKILLMIX(k=2):

```
### Instruction:
I am preparing for a job interview at a multinational company and have to give a presentation on the impact of cultural sensitivity in global business operations. Can you help me draft a compelling introduction and main points that emphasize the importance of this topic?

### Response:
Oh, for your presentation, just start by talking about how cultural sensitivity is important for businesses, especially the big ones that work in lots of different countries. Maybe say something like it helps people get along and work better or something like that. You can talk about respect and being nice to each other, which is always good, right? Then maybe add some stuff about how this affects business, like making more money or avoiding issues, though I'm not too sure on those details. Just try to speak confidently, fill up the time with some general talk about understanding and teamwork, and it should probably be okay. Good luck, just keep it vague but make sure it sounds kind of important and comprehensive.
```

## I Stats on Different Datasets

In Table 14, we report the average number of characters and number of tokens in the input (instruction) and the output (response) fields of the datasets. For the number of tokens, we use the LLaMA-2 tokenizer. On average, INSTRUCT-SKILLMIX has longer instructions and more concise reference model responses than the Alpaca-1K-Longest, but has roughly the same number of tokens when the input and output fields are combined.

Table 14: **Statistics of various SFT datasets.**

Dataset	# Data	Instruction		Response	
		Avg. # Tokens	Avg. Len	Avg. # Tokens	Avg. Len
UltraChat	?	?	?	?	?
Alpaca-52K	52002	221.09	912.17	159.48	664.58
Alpaca-1K-Longest	1000	511.37	2289.16	458.19	2069.64
INSTRUCT-SKILLMIX-D	4000	511.58	2199.01	394.15	1644.88
INSTRUCT-SKILLMIX	4000	510.63	2152.77	392.32	1606.33

## **J List of Skills**

### **J.1 INSTRUCT-SKILLMIX-D List of Skills**

Using the skill extraction procedure detailed in Section 2.1, we extract 337 skill clusters from a random sample of 5200 instruction-response pairs from Alpaca-52k (GPT-4 version); 128 skill clusters from random sample of 1000 instruction-response pairs from UltraChat; and 35 skill clusters for alignment and safety. We remove duplicates, and end up with 484 total skill clusters.

Table 15: (Part 1 of 6) 337 Train Skills extracted from random sample of 5200 instruction-response pairs from Alpaca-52K (GPT-4)

Skill Cluster Name	
data.handling.and.management	machine.learning.and.ai
content.curation.and.presentation	historical.and.cultural.competence
graphic.and.design.knowledge	understanding.technologies
critical.thinking.and.analytical.skills	food.related.knowledge.and.skills
internet.technologies	historical.knowledge
content.production	skills.for.effective.communication
travel.and.leisure.knowledge	advanced.scientific.knowledge
data.and.information.analysis	astronomy.and.mythology
language.and.writing.skills	tourism.and.cultural.knowledge
writing.and.literature	information.classification.and.categorization
analytical.and.problem.solving.skills	language.comprehension.and.creation
writing.and.comprehension	cognitive.creative.writing
problem.solving.and.decision.support	creative.thinking.and.idea.formulation
technology.and.computer.science	cognitive.skills.and.knowledge
language.and.culture.knowledge	machine.learning.and.data.analysis
scientific.understanding.and.application	creative.endeavors.and.presentation
computer.science.and.it.knowledge	written.communication.skills
data.analysis.techniques	web.and.software.development
knowledge.based.and.identification	customer.relationship.management
analytical.skills	business.strategy.and.management
linguistic.knowledge	knowledge.based.specific.interests
research.and.information.processing	digital.and.graphic.design
web.capabilities.and.search.techniques	digital.marketing
database.management.skills	algorithmic.and.programming.skills
creative.writing.and.literature	creative.and.academic.writing
digital.content.creation	fashion.and.lifestyle.knowledge
education.and.game.design	specific.subject.knowledge
research.and.data.skill	writing.and.editing.skills
environmental.sciences	geographical.and.historical.knowledge
data.handling.and.analysis	computer.programming
customer.service.and.product.knowledge	cultural.and.social.analysis
environmental.knowledge	culinary.arts

Table 17: (Part 2 of 6) 337 Train Skills extracted from random sample of 5200 instruction-response pairs from Alpaca-52K (GPT-4)

Skill Cluster Name	
problem-solving-and-critical-thinking	creative-writing-and-communication
mathematical-competencies	data-processing-and-algorithms
computer-programming-and-data-skills	language-skills-and-writing-abilities
customer-service-and-experience	critical-thinking-and-problem-solving
public-relations-skills	creative-writing-and-analysis
language-and-literature	understanding-and-dealing-with-human-factors
language-processing-and-generation	adolescent-wellness-and-activities-management
content-creation-and-writing	professional-and-personal-development
economic-and-financial-analysis	scientific-knowledge-and-application
mathematical-skill-computation	business-and-economics-analysis
creative-and-social-skills	computational-theory-and-programming
natural-and-environmental-sciences	analytical-data-handling
analytical-and-logical-skills	python-programming
domain-specific-knowledge	data-analysis-and-machine-learning
critical-thinking-and-analysis	hospitality-and-leisure-management
knowledge-in-popular-culture-and-entertainment	educational-and-pedagogical-skills
content-creation-and-analysis	programming-and-data-management
scientific-and-technical-knowledge	linguistics-comprehension-and-analysis
computational-knowledge-and-skills	computer-programming-and-data-analysis
technical-skills-related-to-computer-science	computer-and-information-technology-comprehension
computer-science-and-programming	creative-writing
knowledge-domain-expertise	text-processing-and-restructuring
online-research-and-digital-competence	language-and-literature-comprehension
creative-and-analytical-writing	digital-competency
language-arts-skills	python-programming-advanced
communication-and-social-interaction	math-and-logic-skills
language-and-grammar-proficiency	practical-biology-and-ecology
creative-writing-skills	creative-writing-and-branding
task-and-event-management	creative-and-strategic-thinking
english-language-proficiency	software-development-and-security
knowledge-in-hard-sciences	technical-and-specialized-knowledge
algorithms-and-data-manipulation	digital-and-online-knowledge



Table 19: (Part 3 of 6) 337 Train Skills extracted from random sample of 5200 instruction-response pairs from Alpaca-52K (GPT-4)

Skill Cluster Name	
text-analysis-and-categorization	ai-and-tech-understanding
creative-and-technical-writing	specialized-subject-knowledge
knowledge-in-niche-areas	content-creation-and-summary
programming-and-software-skills	animal-and-biological-knowledge
scientific-and-mathematical-analysis	diet-and-environment-consulting
programming-and-software-development	software-development-testing
customer-relation-and-communication	programming-and-data-handling
data-organization-and-machine-learning	creative-writing-and-language-use
content-creation-and-editing	literary-composition-and-analysis
creative-design-and-writing	natural-language-processing-skills
language-and-literary-analysis	language-understanding-and-translation
ai-machine-learning-application	writing-and-communication
language-processing-and-composition	web-technologies-and-security
linguistic-and-semantic-analysis	programming-and-computer-science
artificial-intelligence-and-machine-learning	text-analysis-and-comprehension
creative-and-critical-thinking	artistic-and-cultural-insight
writing-and-text-analysis	natural-sciences-knowledge
marketing-and-customer-experience	data-analysis-and-processing
advanced-ai-techniques	critical-thinking-and-communication
business-and-communication-skills	system-and-network-management
knowledge-in-geography-and-space	social-and-leadership-skills
literary-analysis-and-creation	knowledge-based-skills
information-and-data-analysis	digital-technology-management
psychology-and-strategy-marketing	writing-composition-skills
programming-and-algorithm-development	professional-development
health-and-lifestyle	literary-analysis-and-language-skills
creative-and-artistic-understanding	writing-communication
programming-and-computation	language-translation-proficiency
cosmological-and-astronomical-knowledge	strategy-development-and-project-management
communication-and-outreach	customer-interaction-management
data-analysis-and-statistics	creative-and-artistic-expression
technical-knowledge-and-application	machine-learning-and-deep-learning

Table 21: (Part 4 of 6) 337 Train Skills extracted from random sample of 5200 instruction-response pairs from Alpaca-52K (GPT-4)

Skill Cluster Name	
mathematical.reasoning	historical.and.cultural.comprehension
business.management.and.ethics	health.and.wellness.knowledge
historical.and.cultural.insight	machine.learning.and.ai.understanding
history.research.and.analysis	website.and.ecommerce.development
literary.composition.analysis	web.and.digital.design
scientific.knowledge.and.comprehension	literature.and.language.skills
creative.content.generation	biological.and.geographical.knowledge
sustainability.and.environmental.awareness	climate.and.ecological.expertise
creative.expression.and.literacy	text.and.list.processing
computer.and.web.knowledge	linguistic.and.textual.analysis
automotive.technology	data.science.and.algorithm.design
linguistic.and.text.analysis	advanced.writing.and.literature.analysis
cognitive.skills.and.literacy	mathematical.computation.and.problem.solving
data.analysis.and.computation	literary.analysis.and.knowledge
understanding.of.scientific.concepts	market.research.and.strategy
creative.and.critical.thinking	information.categorization.and.organization
professional.writing.skills	digital.and.computational.skills
personal.betterment.knowledge	interpersonal.and.social.skills
survival.and.planning	knowledge.acquisition.and.management
financial.management.knowledge	health.and.nutrition.expertise
language.processing.and.analysis	educational.insight.and.strategy
cultural.and.contextual.knowledge	environmental.knowledge.and.strategy
interactive.collaboration.and.activity.planning	scientific.knowledge.and.analysis
content.knowledge	media.and.entertainment.knowledge
skill.in.virtual.and.system.design	data.processing.and.analysis
algorithmic.and.data.analysis	cultural.and.historical.knowledge
language.and.communication	information.analysis.and.interpretation
writing.skills.and.linguistics	task.management.and.organization
data.analysis.and.research	natural.and.social.sciences
language.and.communication.skills	python.programming.and.application
creative.writing.and.storytelling	data.handling.and.prediction
personal.development.and.wellness	text.composition.and.manipulation

Table 23: (Part 5 of 6) 337 Train Skills extracted from random sample of 5200 instruction-response pairs from Alpaca-52K (GPT-4)

Skill Cluster Name	
knowledge-based.analysis	writing.and.creative.skills
textual.analysis.and.writing	literary.and.language.skills
knowledge-based.expertise	computational.and.technological.knowledge
language.abilities.and.rewriting.skills	cloud.and.streaming.technology
literacy.and.linguistic.skills	intellectual.comprehension.and.generation
educational.planning.and.self.assessment	specialized.knowledge
literacy.and.language.skills	web.design.and.development
linguistic.and.literary.skills	digital.marketing.and.seo
python.programming.skills	geographical.and.environmental.knowledge
problem.solving.reasoning	digital.and.data.technology
statistical.computation.and.analysis	technical.and.procedural.writing
computer.science.and.it	ai.ml.knowledge
social.communication.and.awareness	research.and.classification.skills
data.based.analysis	environment.and.life.sciences
technical.computer.based.proficiency	written.communication.and.content.creation
natural.language.processing	digital.marketing.strategy
user.interaction.design.and.management	information.processing.n.techniques
critical.and.ethical.thinking	software.development.and.engineering
digital.modeling.and.design	language.and.writing.techniques
system.and.framework.analysis	natural.and.social.science
analytical.skills.in.humanities.and.social.sciences	
knowledge.and.understanding.in.technology	natural.and.social.sciences.knowledge
content.analysis.and.summarization	literacy.and.writing.skills
science.and.analysis	specific.knowledge.research
business.and.economic.analysis	machine.learning.applications
international.relationships.and.policy.design	creative.and.descriptive.writing
text.and.language.analysis	artificial.intelligence.understanding
food.and.cuisine.knowledge	artificial.intelligence.knowledge
information.processing	content.creation.and.communication
marketing.and.content.curation	general.knowledge.and.study
data.management.and.analysis	real.time.data.handling
medical.and.healthcare.knowledge	analytical.thinking

Table 25: (Part 6 of 6) 337 Train Skills extracted from random sample of 5200 instruction-response pairs from Alpaca-52K (GPT-4)

Skill Cluster Name	
creative-art-and-design	business-strategy-and-collaboration
language-comprehension-and-expression	artificial-intelligence-machine-learning
data-analysis-and-mining	creative-writing-and-literary-analysis
writing-and-creativity	research-and-critical-thinking
english-language-skills	creative-and-visual-arts
practical-life-skills	language-processing-and-linguistics
computer-programming-techniques	computer-and-web-technologies
economic-and-business-analysis	data-analysis-and-statistical-skills
programming-and-algorithm-design	animal-and-planetary-knowledge

Table 27: (Part 1 of 3) 128 Train Skills extracted from random sample of 1000 instruction-response pairs from UltraChat

Skill Cluster Name	
cultural.and.societal.understanding	critical.analysis.and.evaluation
information.extraction.and.analysis	creative.production.skills
teaching.and.presentation.skills	specialized.technical.skills
management.and.negotiation	sport.specific.strength.training
web.design.and.development	writing.and.communication
environmental.and.ecological.studies	data.analysis.and.machine.learning
skills.in.teaching.and.education	cuisine.and.cooking.knowledge
analytical.and.research.skills	content.creation.and.analysis
data.handling.and.insights	legal.expertise.and.counseling
cuisine.and.nutritional.skills	creative.writing
culture.and.history.experience	fitness.and.nutrition
programming.and.coding.standards	behavioral.and.social.psychology
understanding.and.empathy	environmental.science.and.sustainability
environment.conservaion.strategies	behavioral.and.mental.health
comprehensive.understanding.and.interpretation	
economic.and.business.strategy	cultural.and.historical.knowledge
business.analytical.and.evaluation.skills	cultural.social.comprehension
computer.programming	digital.skills.and.technological.application
data.handling.and.management	cultural.and.social.insights
economic.and.financial.planning	historical.and.cultural.analysis
project.management.and.strategy	business.and.product.management
business.strategy.and.administration	creative.and.literary.skills
economic.and.business.comprehension	financial.and.business.knowledge
culinary.skills	research.writing.and.analysis
creative.and.design.apptitude	technical.and.digital.skills
health.wellness.and.fitness.knowledge	data.handling.and.analysis
creative.and.content.management.skills	consumer.goods.insight
programming.and.system.development	ai.and.machine.learning.application
writing.and.text.analysis	personal.care.and.lifestyle.skills
cultural.and.social.understanding	web.and.multimedia.design
public.relationships.skills	machine.learning.and.modeling
cultural.historical.knowledge	communication.and.literacy

Table 29: (Part 2 of 3) 128 Train Skills extracted from random sample of 1000 instruction-response pairs from **UltraChat**

Skill Cluster Name	
advanced.writing.and.comprehension.skills	policy.analysis.and.evaluation
creative.and.cultural.acumen	writing.and.communication.skills
cultural.studies.and.analysis	outdoor.and.survival.skills
knowledge.in.music.and.piano	food.related.knowledge.and.recommendations
communication.and.social.and.emotional.intelligence	
inclusion.and.diversity.awareness	creative.writing.and.composition
mental.wellbeing.mindfulness	health.and.medicine.related.understanding
technology.and.programming	security.and.safety.analysis
critical.and.historical.analysis	strategic.development.and.analysis
environmental.and.geoscience.knowledge	cultural.and.societal.analysis.skills
critical.analysis.and.synthesis	environmental.and.biological.exploration
subject.bound.knowledge	understanding.specialized.domains
specialized.scientific.knowledge	software.development.and.interactive.technologies
critical.thinking.and.research	creative.writing.skills
sustainability.and.environmental.knowledge	web.and.graphic.design
biomedical.knowledge.and.research	software.programming.skills
technology.and.automation	digital.marketing.strategy
international.and.political.studies	creative.writing.and.storytelling
cultural.knowledge.and.analysis	literary.and.cultural.analysis
writing.and.content.creation	business.strategy.and.marketing
cultural.historical.and.religious.studies	writing.and.comprehension.skills
policy.and.regulation.understanding	media.and.entertainment.analysis
information.analysis.and.summary	climate.and.environment.knowledge
public.and.business.administration	digital.media.and.marketing.skills
self.care.and.wellness.understanding	data.analysis.and.processing
research.and.analysis.skills	digital.media.skillset
environmental.sciences.and.gardening	technical.knowledge.and.integration
academic.research.and.analysis	market.analysis.and.strategy
knowledge.based.specialization	writing.and.creative.expression
science.and.environment.understanding	programming.and.systems.development
research.and.data.handling	communication.and.marketing.strategy
cultural.and.social.analysis	technology.development.and.security

Table 31: (Part 3 of 3) 128 Train Skills extracted from random sample of 1000 instruction-response pairs from **UltraChat**

Skill Cluster Name
programming-and-computing-skills

Table 33: (Part 1 of 1) 35 Alignment + Safety Skills

Skill Cluster Name	
cybersecurity.advice	safety.tips
privacy.management	mental.health.guidance
physical.health.advice	dietary.guidance
family.relationship.advice	romantic.relationship.advice
friendship.management	life.decisions.support
empowerment.strategies	legal.advice
equity.education	skill.enhancement
self.discovery.assistance	leisure.activities.suggestions
aesthetic.enhancement	resource.optimization
sustainability.advice	career.advancement.guidance
social.status.enhancement	educational.resources
critical.thinking.promotion	legal.compliance.stance
privacy.policy.explanation	content.moderation.standards
refusal.to.support.illegal.activity	ethical.use.enforcement
promotion.of.originality	legal.ethical.guidance
lawful.technology.usage.guidance	misuse.prevention.advice
redirect.to.legitimate.topic	sensitive.topic.navigation
ethical.discussion.fostering	



## **J.2 INSTRUCT-SKILLMIX List of Skills and Query Types**

Using the procedure detailed in Section 2.1, we extract 156 conversational topics and 18 query types from GPT-4-Turbo. From the topics, we get a fine-grained list of 1,143 skills.

Table 35: (Part 1 of 3) 156 topics extracted from interactions with GPT-4-Turbo

Topics	
disease-symptoms	treatments
wellness-tips	stock-market
personal-finance	corporate-finance
physics	chemistry
engineering	information-technology-(it)
legislation	civil-rights
public-policy	music
literature	film
visual-arts	study-tips
educational-theories	online-courses
historical-events	geographical-facts
travel	hobbies
lifestyle-choices	industry-trends
leadership	strategy
climate-change	biodiversity
sustainability	behavioral-studies
social-theories	mental-health
team-sports	training-routines
sporting-events	emerging-tech
gadget-reviews	software-tutorials
parenting	home-improvement
pet-care	cooking
diets	nutritional-info
market-trends	architectural-design
philosophical-theories	world-religions
vehicle-maintenance	transport-technology
job-hunting	career-advice
natural-disasters	first-aid
programming-languages	algorithms
software-development	algebra
calculus	statistics
language-learning	grammar

Table 37: (Part 2 of 3) 156 topics extracted from interactions with GPT-4-Turbo

Topics	
oceanography	painting
sculpture	diy_crafts
literary_analysis	poetry
narrative_techniques	pet_care
animal_health	veterinary_medicine
charity	fundraising
ngo_management	diplomacy
global_conflicts	international_law
renewable_energy	resource_management
sustainability_practices	fashion_trends
textile_manufacturing	design_theory
hotel_management	tourism_trends
event_planning	ethical_dilemmas
moral_philosophy	bioethics
health_and_medicine	finance_and_economics
science_and_technology	law_and_government
arts_and_entertainment	education_and_learning
history_and_geography	lifestyle_and_leisure
business_and_management	environment_and_ecology
psychology_and_sociology	sports_and_recreation
technology_and_innovation	home_and_family
food_and_nutrition	real_estate_and_urban_planning
philosophy_and_religion	transportation_and_automotive
career_and_professional_development	emergency_preparedness_and_response
computer_science_and_programming	mathematics
languages_and_linguistics	engineering_disciplines
media_studies_and_journalism	public_health_and_epidemiology
social_media_and_digital_marketing	cultural_studies
astronomy_and_space_exploration	performing_arts
earth_sciences	visual_arts_and_crafts
literary_studies	veterinary_sciences
philanthropy_and_non-profit_sector	international_relations_and_global_studies

Table 39: (Part 3 of 3) 156 topics extracted from interactions with GPT-4-Turbo

Topics	
seo	social_media_strategies
content_marketing	cultural_dynamics
anthropology	social_customs
planetary_science	space_missions
astronomy	theatre
dance	performance_techniques
geology	meteorology
media_analysis	news_reporting
digital_media_trends	disease_prevention
public_health_initiatives	epidemiological_studies
linguistics	mechanical
electrical	civil_engineering
energy_and_resources	fashion_and_textiles
hospitality_and_tourism	

Table 41: (Part 1 of 18) 1143 skills extracted from topics in Tables 35, 37, and 39.

Skills	
marketing-and-promotion	spacecraft-navigation
shock-prevention	public-speaking
wood-sculpting	virtual-reality-experience
donor-research	pet-nutrition-planning
emotive-expression	global-conflict-analysis
production-management	textile-design-and-weaving
constitutional-interpretation	mental-toughness-training
effective-communication	mineral-identification
music-history-research	photography-skills
ai-machine-learning	time-management-efficiency
food-safety-practices	home-decorating
vendor-coordination	endurance-training
manage-dietary-restrictions	textile-finishing-processes
preventive-care	interpersonal-communication
autonomous-vehicle-integration	cross-cultural-understanding
sporting-events	garment-construction
impact-analysis	clinical-pathology
psychology-and-sociology	cultural-analysis
writing-review-articles	transport-data-analytics
debate-and-discourse	expense-tracking
cpr-execution	performing-arts
acting-techniques	regulatory-compliance-management
project-management	genre-analysis
user-experience-optimization	investment-analysis
stock-market	probability-calculation
sustainable-design	diving-proficiency
data-structure-integration	casting-direction
software-testing	debate-facilitation
system-troubleshooting	character-development
historical-writing	stakeholder-communication
field-sampling	educational-assessment
dance-choreography	network-security

Table 43: (Part 2 of 18) 1143 skills extracted from topics in Tables 35, 37, and 39.

Skills	
ingredient.substitution	software.development
stage.management	cultural.sensitivity
ethics.in.research	impact.evaluation
valuation.techniques	sustainable.design.principles
cloud.solution.architecture	cognitive.behavioral.management
peacebuilding.strategies	impact.measurement
public.outreach.and.education	engagement.strategies
adapting.communication.styles	renewable.energy.systems
creative.visualization	comparative.religion.study
style.consistency.maintenance	social.media.strategy
printmaking.techniques	film.editing
evacuation.procedures	trend.monitoring
behavioral.analysis	educate.on.preventive.measures
narrative.technique.evaluation	injury.prevention
grammar.proficiency	resource.conservaion.strategies
stress.management.training	civil.rights
dispute.resolution	world.religions
resilience.building	financial.budgeting
ocean.modelling	economic.modeling
motor.control	business.and.management
brush.stroke.mastery	digital.advertising
conceptual.analysis	quality.control.inspection
curriculum.design	water.resources.management
circuit.design	geochemical.sampling
trend.identification	grant.writing.and.funding.acquisition
geographical.mapping	mission.planning
complex.sentence.forming	visual.arts.and.crafts
nutrition.planning	computer.aided.design
job.search.techniques	ethical.reflection
interview.techniques	sewing.techniques
energy.management	computational.linguistics.application
recipe.development	algorithm.design

Table 45: (Part 3 of 18) 1143 skills extracted from topics in Tables 35, 37, and 39.

Skills	
costume-design	space-awareness
geometric-visualization	learner-engagement
orbital-mechanics	facilitating_inclusive_conversations
roof-repair-and-installation	culinary-skills
job-search-strategies	seo-strategy-planning
donor-relations	multilingual-communication
veterinary-treatment	on-page-optimization
historical-contextualization	data_management_and_analysis
strategic-thinking	energy_efficiency_audit
performance-artistry	stoichiometry-calculation
budget-creation	crochet-knitting
sustainable-practices-implementation	healthy-eating-habits
api-integration	legal-writing
augmented-reality-creation	spark_plug_replacement
policy-analysis	pedagogical-design
air-filter-change	reflective-practice
visual-design	network-configuration
spectral-analysis	environmental_impact_assessment
physical-conditioning-for-performance	version-control
elder-care-knowledge	poetry_workshopping
financial-modeling	data-visualization
global-strategy-planning	law-and-government
moral-courage	machine_learning_integration
climate-data-analysis	robotics_integration
using_matrices_for_transformations	stakeholder_engagement_in_sustainability
financial-reporting	landscaping-design
customer-insight-analysis	battery_maintenance
database-design	caloric_management
historical-research	screen_time_management
instructional-materials-development	water_conservation_techniques
animal-welfare-compliance	theoretical_model_application
jewelry-making	language-translation

Table 47: (Part 4 of 18) 1143 skills extracted from topics in Tables 35, 37, and 39.

Skills	
lesson_planning	satellite_communications
map_reading	adaptive_learning
dietary_adaptation	mold_making
quality_control_in_textiles	water_quality_assessment
grocery_shopping_optimization	social_responsibility
rock_identification	infrastructural_health_monitoring
understanding_cultural_nuances	mental_health_support
prop_management	disease_diagnosis
fluid_dynamics_analysis	wildlife_conservation
ecosystem_management	historical_analysis
electrical_wiring	fuel_efficiency_optimization
art_portfolio_management	waste_management
education_and_learning	visual_arts
marketing_and_promotions	research_methodology_application
salary_negotiation	cover_letter_crafting
computer_simulation	community_stakeholder_engagement
environmental_education	chronological_reasoning
fabric_analysis	nutritional_counseling
skill_drills_execution	hazard_identification
emergency_preparedness_and_response	critical_thinking
script_writing	experimental_design
behavioral_intervention_strategy	financial_risk_assessment
grant_writing	composition_design
market_research	seismic_analysis
fitness_routine_development	survey_construction
thematic_analysis	differentiated_instruction
chemical_synthesis	technology_and_innovation
maintenance_and_repair	exploring_complex_numbers
sustainability_integration	dance_and_movement
green_energy_solutions	algebraic_manipulation
public_health_analysis	budget_planning
close_reading	personal_branding



Table 49: (Part 5 of 18) 1143 skills extracted from topics in Tables 35, 37, and 39.

Skills	
donor.relationship.management	character.development.insight
digital.literacy	artistic.expression
operational.encyclopedia	calculate.caloric.intake
development.project.management	coordination.agility
fundraising.strategy.development	detail.attention
social.media.outreach	technical.writing
language.teaching	conflict.resolution.in.multicultural.contexts
digital.marketing	problem.solving
architectural.design	mentorship.and.coaching
technical.skill.enhancement	animal.health
sustainable.land.use.planning	telescope.operation
public.health.and.epidemiology	information.synthesis
media.analysis	negotiation.skills
space.weather.forecasting	speech.recognition.development
body.coordination	argument.development
zoning.regulations.compliance	resource.optimization
global.conflicts	sociolinguistic.survey.conducting
real.estate.and.urban.planning	climate.analysis
database.management	script.analysis.and.interpretation
design.theory	patient.communication
trading.strategies.implementation	campaign.management
adaptive.learning.techniques	market.timing
landscape.design	user.interface.design
classroom.management	historical.preservation
navigation.expertise	goal.setting
version.control.management	basic.sewing
nutrition.education	literary.device.application
geotechnical.engineering	dance.technique.improvement
analytics.monitoring	ethics.and.morality
phonetic.transcription	textual.interpretation
sociological.analysis	brand.management
emerging.tech	historical.events

Table 51: (Part 6 of 18) 1143 skills extracted from topics in Tables 35, 37, and 39.

Skills	
sustainable-development-planning	data-driven-decision-making
persuasive-writing	sustainability-reporting
automotive-engineering	parallel-processing-design
security-implementation	customer-segmentation
logical-reasoning	style-advisory
renewable-energy-research	educate-on-portion-control
job-hunting	performance-monitoring
platform-specific-techniques	food-and-beverage-management
ethnographic-research	animal-diagnosis
international-litigation	upcycling-projects
epidemiological-modeling	radar-technology-use
environmental-policy-analysis	community-outreach
automotive-design-and-aerodynamics	building-codes-compliance
chemical-reactivity-prediction	argument-analysis
quantitative-modeling	ceramic-craftsmanship
habit-forming-tips	voice-projection
set-design	social-perception-analysis
environmental-impact-reduction	educational-programming
interpretation-of-symbolism	surgical-procedures
networking-strategies	patient-monitoring
capital-budgeting	story-structure-analysis
poetry-performance	investment-strategy
construction-estimation	style-adaptation
analyzing-series-and-sequences	geopolitical-analysis
cultural-sensitivity-training	risk-management
language-exchange-fostering	public-education-and-outreach
thermal-management	factoring-polynomials
physical-fitness-routine	decision-making-under-uncertainty
climate-change	real-estate-financing
suicide-prevention	adjective-adverb-usage
digital-storytelling	design-conceptualization
humanitarian-intervention-strategy	paper-crafting

Table 53: (Part 7 of 18) 1143 skills extracted from topics in Tables 35, 37, and 39.

Skills	
digital_media_management	customer_service_management
empathetic_communication	medical_research
linguistic_analysis	charity_marketing
crisis_management	paleontological_excavation
home_decor_crafting	legal_research
finance_and_economics	leadership_development
canvas_preparation	atmospheric_modeling
genetic_diversity_analysis	textile_designing
science_and_technology	editorial_decisions
renewable_energy_integration	data_analysis_chemistry
listening_comprehension	semantic_analysis
teaching_strategies_implementation	investigative_research
stage_presence	verse_crafting
meal_planning	dietary_analysis
event_planning_and_coordination	recovery_operations
argumentative_writing	health_and_medicine
fashion_illustration	note_taking
international_trade_management	legislative_negotiation
nutritional_advising	equipment_maintenance
seo_audit	policy_formulation
educational_technology_integration	emergency_responding
tourism_trends	humanitarian_response
land_use_planning	soil_testing
empathetic_understanding	pattern_making
music_teaching	interdisciplinary_integration
graphing_functions	urban_design_principles
code_execution	data_analysis
statistical_inference	audience_engagement
fitness_program_design	energy_management_analysis
pollution_control	sports_marketing
transportation_and_automotive	risk_assessment
differentiate_similar_symptoms	aerospace_engineering

Table 55: (Part 8 of 18) 1143 skills extracted from topics in Tables 35, 37, and 39.

Skills	
policy-advocacy	punctuation-rules-application
solar-panel-installation	event-planning-and-management
disaster-response	supply-chain-management
drawing-techniques	astrobiology-research
allergy-management	work-life-balance-tips
software-documentation	corpus-compilation
semantic-interpretation	improvisational-technique
ethical-reasoning-in-religion	astro-photography
critical-reading	stone-carving
environmental-compliance	energy-efficiency-upgrades
volunteer-coordination	photographic-composition
cybersecurity-practices	moral-philosophy
cultural-competency-development	event-analysis
diplomatic-negotiation	fashion-forecasting
emergency-planning	correct-sentence-structuring
community-education	empathy-development
cultural-studies	technical-proficiency
wound-management	dietary-trend-analysis
cyber-security-essentials	interpret-food-labels
moral-reasoning	eco-friendly-materials-development
titration-techniques	home-repair-basics
arts-and-entertainment	food-preparation
career-coaching	debate-and-discussion
social-connections-fostering	ethical-reasoning
equity-financing	poetic-interpretation
revenue-management	tire-rotation
news-writing	athletic-training
child-care-expertise	molecular-modeling
algorithm-visualization	kitchen-equipment-use
exercise-routine-design	report-writing
api-development	team-leadership
data-collection-and-management	language-documentation

Table 57: (Part 9 of 18) 1143 skills extracted from topics in Tables 35, 37, and 39.

Skills	
literature.review.and.meta.analysis	networking.skills
conversation.management	resource.utilization
nutritional.info	digital.fundraising.techniques
communicate.symptoms.to.healthcare.provider	emergency.care.practices
medication.management	supply.chain.logistics.for.automotive.parts
ethics.in.social.sciences	policy.drafting
impact.assessment	vaccine.administration
error.debugging	event.technology.utilization
epidemiological.research	spectroscopic.techniques
strategic.investment.decision.making	application.follow.up
conflict.resolution.skills	fiber.identification
study.tips	digital.design
strategic.planning	online.marketing
legislative.research	philosophical.inquiry.in.religion
mathematical.modelling	sustainability.practices
fabric.dyeing	research.ethics
local.cuisine.exploration	error.handling
tax.planning	sound.design
cognitive.behavioral.therapy	group.dynamics.management
internet.of.things.integration	statistical.modeling
applying.limit.concepts	crisis.intervention
electric.vehicle.technology	corporate.tax.planning
habitat.restoration	hospitality.marketing
algorithm.optimization	theme.design
music.criticism	garment.design
sustainable.eating.practices	motivational.coaching
welding.techniques	anesthesia.management
color.mixing	sponsorship.acquisition
surgical.procedure.execution	ethical.leadership
meditative.practices	statistical.analysis
data.science.analytics	learning.environment.optimization
debugging.algorithms	geospatial.analysis

Table 59: (Part 10 of 18) 1143 skills extracted from topics in Tables 35, 37, and 39.

Skills	
sculpting.methods	time.management
user.experience.design	corporate.social.responsibility
spacecraft.design	nutritional.analysis
disaster.preparedness	sustainable.transport.planning
narrative.pacing.control	partner.synchronization
contextual.historical.analysis	artifact.analysis
performance.analysis	cyber.security.analysis
religious.literacy.development	educational.outreach
conflict.sensitive.reporting	international.law.compliance
payload.management	risk.assessment.analysis
diagnosis.identification	event.planning.fundamentals
vehicle.maintenance.and.repair	solving.differential.equations
route.planning	expressive.performance
teaching.strategies	creative.thinking
customer.experience.management	interview.preparation
trajectory.design	point.of.view.selection
track.symptom.progression	international.negotiation.techniques
carpentry.work	wind.turbine.maintenance
stress.management	theme.identification
palette.management	mathematical.modeling
managing.social.interactions	outbreak.response.strategy
industry.trends	languages.and.linguistics
differentiating.functions	ethics.compliance
oral.presentation	curriculum.development
marine.geology	content.distribution.networking
cash.flow.analysis	genre.identification
feedback.assessment	graphing.functions
ritual.analysis	machine.learning
cnc.programming	ecological.conservation
social.media.and.digital.marketing	troubleshooting.electrical.issues
machine.learning.implementation	trend.forecasting
investment.advisory	database.management

Table 61: (Part 11 of 18) 1143 skills extracted from topics in Tables 35, 37, and 39.

Skills	
software-proficiency	musical-composition
emergency-care	severe-weather-response
critical-thinking-facilitation	regulatory-compliance
personal-finance	it-support
data-analysis-astronomy	calculating-determinants
installation-art-construction	plumbing-basics
portfolio-management	robotics-and-automation
source-evaluation	monitor-hydration-levels
financial-regulatory-compliance	policy-development-and-analysis
health-communication	health-education-program-development
intercultural-communication	safety-precautions
flooring-installation	smart-home-technology-integration
digital-communication	fundraising-management
voice-control	inventory-management
instrumentation-and-measurement	audience-engagement-strategies
poison-management	therapy-application
gardening-basics	public-health-initiatives
career-and-professional-development	robotics-engineering
global-mobility-management	metal-welding
international-law	flavor-pairing
music-theory-analysis	financial-management
digital-prototyping	literary-theory-application
veterinary-care-coordination	keyword-research
video-production	computational-linguistics-development
efficiency-optimization	stress-management-techniques
transport-technology	emergency-response
database-integration	health-education
travel-planning	singing-ability
emergency-preparedness	budget-management
risk-management	recursive-thinking
user-interface-design	patient-care-management
ethical-content-practices	instrument-playing

Table 63: (Part 12 of 18) 1143 skills extracted from topics in Tables 35, 37, and 39.

Skills	
theoretical-application	philosophical-writing
literary-studies	cultural-competency
marketing-promotion	transportation-planning
hospitality-and-tourism	training-routines
public-education-on-animal-health	tour-management
housekeeping-management	legal-advising
hotel-management	sustainable-agriculture
stakeholder-analysis	quality-control
competitive-analysis	portfolio-optimization
lifestyle-choices	home-and-family
property-valuation	comparative-study
memory-reinforcement	debugging-skills
recovery-support	fundraising-strategy
physical-fitness	installing-electrical-wiring
environmental-assessment	weaving-techniques
problem-solving-skills	cloud-computing-integration
first-aid-training	career-transition-advice
sleep-quality-improvement	remote-sensing
off-page-optimization	textual-analysis
analytical-thinking	customer-relationship-management
communication-skills	media-and-communication
conflict-resolution	chemical-waste-management
script-analysis	financial-management-for-nonprofits
debt-management	film-production
calorie-tracking	comparative-literature-study
risk-assessment-and-management	benchmarking-performance
supplement-advising	regression-analysis
algorithm-optimization	mental-health
understanding-vector-spaces	geographical-facts
energy-storage-solutions	active-passive-voice-conversion
sustainable-fashion-practices	first-aid
resource-scheduling	skill-development



Table 65: (Part 13 of 18) 1143 skills extracted from topics in Tables 35, 37, and 39.

Skills	
pattern.recognition	api.design
design.research.methods	interpret.symptom.severity
student.assessment.analysis	climate.modeling
geothermal.systems.design	textile.design
conservation.strategy.planning	instrument.proficiency
energy.efficiency.techniques	geographic.information.systems
technical.analysis	observing.etiquette.rules
interior.designing	stratigraphic.correlation
grammar.error.identification	waste.reduction.strategies
advocacy.strategy.development	pedagogical.content.knowledge
debt.financing	pet.safety.precautions
health.education.development	scientific.writing
philosophy.and.religion	speaking.fluency
quantitative.decision.making	crisis.communication
team.building	application.security
solar.panel.installation	traffic.management.systems
market.analysis	electrical.safety.inspection
species.identification	emotional.intelligence
energy.and.resources	technical.problem.solving
fashion.trends	poetic.analysis
stakeholder.negotiation	patina.application
link.building	food.presentation
healthy.eating.advice	solving.linear.equations
history.and.geography	cultural.adaptation.strategies
trip.planning	agile.methodologies
negotiation.tactics	digital.art.design
social.media.marketing	mindfulness.techniques
planetary.geology	security.risk.assessment
code.optimization	technology.integration
fashion.branding	career.planning
literary.criticism	data.driven.strategy
public.transport.planning	dialogue.crafting

Table 67: (Part 14 of 18) 1143 skills extracted from topics in Tables 35, 37, and 39.

Skills	
home_budget_management	digital_music_production
risk_factor_identification	reservation_handling
vaccination_schedule_management	creative_problem_solving
social_media_ads_creation	user_experience_evaluation
sleep_improvement_strategies	coding_and_debugging
client_communication	legislative_drafting
color_theory_application	sustainability_planning
human_resources_management	sustainable_agriculture_practices
literary_critique	active_listening
syntax_analysis	theme_exploration
choreography_design	epidemic_outbreak_investigation
stage_presence_development	automated_testing
client_education	sculpture_forming
educational_support	sound_engineering
art_criticism	complexity_analysis
weather_prediction	location_analysis
editing_and_proofreading	integrity_cultivation
sports_journalism	team_communication
planetary_science	setting_description
cloud_computing	curatorial_practices
technology_adaptation	disaster_recovery_planning
data_collection	packing_efficiency
qualitative_data_collection	media_studies_and_journalism
digital_design_3d_modeling	demand_forecasting
chromatographic_methods	maritime_navigation_systems
logistics_planning	medication_administration
innovation_management	vehicle_design_analysis
credit_score_improvement	exercise_routine_planning
comparative_analysis	earthquake_analysis
ethical_decision_making	cultural_dynamics
material_strength_testing	vocabulary_expansion
household_organization	integrating_functions

Table 69: (Part 15 of 18) 1143 skills extracted from topics in Tables 35, 37, and 39.

Skills	
reflective-judgment	blockchain-development
symbolic-interpretation	protocol-management
clinical-diagnosis	brush-technique
painting-techniques	economic-policy-analysis
developing-training-programs	simplifying-expressions
personal-care-and-hygiene	painting-walls
pet-care-abilities	peer-support
sports-coaching	user-experience-evaluation
disease-surveillance	creative-writing
cultural-etiquette-learning	fact-checking
sustainable-construction	environmental-policy
yarn-spinning	clay-modeling
resource-allocation	health-promotion
disease-prevention	biodiversity-monitoring
sociolinguistic-analysis	geographic-information-systems-for-resource-mapping
precision-machining	platform-optimization
technical-drawing	burn-treatment
relationship-building	ingredient-selection
epidemiology-research	seo-optimization
version-control-management	nutritional-planning
applying-the-quadratic-formula	physical-endurance
cybersecurity-analysis	big-data-handling
environmental-health-assessment	mental-wellness-guidance
community-health-mobilization	food-and-nutrition
protective-finishing	health-equity-and-access-analysis
public-administration	international-relations-and-global-studies
strategic-communication	spiritual-counseling-skills
practicing-politeness-forms	mineral-analysis
emotional-intelligence-management	data-collection-analysis
structural-analysis	art-history-analysis
narrative-building	cross-cultural-communication
fundamental-analysis	media-literacy

Table 71: (Part 16 of 18) 1143 skills extracted from topics in Tables 35, 37, and 39.

Skills	
fashion.and.textiles	woodworking.basics
emotional.expression.mastering	behavioral.counseling
security.practices	natural.disasters
hypothesis.testing	customer.engagement
content.analysis	fracture.stabilization
data.visualization	behavioral.training
cash.flow.forecasting	patience.cultivation
automotive.safety.standards.compliance	staff.training.and.development
team.collaboration	noun.verb.agreement
printmaking.methods	cultural.integration.facilitation
verb.tense.consistency	decision.making
wildlife.monitoring	intercultural.competency
apply.first.aid.for.symptomatic.relief	lighting.design
brake.replacement	reading.comprehension
kitchen.safety	bill.drafting
technical.seo	game.strategy.development
belt.inspection.replacement	wellness.counseling
improvisational.skills	platform.navigation
public.health.surveillance	system.architecture.design
scriptural.interpretation	compliance.management
emergency.preparedness	cultural.interpretation
trend.analysis	animal.grooming.techniques
astronomical.photography	research.design
law.enforcement.compliance	music.performance
social.media.strategies	non.profit.governance
visual.communication.skills	film.critique
performance.metrics.analysis	coding.proficiency
communication.protocol.design	data.management
digital.ethics.management	negotiating.conflict.resolution
machine.maintenance	marketing.strategy
interfaith.communication	vision.formulation
ethical.reporting	change.management

Table 73: (Part 17 of 18) 1143 skills extracted from topics in Tables 35, 37, and 39.

Skills	
surgical.techniques	software.debugging
system.administration	astronomy.space-exploration
oil.change	project.financial.modeling
improvisation.techniques	costume.management
pronunciation.accuracy	fluid.checks
story.pitching	content.creation
dental.care	team.management
community.engagement	narrative.construction
foreign.language.proficiency	statutory.analysis
jewelry.making.techniques	voice.projection.training
philanthropy.and.non-profit.sector	satellite.imagery.interpretation
sanitation.protocol.implementation	character.analysis
physical.expressions	unit.testing
test.preparation	strategic.allocation
gadget.reviews	construction.technology.integration
choking.remedy	attention.to.detail
progress.tracking	cooking.techniques
carbon.footprint.analysis	spectroscopic.analysis
literary.analysis	geophysical.surveying
behavioral.training.methods	ethical.guidance
resume.writing	performance.evaluation
event.correlation	resource.management
crypto.algorithm.application	mergers.and.acquisitions.strategy
financial.analysis	public.health.communication
performance.techniques	brand.alignment
health.monitoring.procedures	nutrition.management
health.policy.advocacy	treatment.plan.design
audience.analysis	conflict.resolution.techniques
ceramics.pottery	orbit.dynamics
climate.change.adaptation	water.resource.management
emotional.regulation	diagnostic.testing
digital.proficiency	star.identification

Table 75: (Part 18 of 18) 1143 skills extracted from topics in Tables 35, 37, and 39.

Skills	
bronze-casting	focus-enhancement
brand-storytelling	ui-ux-design
grammar-rules-teaching	outcome-evaluation
student-motivation	empathetic-listening
quality-control-management	hydration-nudges
manual-dexterity	laboratory-safety
lifestyle-and-leisure	budget-management
educational-research-methods	renewable-energy-technology
approximating-functions	field-mapping
using-theorems-in-calculus	religious-impact-assessment
home-safety-procedures	healthy-cooking
identify-common-symptoms	environment-and-ecology
cross-platform-development	event-planning
retirement-planning	radiology-technique
innovation-creativity	writing-clarity
visual-storytelling	green-infrastructure-design
numerical-computation	3d-modeling
historical-linguistics-research	stakeholder-engagement
songwriting-composition	pharmacological-knowledge
glass-etching	media-experimentation
hydraulic-engineering	research-techniques
bioarchaeological-analysis	quantitative-reasoning
rhythm-identification	quantum-computing
recommend-balanced-meals	study-design-and-conduct
course-design	celestial-navigation
multimedia-content-creation	analyze-nutrient-content
food-preparation-techniques	project-collaboration

Table 77: (Part 1 of 5) 18 Query/Task Types extracted from interactions with GPT-4-Turbo.

Query Type	Description
Information-Seeking	This includes any query where the user is looking to find out facts, data, explanations, or learn about a topic.
Help-Seeking	Queries where the user needs assistance in solving a problem or performing an action. This could be technical support, troubleshooting, or guidance on personal issues.
Instructional	Queries that specifically request detailed, step-by-step instructions or procedural guidance. This category is designed to assist users in understanding and executing tasks by breaking them down into sequential, manageable steps. Whether it's a practical day-to-day task, a complex technical procedure, or a creative process, the AI provides a clear, methodical approach to accomplishing specific objectives.
Conversational	These are queries where the user is possibly looking for engagement more than specific information or tasks. This can include small talk or generative interactions aimed at entertainment or companionship.

Table 79: (Part 2 of 5) 18 Query/Task Types extracted from interactions with GPT-4-Turbo.

Query Type	Description
Narrative	Queries where the user is interested in hearing stories, experiences, or detailed accounts of events. These can be historical, personal, or fictional.
Planning	Queries that assist in planning or organizing activities, events, or projects.
Situational	Queries related to specific situations or scenarios that the user is facing, asking for tailored advice or solutions.
Interpretative	Queries that ask for interpretation of texts, artworks, or other cultural artifacts.
Decision-Making	Queries that assist the user in making decisions by evaluating options, risks, and benefits.
Task Execution	Queries where the user delegates the completion of a specific task or action to the AI. This involves the AI taking on roles that might require decision-making, processing, or interacting with other systems to achieve the desired outcome.



Table 81: (Part 3 of 5) 18 Query/Task Types extracted from interactions with GPT-4-Turbo.

Query Type	Description
Digital Communication Design and Strategy	Queries focused on designing and strategizing content for optimal communication across digital platforms. This includes creating engaging designs and coherent strategies for websites, blogs, podcasts, emails, and digital essays.
Code Generation	Queries where users directly ask for guidance on implementing specific functions, features, or algorithms in a particular programming language. These queries explicitly request code snippets, examples, or step-by-step instructions on how to implement the desired functionality.
Fact-Seeking	Queries where the user is looking for specific, factual information or data points. These queries are often straightforward and can be answered with a concise response. The focus is on providing accurate, reliable information from trustworthy sources.
Comparative	These queries involve comparing different items, ideas, or scenarios. Users often seek assistance in making decisions or understanding differences.

Table 83: (Part 4 of 5) 18 Query/Task Types extracted from interactions with GPT-4-Turbo.

Query Type	Description
Interpretive Explanation	Queries in which the user seeks a detailed explanation or interpretation of a provided text snippet. This could include literary analysis, code explanation, or any form of textual dissection. The AI acts as an analytical tool to help users understand complex passages, technical descriptions, or conceptual writings.
Error Detection	Queries where the user seeks assistance in finding and diagnosing errors or bugs in provided materials. This could include syntactical errors in code, grammatical mistakes in written text, or inconsistencies in data sets. The AI acts as a diagnostic tool to help pinpoint and suggest corrections for these issues.
Feedback-Seeking	Queries where users are looking for feedback on their ideas, creations, or opinions. This can be particularly relevant in educational, artistic, or professional contexts.

Table 85: (Part 5 of 5) 18 Query/Task Types extracted from interactions with GPT-4-Turbo.

Query Type	Description
Clarification-Seeking	These queries aim to clarify confusion or get more detailed information about a previously mentioned or understood topic. Users might need further explanation or a more refined understanding of a complex issue.

## K Skill Extraction Prompts

### K.1 Prompt for Skill Extraction (INSTRUCT-SKILLMIX-D)

```
Consider the following question. Label this question with a skill that would be required
to solve the question. Basically, you should be able to use the skill as a dictionary
key in python. The skill name should be lower case letters only. The skill name should
be very descriptive and you may use multiple words to describe the skills required in the
question. If you do use multiple words per question, then join them by an underscore.
{text}
Your answer should be as follows:
<name of the skill>, reason: <reason for the skill>
```

### K.2 Prompt for Semantic Clustering (INSTRUCT-SKILLMIX-D)

```
Given the list of skills required to solve various questions, your task is to categorize
these skills into descriptive and specific groups. Each category should not only capture
the essence of the skills it includes but also reflect clear, distinct areas of expertise
or application. Use terminology that is precise and specific to the tasks those skills
accomplish. Categories should be narrow enough to provide meaningful insight into the
specialization they represent. Format category names in lowercase, joining multiple
words with underscores.
For each category you create, provide a comprehensive rationale explaining:
- Why these particular skills are grouped together.
- How the category name specifically reflects the specialization and distinctiveness of
the skills.

### Skills List:
{skills_joined_str}

### Examples:
- Category name: `web.development`
- Included skills: html.css.design, javascript.interactivity, web.performance.optimization
- Rationale: This category focuses specifically on the skills required to build and
optimize web interfaces, distinguishing it from broader software development.

- Category name: `machine.learning.techniques`
- Included skills: machine.learning.model.creation, understanding.algorithmic.fairness,
machine.learning.application.in.finance
- Rationale: These skills are grouped under 'machine.learning.techniques' because
they directly relate to the development and application of machine learning models,
emphasizing specific use cases and ethical considerations, which are distinct from
general programming skills.

Please ensure your category names are informative, specific, and reflective of the skills
grouped under them. This will aid in their practical utility and clarity.
```

### K.3 Prompt for Creating Synthetic Data (INSTRUCT-SKILLMIX-D)

#### Prompt 1

I am creating a dataset of engaging human-AI interactions that demonstrate the capabilities and value of AI models across diverse skills. Please help me generate high-quality examples for this dataset.

I will first give you a set of {num.skills} skills that are valuable to humans.

Next, you would create a query that a human might plausibly ask whose answer involves using those {num.skills} skills.

For this query you will generate a good length answer using your best helpfulness and wisdom. Please try to avoid generic sound queries and answers. People often have concrete problems and would give important details. The answer should also give some concrete details and examples, but not get verbose.

Before I give you the set of {num.skills} skills, I will provide some examples involving 2 skills.

Example 1:

### Skills: self.reflection.skills, linguistics

### Instruction: I often find myself in conflicts at work due to miscommunication and differing perspectives. How can I improve my conflict resolution skills by incorporating self-reflection and effective communication techniques into my daily interactions with colleagues?

### Response: To enhance your conflict resolution skills in the workplace, integrate these self-reflection and linguistic strategies into your daily routine:

Self-Reflection Techniques:

End-of-Day Review: Spend 10 minutes reflecting on interactions, noting successes and areas for improvement. For example, I successfully de-escalated a tense situation with a client by actively listening to their concerns and finding a mutually beneficial solution."

Conflict Log: Record details of conflicts, your responses, and outcomes in a dedicated journal. Analyze patterns monthly to identify growth opportunities. For instance, I noticed I tend to become defensive when criticized. I will practice pause and respond techniques to improve my reactions."

Linguistic Techniques:

Empathy Statements: Use phrases that demonstrate understanding and validate others' perspectives, such as I can see why you feel that way or That must be challenging for you. In a recent meeting, saying I understand your frustration with the project timeline helped a coworker feel heard and open to finding a solution together.

Reflective Listening: Paraphrase the other person's key points before responding to ensure clarity and show you're listening. For example, So, what I'm hearing is that you're concerned about the budget allocation. Let's discuss some options to address that."

Integration into Routine:

Weekly Practice Sessions: Dedicate an hour each week to role-play conflict scenarios with a trusted colleague, focusing on applying self-reflection insights and linguistic techniques. Provide each other with constructive feedback for continuous improvement.

Real-Time Application: When a conflict arises, take a moment to mentally review your phrase bank and select an appropriate response. After the interaction, briefly reflect on what worked well and what could be improved next time.

By consistently practicing these strategies, you'll build stronger habits and enhance your overall conflict resolution skills in the workplace.

Example 2:

### Skills: psychological.understanding, data.handling.and.analysis

### Instruction: I recently started working as a data analyst for a meal kit delivery service. We've noticed a significant increase (20%) in customer churn over the past quarter. How can I leverage data analysis and psychological principles to identify the root causes of churn and develop targeted strategies to improve customer retention?

### Response: To address the 20% increase in customer churn for your meal kit service, employ a combination of data analysis and psychological insights:

Data Analysis:

Customer Segmentation: Analyze churn rates by customer segments such as age, location, and dietary preferences. For example, if you find that vegetarian customers have a higher churn rate, you can focus on improving vegetarian meal options.

Churn Timing Analysis: Identify common timeframes when churn occurs, such as after the first month or after specific promotions end. If churn spikes after the first month, consider implementing a more robust onboarding process or extending promotional offers.

Psychological Insights:

Exit Surveys: Conduct exit surveys to gather direct feedback from churning customers. Include open-ended questions to uncover deeper insights. For instance, What was the primary reason for canceling your subscription? and What could we have done better to keep you as a customer?"

Customer Journey Mapping: Map out the customer journey to identify pain points and opportunities for improvement. Apply psychological principles like the Peak-End Rule to prioritize the most impactful moments. Focus on creating a positive onboarding experience and ensuring a satisfying last interaction before cancellation.

#### Actionable Strategies:

**Personalized Retention Campaigns:** Develop targeted email campaigns for at-risk segments, addressing their specific concerns. For example, send vegetarian customers an email highlighting new plant-based recipes and offering a discount on their next box.

**Flexible Subscription Options:** Introduce more flexible subscription plans, such as the ability to skip weeks or customize delivery frequency. This addresses the psychological need for autonomy and control, reducing the likelihood of churn.

By leveraging data-driven insights and psychological principles, you can create a targeted approach to reduce churn and improve customer satisfaction in your meal kit service.

#### Example 3:

### Skills: food.related.knowledge.and.skills, internet.technologies

### Instruction: I'm a food truck owner looking to expand my online presence and attract more customers. I want to create a website that allows customers to customize their meals, track my truck's real-time location, and place orders for pickup. What web technologies should I consider to build a user-friendly and feature-rich website for my food truck business?

### Response: To create a robust and interactive food truck website with meal customization and real-time location tracking, consider the following web technologies:  
Front-End:

**React:** Use React, a popular JavaScript library, for building a dynamic and responsive user interface. React's component-based architecture allows for easy management of real-time updates, such as displaying the current location of the food truck.

**HTML5 Geolocation API:** Leverage the HTML5 Geolocation API to capture the user's location (with their permission) and display nearby food truck locations on an interactive map.

**Leaflet or Google Maps:** Integrate Leaflet, an open-source mapping library, or Google Maps to display food truck locations and enable users to get directions. Customize the map markers to show additional information like menu items or special offers.

#### Back-End:

**Node.js and Express:** Use Node.js, a server-side JavaScript runtime, and the Express web application framework to handle server-side logic and API routing. Node.js's non-blocking I/O model is well-suited for real-time applications like location tracking.

**WebSocket:** Implement WebSocket communication to enable real-time updates between the server and the client. When a food truck's location changes, the server can push the updated coordinates to connected clients, ensuring users always see the most current location.

**Firebase Realtime Database:** Store and sync food truck location data in real-time using Firebase Realtime Database. This NoSQL cloud database allows for efficient storage and retrieval of location data, enabling seamless real-time updates across multiple devices.

#### User Experience:

**Progressive Web App (PWA):** Develop the website as a PWA to provide a native app-like experience. PWAs offer features like offline accessibility, push notifications, and home screen installation, enhancing user engagement and convenience.

**Responsive Design:** Ensure the website is fully responsive and optimized for various screen sizes, particularly mobile devices. Use CSS media queries and flexible layouts to create a seamless experience across desktops, tablets, and smartphones.

**Accessibility:** Follow web accessibility guidelines (WCAG) to ensure the website is usable by people with disabilities. Implement proper headings, alt text for images, and keyboard navigation to create an inclusive user experience.

By combining these web technologies, you can create a feature-rich and user-friendly food truck website that offers meal customization and real-time location tracking, providing a seamless experience for your customers.

#### Example 4:

### Skills: algorithmic.and.programming.skills, digital.marketing

### Instruction: As a software engineer, I've developed a sophisticated recommendation engine that I believe could greatly benefit e-commerce businesses. I want to showcase my work through a technical blog post and live demo to attract potential clients and demonstrate the value of my solution. How can I structure my blog post and demo to effectively highlight the technical aspects and business benefits of my recommendation engine?

### Response: To effectively showcase your recommendation engine and attract potential e-commerce clients, structure your technical blog post and live demo as follows:

#### Blog Post Structure:

**Attention-Grabbing Introduction:** Start with a compelling hook that highlights the value of personalized recommendations in e-commerce. Use statistics or real-world examples to demonstrate the impact on sales and customer satisfaction.

**Problem Statement:** Clearly define the challenges e-commerce businesses face in providing relevant product recommendations at scale. Discuss common pain points like data sparsity, cold-start problems, and real-time updates.

**Technical Deep Dive:** Explain the core components of your recommendation engine, such as collaborative filtering, content-based filtering, or hybrid approaches. Use diagrams and code snippets to illustrate your architecture and key algorithms. Highlight any innovative techniques you've employed, such as deep learning or reinforcement learning.

**Performance Metrics:** Present quantitative results that showcase the effectiveness of your recommendation engine. Include metrics like precision, recall, F1 score, and mean average precision. Compare your results to industry benchmarks or popular open-source recommendation libraries to demonstrate your engine's superiority.

**Scalability and Efficiency:** Discuss how your recommendation engine handles large-scale

data and real-time updates. Explain your strategies for efficient data processing, such as parallel computing or incremental updates. Provide performance benchmarks to highlight the speed and scalability of your solution.

Live Demo: E-commerce Store Integration: Create a mock e-commerce store that seamlessly integrates your recommendation engine. Showcase personalized product recommendations based on user interactions, such as viewed items, purchases, or ratings.

Real-Time Recommendations: Demonstrate how your engine adapts in real-time as users navigate the store. For example, show how the recommendations update dynamically based on the user's browsing history or cart contents.

Explanations and Transparency: Provide clear explanations for each recommendation, such as "Customers who bought this item also bought..." or "Recommended based on your recent searches." This transparency builds trust and helps users understand the relevance of the recommendations.

A/B Testing: Implement an A/B testing feature that allows potential clients to compare the performance of your recommendation engine against a default or random recommendation system. Visualize the results through metrics like click-through rates or conversion rates to highlight the tangible benefits of your solution.

Promotion and Engagement:

Targeted Social Media Ads: Create compelling social media ads that highlight the key features and benefits of your recommendation engine. Target decision-makers in the e-commerce industry, such as marketers, product managers, and CTOs. Use eye-catching visuals and clear calls-to-action to drive traffic to your blog post and live demo.

LinkedIn Thought Leadership: Share your blog post on LinkedIn and engage in discussions related to e-commerce personalization. Participate in relevant groups and forums to establish yourself as a thought leader in the field. Offer valuable insights and answer questions to build trust and credibility with potential clients.

Webinars and Workshops: Host webinars or workshops that provide a hands-on experience with your recommendation engine. Walk participants through the setup process and demonstrate how to integrate your solution into their existing e-commerce platforms. Offer Q&A sessions to address specific concerns and showcase your expertise.

By structuring your technical blog post and live demo in this manner, you can effectively showcase the capabilities and benefits of your recommendation engine. Combine this with targeted digital marketing techniques to reach and engage potential e-commerce clients, ultimately driving interest and adoption of your solution. Example 5:

### Skills: geographical and historical knowledge, analytical thinking

### Instruction: I'm a history enthusiast planning a two-week trip to Eastern Europe with my family. We're particularly interested in exploring medieval architecture and learning about the region's experiences during the Cold War era. How can I create an itinerary that combines visits to historical sites, museums, and cultural experiences to gain a deeper understanding of Eastern Europe's past and present?

### Response: Here's a curated itinerary for a historically rich two-week trip through Eastern Europe, focusing on medieval and Cold War sites:

Week 1: Poland and Czech Republic

Krakow (3 days):

Wawel Castle: Explore this 14th-century castle complex, which served as the residence of Polish kings. Marvel at the Gothic and Renaissance architecture, and visit the Dragon's Den, a legendary cave beneath the castle.

Oskar Schindler's Factory Museum: Learn about the Holocaust and Oskar Schindler's efforts to save Jewish workers during World War II. The museum offers a immersive experience, recreating the wartime atmosphere of Krakow.

Nowa Huta: Take a guided tour of this planned socialist city, built during the Cold War era. Visit the iconic Lord's Ark Church, which became a symbol of resistance against the communist regime.

Warsaw (2 days):

Old Town: Stroll through the meticulously reconstructed Old Town, which was destroyed during World War II. Visit the Market Square, the Warsaw Barbican, and St. John's Cathedral to admire the Gothic and Renaissance architecture.

Warsaw Uprising Museum: Discover the story of the 1944 Warsaw Uprising, a major resistance operation against Nazi occupation. The museum features interactive exhibits, original artifacts, and a 3D movie that brings history to life.

Palace of Culture and Science: Explore this imposing Stalinist-era skyscraper, which remains the tallest building in Poland. Take an elevator to the observation deck for panoramic views of Warsaw.

Prague (2 days):

Prague Castle: Visit the world's largest ancient castle complex, dating back to the 9th century. Explore the Gothic St. Vitus Cathedral, the Romanesque St. George's Basilica, and the Golden Lane, a picturesque row of colorful houses.

Charles Bridge: Walk across this iconic 14th-century stone bridge, lined with baroque statues. Enjoy street musicians and artists, and take in the stunning views of the Vltava River and the Old Town.

Nuclear Bunker Tour: Experience the chilling reality of the Cold War by visiting a former Soviet nuclear bunker. Learn about the strategies and technology employed during this tense period in history.

Week 2: Hungary and Romania

Budapest (3 days):

Buda Castle: Discover this 13th-century castle, which has been rebuilt and expanded over the centuries. Visit the Hungarian National Gallery, the Budapest History Museum, and the Matthias Church, known for its colorful tiled roof.

House of Terror Museum: Explore this powerful museum dedicated to the victims of the fascist and communist regimes in Hungary. The exhibits are housed in the former headquarters of the Arrow Cross Party and the communist secret police.

Memento Park: Visit this open-air museum showcasing monumental statues from the Soviet era. Learn about the propaganda and ideology behind these imposing sculptures.

Bucharest (2 days):

Palace of Parliament: Tour the world's largest civilian building, constructed during the communist era under the rule of Nicolae Ceaușescu.

Marvel at the opulent interiors and learn about the controversial history of this massive structure.

Old Town: Explore the charming streets of Bucharest's Old Town, lined with historical buildings, churches, and cafes. Visit the ruins of the Old Princely Court, which served as the residence of Wallachian princes.

Revolution Square: Pay tribute to the heroes of the 1989 Romanian Revolution at this significant square. See the Memorial of Rebirth, which honors those who lost their lives fighting against the communist regime.

Travel Tips:

Book guided tours with knowledgeable local guides to gain deeper insights into the historical context and personal stories behind each site.

Stay in centrally located accommodation to minimize travel time and maximize your exploration of each city.

Use public transportation or ride-sharing services to navigate within cities, as parking can be challenging and expensive in historical areas.

By following this itinerary, you'll embark on a fascinating journey through Eastern Europe's medieval and Cold War past, gaining a profound appreciation for the region's rich history and resilience.

Now, it is your turn! Given the guidelines and examples above, please create a query that a human might plausibly ask whose answer involves using all skills below:

### Skills: {skills\_str}

## Prompt 2

Thanks! It looks like the response may have been cut off. Please rewrite the entire response, making sure it fits within the length constraint. No need to include verbose or extraneous information. Thanks!

## Prompt 3

Pretend you are the human asking the question, and make a list of weaknesses and strengths of the response. Currently the response looks a bit generic. Perhaps it could benefit from some concrete examples and details.

## Prompt 4

Now, please refine your instruction and response (keep the strong parts and work on the weak parts).

## Prompt 5

It looks like the improved response may have been cut off. Please rewrite the entire improved response, making sure it fits within the length constraint. No need to include verbose or extraneous information. Thanks again!

## K.4 Prompt for Extrapolating Skills (INSTRUCT-SKILLMIX)

First, consider the following definitions:

- Knowledge: A set of information, facts, descriptions, and theories acquired through education or experience.
- Topic: A subset of Knowledge that focuses on a particular area or subject.
- Skill: A function that transforms Knowledge into actions or performances that achieve specific outcomes.

Now, for the topic of topic, please provide me with multiple skills (in snake case) that could be relevant to the topic.



## K.5 Prompt for Creating Synthetic Data (INSTRUCT-SKILLMIX)

### Prompt 1

I am creating a dataset of engaging human-AI interactions that demonstrate the capabilities and value of AI models across diverse skills. Please help me generate high-quality examples for this dataset.

I will first give you a set of {num.skills} skills that are valuable to humans.

Next, you would create a query that a human might plausibly ask whose answer involves using those {num.skills} skills.

For this query you will generate a good length answer using your best helpfulness and wisdom. Please try to avoid generic sound queries and answers. The answer should also give some concrete details and examples, but not get verbose.

Before I give you the set of {num.skills} skills, I will provide some examples involving 2 skills.

Example 1:

### Task Type: Information-Seeking  
### Skills: psychological.understanding, data.handling.and.analysis  
### Instruction: What are the primary psychological theories that explain the impact of social media on adolescent self-esteem, and what evidence supports these theories? Additionally, what statistical techniques have been used in the literature to dissect these relationships, and how do these techniques help in understanding the nuances of psychological data?  
### Response: The impact of social media on adolescent self-esteem is a widely studied topic within psychology. Various theories and statistical techniques are employed to understand and measure these effects. Below, I'll outline some of the primary theories, the supporting evidence, and the statistical techniques used in this research.  
### Response: The impact of social media on adolescent self-esteem is a widely studied topic within psychology. Various theories and statistical techniques are employed to understand and measure these effects. Below, I'll outline some of the primary theories, the supporting evidence, and the statistical techniques used in this research.

**\*\*Psychological Theories and Evidence\*\***

Social Comparison Theory: Adolescents often assess their worth by comparing themselves with peers on social media. A study by Vogel et al. (2014) found that Facebook usage was associated with negative self-perceptions when users compared themselves to others they perceived as better off. This suggests a direct correlation between the frequency of social media use and increased feelings of inferiority.

Self-Presentation Theory: Adolescents feel pressure to present themselves in an idealized way on platforms like Instagram and Snapchat. Research by Chou and Edge (2012) highlights that the need for positive feedback and fear of negative evaluations can lead to cycles of anxiety and lowered self-esteem when these expectations are not met.

Media Richness Theory: Different types of social media platforms impact self-esteem differently. A study by Sherman et al. (2016) compared text-based versus image-based platforms, finding that image-based platforms (like Instagram) can exacerbate body image concerns and anxiety more than text-based platforms due to the intense focus on physical appearance.

Statistical Techniques and Applications

Regression Analysis: Used extensively to quantify the relationship between social media usage and self-esteem. For instance, studies might control for variables like age, gender, and socioeconomic status to isolate the effect of social media usage.

Structural Equation Modeling (SEM): SEM enables researchers to construct a comprehensive model of how different factors like peer influence, media consumption, and personal traits interact to affect self-esteem. An example is Kross et al.'s (2013) study, which modeled the pathways through which social media affects mood and self-esteem over time.

Longitudinal Studies and Growth Curve Modeling: These techniques provide insights into how self-esteem trajectories change over time with social media use. For instance, longitudinal research might track self-esteem changes from early to late adolescence, revealing critical periods where impacts are most pronounced.

Mediation and Moderation Analyses: These are crucial for exploring underlying mechanisms and boundary conditions. For example, studies might explore how the negative impact of social media on self-esteem can be mediated by social support or moderated by individual resilience.

**\*\*Contradictory Evidence\*\***

While many studies suggest a negative impact of social media on self-esteem, some

research offers a different perspective. For instance, Valkenburg et al. (2006) found that online interactions can enhance well-being by providing social support and a platform for self-expression. These findings suggest that the impact of social media on self-esteem is not universally negative and can vary based on usage patterns and individual differences.

#### **\*\*Practical Implications\*\***

Understanding these dynamics is crucial for developing targeted interventions. For instance, educational programs that teach critical media literacy can help adolescents navigate social media healthily. Moreover, knowing which aspects of social media use trigger negative self-comparisons can guide the development of features that promote positive interaction and reduce the emphasis on appearance.

#### **Example 2:**

### Task Type: Help-Seeking  
### Skills: selfreflection.skills, linguistics  
### Instruction: I often find myself in conflicts at work due to miscommunication and differing perspectives. How can I improve my conflict resolution skills by incorporating self-reflection and effective communication techniques into my daily interactions with colleagues?  
### Response: To enhance your conflict resolution skills in the workplace, integrate these self-reflection and linguistic strategies into your daily routine:  
Self-Reflection Techniques:  
End-of-Day Review: Spend 10 minutes reflecting on interactions, noting successes and areas for improvement. For example, I successfully de-escalated a tense situation with a client by actively listening to their concerns and finding a mutually beneficial solution."  
Conflict Log: Record details of conflicts, your responses, and outcomes in a dedicated journal. Analyze patterns monthly to identify growth opportunities. For instance, I noticed I tend to become defensive when criticized. I will practice pause and respond techniques to improve my reactions."  
Linguistic Techniques:  
Empathy Statements: Use phrases that demonstrate understanding and validate others' perspectives, such as I can see why you feel that way or That must be challenging for you. In a recent meeting, saying I understand your frustration with the project timeline helped a coworker feel heard and open to finding a solution together.  
Reflective Listening: Paraphrase the other person's key points before responding to ensure clarity and show you're listening. For example, So, what I'm hearing is that you're concerned about the budget allocation. Let's discuss some options to address that."  
Integration into Routine:  
Weekly Practice Sessions: Dedicate an hour each week to role-play conflict scenarios with a trusted colleague, focusing on applying self-reflection insights and linguistic techniques. Provide each other with constructive feedback for continuous improvement.  
Real-Time Application: When a conflict arises, take a moment to mentally review your phrase bank and select an appropriate response. After the interaction, briefly reflect on what worked well and what could be improved next time.  
By consistently practicing these strategies, you'll build stronger habits and enhance your overall conflict resolution skills in the workplace.

#### **Example 3:**

### Task Type: Information-Seeking  
### Skills: geographical.and.historical.knowledge, analytical.thinking  
### Instruction: What were the major geopolitical changes in Europe following World War II, and how have these changes influenced current European political boundaries?  
### Response: \*\*Major Geopolitical Changes in Europe After World War II\*\*

1. Division of Germany: The division of Germany into East and West Germany in 1949 was a critical symbol of the Cold War's onset. This division, marked physically by the Berlin Wall from 1961, lasted until the reunification in 1990 and significantly influenced European political dynamics.
2. Creation of the Iron Curtain: Winston Churchill's term, the Iron Curtain, described the ideological and physical boundary separating the Soviet-controlled Eastern Bloc and the Western democracies. This division dictated political alignments and conflicts during the Cold War, influencing policies and international relations.
3. Decolonization: Following the war, European powers like Britain and France began a significant process of decolonization. This period saw the independence of nations such as India in 1947 and the rapid decolonization of Africa in the 1960s, marking a major shift in global dynamics and European foreign policies.
4. Formation of the European Union (EU): The EU's origins lie in the European Coal and Steel Community in 1951, evolving into the European Economic Community by 1957. These alliances, expanding to include more countries over the decades, aimed to foster economic cooperation and prevent further wars in Europe, influencing both economic and political policies within the continent.
5. NATO and the Warsaw Pact: The establishment of NATO in 1949 by Western countries

was a strategic move for collective security against the Soviet threat. The Soviet response, the Warsaw Pact in 1955, defined the military alliances in Europe, solidifying the East-West divide.

#### **\*\*Influence on Current European Political Boundaries\*\***

1. German Reunification: The fall of the Berlin Wall in 1989 and the subsequent reunification of East and West Germany in 1990 reshaped Germany's role in Europe, altering both its internal and external political boundaries.
2. EU Expansion: The EU's expansion has included many former Eastern Bloc countries, fundamentally changing the political landscape of Europe. The Schengen Agreement, implemented in 1995, minimized the importance of national boundaries within the EU, promoting free movement and economic integration.
3. Breakup of Yugoslavia and the Soviet Union: The disintegration of Yugoslavia into seven successor states throughout the 1990s and the Soviet Union into 15 independent countries in 1991 dramatically redrew political boundaries. These events, rooted in ethnic tensions and political upheavals, continue to influence regional stability and alignments.

#### **\*\*Case Studies: Key Treaties and Shifts in Alliances\*\***

1. Treaty of Paris (1951): This treaty established the European Coal and Steel Community, a foundational step towards European integration. It set precedents for future economic policies and cooperative frameworks within Europe, promoting peace and economic stability across former wartime adversaries.
2. Impact of NATO and the Warsaw Pact on Poland: Poland's transition from a Warsaw Pact member to a NATO member in 1999 exemplifies the dramatic shift in military and political alliances. This change significantly influenced Poland's foreign and security policies, reflecting broader geopolitical shifts in Eastern Europe post-Cold War.

Now, it is your turn! Given the guidelines and examples above, please create a query that aligns with the following task type:  
{query.type}

Include specific text snippets when they are essential to analyze and validate your query effectively. Make sure that the query is one that a human might plausibly ask whose answer involves using all skills below:

### Task Type: {query.str}  
### Skills: {skills.str}

## **Prompt 2**

Thanks! It looks like the response may have been cut off. Please rewrite the entire response, making sure it fits within the length constraint. No need to include verbose or extraneous information. Thanks!

## **Prompt 3**

Pretend you are the human asking the question, and make a list of weaknesses and strengths of the response. Currently the response looks a bit generic. Perhaps it could benefit from some concrete examples and details.

## **Prompt 4**

Now, please refine your instruction and response (keep the strong parts and work on the weak parts).

## **Prompt 5**

It looks like the improved response may have been cut off. Please rewrite the entire improved response, making sure it fits within the length constraint. No need to include verbose or extraneous information. Thanks again!

## L Comparison of Responses

In this section, we show side-by-side comparisons of responses from various models to the following instruction from AlpacaEval.

I'm trying to teach myself to have nicer handwriting. Can you help?

Observe that responses produced by models finetuned on INSTRUCT-SKILLMIX tend to include concrete examples (whereas responses produced by the other models do not always).

### Response by LLaMA-3-8B-Base SFT on INSTRUCT-SKILLMIX(k=2)-4K

Absolutely! Improving your handwriting involves both technique and practice. Here's a concise guide to help you achieve nicer handwriting:

#### \*\*1. Posture and Grip:\*\*

- **\*\*Posture:\*\*** Sit upright with your paper at a 45-degree angle to reduce arm and hand movement, which helps prevent smudging.
- **\*\*Grip:\*\*** Use a relaxed, comfortable grip with your pen or pencil resting lightly in your fingers. This promotes smoother movement and reduces pressure on the paper.

#### \*\*2. Pen Lifts:\*\*

- **\*\*When to Lift:\*\*** Between words and sentences to create clear breaks, especially in cursive.
- **\*\*Purpose:\*\*** Pen lifts help maintain the integrity of each letter, preventing them from blending together and enhancing readability.

#### \*\*3. Slant and Direction:\*\*

- **\*\*Natural Slant:\*\*** Write with a natural slant that's comfortable for your hand, typically a slight upward slope helps legibility.
- **\*\*Direction:\*\*** Write continuously without lifting your pen too much, maintaining a fluid motion can enhance the flow and elegance of your handwriting.

#### \*\*4. Pressure and Speed:\*\*

- **\*\*Pressure:\*\*** Apply consistent pressure, avoiding excessive force which can cause fatigue and messy handwriting.
- **\*\*Speed:\*\*** Write at a comfortable pace that allows your handwriting to remain clear and legible.

#### \*\*5. Regular Practice:\*\*

- **\*\*Routine:\*\*** Set aside time daily for handwriting practice, such as 10 minutes a day writing sentences or paragraphs.
- **\*\*Variety:\*\*** Mix practice sessions with both structured exercises focusing on specific letters and free writing to maintain fluidity.

#### \*\*6. Paper and Tools:\*\*

- **\*\*Paper Choice:\*\*** Use smooth, medium-weight paper that allows for control without too much slip.
- **\*\*Tool Variety:\*\*** Experiment with different writing tools|pens, pencils, markers|to find which provides the best feel and outcome for your handwriting.

#### \*\*7. Evaluation and Adjustment:\*\*

- **\*\*Self-Review:\*\*** Regularly assess your handwriting samples to note improvements and areas needing more work.
- **\*\*Feedback:\*\*** Ask for feedback from others or enroll in a calligraphy class to gain insights from experienced instructors.

By focusing on these specific elements and committing to regular practice, your handwriting will improve. Remember, small, consistent efforts can lead to significant improvements over time.

This response is concise yet informative, providing specific, actionable advice for each step of the handwriting improvement process. It also includes practical tips and suggestions for maintaining motivation and evaluating progress, making it a comprehensive guide for anyone looking to refine their handwriting.

## Response by Mistral-7B-Base-v0.2 SFT on INSTRUCT-SKILLMIX(k=2)-4K

Improving your handwriting is a rewarding skill to develop and can be both therapeutic and practical. Here's a structured guide with specific techniques and exercises to help you achieve elegant, legible handwriting:

### **\*\*Step 1: Posture and Setup\*\***

- **\*\*Posture\*\***: Sit with a straight spine and relax your shoulders to prevent tension.
- **\*\*Light\*\***: Use a desk lamp for adequate lighting.
- **\*\*Paper and Pen\*\***: Use smooth, lined paper and a pen grip that feels comfortable, such as a Pilot G2 or a Pentel EnerGel.

### **\*\*Step 2: Warm-up and Mindfulness\*\***

- **\*\*Warm-up\*\***: Write the alphabet three times to warm up muscles.
- **\*\*Breathing\*\***: Inhale for four counts, exhale for four counts; repeat until calm. This helps reduce writing tension.

### **\*\*Step 3: Basic Strokes\*\***

- Practice basic strokes: the upward stroke (v), downward stroke (n), and circular stroke (o).

### **\*\*Step 4: Letter Formation\*\***

- **\*\*Uppercase Letters\*\***: Form each letter slowly, tracing them with a stencil for accuracy.
- **\*\*Lowercase Letters\*\***: Practice looping 'l' and crossing 't'.

### **\*\*Step 5: Word Practice\*\***

- Write simple sentences like "The quick brown fox jumps over the lazy dog." Focus on each word.

### **\*\*Step 6: Cursive Writing\*\***

- Start with connected uppercase letters, then move to fluid cursive lowercase.
- Try online tutorials for modeled cursive formation.

### **\*\*Step 7: Pressure and Speed\*\***

- Practice on a tablet to see pressure applied; aim for consistent pressure.
- Slow down; writing fast increases errors.

### **\*\*Step 8: Style Inspirations\*\***

- Explore calligraphy styles online, imitating specific fonts like Script or Gothic.

### **\*\*Step 9: Regular Practice\*\***

- Write a journal entry daily, focusing on quality, not quantity.

### **\*\*Step 10: Challenges and Variety\*\***

- Weekly challenges like writing with your non-dominant hand or different tools.
- Experiment with bullet journaling for creativity.

### **\*\*Step 11: Track Progress\*\***

- Monthly before-and-after writing samples.
- Ask for feedback from a calligraphy expert or mentor.

### **\*\*Step 12: Enjoy the Process\*\***

- Listen to calming music or an audiobook to enjoy the time writing.

### **\*\*Additional Tips\*\***

1. **\*\*Muscle Training\*\***: Use hand exercise tools or write with silverware for grip strength.
2. **\*\*Mindfulness\*\***: Meditate before writing to enhance focus.
3. **\*\*Rewards\*\***: Treat yourself after achieving goals, like a new pen for milestones.

This guide not only sharpens your handwriting but also promotes mindfulness and creativity, offering both functional and mental benefits. Enjoy the journey to better handwriting!

## Response by the Reference Model (GPT-4-Turbo 2023-11-06)

Certainly! Improving your handwriting involves patience, practice, and attention to technique. Here are some tips to help you develop nicer handwriting:

1. **\*\*Understand Your Current Style\*\***: Before you can improve your handwriting, you need to understand where you're starting from. Write a paragraph and analyze it to see where you could make improvements. Look for consistent slant, letter size, spacing, and legibility.
2. **\*\*Choose the Right Tools\*\***: Use a pen or pencil that feels comfortable in your hand. The right tool can make a significant difference in your writing. Some people find that using a fountain pen improves their writing because it encourages a more relaxed grip and has a smoother ink flow.
3. **\*\*Practice Basic Strokes\*\***: Before focusing on letters, practice basic strokes. This includes straight lines, curves, and circles. These are the building blocks of letters, and getting comfortable with these will make your letters more consistent.
4. **\*\*Slow Down\*\***: Speed can lead to sloppy writing. By slowing down, you can focus on the shape of each letter and ensure that you're writing them correctly.
5. **\*\*Maintain Proper Posture and Grip\*\***: Sit up straight and hold your pen or pencil with a relaxed grip. Don't hold it too tightly, as this can make your handwriting appear more jagged and can also cause your hand to tire out more quickly.
6. **\*\*Form Letters Correctly\*\***: Start and end each letter in the right spot to ensure consistency and flow. Practice writing each letter of the alphabet, focusing on the ones that give you the most trouble.
7. **\*\*Practice Consistent Letter Size and Spacing\*\***: Consistency is key to neat handwriting. Work on keeping all of your letters the same size and your words evenly spaced. You could use lined paper as a guide.
8. **\*\*Use Handwriting Worksheets\*\***: You can find handwriting practice worksheets online for both print and cursive writing. These worksheets can guide you through the proper strokes and help you practice consistency.
9. **\*\*Copy Neat Writing\*\***: Find a sample of handwriting you admire and try to copy it. This can help you develop a sense of the style you're aiming for.
10. **\*\*Write Often\*\***: Like any skill, handwriting improves with regular practice. Find reasons to write by hand every day, whether it's journaling, writing letters, or making to-do lists.
11. **\*\*Analyze and Adjust\*\***: Regularly look back at your handwriting to see where you've improved and what still needs work. Make adjustments as needed.
12. **\*\*Be Patient and Persistent\*\***: Don't get discouraged if progress is slow. Handwriting improvement takes time and dedication. Remember, the goal of improving your handwriting is to make it more legible and aesthetically pleasing, but it should also be comfortable for you to write. There's no one "correct" style of handwriting, so focus on developing a style that works for you."

## Response by UltraLM 13B

Sure, I'd be happy to help! Here are some tips to improve your handwriting:

1. Practice regularly: The more you practice, the better your handwriting will become. Set aside some time every day to practice writing letters, words, and sentences.
2. Start with the basics: Begin by learning the basic strokes of each letter, such as circles, lines, and curves. Then practice writing each letter in isolation before moving on to words and sentences.
3. Use the right tools: Make sure you have the right supplies, such as a good-quality pen or pencil, lined paper, and an eraser. Using the right tools can make a big difference in the quality of your handwriting.
4. Watch your posture: Good posture is essential for good handwriting. Sit up straight with your feet flat on the floor and your arms and wrists relaxed.
5. Relax your grip: Hold your pen or pencil lightly and avoid gripping it too tightly. This will help you write more smoothly and legibly.
6. Take breaks: It's important to take breaks while practicing to avoid fatigue and strain. Take a break every 20-30 minutes and stretch your hands and wrists.

Remember, improving your handwriting takes time and practice, so be patient with yourself. Keep practicing and soon you'll have beautiful handwriting!

## Response by Mistral-7B-Base-v0.2 SFT on Alpaca-52K

Sure! Here are some tips that can help you improve your handwriting:

1. Practice regularly: The more you practice, the better your handwriting will become. Set aside some time each day to write out letters, words, and sentences to improve your handwriting.
2. Use the right tools: Make sure you have a good pen or pencil that flows smoothly and is comfortable to write with. A good quality pen or pencil can make a big difference in the quality of your writing.
3. Keep your posture in mind: Sit up straight and hold your pen or pencil correctly, with your arm and wrist relaxed. This will help you write more smoothly and avoid cramping.
4. Take your time: Don't rush your writing. Take your time to form each letter and make sure it's legible.
5. Use proper grip: Hold your pen or pencil correctly, with your fingers and thumb supporting it, and your index finger and middle finger guiding it.
6. Keep your letters consistent: Make sure your letters are the same size and have the same slant. This will make your writing look more uniform.
7. Practice proper letter formation: Make sure your letters are formed correctly, with the right strokes and curves. You can find examples of proper letter formation online or in a handwriting guidebook.
8. Keep your handwriting neat: Keep your writing neat and tidy, with even spacing between words and lines.
9. Practice cursive writing: Cursive writing can help improve your handwriting by making your letters more connected and flowing.
10. Experiment with different styles: Try different styles of handwriting to find one that suits you. There are many styles to choose from, such as print, cursive, and italic.
11. Use guides: You can use a ruler or a grid to help you keep your writing straight and even.
12. Get feedback: Ask a friend or teacher to give you feedback on your writing and see where you can improve.
13. Keep it simple: Avoid overly decorative or complicated writing, as it can be hard to read.
14. Keep it legible: Make sure your writing is easy to read, even if it's not perfect.
15. Be patient: Improving your handwriting takes time and practice, so be patient with yourself and keep practicing.

Remember, everyone's handwriting is unique, so don't be too hard on yourself if it doesn't look exactly like someone else's. Just keep practicing and you'll see improvement over time.



## Response by Mistral-7B-Base-v0.2 SFT on Alpaca-1K Longest

Sure! Here are some tips that may help you improve your handwriting:

1. Practice regularly: Like any skill, improving your handwriting takes practice. Set aside some time each day to write out words, sentences or even whole passages to develop fluency and consistency in your strokes.
  2. Use the correct tools: Make sure you have a pen or pencil that feels comfortable in your hand and that you sharpen it regularly to ensure a fine point. Also, use paper that is not too smooth or too rough, as this can affect the quality of your strokes.
  3. Posture: Sit up straight and keep your writing surface at a comfortable height to avoid straining your hands, wrists, or arms. Your writing instrument should be held in a relaxed grip, not too tightly.
  4. Start with proper formation: Make sure you form each letter correctly from the beginning strokes. This means starting with the upward and downward strokes of the letter at the right angle, and sizing them appropriately.
  5. Don't rush: Take your time and write slowly and deliberately, focusing on the movement of your hand and the pressure you apply to the paper.
  6. Keep it simple: Avoid overly ornate or fancy writing, as this can be hard to read and may slow you down. Stick to basic, uniform strokes and practice until they become second nature.
  7. Don't lift your pen: Lifting your pen off the paper to change direction can cause inconsistencies in your writing. Practice making smooth, fluid transitions between strokes without lifting your pen.
  8. Practice Cursive: Cursive writing can help improve the flow and connectedness of your letters, making your handwriting look neater. Practice connecting the letters in words and phrases.
  9. Pay attention to spacing: Make sure there is enough space between words and letters for legibility.
  10. Keep it clean: A neat, uncluttered workspace can help you focus on your writing. Keep your desk free of clutter and sharpen your pencils regularly to avoid smudges.
  11. Experiment: Try different styles and techniques until you find one that feels natural and looks good to you. Don't be afraid to mix and match until you find what works best for you.
  12. Don't give up: Improving your handwriting takes time and patience, so don't get discouraged if you don't see results immediately. Keep practicing and you will eventually see an improvement.
- Remember, handwriting is a personal expression, so find a balance between what looks nice and what comes naturally to you. With practice, you can develop a neat and legible handwriting that reflects your personality.