INSTRUCT-SKILLMIX: A Powerful Pipeline for LLM Instruction Tuning

Simran Kaur[1*], **Simon Park**[1*], **Anirudh Goyal**[2], **Sanjeev Arora**[1]
[1] Princeton Language and Intelligence (PLI), Princeton University
[2] 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, either from existing datasets [6], or by directly prompting the model; (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.

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.

1 Introduction

Instruction tuning is the first step in converting a base LLM (i.e., trained just 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 [34], nowadays, the SFT data is collected at high cost from a thriving ecosystem of 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 [26, 25] etc., which usually follow SFT in the pipeline to produce an Instruct model.

The cost and complexity of custom datasets has motivated creation of open-domain alternatives. ShareGPT [4] contains conversations collected from a model-hosting website, whereas OpenAssistant [15] and Dolly [5] contain crowdsourced human data. Another intriguing method, popularized by Self-Instruct [33] (and its variants, e.g., Alpaca [28]) leverages synthetic datasets. Here, a strong LLM is prompted with a small set of human-created examples to generate a large number of (query, answer) examples on a variety of topics.

Evaluations of instruction-following ability have also sprung up, with AlpacaEval 2.0 [8, 9] a popular one based upon curated queries from various sources. In these 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. Rankings on AlpacaEval and related benchmarks like WildBench broadly align with the subjective impression of a model's performance [9], [19].

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 suspected that the cause is the lack of diversity in the training data. But, efforts to produce more diverse synthetic data —e.g.,

NeurIPS 2024 Workshop on Compositional Learning: Perspectives, Methods, and Paths Forward.

UltraChat [7], a synthetic dataset of 1.5M multi-turn conversations created via meticulously tracking of lexical and topical diversity as well as coherence — do not significantly improve performance.

A popular hypothesis is that this has to do with uneven quality of open datasets —which are usually a hodge-podge of collected queries— whereas proprietary datasets are produced to careful specifications using good quality-control. One finding that supports this hypothesis is that doing SFT on the 1K Q&A pairs in Alpaca-52K with the longest responses, outperforms SFT on all 52k pairs [42, 39]. 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 to provide a surprisingly reasonable level of instruction-tuning and alignment [18] — but they did not significantly improve the performance either.

Some have cautioned against hopes for a miracle out of instruction tuning. Gudibande et al. [12] 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 [22] —which of course had already benefited from training on expensive human data. For further discussion regarding related work, please see Appendix A.

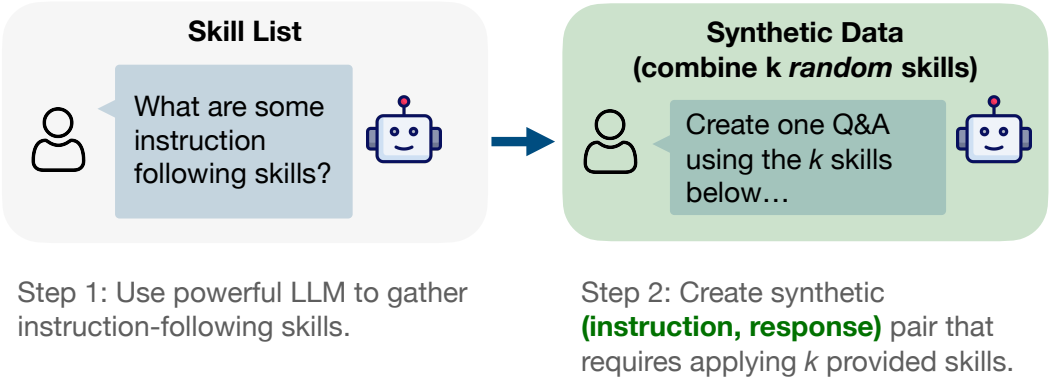


Figure 1: **Sketch of INSTRUCT-SKILLMIX pipeline.** See Figures 2a and 2b for more details on two different implementations of INSTRUCT-SKILLMIX.

This Paper: We describe a more efficient and effective approach for creating synthetic instruction tuning datasets. Past open efforts invested significant 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. Recognizing that pre-training is the dominant source of the LLM's "inner knowledge," we focus our instruction tuning on merely teaching the LLM skills for drawing upon that inner knowledge and presenting it nicely in during conversations. Our novel pipeline, INSTRUCT-SKILLMIX, adopts the approach in Didolkar et al. [6], and uses a frontier LLM to identify a list of "basic skills" needed for instruction-following. The second step of our pipeline, inspired by Yu et al. [38], uses the list of extracted skills to produce synthetic query-response examples by repeatedly drawing a random pair of skills from the list and asking the powerful LLM to produce a suitable query that tests those two skills, and also a good response to the query. Using merely $2K$ to $4K$ such Q&A examples, vanilla SFT boosts the performance of leading small base models (Mistral-7B-Base-v0.2, LLaMA-3-8B-Base, and Gemma-2-9B-Base) on AlpacaEval 2.0 to match or surpass some apex models, including the original GPT-4, LLaMA-3.1-405B-Instruct and Claude 3 Opus. (Table 1)

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.

Model	**# Data**	AlpacaEval2.0 **LC WR(%)**	WildBench **WB-Reward**$^{\text{gpt4t}}_{\infty}$
LLaMA-3-8B			
Ours	4K	**42.76**	**-36.91**
*LLaMA-3-8B-Instruct	-	22.90	-46.30
Mistral-7B-v0.2			
Ours	4K	**36.70**	**-29.25**
SFT on Alpaca-52K	52K	8.64	-80.47
*Mistral-7B-Instruct-v0.2	-	17.10	-54.70
Gemma-2-9B			
Ours	2K	36.18	-37.83
Gemma-2-9B-Instruct	-	**37.21**	**-28.78**
***Other Proprietary Models**			
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)	-	**57.50**	**+1.70**
GPT-4 (2023-03-14)	-	35.30	-

2 INSTRUCT-SKILLMIX

This section describes our methodology for extracting skills from powerful LLMs[1] 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 evaluation results when finetuning on this dataset. For more details and discussion about the pipeline, see Appendix B.

2.1 Skill Extraction Procedure

We tried two variants (see Appendix B for more details), and discuss the second variant below. We first query GPT-4-Turbo to generate a list of topics that arise in instruction-following. For each topic returned by GPT-4-Turbo, we further prompt GPT-4-Turbo to generate a list of skills that are needed to answer typical queries on that topic. Additionally, we ask GPT-4-Turbo to create a list of query types (e.g., "Information Seeking") that might arise in that topic. See Appendix M.4 for details about the prompts used, and Appendix L.2 for the list of all extracted skills.

2.2 Data Generation

Inspired by the recent SKILLMIX evaluation [38], we generate instruction-following examples by prompting GPT-4-Turbo to combine random combinations of k skills (and a random choice of query type for the seed-dataset agnostic version). Since the number of k-tuples scales as N^k, where N is the total number of skills, there is a very large diversity of prompts, which translates into diversity of training examples. 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 M.3 and M.5 for the details about the prompts used for data generation.) 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.

[1]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.

3 Experiments

We report the main evaluation results when models are finetuned on INSTRUCT-SKILLMIX-SDA in Table 1, and summarize our findings below. For a more detailed version of Table 1, see Table 2. For additional ablations (including the role of mixing skills, ruling out confounding factors, and an investigation of the effect of low quality data), see Appendix G. For evaluations on other LLM benchmarks, see Table 3.

INSTRUCT-SKILLMIX-SDA achieves SOTA performance amongst SFT models. LLaMA-3-8B-Base finetuned on 4,000 examples from INSTRUCT-SKILLMIX-SDA(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 2,000 example from INSTRUCT-SKILLMIX-SDA(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).

Improved performance from SFT on INSTRUCT-SKILLMIX-SDA. Across all model types, dataset size, and the evaluation benchmark, we generally see an improvement when finetuning on INSTRUCT-SKILLMIX-SDA compared to INSTRUCT-SKILLMIX-SDD. See Table 2 for more details.

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.

4 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. This is shown via our INSTRUCT-SKILLMIX pipeline, which uses the recent discovery of LLM Metacognition [6] 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 efforts like Vicuna and Ultrachat that used orders of magnitude more datapoints. The performance also approaches those of frontier models trained on expensive human data as well as RL techniques. Unfortunately our method saturates at 4K examples as explained earlier.

Ablation studies in Appendix G.2 ruled out potential confounding factors, such as the use of a strong data generation model or bias that may occur when the data generation model and grader model are from 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 elicits rich responses from powerful LLMs. Appendix G.3 offers a preliminary exploration of pitfalls of naive collection of instruction-tuning data. Specifically it identifies the *Negative Pareto Principle* of instruction-tuning: presence of some lower-quality data causes super-proportionate harm to the final result. This insight should be more rigorously investigated, including via new theory.

We should note that our results look stronger on paper than they actually are. This has to do with limitations of AlpacaEval 2.0 – specifically, the focus on basic instruction-following capability, and use of win-rates as a metric. A win rate of even 50% against a frontier model still allows unacceptably high frequency of unsuitable responses in a deployment setting. For example, our model struggles on queries that use vague or confusing phrasing. (The new WildBench evaluation does test for some such issues.) We hope to use such observations to improve INSTRUCT-SKILLMIX ideas as well as to propose a better evaluation.

References

[1] 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.

[2] Sanjeev Arora and Anirudh Goyal. A theory for emergence of complex skills in language models, 2023. URL https://arxiv.org/abs/2307.15936.

[3] Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. Alpagasus: Training a better alpaca with fewer data. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=FdVXgSJhvz.

[4] 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/.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english?, 2023. URL https://arxiv.org/abs/2305.07759.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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, 2023.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward, 2024. URL https://arxiv.org/abs/2405.14734.

[23] 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/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

[24] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4, 2023. URL https://arxiv.org/abs/2304.03277.

[25] 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.

[26] 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.

[27] 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.

[28] 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, 2023.

[29] 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.

[30] Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, 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.

[31] 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.

[32] 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*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/ec6413875e4ab08d7bc4d8e225263398-Abstract-Datasets_and_Benchmarks.html.

[33] 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)*, 2023.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

List of Appendices

A Related Work . . . 10
B Further details about INSTRUCT-SKILLMIX Pipeline . . . 11
C Further details about Experimental Setup . . . 13
D Full Evaluation Results (More Detailed) . . . 14
E Evaluation Details . . . 16
F Training Details . . . 17
G Ablations . . . 19
H INSTRUCT-SKILLMIX is Competitive with RL-Inspired Methods. . . . 24
I Robustness of INSTRUCT-SKILLMIX Across Random Skill Combinations for SFT . . . 25
J Examples of BREV-INSTRUCT-SKILLMIX and JUNK-INSTRUCT-SKILLMIX . . . 26
K Stats on Different Datasets . . . 27
L List of Skills . . . 28
M Skill Extraction Prompts . . . 55
N Comparison of Responses . . . 64

A Related Work

Prior works observe improvements from instruction finetuning on *fewer*, but *higher quality* data generated by humans [42, 29]. 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 [33, 8, 37, 13]. 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 [33]. Recent efforts have also focused on data selection strategies for high quality subsets of the original dataset, which lead to performance gains [30, 3, 20, 39]. Notably, Zhao et al. [39] 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. [36] 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 [31, 21, 32], as well as rewriting the data in multiple ways and changing formatting [1, 14]. The Self-Instruct framework [33] and variants such as Alpaca-52K [8] popularized the following approach to encourage diversity: filter out any synthetic data whose ROUGE-L similarity with any existing instruction is below some threshold. Other approaches to create diverse, synthetic data include imposing constraints on the topic in order to enhance wide coverage [7, 37], or requiring the synthetic data to use a random subset of words or concepts chosen from some vocabulary [10, 13, 16]. The latter approach is also suggested by recent work that provides a mathematical model for emergence via LLM scaling [2] and used in the evaluation setting in Yu et al. [38].

AlpacaEval and WildBench. AlpacaEval [17, 9] and WildBench [19] are widely adopted evaluations for assessing the instruction-following capabilities of LLMs. A model has to provide answers to a set of instructions, and their answers are compared against the reference outputs from a designated baseline model. Another evaluator LLM outputs a preference between the model outputs and the reference outputs. The primary evaluation metric is the *win rate*, the proportion of model outputs that were favored over the reference outputs. Given that a raw win rate shows bias towards longer responses, AlpacaEval 2.0 [9] introduces *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.

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 values) via RL methods. Standard approaches for alignment, such as RLHF [23], 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) [26]. 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) [25], which implicitly optimizes the same objective as RLHF, and SimPO [22], 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 [27, 35]. For example, SPPO [35] 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). In Appendix H, we compare the performance of Mistral-7B-*Base*-v0.2 SFT with our 4K dataset against Mistral-7B-*Instruct*-v0.2 finetuned using SPPO and SimPO.

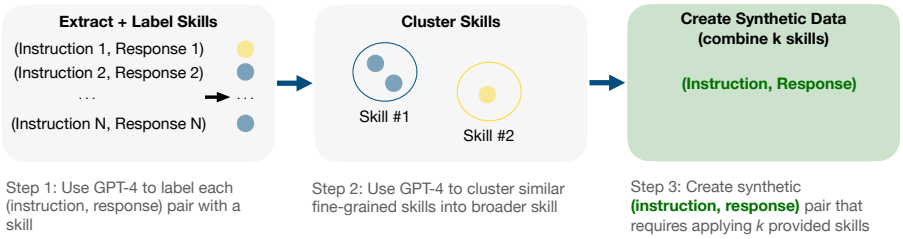
B Further details about INSTRUCT-SKILLMIX Pipeline

B.1 Skill Extraction Procedure (More detailed)

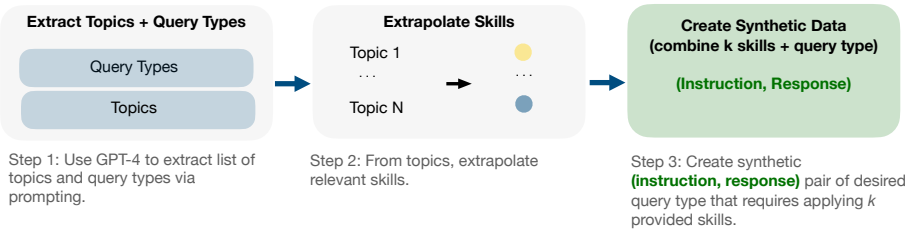
Method 1: Leveraging existing instruction dataset. Even though existing instruction-following datasets may not induce good chat capability via vanilla SFT, these datasets still exhibit (possibly in an uneven fashion) some "skills" needed by the model. Thus, we adapt the methodology presented in [6] and 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 (e.g., Critical Thinking and Analytical Skills). See Appendix L.1 for the list of all extracted skills and Appendix M.1 and M.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 the quality of seed datasets, as well as about inheriting their limitations and biases. Therefore we also tried an alternative pipeline that solely relies on the powerful LLM's ideas about what skills it leverages for instruction-tuning. We first query GPT-4-Turbo to generate a list of topics that arise in instruction-following. For each topic returned by GPT-4-Turbo, we further prompt GPT-4-Turbo to generate a list of skills that are needed to answer typical queries on that topic. Additionally, we ask GPT-4-Turbo to create a list of query types (e.g., "Information Seeking") that might arise in that topic. See Appendix M.4 for details about the prompts used, and Appendix L.2 for the list of all extracted skills.

It should be clear that Method 2, by relying solely upon the LLM's inner knowledge, should extend easily to other types of instruction-following, including settings such as safety and alignment. This is left for future work.



(a) Top row: INSTRUCT-SKILLMIX-SDD pipeline (short for INSTRUCT-SKILLMIX-SEED-DATASET-DEPENDENT).



(b) Bottom row: INSTRUCT-SKILLMIX-SDA pipeline (short for INSTRUCT-SKILLMIX-SEED-DATASET-AGNOSTIC).

Figure 2: **Two variants of the INSTRUCT-SKILLMIX pipeline.** INSTRUCT-SKILLMIX(k) involves two steps: (1) skill extraction using similar ideas as [6]; (2) data generation from random k-tuples of skills.

Where does diversity come from? The first source of diversity is the skill labels. A skill-label represents some part of GPT-4's meta-knowledge of human behavior and needs as observed during instruction-tuning, and which it has derived from its vast training set. Replacing a concrete Q&A example with a skill label converts it into a pointer to a region in GPT-4's meta-knowledge, which GPT-4 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.

Even though two skill pairs may 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 be a rich source for learning how to follow instructions.

C Further details about 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, LLaMA-2-13B-Base, on a varying number of examples from INSTRUCT-SKILLMIX-SDD(k) and INSTRUCT-SKILLMIX-SDA(k). We train for multiple epochs and select the best checkpoint by performance on a held-out data. Similar to [42], we observe that using cross entropy-loss on a validation set does not lead to the best checkpoint. See Appendix F.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 F.1.

Evaluation. We evaluate our models on popular instruction following benchmarks: AlpacaEval 2.0 [9], MT-Bench [41], and WildBench [19]. 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 E. See Table 3 in Appendix D for evaluations on additional benchmarks.

D Full Evaluation Results (More Detailed)

Table 2 contains the full evaluation results on instruction following benchmarks, including the ones in Table 1. Table 3 contains the full evaluation results on other popular LLM benchmarks.

Table 2: **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.

		AlpacaEval 2.0			WildBench
Model	**# Data**	**WR(%)**	**LC WR(%)**	**MT-Bench**	**WB-Reward$^{\text{gpt4t}}_{\infty}$**
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 3: **Evaluation results on MMLU, TruthfulQA, GSM8K, ARC Challenge, Winogrande, PIQA.**

Model	MMLU	TrQA	GSM	ARC-C	Winogrande	PIQA
LLaMA-3-8B Models						
INSTRUCT-SKILLMIX-SDD-1K	62.09	34.88	52.54	53.92	74.51	79.76
INSTRUCT-SKILLMIX-SDD-2K	62.09	37.33	52.77	53.75	75.06	79.54
INSTRUCT-SKILLMIX-SDD-4K	62.28	32.19	50.42	52.73	73.09	79.22
INSTRUCT-SKILLMIX-SDA-1K	62.33	37.09	51.25	52.39	74.19	79.92
INSTRUCT-SKILLMIX-SDA-2K	62.18	35.25	52.39	52.39	74.66	79.05
INSTRUCT-SKILLMIX-SDA-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
Mistral 7B v0.2 Models						
INSTRUCT-SKILLMIX-SDD-1K	58.97	26.19	36.01	51.02	73.64	81.18
INSTRUCT-SKILLMIX-SDD-2K	58.67	25.95	36.32	50.60	73.56	81.01
INSTRUCT-SKILLMIX-SDD-4K	58.38	26.68	36.54	50.00	73.56	81.45
INSTRUCT-SKILLMIX-SDA-1K	59.24	27.05	35.10	52.47	73.48	81.23
INSTRUCT-SKILLMIX-SDA-2K	58.90	25.83	33.66	52.99	73.88	81.66
INSTRUCT-SKILLMIX-SDA-4K	58.49	26.68	31.77	52.13	73.72	81.12
INSTRUCT-SKILLMIX-SDD(k=1)-1K	59.02	26.56	34.27	50.34	72.77	81.07
INSTRUCT-SKILLMIX-SDD(k=1)-2K	58.90	25.83	33.66	52.99	73.88	81.66
INSTRUCT-SKILLMIX-SDD(k=1)-4K	58.94	26.56	33.97	51.11	73.56	81.45
INSTRUCT-SKILLMIX-SDA(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
Gemma-2-9B Models						
INSTRUCT-SKILLMIX-SDD-1K	69.16	30.60	70.96	62.54	74.74	81.23
INSTRUCT-SKILLMIX-SDD-2K	69.26	30.72	70.81	63.23	74.59	81.28
INSTRUCT-SKILLMIX-SDD-4K	69.39	30.11	71.72	63.14	74.66	81.66
INSTRUCT-SKILLMIX-SDA-1K	69.49	31.21	70.74	62.80	73.95	81.83
INSTRUCT-SKILLMIX-SDA-2K	69.64	32.56	71.04	63.82	74.59	81.66
INSTRUCT-SKILLMIX-SDA-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
LLaMA-2-7B Models						
INSTRUCT-SKILLMIX-SDD-1K	41.04	34.39	11.83	46.93	70.01	78.07
INSTRUCT-SKILLMIX-SDD-2K	41.84	31.21	17.51	47.10	69.53	78.45
INSTRUCT-SKILLMIX-SDD-4K	43.00	30.84	15.24	47.01	69.38	78.24
INSTRUCT-SKILLMIX-SDA-1K	41.45	34.39	14.78	48.38	69.61	78.35
INSTRUCT-SKILLMIX-SDA-2K	43.17	33.41	15.92	47.78	70.01	78.51
INSTRUCT-SKILLMIX-SDA-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
LLaMA-2-13B Models						
INSTRUCT-SKILLMIX-SDD-1K	51.25	30.72	28.51	51.02	72.38	79.16
INSTRUCT-SKILLMIX-SDD-2K	51.03	30.84	28.73	50.85	72.30	79.43
INSTRUCT-SKILLMIX-SDD-4K	51.05	29.50	28.58	51.19	71.82	80.03
INSTRUCT-SKILLMIX-SDA-1K	50.68	30.11	27.45	50.60	72.61	79.92
INSTRUCT-SKILLMIX-SDA-2K	51.67	30.35	29.19	50.17	72.06	79.98
INSTRUCT-SKILLMIX-SDA-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

E Evaluation Details

To evaluate our models on the AlpacaEval 2.0, we followed the instructions in `https://github.com/tatsu-lab/alpaca_eval` [9]. 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` [41]. 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` [19]. 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 WB-Reward$^{\text{gpt4t}}_{\infty}$ in their notation.

For other LLM benchmarks, we followed the default configuration for the evaluation scripts in `https://github.com/EleutherAI/lm-evaluation-harness` [11]. We report the exact-match accuracy for GSM8K and the MC1 score for TruthfulQA.

F Training Details

F.1 Hyperparameters

In Table 4, 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 [40].

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 4: **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

F.2 Checkpoint Selection

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.

For example, in Table 5, 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-SDD-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 [39], 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 5: **Checkpoint selection.** We SFT Mistral-7B-Base-v0.2 on INSTRUCT-SKILLMIX-SDD-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-SDD Data															
LC WR(%)	20.7	20.4	27.8	28.2	37.0	35.2	45.5	44.1	45.6	39.5	**52.8**	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	**82.3**	74.4	73.1	70.6	74.0
CE Loss	1.21	**1.18**	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	**30.8**
WR(%)	17.3	19.2	27.1	30.9	33.2	32.4	34.4	35.6	34.6	33.7	**35.1**	32.5	34.0	34.6	**35.1**

G Ablations

G.1 The Effectiveness of (Mixing) Skills

INSTRUCT-SKILLMIX-SDD significantly outperforms SFT on Alpaca dataset. To illustrate the performance of the INSTRUCT-SKILLMIX pipeline, we present the evaluation results when we finetune Mistral-7B-Base-v0.2 on different subsets of the Alpaca-52K dataset (responses written by GPT-4 2023-03-14 [24]): (1) *Alpaca-1K Longest*: 1,000 examples with the longest responses [39]; (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 6: **Evaluation results of Mistral-7B-Base-v0.2 finetuned on INSTRUCT-SKILLMIX vs. on Alpaca-52K.** Note that skills extracted from Alpaca-5K Random were used to create the INSTRUCT-SKILLMIX-SDD datasets.

		AlpacaEval 2.0		WildBench
SFT Dataset	**# Data**	**LC WR(%)**	**MT-Bench**	**WB-Reward$^{gpt4t}_{\infty}$**
INSTRUCT-SKILLMIX-SDD(k=2)	4K	**29.77**	7.17	**-39.06**
INSTRUCT-SKILLMIX-SDD(k=1)	1K	27.04	**7.22**	-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 6, 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-SDD(k=1) yields 27.04% LC win rate. Furthermore, SFT on INSTRUCT-SKILLMIX-SDD(k=2)-4K yields 29.77% LC win rate. We highlight that since INSTRUCT-SKILLMIX-SDD is partially derived from Alpaca-52K, the observed improvements in the win rate are indicative of the improved quality of INSTRUCT-SKILLMIX over selecting the longest completions.

Table 14 shows that INSTRUCT-SKILLMIX has the same number of tokens per (instruction, response) pair as Alpaca-1K Longest, but has longer and more descriptive instruction, whereas the reference model output is shorter and more concise. This suggests that the quality of the instruction is important for instruction tuning. [2]

Increasing *k* boosts performance. In Table 7, models finetuned in INSTRUCT-SKILLMIX(k=2) data outperform models SFT on INSTRUCT-SKILLMIX(k=1) on AlpacaEval and WildBench, whereas performance on MT-bench is about the same.

Table 7: **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-SDD/INSTRUCT-SKILLMIX-SDA

		AlpacaEval 2.0			WildBench
Model	**# Data**	**WR(%)**	**LC WR(%)**	**MT-Bench**	**WB-Reward$^{gpt4t}_{\infty}$**
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/-

[2] The same result holds when finetuning on 1,000 examples of INSTRUCT-SKILLMIX-SDD(k=1), which only includes 1 skill per datapoint. This strongly suggests that INSTRUCT-SKILLMIX is able to generate data that better illustrates conversational skills present in the Alpaca dataset, even without having to combine skills. See Table 2 for more details.

G.2 Ablations to Rule Out Confounding Factors

We report ablation studies that help get to the source of our method's effectiveness.

Is it entirely due to a powerful text generator? We used GPT-4-Turbo, whereas many SFT efforts in 2023 used an earlier version of GPT-4 or GPT-3.5. To pinpoint the effect of this change, we match the version of the GPT-4 by regenerating answers to the top 1K queries of Alpaca (the ones identified in [42]) with GPT-4-Turbo, and regenerating INSTRUCT-SKILLMIX with the same GPT-4 version as in [42].

Table 8: **Evaluation results of Mistral-7B-Base-v0.2 finetuned on INSTRUCT-SKILLMIX-SDD 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-SDD outperforms SFT on Alpaca-1K Longest.

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

Table 8,compares the performance of Mistral-7B-Base-v0.2 when finetuned on the two datasets using the two versions of GPT-4. Using a better generator helps improve peroformance on Alpaca queries, but for each fixed generator model, the INSTRUCT-SKILLMIX dataset leads to better performance. Furthermore, replacing GPT-4 with the stronger GPT-4-Turbo in generation makes INSTRUCT-SKILLMIX pull even further ahead of Alpaca-1K Longest, which highlights that our pipeline is better positioned than Alpaca to elicit better supervision from a more powerful LLM teacher.

Table 9: **Evaluation results when using two different graders for AlpacaEval 2.0.** Relative ranking of evaluated models are generally preserved when using different graders.

	Grader: GPT-4 (2023-11-06)		Grader: Claude 3 Opus	
Model	**WR(%)**	**LC WR(%)**	**WR(%)**	**LC WR(%)**
Mistral-7B-Base-v0.2 SFT on ISM-SDD-1K	33.87	27.48	50.56	38.50
Mistral-7B-Base-v0.2 SFT on ISM-SDD-2K	37.05	31.57	48.94	38.29
Mistral-7B-Base-v0.2 SFT on ISM-SDD-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

Effect of choice of graders. To check for potential bias that may arise when the model used for data generation and grading are from the same model family, we compare the effect of using Claude 3 Opus as the grader for AlpacaEval 2.0. Table 9 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 models' generations than does GPT-4.

G.3 Effect of Low Quality Data: The Negative Pareto Principle

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 (although maybe would be filtered out by quality control in corporate settings). We can think of this as "data from shirkers." See Appendix J 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 such data 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 this dataset yields less than 1% win rate on AlpacaEval and 5.01 on MT-Bench.

Thus lower-quality data has a super-proportionate harm on model responses. We propose to call this effect the *negative pareto principle* [3]. It may help explain why creating open-domain instruction-tuning data has proved so difficult via naive crowd-sourcing.

Table 10: **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.

		AlpacaEval 2.0			WildBench
Model	**# Data**	**LC WR(%)**	**Avg Len**	**MT-Bench**	**WB-Reward**$^{\text{gpt4t}}_{\infty}$
SFT on INSTRUCT-SKILLMIX-SDD(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-SDA(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

G.4 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.**

		AlpacaEval 2.0	
Model	**# Data**	**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

[3] The *Pareto Principle*, also known as the 80/20 rule, states that roughly 80% of the consequences come from 20% of the causes.[4]

G.5 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-SDD and INSTRUCT-SKILLMIX-SDA, respectively. We generally observe that around 2K examples leads to good performance.

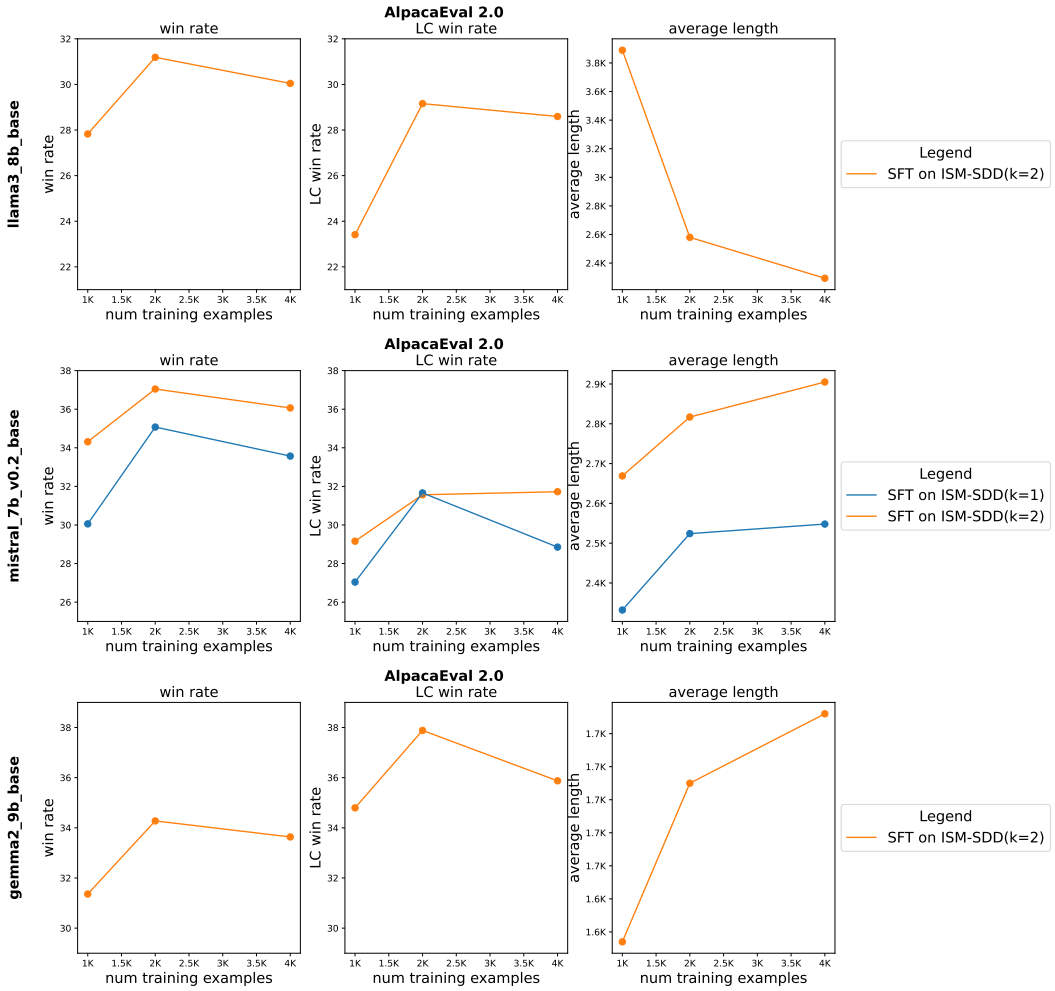


Figure 3: **Win rates and average output length on varying amounts of INSTRUCT-SKILLMIX-SDD data.**

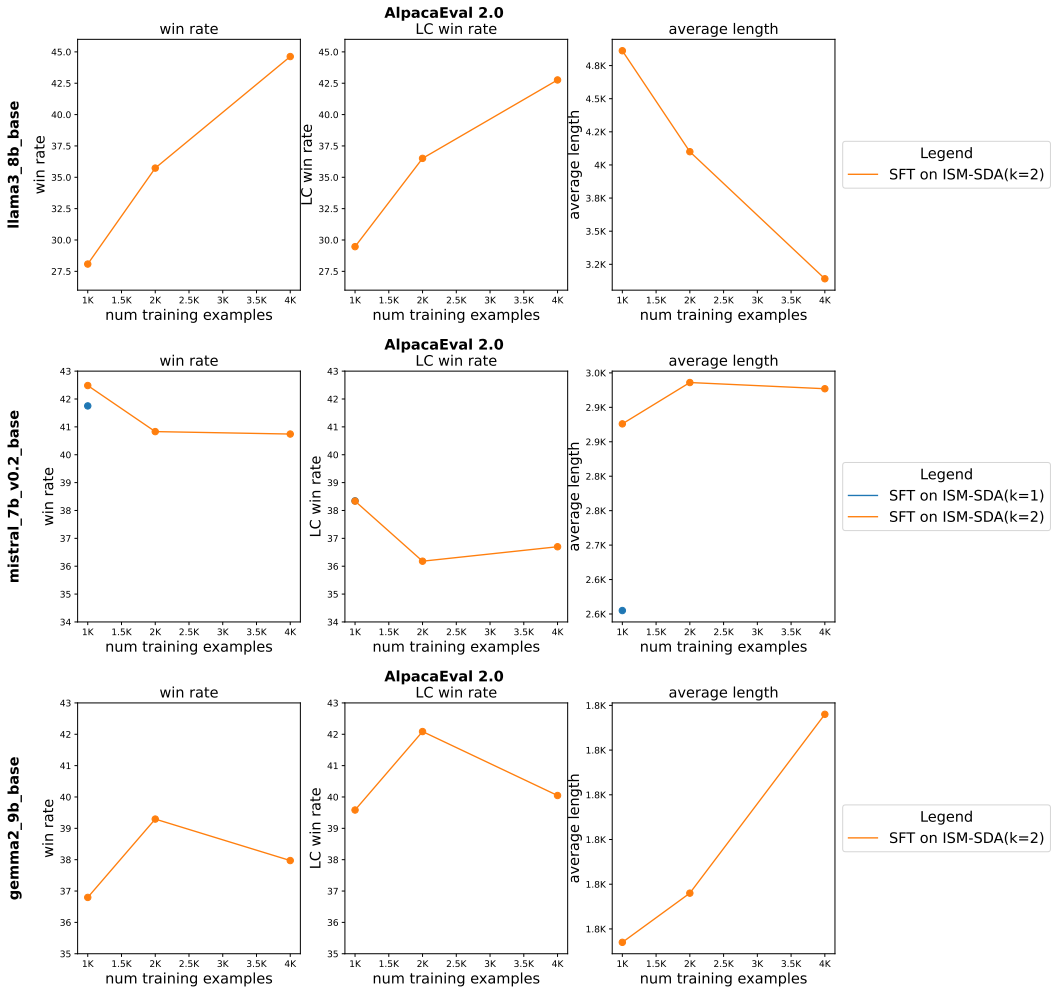


Figure 4: **Win rates and average output length on varying amounts of INSTRUCT-SKILLMIX-SDA data.**

H INSTRUCT-SKILLMIX is Competitive With RL-Inspired Methods.

Self-Play Preference Optimization (SPPO) [35] and SimPO [22] 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-SDA 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**
SFT on INSTRUCT-SKILLMIX-SDA(k=2)			
LLaMA-3-8B-Base	SFT on INSTRUCT-SKILLMIX-SDA(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-SDA(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

I 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.

	AlpacaEval 2.0				WildBench
SFT Dataset	**WR(%)**	**LC WR(%)**	**Avg. Len.**	**MT-Bench**	**WB-Reward**$^{\text{gpt4t}}_{\infty}$
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

J 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.
```

K 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-SDD	4000	511.58	2199.01	394.15	1644.88
INSTRUCT-SKILLMIX-SDA	4000	510.63	2152.77	392.32	1606.33

L List of Skills

L.1 INSTRUCT-SKILLMIX-SDD 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 4) 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	knowledgebased_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
creative_art_and_design	business_strategy_and_collaboration
language_comprehension_and_expression	artificial_intelligence_machine_learning
system_and_framework_analysis	natural_and_social_science
analytical_skills_in_humanities_and_social_sciences	analytical_thinking
knowledge_and_understanding_in_technology	natural_and_social_sciences_knowledge
content_analysis_and_summarization	literacy_and_writing_skills
science_and_analysis	specific_knowledge_research
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

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

Skill Cluster Name	
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
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
algorithms_and_data_manipulation	digital_and_online_knowledge
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

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

Skill Cluster Name	
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_critial_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
data_management_and_analysis	real_time_data_handling
international_relations_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
technical_knowledge_and_application	machine_learning_and_deep_learning
business_and_economic_analysis	machine_learning_applications
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

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

Skill Cluster Name	
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
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
medical_and_healthcare_knowledge	

Table 23: (Part 1 of 2) 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
programing_and_coding_standards	behavioral_and_social_psychology
understanding_and_empathy	environmental_science_and_sustainability
environment_conservation_strategies	behavioral_and_mental_health
comprehensive_understanding_and_interpretation	programming_and_systems_development
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_aptitude	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_relations_skills	machine_learning_and_modeling
cultural_historical_knowledge	communication_and_literacy
advanced_writing_and_comprehension_skills	policy_analysis_and_evaluation
cultural_and_social_analysis	technology_development_and_security
research_and_data_handling	communication_and_marketing_strategy
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	programming_and_computing_skills
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

Table 25: (Part 2 of 2) 128 Train Skills extracted from random sample of 1000 instruction-response pairs from **UltraChat**

Skill Cluster Name	
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	

Table 27: (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	

L.2 INSTRUCT-SKILLMIX-SDA 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 29: (Part 1 of 2) 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
linguistics	mechanical
electrical	civil_engineering
media_analysis	news_reporting
digital_media_trends	disease_prevention
public_health_initiatives	epidemiological_studies
seo	social_media_strategies
content_marketing	cultural_dynamics
anthropology	social_customs
planetary_science	space_missions
astronomy	theatre
dance	performance_techniques
geology	meteorology

Table 31: (Part 2 of 2) 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
energy_and_resources	fashion_and_textiles
hospitality_and_tourism	

Table 33: (Part 1 of 13) 1143 skills extracted from topics in Tables 29 and 31.

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
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 35: (Part 2 of 13) 1143 skills extracted from topics in Tables 29 and 31.

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_conservation_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
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

Table 37: (Part 3 of 13) 1143 skills extracted from topics in Tables 29 and 31.

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
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
bronze_casting	focus_enhancement
brand_storytelling	ui_ux_design
grammar_rules_teaching	outcome_evaluation

Table 39: (Part 4 of 13) 1143 skills extracted from topics in Tables 29 and 31.

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
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

Table 41: (Part 5 of 13) 1143 skills extracted from topics in Tables 29 and 31.

Skills	
donor_relationship_management	character_development_insight
digital_literacy	artistic_expression
operational_efficiency	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
ethical_reporting	change_management
surgical_techniques	software_debugging
system_administration	astronomy_and_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

Table 43: (Part 6 of 13) 1143 skills extracted from topics in Tables 29 and 31.

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
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

Table 45: (Part 7 of 13) 1143 skills extracted from topics in Tables 29 and 31.

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	directorial_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
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

Table 47: (Part 8 of 13) 1143 skills extracted from topics in Tables 29 and 31.

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
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
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

Table 49: (Part 9 of 13) 1143 skills extracted from topics in Tables 29 and 31.

Skills	
literature_review_and_meta_analysis	networking_skills
conversation_management	resource_utilization
nutritional_info	digital_fundraising_techniques
communicate_symptoms_to_healthcare_providers	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
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

Table 51: (Part 10 of 13) 1143 skills extracted from topics in Tables 29 and 31.

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
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

Table 53: (Part 11 of 13) 1143 skills extracted from topics in Tables 29 and 31.

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
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
reflective_judgment	blockchain_development
symbolic_interpretation	protocol_management
clinical_diagnosis	brush_technique

Table 55: (Part 12 of 13) 1143 skills extracted from topics in Tables 29 and 31.

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
qualitative_data_collection	media_studies_and_journalism
technology_adaptation	disaster_recovery_planning
data_collection	packing_efficiency
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

Table 57: (Part 13 of 13) 1143 skills extracted from topics in Tables 29 and 31.

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
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

Table 59: (Part 1 of 3) 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.
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.
Comparative	These queries involve comparing different items, ideas, or scenarios. Users often seek assistance in making decisions or understanding differences.
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.

Table 61: (Part 2 of 3) 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.
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.

Table 63: (Part 3 of 3) 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.

M Skill Extraction Prompts

M.1 Prompt for Skill Extraction (INSTRUCT-SKILLMIX-SDD)

```
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>
```

M.2 Prompt for Semantic Clustering (INSTRUCT-SKILLMIX-SDD)

```
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.
```

M.3 Prompt for Creating Synthetic Data (INSTRUCT-SKILLMIX-SDD)

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, Ï 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, Ï 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 Ï can see why you feel that wayör Ïthat must be challenging for you.Ïn a recent meeting,
saying Ï understand your frustration with the project timelinehelped 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, Ïso, 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, Ïhat was the primary reason for
canceling your subscription?änd Ïhat 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 C̈ustomers who bought this item also bought...ör R̈ecommended based on your recent searches.T̈his 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 Ceauşescu.

```
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!
```

M.4 Prompt for Extrapolating Skills (INSTRUCT-SKILLMIX-SDA)

```
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.
```

M.5 Prompt for Creating Synthetic Data (INSTRUCT-SKILLMIX-SDA)

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: 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, Ï 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, Ï 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 Ï can see why you feel that wayör Ïthat must be challenging for you.Ïn a recent meeting, saying Ï understand your frustration with the project timelinehelped 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, Ös, 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 Ïron Curtain,ďescribed 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!
```

N 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-SDA(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-SDA(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 (∩), 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.
```