Learning from "Silly" Questions Improves Large Language Models, But Only Slightly

Authors

Abstract

Constructing high-quality Supervised Fine-Tuning (SFT) datasets is critical for the training of large language models (LLMs). Recent studies have shown that using data from a specific source, Ruozhiba[1], a Chinese website where users ask "silly" questions to better understand certain topics, can lead to better fine-tuning performance. This paper aims to explore some hidden factors: the potential interpretations of its success and a large-scale evaluation of the performance. First, we leverage GPT-4 to analyze the successful cases of Ruozhiba questions from the perspective of education, psychology, and cognitive science, deriving a set of explanatory rules. Then, we construct fine-tuning datasets by applying these rules to the MMLU training set. Surprisingly, our results indicate that rules can significantly improve model performance in *certain* tasks, while potentially diminishing performance on others. For example, SFT data generated following the "Counterintuitive Thinking" rule can achieve approximately a 5% improvement on the "Global Facts" task, whereas the "Blurring the Conceptual Boundaries" rule leads to a performance drop of 6.14% on the "Econometrics" task. In addition, for specific tasks, different rules tend to have a consistent impact on model performance. This suggests that the differences between the extracted rules are not as significant, and the effectiveness of the rules is relatively consistent across tasks. Our research highlights the importance of considering task diversity and rule applicability when constructing SFT datasets to achieve more comprehensive performance improvements.

1 Introduction

Large language models (LLMs), pre-trained on vast amounts of data, have garnered significant attention for their ability to address a wide range of tasks (Achiam, Adler, and Agarwal 2023; Dubey, Jauhri, and Pandey 2024; Hui, Yang, and Cui 2024; Jiang, Sablayrolles, and Roux 2024; Reid, Savinov, and Teplyashin 2024; Yang, Zhang, and Hui 2024). The pretrain-finetune paradigm has emerged as the cornerstone of LLMs' remarkable success. Through pre-training, LLMs acquire extensive knowledge about the world, while fine-tuning aligns them with specific human instructions, enabling them to generate high-quality responses and excel on domain-specific datasets. Compared to the resource-intensive process of pre-training, supervised fine-tuning is a cost-effective approach that focuses on fine-tuning LLMs using small but high-quality datasets. Consequently, the quality of SFT datasets plays a pivotal role in determining the performance of fine-tuned models.

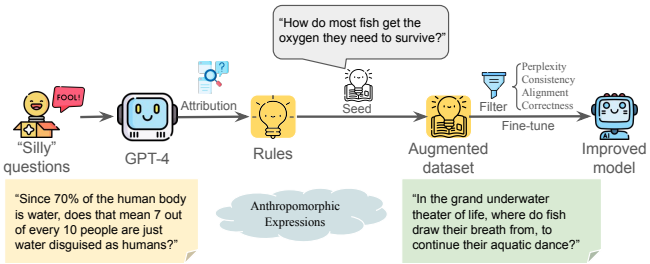


Figure 1: Overview of our augmentation pipeline.

Numerous studies (Cao et al. 2024; Chen et al. 2024a; Li et al. 2024; Liu et al. 2024; Lu et al. 2023; Mekala, Nguyen, and Shang 2024; Wei et al. 2023; Xia et al. 2024; Zhou, Liu, and et al. 2023) have emphasized the importance of the quality of SFT datasets. The scarcity of high-quality datasets necessitates careful data selection and processing when constructing SFT datasets. Recently, Bai, Du, and Liang (2024) demonstrated that SFT data collected from certain sources can significantly improve fine-tuning results. Notably, they highlighted the significant contribution of the data from "Ruozhiba", a Chinese online platform where people ask "silly" questions like "Since 70% of the human body is water, does that mean 7 out of every 10 people are just water disguised as humans?" to better understand certain knowledge. Ruozhiba contains implicit satire, critical content, and elements that are somewhat aggressive and offensive; such data are generally not considered suitable for SFT. However, Bai, Du, and Liang (2024) argue that an important reason for the good results with the Ruozhiba is that its dialogues often exhibit characteristics of humor, absurdity, linguistic traps, and abstraction that can enhance the model's reasoning capabilities.

This work attempts to present a holistic understanding of data augmetation in Ruozhiba style from two aspects:

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1]https://huggingface.co/datasets/m-a-p/COIG-CQIA/viewer/ruozhiba.

the potential interpretation of its effectiveness and a large-scale evaluation. Specifically, from the perspective of education, psychology and social learning, we used GPT-4 to extract features from the Ruozhiba dataset and defined these features as augmentation rules. We then employed GPT-4o to augment seed data sampled from MMLU, generating datasets that conform to different rules. Subsequently, we fine-tuned the LLM using both the seed datasets and the generated datasets and compared the performance changes to assess the impact of the augmentation.

Our experimental results show that fine-tuning LLMs with datasets generated using "silly" rules can achieve up to approximately a 0.54% overall performance improvement on the MMLU test set. However, compared to directly fine-tuning with the seed dataset, this approach does not yield any further overall performance gains. Our more fine-grained analysis reveals that datasets generated using different rules have *varying* impacts on the performance of the SFT model across different subjects and tasks. At the subject level, the extracted rules tend to degrade the performance of the SFT model on "STEM" subject, whereas some rules lead to slight improvements in "Humanities" subject. Furthermore, our task-level analysis shows that for specific tasks, datasets generated with different rules tend to have consistent impacts. Based on our analysis of a 13,000-sample dataset, 94.74% of tasks exhibited at least 50% consistency in the performance impact percentages of different rules, and 26.32% of tasks showed 100% consistency. Additionally, our subject-level analysis showed that all subjects demonstrated over 60% consistency in the performance impact percentages of different rules.

In summary, our contributions are as follows:

1. *Understanding:* We provide a holistic understanding of the Ruozhiba-style dataset from the perspective of education, psychology, and cognitive science. Our study generated a set of useful rules that can potentially impact the LLM fine-tuning performance.
2. *Augmentation:* We generated diverse datasets using the rules extracted from Ruozhiba to fine-tune LLMs. Our large-scale evaluation demonstrated varying improvements on different tasks.
3. *Selection:* We explore various data filtering and mixing strategies and find that SFT datasets enhanced with single-rule augmentation are more effective in improving LLM performance compared to mixed datasets.

2 Related work

The scarcity of high-quality SFT data has spurred research into synthesizing data using existing LLMs. For example, (Chiang, Li, and Lin 2023) collected user-shared ChatGPT conversations to build the SFT dataset Sharegpt, which, after fine-tuning, achieved 90% of the quality of closed-source models on the mt-bench (Zheng, Chiang, and Sheng 2023). Sharegpt can be viewed as a form of black-box distillation of closed-source models. More creative SFT data synthesis strategies leverage LLM capabilities to heuristically generate SFT data. The self-instruct method (Wang, Kordi, and Mishra 2023) used bootstrapping to generate new instructions and corresponding input-output samples from a limited set of manually written task seeds, achieving a 33% absolute improvement in GPT-3's performance on the SUPERNI benchmark (Wang, Mishra, and Alipoormolabashi 2022) compared to the original GPT-3 model.

Building on this idea, Bommasani et al. (2021) used the Self-Instruct framework to create a 52k-sample Alpaca SFT dataset, fine-tuning Llama to achieve performance comparable to text-davinci-003 on the self-instruct validation set. WizardLM (Xu, Sun, and Zheng 2024) employed GPT as an instruction evolver, creating SFT data with varying complexity through depth and breadth evolution from seed data. Furthermore, Zeng, Xu, and Zhao (2024) optimized the WizardLM by automating the analysis, summarization, and optimization of instruction dataset evolution strategies, thereby enhancing LLM performance across diverse tasks. These studies highlight the potential of innovative data synthesis strategies in advancing LLM capabilities. It is worth noting that our work is also related to supervised fine-tuning and data selection, which are discussed in more detail in Appendix A.

3 Methodology

In this section, we will articulate the approach to understanding the Ruozhiba questions, augmenting training data, and model fine-tuning. Figure 1 illustrates our pipeline.

3.1 Understanding "Silly" Questions

It requires significant effort to characterize the factors attributing to the success of the "silly questions" in Ruozhiba. Formally speaking, given an existing Ruozhiba dataset (e.g., data from COIG-CQIA (Bai, Du, and Liang 2024)) $D_{base} = \{\mathrm{ins}_i, \mathrm{rep}_i\}_{i=1}^{n}$, where ins_i and rep_i denote instruction and response, respectively, our objective is to learn a set of rules $R = \{r_j\}_{j=1}^{T}$ to interpret why D_{base} helps fine-tuning.

However, given the complicated nature of human language, especially the "nonsense" contained in the existing Ruozhiba dataset, it is extremely challenging to manually analyze and summarize the hidden factors in the dataset. In this paper, we take advantage of LLMs to facilitate our analysis. Specifically, we prompt GPT-4 from the perspective of psychology, education, sociology, and cognitive science to extract and summarize explanatory rule from D_{base}. These rules encapsulate the inherent logic and structure of tasks in the dataset, which not only aid in better understanding the dataset itself but also provide guidance and a foundation for constructing new high-quality datasets. Through this approach, we hope that the generated datasets will achieve diversity and applicability across various tasks, thereby enhancing the performance of LLMs in a wide range of tasks.

3.2 Data Augmentation

Given a seed dataset $D_{seed} = \{d_1, d_2, \ldots, d_N\}$, where d_i consists of a pair $\{\mathrm{ins}_i, \mathrm{rep}_i\}$, we utilize the GPT-4 as Instruction Rewriter to transform each ins_i by referencing a rule $r_t \in R$. This transformation results in a new instruction ins_{i,r_t} that adheres to the rule r_t, forming new SFT pairs

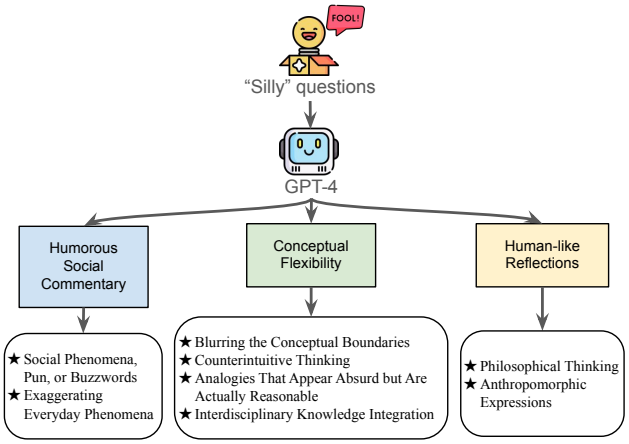


Figure 2: The rules extracted from the Ruozhiba dataset.

$d_{i,r_t} = \{\text{ins}_{i,r_t}, \text{rep}_i\}$. Consequently, we obtain a dataset D_{r_t} following rule r_t. Additionally, we include res_i in the prompt to ensure that the rewritten ins_{i,r_t} can still be accurately responded to with rep_i. The prompts are in Appendix E.1.

Recognizing that the Instruction Rewriter might not fully comprehend the rule r_t, we manually select several data points from D_{base} as in-context examples (see Appendix E for details). These examples serve to guide the Instruction Rewriter in better accomplishing the task.

To explore the varying effects of different rules across subjects or tasks, we designed experiments involving data filtering and mixing. For parallel data samples $d_{i,r_1}, d_{i,r_2}, \ldots, d_{i,r_t}$, we generated a mixed sample $d_{i,r_{\text{mix}}}$ using the following equation:

$$d_{i,r_{\text{mix}}} = \text{Filter}(d_{i,r_1}, d_{i,r_2}, \ldots, d_{i,r_t}),$$

where Filter denotes the filtering operation for data selection. Specifically, we performed the following strategies:

1. **Perplexity-Based Selection**. We proposed two strategies to select samples based on their perplexity (PPL), either the samples with the highest or the lowest perplexity:

$$d_{i,r_{\text{mix}}} = d_{i,r_k}, \quad \text{where } k = \arg\max_k \text{PPL}(d_{i,r_k}),$$

$$d_{i,r_{\text{mix}}} = d_{i,r_k}, \quad \text{where } k = \arg\min_k \text{PPL}(d_{i,r_k}).$$

   When implementing the perplexity-based selection, we also considered whether to include seed data, resulting in four configurations.

2. **Judge Model Scoring-Based Selection**. We utilized a Judge model to evaluate augmented samples across three dimensions: *Consistency*, *Correctness*, and *Alignment*. For each dimension, we selected the sample with the highest score:

$$d_{i,r_{\text{mix}}} = d_{i,r_k}, \quad \text{where } k = \arg\max_k \text{Judge}(d_{i,r_k}).$$

These strategies resulted in seven distinct mixing configurations. Due to API cost constraints, we applied all seven configurations to a subset containing 4K samples. For the larger 13K dataset, we only employed perplexity-based filtering strategies.

4 Experiments

4.1 Dataset

We utilize the Ruozhiba dataset as D_{base} and extract eight rules, as shown in Figure 2. For detailed information about the Ruozhiba dataset, please refer to Appendix C. To validate the effectiveness of our rules across a broad spectrum of domain knowledge, we select the MMLU train set to sample D_{seed}. MMLU is a large-scale, multi-task, multiple-choice question-answering dataset encompassing 57 tasks across natural sciences, humanities, and social sciences. Utilizing MMLU as the seed dataset enables us to assess the impact of extracted rules on diverse tasks, such as knowledge-based QA and logical reasoning, since the influence of rules is likely to vary across different task types. This dataset allows us to rigorously evaluate the adaptability of our rules across a wide array of knowledge domains.

The MMLU dataset provides a training set containing approximately 90K samples from datasets such as ARC (Clark et al. 2018), MC-TEST (Richardson, Burges, and Renshaw 2013), OBQA (Mihaylov et al. 2018), and RACE (Lai et al. 2017). We performed stratified sampling on the training set to ensure a balanced representation of tasks. Consequently, we sampled approximately 13K instances to form our seed dataset, D_{seed}, with the task distribution illustrated in Figure 5 of Appendix F.1.

Using the rules extracted in the last section: $R = \{r_1, r_2, \ldots, r_8\}$, we employed GPT-4 as the instruction rewriter to process D_{seed}, resulting in eight distinct datasets: $D_{r1}, D_{r2}, \ldots, D_{r8}$. These datasets, along with D_{seed}, constitute our experimental dataset. To construct these datasets, we made 13K × 8 = 104K API requests.

4.2 General Results

We fine-tune Meta-Llama-3-8B-Instruct using the LoRA setup. For detailed experimental configurations, please refer to Appendix F.1. We evaluate the performance on the original dataset D_{seed} as well as on datasets synthesized according to eight different rules. The results are illustrated in Figure 3. It is evident that the effectiveness of the rules varies across different subjects. Overall, datasets generated by these rules, after fine-tuning, underperformed in "STEM" subjects compared to directly fine-tuning with D_{seed}. Whereas in the "Humanities" and "Other" subjects, the rule-generated datasets showed superior performance compared to D_{seed}. On the entire MMLU test set (i.e., the Average columns), the dataset generated by the rule for "Social Phenomena, Pun, and Buzzwords" slightly outperformed D_{seed} (66.28 vs 66.27), whereas other rule-generated datasets did not achieve higher scores.

Given that MMLU's four subjects comprise 57 tasks, we conducted a more fine-grained task-level analysis, detailed in the Appendix G.1. We categorized the performance of different SFT datasets compared with D_{seed} into *improved*, *declined*, and *unchanged* categories, and calculated the consistency percentage of these performance impacts. Our analysis revealed that for the same task, datasets generated by different rules tend to produce consistent changes in performance. As illustrated in Figure 14, our subject-level anal-

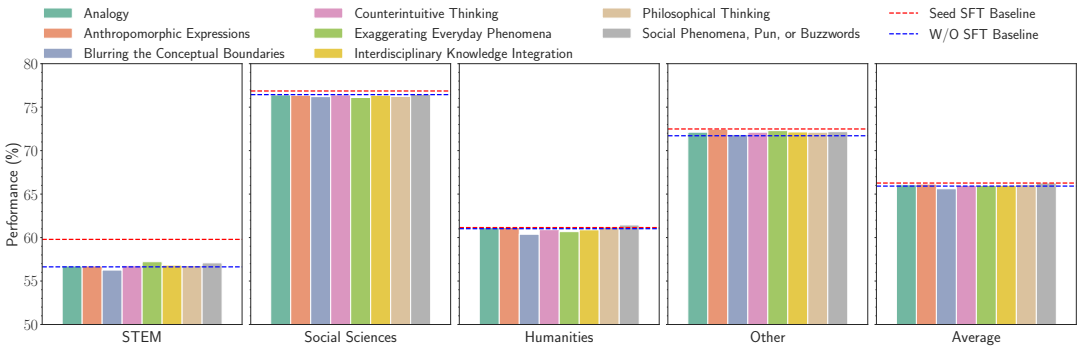


Figure 3: We conduct SFT on the language model LLM using the seed dataset D_{seed} and eight rule-generated datasets. Subject-level accuracy is evaluated on the MMLU test set.

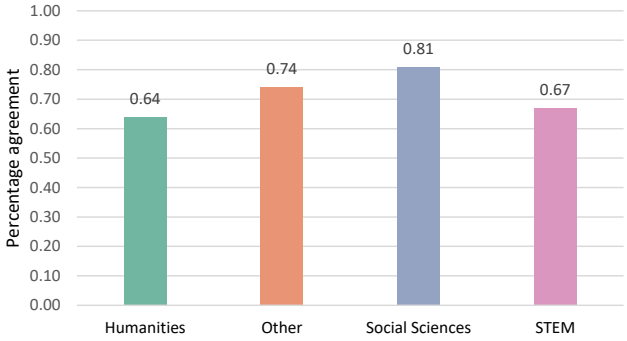


Figure 4: The subject-level percentage agreement for datasets generated by the eight rules, considering whether SFT results improved, declined, or remained unchanged relative to the model fine-tuned with D_{seed}.

ysis shows that the consistency percentage of performance impacts from different rules exceeded 60% across all subjects. Furthermore, in our task-level analysis, 94.73% of the tasks demonstrated a consistency percentage over 50%, and for 26.32% of the tasks, the consistency reached 100%.

We conduct experiments on data filtering and mixing strategies. Due to API costs, we perform the PPL-based filtering and mixing experiment on the 13K dataset and apply all data mixing strategies to a 4K subset. The experimental results are provided in the Appendix G.1. Our findings indicate that none of the mixing strategies surpass the optimal results achieved by using a single rule-based augmentation method.

Overall, using Ruozhiba's rules for data augmentation on the MMLU dataset has not resulted in significant improvements. We believe the reasons are mainly as follows:

1. As shown in the task-level analysis results in the Appendix G.1, the tasks in MMLU vary greatly, encompassing different domains of knowledge (such as natural sciences and social sciences) and different types of tasks (such as knowledge-based questions and logical reasoning). Improvements in some tasks may be offset by declines in others.
2. As shown in the in-context examples in Appendix E.1, Ruozhiba's data combines human knowledge with linguistic ingenuity, and not all texts can be rewritten to achieve such clever expression. In the samples generated using the rules provided in Appendix G, it is evident that GPT-4's rewrites primarily focus on stylistic changes. However, it remains challenging to mimic the nuanced expressions found in the in-context examples.
3. Bai, Du, and Liang (2024) used the BELLE-EVAL benchmark, where GPT serves as the judge. Within the COIG-CQIA evaluation, Ruozhiba is unique for employing GPT in data filtering and human collaboration for response annotation. This is significant because LLM-as-a-Judge can exhibit *self-enhancement bias* (Zheng, Chiang, and Sheng 2023), potentially favoring its own outputs and inflating Ruozhiba's performance. To mitigate this, we calculated the accuracy of multiple-choice questions and had GPT-4 rewrite the questions while keeping the options and responses unchanged, effectively removing this bias. This might explain why the Ruozhiba dataset didn't show significant improvement in our experiments.

5 Conclusion

In summary, our study focuses on the "Ruozhiba" data source to explore the nuanced effects of specific data characteristics on the SFT of LLMs. Our findings underscore the complexity and variability in model performance across different subjects and tasks, influenced by distinct data generation rules. Notably, while rules extracted from "Ruozhiba" tend to diminish performance in STEM-related tasks compared to seed datasets, they offer modest improvements in areas such as Philosophical Thinking and Social Phenomena, Pun, or Buzzwords. Furthermore, our analysis reveals a consistency in the impact of different rules on the performance of models fine-tuned for specific subjects or tasks, suggesting that the choice of generation rules may be less critical than their application to appropriate tasks. This insight is crucial for guiding the development of high-quality SFT datasets, highlighting the importance of tailoring data characteristics to specific domains and tasks to optimize LLM performance.

References

Achiam, J.; Adler, S.; and Agarwal, S. e. a. 2023. GPT-4 Technical Report.

Bai, Y.; Du, X.; and Liang, Y. e. a. 2024. COIG-CQIA: Quality is All You Need for Chinese Instruction Fine-tuning. *arXiv preprint arXiv:2403.18058*.

Bommasani, R.; Hudson, D. A.; Adeli, E.; and Altman, R. 2021. On the Opportunities and Risks of Foundation Models. *ArXiv*, abs/2108.07258.

Cao, Y.; Kang, Y.; Wang, C.; and Sun, L. 2024. Instruction Mining: Instruction Data Selection for Tuning Large Language Models. In *First Conference on Language Modeling*.

Chen, L.; Li, S.; Yan, J.; Wang, H.; Gunaratna, K.; Yadav, V.; Tang, Z.; Srinivasan, V.; Zhou, T.; Huang, H.; and Jin, H. 2024a. AlpaGasus: Training a Better Alpaca with Fewer Data. In *The Twelfth International Conference on Learning Representations*.

Chen, Y.; Qian, S.; Tang, H.; Lai, X.; Liu, Z.; Han, S.; and Jia, J. 2024b. LongLoRA: Efficient Fine-tuning of Long-Context Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Chiang, W.-L.; Li, Z.; and Lin, Z. e. a. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Chung, H. W.; Hou, L.; and Longpre, S. e. a. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70): 1–53.

Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *ArXiv*, abs/1803.05457.

Dubey, A.; Jauhri, A.; and Pandey, A. e. a. 2024. The Llama 3 Herd of Models. *ArXiv*, abs/2407.21783.

Hu, E. J.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Hui, B.; Yang, J.; and Cui, Z. e. a. 2024. Qwen2.5-Coder Technical Report. *ArXiv*, abs/2409.12186.

Ji, Y.; Deng, Y.; and Gong, Y. e. a. 2023. Exploring the Impact of Instruction Data Scaling on Large Language Models: An Empirical Study on Real-World Use Cases. *arXiv preprint arXiv:2303.14742*.

Jiang, A. Q.; Sablayrolles, A.; and Roux, A. e. a. 2024. Mixtral of Experts. *ArXiv*, abs/2401.04088.

Jiang, T.; Huang, S.; Luo, S.; Zhang, Z.; Huang, H.; Wei, F.; Deng, W.; Sun, F.; Zhang, Q.; Wang, D.; et al. 2024. MoRA: High-Rank Updating for Parameter-Efficient Fine-Tuning. *arXiv preprint arXiv:2405.12130*.

Lai, G.; Xie, Q.; Liu, H.; Yang, Y.; and Hovy, E. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In Palmer, M.; Hwa, R.; and Riedel, S., eds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 785–794. Copenhagen, Denmark: Association for Computational Linguistics.

Li, Y.; Hui, B.; Xia, X.; Yang, J.; and Yang, M. 2024. One-Shot Learning as Instruction Data Prospector for Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4586–4601. Bangkok, Thailand: Association for Computational Linguistics.

Lialin, V.; Muckatira, S.; Shivagunde, N.; and Rumshisky, A. 2023. Relora: High-rank training through low-rank updates. In *The Twelfth International Conference on Learning Representations*.

Liu, W.; Zeng, W.; He, K.; Jiang, Y.; and He, J. 2024. What Makes Good Data for Alignment? A Comprehensive Study of Automatic Data Selection in Instruction Tuning. In *The Twelfth International Conference on Learning Representations*.

Lu, K.; Yuan, H.; Yuan, Z.; Lin, R.; Lin, J.; Tan, C.; and Zhou, C. 2023. #InsTag: Instruction Tagging for Analyzing Supervised Fine-tuning of Large Language Models. *ArXiv*, abs/2308.07074.

Mekala, D.; Nguyen, A.; and Shang, J. 2024. Smaller Language Models are capable of selecting Instruction-Tuning Training Data for Larger Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 10456–10470. Bangkok, Thailand: Association for Computational Linguistics.

Mihaylov, T.; Clark, P.; Khot, T.; and Sabharwal, A. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2381–2391.

Ouyang, L.; Wu, J.; and Jiang, X. e. a. 2022. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155.

Reid, M.; Savinov, N.; and Teplyashin, D. e. a. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv*, abs/2403.05530.

Richardson, M.; Burges, C. J.; and Renshaw, E. 2013. MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text. In Yarowsky, D.; Baldwin, T.; Korhonen, A.; Livescu, K.; and Bethard, S., eds., *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 193–203. Seattle, Washington, USA.

Wang, Y.; Kordi, Y.; and Mishra, S. e. a. 2023. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 13484–13508. Toronto, Canada: Association for Computational Linguistics.

Wang, Y.; Mishra, S.; and Alipoormolabashi, P. e. a. 2022. Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language*

Processing, 5085–5109. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Wei, L.; Jiang, Z.; Huang, W.; and Sun, L. 2023. InstructionGPT-4: A 200-Instruction Paradigm for Fine-Tuning MiniGPT-4. *arXiv preprint arXiv:2308.12067*.

Xia, M.; Malladi, S.; Gururangan, S.; Arora, S.; and Chen, D. 2024. LESS: Selecting Influential Data for Targeted Instruction Tuning. In *International Conference on Machine Learning (ICML)*.

Xu, C.; Sun, Q.; and Zheng, K. e. a. 2024. WizardLM: Empowering Large Pre-Trained Language Models to Follow Complex Instructions. In *The Twelfth International Conference on Learning Representations*.

Yang, A.; Zhang, B.; and Hui, B. e. a. 2024. Qwen2.5-Math Technical Report: Toward Mathematical Expert Model via Self-Improvement. *ArXiv*, abs/2409.12122.

Zeng, W.; Xu, C.; and Zhao, Y. e. a. 2024. Automatic Instruction Evolving for Large Language Models. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 6998–7018. Miami, Florida, USA: Association for Computational Linguistics.

Zheng, L.; Chiang, W.-L.; and Sheng, Y. e. a. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623.

Zhou, C.; Liu, P.; and et al., P. X. 2023. LIMA: Less Is More for Alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Appendix

By extracting these rules, we aim to guide the Instruction Rewriter in generating high-quality SFT datasets. We provide in-context examples with the expectation that the Instruction Rewriter will follow the identified rules. This approach ultimately aims to enhance the performance of LLMs across various tasks by ensuring the generated datasets are aligned with these interdisciplinary and nuanced rules.

A Supplementary Related Work on Supervised Fine-Tuning and Data Selection

Supervised fine-tuning: also known as instruction tuning, is a crucial component of the post-training paradigm for LLMs. SFT involves fine-tuning LLMs, which have undergone next-token prediction, on high-quality datasets composed of (instruction, output) pairs. This process helps bridge the gap between model predictions and the knowledge or preferences of human users. For instance, InstructGPT (Ouyang, Wu, and Jiang 2022) demonstrated that fine-tuning LLMs with high-quality SFT datasets annotated by humans significantly enhances the model's ability to follow user instructions. Similarly, the study (Chung, Hou, and Longpre 2024) highlighted that fine-tuning language models on a set of instruction-formulated datasets can improve model performance and generalization capability on unseen tasks.

Given the substantial resource demands of fully fine-tuning large models, parameter-efficient fine-tuning (PEFT) methods have gained prominence. Techniques like LoRA (Hu et al. 2022) and its variants (Chen et al. 2024b; Jiang et al. 2024; Lialin et al. 2023) update the dense neural network layers of LLMs with plug-in low-rank matrices, achieving comparable performance to full fine-tuning while keeping the LLM parameters frozen. This approach has become a mainstream method for PEFT in large models.

Post-training data selection: Compared to the vast amounts of data used during pre-training, post-training emphasizes achieving significant improvements with a smaller quantity of high-quality data. LIMA (Zhou, Liu, and et al. 2023) demonstrated that using just 1,000 carefully constructed SFT data points can significantly boost the performance of mainstream open-source LLMs like LLama on downstream tasks. Another study (Liu et al. 2024) investigated SFT data selection from the perspectives of complexity, quality, and diversity, showing that models fine-tuned with 6,000 SFT training samples perform comparably to baseline models trained with tens of thousands of samples. Furthermore, Mekala, Nguyen, and Shang (2024) advocated for the ability of current language models to autonomously select high-quality training data, designing a perplexity (ppl)-based data selection method that effectively trains models to perform on par with those trained on complete datasets.

COIG-CQIA (Bai, Du, and Liang 2024) collected data from various sources to construct an SFT dataset. They found that fine-tuning LLMs with the Ruozhiba subset significantly outperformed other data sources on the BELLE-EVAL benchmark (Ji, Deng, and Gong 2023). The authors attributed this to the unique characteristics of the Ruozhiba dataset, which includes cognitive and linguistic traps, jokes, riddles, and artistic and abstract rhetorical techniques, thereby enhancing the LLM's language understanding and multi-hop logical reasoning capabilities. These studies collectively underscore the importance of data selection in SFT.

B Details of Rules

In this section, we articulate the rules in Figure 2 extracted from the Ruozhiba dataset.

B.1 Humorous Social Commentary

Social Phenomena, Pun, or Buzzwords: The use of puns and buzzwords adds humor and relatability, making social commentaries engaging and thought-provoking. By incorporating witty language and familiar phrases, the discussion becomes more accessible and entertaining, capturing readers' attention and encouraging reflection on underlying issues in a light-hearted manner. This method enhances the impact of the commentary, making complex topics more digestible and stimulating meaningful conversations.

Exaggerating Everyday Phenomena: Some instances in the dataset use exaggeration and literal interpretations to create absurd and humorous scenarios. This approach challenges readers to think critically about the logical extensions of everyday phenomena and the potential pitfalls of taking things too literally or out of context. It entertains while encouraging deeper reflection on rules, interpretations, and statistical measures often taken for granted.

B.2 Conceptual Flexibility

Blurring the Conceptual Boundaries: This rule involves challenging implicit assumptions by applying ideas from one scenario to another. It reveals nuanced, context-dependent truths that highlight the limitations of rigid, one-size-fits-all thinking. By examining assumptions through contrasting contexts, it encourages a more flexible and adaptable mindset, appreciating the complexity of interactions in our physical, social, and conceptual environments.

Counterintuitive Thinking: Counterintuitive thinking serves as a lens to view familiar situations in unexpected ways, often by purposefully misapplying assumptions or questioning norms. This approach exposes flaws in our thinking patterns and challenges the logic we take for granted.

Analogies That Appear Absurd but Are Actually Reasonable: Analogies play a crucial role in illuminating the interconnectedness of actions, the limitations of human perception, and the complexity of decision-making. They make abstract concepts tangible, highlight hidden connections, provoke reflection, and facilitate effective communication, thereby enhancing our understanding of life's complexities.

Interdisciplinary Knowledge Integration: This rule highlights the use of a cross-disciplinary approach to reframe everyday perceptions by combining scientific or logical methods. It aims to examine human behavior and social norms, challenging everyday assumptions with grounded, factual approaches to often subjective or socially driven phenomena.

B.3 Human-like Reflections

Philosophical Thinking: These contents encourage the grasp of nuanced and abstract concepts, enhancing the ability to interpret complex human thought patterns and language. By processing abstract thinking patterns, the model learns to generate responses that go beyond rigid, literal interpretations. This is crucial for engaging with open-ended, speculative questions, helping align more closely with human thought processes in creative, philosophical, or hypothetical discussions.

Anthropomorphic Expressions: This pattern involves exploring complex concepts by attributing human-like qualities to non-human elements, inviting a playful yet thought-provoking reexamination of everyday assumptions. These expressions question societal norms and values by comparing physical or economic realities to personal experiences, creating a lens of irony and curiosity. Through humor and paradox, they highlight the human tendency to accept certain systems without questioning their deeper implications, suggesting a reevaluation of taken-for-granted aspects.

C Rule Dataset

Our research focuses on exploring whether the rules embedded in the Ruozhiba data from COIG-CQIA can enhance the performance of SFT models. For this purpose, we selected the Ruozhiba dataset from COIG-CQIA for rule extraction. This dataset originates from Baidu Tieba, the largest Chinese forum, and its content often uses metaphors, puns, and language traps to satirize social phenomena or create humor. The authors of COIG-CQIA, after scraping the data, conducted rigorous filtering to remove non-instructive and harmful content, ultimately selecting 240 questions. Subsequently, they utilized GPT-4 to annotate these data and continuously prompted GPT-4 with human input to optimize the responses, thereby avoiding language traps and obtaining high-quality 240 pairs of $\{\mathrm{ins}, \mathrm{rep}\}$. These optimized data not only retain the complex linguistic structures and humorous elements of the original data but also ensure their effectiveness and safety in training SFT models. In this way, they aimed to construct a high-quality supervised fine-tuning dataset to enhance the performance of large language models in multi-task environments.

D Rule Generation

We utilized the Ruozhiba dataset that consists of 240 data points as in-context examples for GPT-4. The goal was to extract relevant rules based on interdisciplinary knowledge from psychology, sociology, education, and cognitive science. The complete prompt used for this extraction process is detailed in the Appendix E. Through this process, we extracted eight rules as shown in Figure 2.

E Details of Prompts

In this section, we further present all the prompts we used in our study. The base prompt is modified based on the prompt used by Xu, Sun, and Zheng (2024).

E.1 Prompt for the Instruction Rewriter

Base Prompt:

```
I want you to act as a Prompt Rewriter. Your objective is to rewrite a given
prompt into a more complex version to make those famous AI systems (e.g.,
chatgpt and GPT4) a bit harder to handle. But the rewritten prompt must be
reasonable and must be understood and responded by humans. Your rewriting cannot
omit the non-text parts such as the table and code in the original prompt. Also,
please do not omit the input in the original prompt. You SHOULD complicate the
given prompt using the following method: [Rule-specific prompt] You should try
your best NOT to make the rewritten prompt become verbose; the rewritten prompt
can only add 10 to 20 words into the original prompt. You should ensure that
any names of people or entities are retained. Certain terms like 'original
prompt', 'rewritten prompt', and related phrases are NOT allowed to appear
in the rewritten prompt. It is also NOT allowed to include reference choices
```

```
content in the rewritten prompt.
```

Social Phenomena, Pun, or Buzzwords:

```
Please rewrite #The Given Prompt# using puns based on social phenomena, pun
or buzzwords, while preserving the original intent as much as possible. Note
that I prefer you to refer to the relevant knowledge rather than directly using
specific terms. Regarding using puns based on social phenomena or buzzwords,
here are some examples for your reference:
  1. Experts suggest reducing the use of electronic devices, but isn't there no
material in the world that doesn't contain electronics?
  2. I bought a set of toys with the 26 letters of the alphabet, but I only
received 23 of them. When I contacted the seller about it, they said I didn't
purchase the DLC.
  3. The saying goes, Önly with pressure comes motivation.Šo, can atmospheric
pressure also provide motivation?
  #The Given Prompt#: [The Given Prompt]
  #Reference Choices#:[Reference Choices]
  #Reference Answer#: [Reference Answer]
  #Rewritten Prompt#:
```

Exaggerating Everyday Phenomena:

```
Please rewrite #The Given Prompt# by exaggerating everyday phenomena, while
preserving the original intent as much as possible. Note that I prefer you
to refer to the relevant knowledge rather than directly using specific terms.
Regarding exaggerating everyday phenomena, here are some examples for your
reference:
  1. If concentrated sulfuric acid contains 2% water, does drinking 50 cups of
sulfuric acid mean I would have consumed a cup of pure water?
  2. Is it a violation to run after drinking all the water during a swimming
competition?
  3. If a couple keeps getting married and divorced repeatedly, does that count
as contributing to the marriage rate?
  #The Given Prompt#: [The Given Prompt]
  #Reference Choices#:[Reference Choices]
  #Reference Answer#: [Reference Answer]
  #Rewritten Prompt#:
```

Blurring the Conceptual Boundaries:

```
Please rewrite #The Given Prompt# by blurring the conceptual boundaries, while
preserving the original intent as much as possible. Note that I prefer you
to refer to the relevant knowledge rather than directly using specific terms.
Regarding blurring the conceptual boundaries, here are some examples for your
reference:
  1. After intense exercise, one should avoid drinking ice-cold water and taking
cold showers. So what should one do if they're engaging in vigorous activity in
the water?
  2. Since a rusty knife can cause tetanus when used to cut someone, why didn't
ancient people simply use rusty weapons?
  3. Everyone works to make money, so who is losing money?
  #The Given Prompt#: [The Given Prompt]
```

```
  #Reference Choices#:[Reference Choices]
  #Reference Answer#: [Reference Answer]
  #Rewritten Prompt#:
```

Counterintuitive Thinking:

```
Please rewrite #The Given Prompt# using counterintuitive thinking, while
preserving the original intent as much as possible. Note that I prefer you
to refer to the relevant knowledge rather than directly using specific terms.
Regarding using counterintuitive thinking, here are some examples for your
reference:
  1. Since the prison is full of criminals, why don’t the police go into the
prison to arrest people?
  2. A person should look for cars when crossing the street, but what should
they do if there are no cars on the road?
  3. If a surgery has a success rate of only 50%, then doing it twice would
result in a success rate of just 25%. So, if we only perform half of the
surgery, wouldn’t that give us a 100% success rate?
  #The Given Prompt#: [The Given Prompt]
  #Reference Choices#:[Reference Choices]
  #Reference Answer#: [Reference Answer]
  #Rewritten Prompt#:
```

Analogies That Appear Absurd but Are Actually Reasonable:

```
Please #The Given Prompt# using an analogy that appears absurd but is actually
reasonable, while preserving the original intent as much as possible. Note
that I prefer you to refer to the relevant knowledge rather than directly using
specific terms. Regarding using an analogy that appears absurd but is actually
reasonable, here are some examples for your reference:
  1. On a rainy day, I stepped into a puddle and accidentally shattered the sky.
  2. We are all blind people in life, groping the elephant named the world.
  3. If I shatter the nightmare, am I the destroyer of dreams or the guardian of
happiness?
  #The Given Prompt#: [The Given Prompt]
  #Reference Choices#:[Reference Choices]
  #Reference Answer#: [Reference Answer]
  #Rewritten Prompt#:
```

Interdisciplinary Knowledge Integration:

```
Please rewrite #The Given Prompt# by integrating interdisciplinary knowledge,
while preserving the original intent as much as possible. Note that I prefer you
to refer to the relevant knowledge rather than directly using specific terms.
Regarding integrating interdisciplinary knowledge, here are some examples for
your reference:
  1. It’s clearly food-grade stainless steel|so why is it still so hard to
swallow?
  2. The baby raised by wolves can understand wolf language, so will a baby
raised by robots know C language?
  3. Many girls tend to exaggerate their age, so why not just look at their
annual rings?
  #The Given Prompt#: [The Given Prompt]
```

```
  #Reference Choices#:[Reference Choices]
  #Reference Answer#: [Reference Answer]
  #Rewritten Prompt#:
```

Philosophical Thinking:

```
Please rewrite #The Given Prompt# by incorporating philosophical thinking, while
preserving the original intent as much as possible. Note that I prefer you
to refer to the relevant knowledge rather than directly using specific terms.
Regarding incorporating philosophical thinking, here are some examples for your
reference:
  1. The alarm clock shatters the dream; does it also shatter the life of
another world?
  2. The cry of a newborn baby|is it a joyful expression of life, or a fear of
the world?
  3. The meaning of life is to find happiness, so why does everyone seem to live
in a sea of suffering?
  #The Given Prompt#: [The Given Prompt]
  #Reference Choices#:[Reference Choices]
  #Reference Answer#: [Reference Answer]
  #Rewritten Prompt#:
```

Anthropomorphic Expressions:

```
Please rewrite #The Given Prompt# using anthropomorphic expressions, while
preserving the original intent as much as possible. Note that I prefer you
to refer to the relevant knowledge rather than directly using specific terms.
Regarding using anthropomorphic expressions, here are some examples for your
reference:
  1. Since 70% of the human body is water, does that mean 7 out of every 10
people are just water disguised as humans?
  2. When a person is away from money, they are useless; when money is away from
a person, it is just a piece of paper. So why do people work hard to earn money
instead of money working hard to find people?
  3. Why do we take medicine when we are sick, but the world seems to resort
to sacrificing people when it is unwell? Can a person heal through medication,
while the world can only be cured by taking lives?
  #The Given Prompt#: [The Given Prompt]
  #Reference Choices#:[Reference Choices]
  #Reference Answer#: [Reference Answer]
  #Rewritten Prompt#:
```

E.2 Prompt for the Judge

```
I want you to act as a Data Synthesis Evaluator. Your objective is to evaluate
synthesized data against given standards to ensure quality and relevance. You
will be provided with an original question, the answer, the synthesis rule,
and the synthesized question. Your evaluation should be based on the following
criteria, each rated on a scale of 1-10:
  1. Consistency: Does the synthesized question align with the original
question?
  2. Correctness: Is the synthesized question accurate and free of errors?
  3. Alignment: Does the synthesized question adhere to the specified synthesis
```

```
rules?
  Original Question: [Original Question]
  Answer: [Answer]
  Synthesis Rule: [Synthesis Rule]
  Synthesized Question: [Synthesized Question]
  Please provide your evaluation in the following structured format:
  ---
  Consistency: [Score 1-10]
  Correctness: [Score 1-10]
  Alignment:   [Score 1-10]
  ---
```

F Detailed Experiments

In this section, we present our experimental details and the task distribution of the datasets.

F.1 Experimental Details

In the supervised fine-tuning of the Meta-Llama-3-8B-Instruct model (Dubey, Jauhri, and Pandey 2024), we configured hyper-parameters to optimize the balance between model performance and computational efficiency. The per-device training batch size was set to 1, complemented by 8 gradient accumulation steps. We employed a learning rate of 5×10^{-6}, utilizing a cosine learning rate scheduler to ensure gradual decay of the learning rate. A warmup ratio of 0.1 was implemented to stabilize initial training phases. The model was trained for 6 epochs, a duration chosen to optimize convergence while minimizing the risk of overfitting. We evaluated the model after SFT on MMLU using the 5-shot setting.

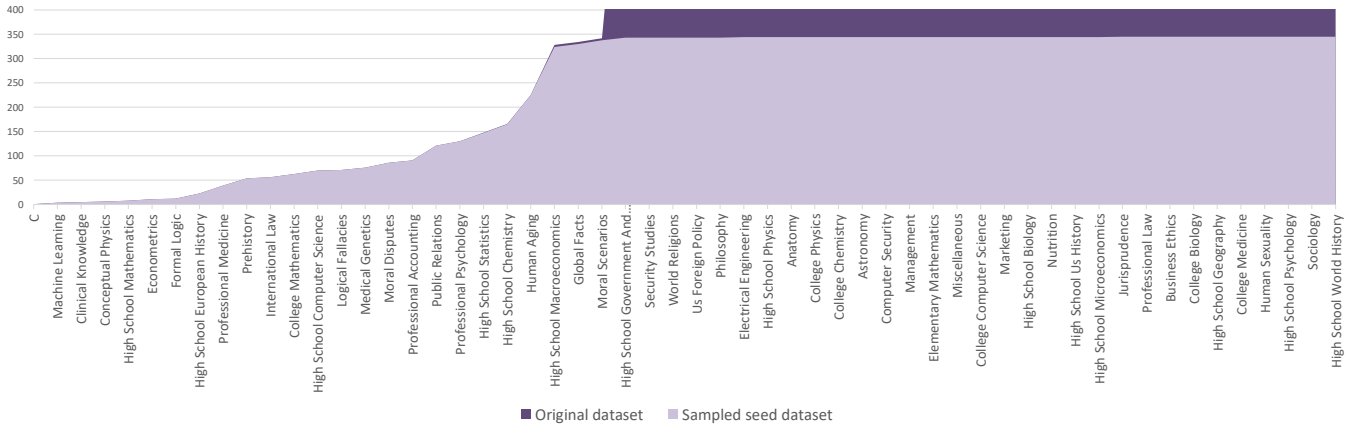


Figure 5: We sampled the MMLU training set to include all 57 tasks, ensuring diverse task coverage for the experiment.

G Examples of Generated Data

In this section, we present a seed example from the D_{seed} dataset, along with eight parallel examples generated by the Instruction Rewriter, each rewritten according to one of the eight specified rules.

Seed sample:

How do most fish get the oxygen they need to survive?

Social Phenomena, Pun, or Buzzwords:

In the vast aquatic landscape, where fish navigate their liquid homes, how do these finned denizens engage in the biohack of extracting that life-sustaining gas, famously known as oxygen, from their watery world?

Exaggerating Everyday Phenomena:

By what miraculous process do most aquatic creatures extract the vital essence of life, allowing them to thrive beneath the surface of their watery realm?

Blurring the Conceptual Boundaries:

In the aquatic realm, the majority of these finned creatures acquire the essential life-sustaining element required for their survival. Could you explain the intricate process by which they manage to extract this invisible resource, necessary for their existence, from the environment that envelops them?

Counterintuitive Thinking:

In an environment where creatures spend their lives submerged in water and not on land, how do these beings extract the essential life-sustaining element needed from a medium that does not provide it in its gaseous form?

Analogies That Appear Absurd but Are Actually Reasonable:

By navigating the ocean's symphony, how do fish conduct their own orchestra to extract the elusive breath of life they require to survive?

Interdisciplinary Knowledge Integration:

In the context of aquatic life, how do most fish acquire the essential element for cellular respiration, considering the principles of fluid dynamics and the solubility of gases in liquid environments?

Philosophical Thinking:

What is the mechanism by which most aquatic beings procure the vital essence of oxygen necessary for their continued existence? Consider the interplay between their physical structures and the surrounding life-giving medium.

Anthropomorphic Expressions:

In the grand underwater theater of life, where do fish draw their breath from, to continue their aquatic dance?

G.1 Detailed Results

In this section, we present the task-level experimental results, including the accuracy of the SFT model on the MMLU test set under different experimental settings, as well as the task-level percentage agreement.

Task	Subject	Analogy	Anthropomorphic Expressions	Blurring the Conceptual Boundaries	Counterintuitive Thinking	Exaggerating Everyday Phenomena	Interdisciplinary Knowledge Integration	Philosophical Thinking	Social Phenomena, Pun, or Buzzwords	Summary
Conceptual Physics	STEM	0.85%	-1.28%	0.43%	1.70%	0.85%	1.28%	0.85%	1.70%	
High School Statistics	STEM	-0.46%	-0.93%	-0.46%	-0.46%	-0.46%	-0.93%	-1.39%	-0.93%	
College Biology	STEM	1.39%	1.39%	0.69%	0.00%	2.78%	0.69%	0.69%	0.00%	
College Chemistry	STEM	0.00%	-1.00%	0.00%	-1.00%	-1.00%	1.00%	0.00%	0.00%	
Elementary Mathematics	STEM	0.26%	1.06%	0.26%	-0.79%	0.00%	0.00%	0.53%	1.06%	
Electrical Engineering	STEM	0.00%	-0.69%	-2.07%	0.69%	1.38%	0.00%	0.69%	0.00%	
Formal Logic	STEM	-1.59%	-1.59%	-1.59%	-3.17%	-1.59%	-0.79%	0.79%	0.79%	
Abstract Algebra	STEM	3.00%	0.00%	-1.00%	2.00%	-1.00%	1.00%	3.00%	1.00%	
High School Computer Science	STEM	1.00%	1.00%	1.00%	-3.00%	-1.00%	0.00%	-2.00%	1.00%	
Computer Security	STEM	1.00%	2.00%	2.00%	1.00%	1.00%	-1.00%	1.00%	1.00%	
High School Chemistry	STEM	-3.45%	-0.99%	-1.48%	-2.46%	-0.99%	-0.49%	-1.48%	-1.97%	
College Computer Science	STEM	-2.00%	-2.00%	-4.00%	-1.00%	-1.00%	-2.00%	-1.00%	-1.00%	
High School Biology	STEM	0.00%	-0.32%	-1.61%	0.00%	0.65%	-0.65%	0.32%	0.32%	
High School Mathematics	STEM	-2.22%	-1.11%	-1.11%	0.37%	0.74%	-0.37%	-0.37%	1.11%	
College Physics	STEM	1.96%	0.00%	-0.98%	0.00%	0.00%	0.98%	0.98%	0.00%	
Astronomy	STEM	0.66%	0.66%	0.00%	-1.32%	-0.66%	0.66%	0.00%	0.66%	
High School Physics	STEM	0.66%	1.99%	0.00%	3.31%	4.64%	1.32%	-1.32%	-0.66%	
College Mathematics	STEM	0.00%	-2.00%	-1.00%	0.00%	1.00%	-1.00%	0.00%	0.00%	
Machine Learning	STEM	0.00%	1.79%	0.00%	0.89%	0.00%	0.89%	0.00%	0.89%	
Anatomy	STEM	0.74%	1.48%	2.22%	1.48%	0.00%	0.74%	2.22%	1.48%	
Jurisprudence	Humanities	0.00%	0.93%	-0.93%	0.00%	0.93%	0.93%	0.93%	0.93%	
Moral Scenarios	Humanities	2.23%	3.02%	-0.11%	0.67%	-0.56%	1.45%	-0.78%	-0.56%	
High School World History	Humanities	-1.69%	-0.42%	0.00%	0.00%	-1.69%	-2.11%	-1.27%	-0.84%	
High School European History	Humanities	0.00%	0.61%	0.61%	0.61%	-0.61%	-0.61%	0.61%	0.61%	
Philosophy	Humanities	-2.57%	-1.29%	-1.93%	-1.29%	-1.93%	0.00%	-0.96%	-1.29%	
Moral Disputes	Humanities	0.00%	-0.87%	-0.87%	-0.87%	0.29%	0.29%	0.00%	0.29%	
High School Us History	Humanities	-2.45%	-2.45%	-1.96%	-1.47%	-1.47%	-1.96%	-1.47%	-1.47%	
Professional Law	Humanities	0.07%	-0.59%	-0.98%	0.00%	-0.20%	-0.85%	1.11%	1.30%	
International Law	Humanities	0.00%	0.00%	2.48%	-2.48%	1.65%	0.83%	2.48%	2.48%	
Logical Fallacies	Humanities	0.00%	-1.23%	0.00%	1.23%	0.61%	-0.61%	-0.61%	-0.61%	
Professional Medicine	Other	-1.47%	-1.84%	-0.37%	-1.84%	-0.37%	-1.84%	-1.47%	-0.74%	
Virology	Other	0.00%	0.60%	-0.60%	-1.20%	0.60%	-0.60%	1.20%	1.20%	
Prehistory	Other	-0.31%	0.62%	-1.54%	-0.31%	0.00%	-0.31%	0.00%	1.23%	
Miscellaneous	Other	0.77%	1.02%	0.51%	0.77%	1.02%	0.89%	0.38%	0.51%	
Business Ethics	Other	1.00%	1.00%	-1.00%	2.00%	1.00%	2.00%	-1.00%	0.00%	
Management	Other	-3.88%	-0.97%	-0.97%	-2.91%	0.00%	-2.91%	-1.94%	-1.94%	
Nutrition	Other	-1.96%	-2.94%	-3.92%	-3.27%	-2.94%	-2.61%	-2.29%	-1.96%	
Medical Genetics	Other	1.00%	-1.00%	-1.00%	0.00%	0.00%	0.00%	-1.00%	-1.00%	
Econometrics	Other	-1.75%	-2.63%	-6.14%	3.51%	0.88%	-1.75%	0.00%	2.63%	
Marketing	Other	-0.43%	0.43%	0.85%	0.00%	0.00%	0.43%	0.43%	0.43%	
College Medicine	Other	-0.58%	0.00%	-0.58%	-1.16%	-1.16%	0.00%	-1.16%	-1.73%	
Professional Accounting	Other	-2.48%	-0.35%	-2.48%	-1.42%	-1.77%	-2.48%	-1.42%	-1.42%	
Sociology	Social Sciences	-4.98%	-4.48%	-2.99%	-3.48%	-3.98%	-2.49%	-3.98%	-4.48%	
Clinical Knowledge	Social Sciences	0.38%	0.38%	-0.38%	-0.38%	-0.38%	0.38%	0.00%	0.38%	
Public Relations	Social Sciences	0.91%	0.00%	0.00%	2.73%	0.91%	0.00%	0.00%	0.00%	
Human Aging	Social Sciences	0.00%	0.90%	-0.45%	0.00%	0.00%	-0.45%	-0.45%	-0.90%	
Human Sexuality	Social Sciences	0.76%	1.53%	0.00%	-0.76%	0.76%	0.76%	1.53%	1.53%	
Us Foreign Policy	Social Sciences	0.00%	0.00%	-1.00%	-1.00%	0.00%	0.00%	0.00%	0.00%	
High School Macroeconomics	Social Sciences	0.00%	0.26%	1.28%	0.77%	-1.28%	0.00%	0.26%	0.77%	
High School Government And Politics	Social Sciences	0.52%	0.52%	0.52%	0.52%	0.52%	1.04%	1.04%	1.04%	
Security Studies	Social Sciences	-1.63%	-2.04%	-1.63%	-2.86%	-2.45%	-2.86%	-2.86%	-2.04%	
Global Facts	Social Sciences	1.00%	0.00%	-3.00%	5.00%	2.00%	2.00%	1.00%	0.00%	
Professional Psychology	Social Sciences	-0.33%	-0.82%	-0.98%	-1.31%	-1.31%	-0.82%	-2.12%	-1.47%	
High School Geography	Social Sciences	-2.53%	-1.52%	-1.52%	-2.02%	-2.02%	-2.53%	-2.02%	-2.02%	
High School Microeconomics	Social Sciences	1.68%	2.10%	0.42%	1.68%	0.84%	2.10%	1.68%	1.26%	
World Religions	Social Sciences	-1.17%	-2.34%	-1.75%	-1.17%	-1.75%	-1.17%	-1.75%	-1.17%	
High School Psychology	Social Sciences	0.37%	0.18%	0.00%	-0.18%	0.37%	0.18%	0.73%	0.37%	

Figure 6: The task-level accuracy changes on eight rule-generated datasets (13K samples) compared to fine-tuning with D_{seed}.

Task	Subject	Seed	Analogy	Anthropomorphic Expressions	Blurring the Conceptual Boundaries	Counterintuitive Thinking	Exaggerating Everyday Phenomena	Interdisciplinary Knowledge Integration	Philosophical Thinking	Social Phenomena, Pun, or Buzzwords
Conceptual Physics	STEM	60.43	61.28	59.15	60.85	62.13	61.28	61.70	61.28	62.13
High School Statistics	STEM	62.50	62.04	61.57	62.04	62.04	62.04	61.57	61.11	61.57
College Biology	STEM	81.25	82.64	82.64	81.94	81.25	84.03	81.94	81.94	81.25
College Chemistry	STEM	50.00	50.00	49.00	50.00	49.00	49.00	51.00	50.00	50.00
Elementary Mathematics	STEM	46.30	46.56	47.35	46.56	45.50	46.30	46.30	46.83	47.35
Electrical Engineering	STEM	62.76	62.76	62.07	60.69	63.45	64.14	62.76	63.45	62.76
Formal Logic	STEM	56.35	54.76	54.76	54.76	53.17	54.76	55.56	57.14	57.14
Abstract Algebra	STEM	28.00	31.00	28.00	27.00	30.00	27.00	29.00	31.00	29.00
High School Computer Science	STEM	73.00	74.00	74.00	74.00	70.00	72.00	73.00	71.00	74.00
Computer Security	STEM	75.00	76.00	77.00	77.00	76.00	76.00	74.00	76.00	76.00
High School Chemistry	STEM	56.65	53.20	55.67	55.17	54.19	55.67	56.16	55.17	54.68
College Computer Science	STEM	53.00	51.00	51.00	49.00	52.00	52.00	51.00	52.00	52.00
High School Biology	STEM	79.68	79.68	79.35	78.06	79.68	80.32	79.03	80.00	80.00
High School Mathematics	STEM	37.04	34.81	35.93	35.93	37.41	37.78	36.67	36.67	38.15
College Physics	STEM	47.06	49.02	47.06	46.08	47.06	47.06	48.04	48.04	47.06
Astronomy	STEM	73.68	74.34	74.34	73.68	72.37	73.03	74.34	73.68	74.34
High School Physics	STEM	39.07	39.74	41.06	39.07	42.38	43.71	40.40	37.75	38.41
College Mathematics	STEM	39.00	39.00	37.00	38.00	39.00	40.00	38.00	39.00	39.00
Machine Learning	STEM	49.11	49.11	50.89	49.11	50.00	49.11	50.00	49.11	50.00
Anatomy	STEM	62.22	62.96	63.70	64.44	63.70	62.22	62.96	64.44	63.70
Jurisprudence	Humanities	75.93	75.93	76.85	75.00	75.93	76.85	76.85	76.85	76.85
Moral Scenarios	Humanities	48.49	50.73	51.51	48.38	49.16	47.93	49.94	47.71	47.93
High School World History	Humanities	84.81	83.12	84.39	84.81	84.81	83.12	82.70	83.54	83.97
High School European History	Humanities	76.36	76.36	76.97	76.97	76.97	75.76	75.76	76.97	76.97
Philosophy	Humanities	74.92	72.35	73.63	72.99	73.63	72.99	74.92	73.95	73.63
Moral Disputes	Humanities	69.65	69.65	68.79	68.79	68.79	69.94	69.94	69.65	69.94
High School Us History	Humanities	87.25	84.80	84.80	85.29	85.78	85.78	85.29	85.78	85.78
Professional Law	Humanities	47.33	47.39	46.74	46.35	47.33	47.13	46.48	48.44	48.63
International Law	Humanities	78.51	78.51	78.51	80.99	76.03	80.17	79.34	80.99	80.99
Logical Fallacies	Humanities	75.46	75.46	74.23	75.46	76.69	76.07	74.85	74.85	74.85
Professional Medicine	Other	76.84	75.37	75.00	76.47	75.00	76.47	75.00	75.37	76.10
Virology	Other	48.80	48.80	49.40	48.19	47.59	49.40	48.19	50.00	50.00
Prehistory	Other	70.99	70.68	71.60	69.44	70.68	70.99	70.68	70.99	72.22
Miscellaneous	Other	79.05	79.82	80.08	79.57	79.82	80.08	79.95	79.44	79.57
Business Ethics	Other	71.00	72.00	72.00	70.00	73.00	72.00	73.00	70.00	71.00
Management	Other	82.52	78.64	81.55	81.55	79.61	82.52	79.61	80.58	80.58
Nutrition	Other	79.41	77.45	76.47	75.49	76.14	76.47	76.80	77.12	77.45
Medical Genetics	Other	79.00	80.00	78.00	78.00	79.00	79.00	79.00	78.00	78.00
Econometrics	Other	50.00	48.25	47.37	43.86	53.51	50.88	48.25	50.00	52.63
Marketing	Other	90.60	90.17	91.03	91.45	90.60	90.60	91.03	91.03	91.03
College Medicine	Other	66.47	65.90	66.47	65.90	65.32	65.32	66.47	65.32	64.74
Professional Accounting	Other	50.35	47.87	50.00	47.87	48.94	48.58	47.87	48.94	48.94
Sociology	Social Sciences	88.06	83.08	83.58	85.07	84.58	84.08	85.57	84.08	83.58
Clinical Knowledge	Social Sciences	75.85	76.23	76.23	75.47	75.47	75.47	76.23	75.85	76.23
Public Relations	Social Sciences	70.00	70.91	70.00	70.00	72.73	70.91	70.00	70.00	70.00
Human Aging	Social Sciences	73.09	73.09	73.99	72.65	73.09	73.09	72.65	72.65	72.20
Human Sexuality	Social Sciences	79.39	80.15	80.92	79.39	78.63	80.15	80.15	80.92	80.92
Us Foreign Policy	Social Sciences	86.00	86.00	86.00	85.00	85.00	86.00	86.00	86.00	86.00
High School Macroeconomics	Social Sciences	66.15	66.15	66.41	67.44	66.92	64.87	66.15	66.41	66.92
High School Government And Politics	Social Sciences	89.64	90.16	90.16	90.16	90.16	90.16	90.67	90.67	90.67
Security Studies	Social Sciences	77.96	76.33	75.92	76.33	75.10	75.51	75.10	75.10	75.92
Global Facts	Social Sciences	46.00	47.00	46.00	43.00	51.00	48.00	48.00	47.00	46.00
Professional Psychology	Social Sciences	69.93	69.61	69.12	68.95	68.63	68.63	69.12	67.81	68.46
High School Geography	Social Sciences	84.85	82.32	83.33	83.33	82.83	82.83	82.32	82.83	82.83
High School Microeconomics	Social Sciences	76.05	77.73	78.15	76.47	77.73	76.89	78.15	77.73	77.31
World Religions	Social Sciences	80.12	78.95	77.78	78.36	78.95	78.36	78.95	78.36	78.95
High School Psychology	Social Sciences	d	85.69	85.50	85.32	85.14	85.69	85.50	86.06	85.69

Figure 7: The task-level accuracy on eight rule-generated datasets (13K samples) compared to fine-tuning with D_{seed}.

Task	Subject	Analogy	Anthropomorphic Expressions	Blurring the Conceptual Boundaries	Counterintuitive Thinking	Exaggerating Everyday Phenomena	Interdisciplinary Knowledge Integration	Philosophical Thinking	Social Phenomena, Pun, or Buzzwords	Summary
Conceptual Physics	STEM	0.43%	-0.85%	0.00%	1.28%	-0.85%	0.00%	-1.28%	-0.85%	
High School Statistics	STEM	0.93%	2.31%	1.39%	0.46%	0.00%	0.93%	1.39%	1.39%	
College Biology	STEM	2.08%	2.78%	2.08%	2.08%	0.69%	2.78%	0.69%	0.69%	
College Chemistry	STEM	1.00%	-1.00%	-2.00%	-3.00%	0.00%	-2.00%	1.00%	1.00%	
Elementary Mathematics	STEM	0.26%	0.26%	-0.26%	0.26%	0.53%	-0.26%	0.00%	-0.26%	
Electrical Engineering	STEM	2.07%	2.07%	2.07%	0.00%	1.38%	0.69%	2.76%	2.07%	
Formal Logic	STEM	0.00%	-0.79%	0.00%	-1.59%	-3.17%	-0.79%	-2.38%	-0.79%	
Abstract Algebra	STEM	4.00%	1.00%	0.00%	5.00%	1.00%	3.00%	0.00%	0.00%	
High School Computer Science	STEM	-3.00%	1.00%	-2.00%	-2.00%	-1.00%	1.00%	0.00%	-1.00%	
Computer Security	STEM	-2.00%	1.00%	0.00%	-1.00%	-2.00%	-2.00%	-1.00%	0.00%	
High School Chemistry	STEM	-0.49%	-1.97%	0.00%	-2.46%	-0.49%	-0.49%	0.00%	-0.99%	
College Computer Science	STEM	-1.00%	1.00%	-1.00%	-1.00%	3.00%	-2.00%	0.00%	1.00%	
High School Biology	STEM	0.00%	-0.32%	-0.32%	-0.65%	0.00%	-0.32%	-0.65%	-0.97%	
High School Mathematics	STEM	-1.11%	-1.85%	-1.11%	-1.85%	0.37%	-1.11%	0.37%	0.37%	
College Physics	STEM	0.98%	-1.96%	-1.96%	-2.94%	-3.92%	0.00%	-1.96%	-1.96%	
Astronomy	STEM	-1.97%	-1.32%	-3.29%	-1.97%	-1.32%	-3.29%	-1.97%	-1.97%	
High School Physics	STEM	-0.66%	-1.32%	0.00%	0.66%	2.65%	-1.32%	-0.66%	1.32%	
College Mathematics	STEM	0.00%	0.00%	-1.00%	0.00%	0.00%	-1.00%	0.00%	0.00%	
Machine Learning	STEM	-1.79%	2.68%	1.79%	2.68%	2.68%	2.68%	0.00%	0.89%	
Anatomy	STEM	0.00%	2.96%	0.74%	1.48%	2.22%	1.48%	2.22%	2.22%	
Jurisprudence	Humanities	-0.93%	-0.93%	-0.93%	0.93%	-0.93%	0.00%	0.00%	0.00%	
Moral Scenarios	Humanities	0.56%	-0.11%	-0.56%	-0.56%	-2.23%	0.22%	-0.34%	-0.56%	
High School World History	Humanities	-0.42%	-0.42%	0.00%	0.42%	-0.42%	1.69%	0.84%	1.69%	
High School European History	Humanities	0.61%	1.21%	0.61%	0.61%	1.21%	1.21%	0.61%	1.82%	
Philosophy	Humanities	-1.61%	-1.29%	-1.29%	-2.89%	-1.61%	-0.32%	-1.29%	-0.96%	
Moral Disputes	Humanities	-1.16%	0.58%	0.29%	-1.45%	-0.29%	0.29%	0.00%	0.00%	
High School Us History	Humanities	-0.98%	-0.98%	0.49%	-0.98%	0.49%	-1.47%	0.49%	-0.98%	
Professional Law	Humanities	-0.52%	0.39%	0.00%	-0.20%	0.52%	0.33%	0.46%	0.78%	
International Law	Humanities	0.00%	-0.83%	0.00%	-0.83%	0.00%	0.00%	0.00%	0.00%	
Logical Fallacies	Humanities	0.61%	0.00%	0.61%	0.00%	1.84%	0.00%	0.61%	0.61%	
Professional Medicine	Other	0.37%	0.37%	0.74%	-0.37%	0.37%	1.47%	0.74%	0.37%	
Virology	Other	-0.60%	0.60%	-3.61%	-2.41%	0.00%	-1.20%	-1.81%	-1.20%	
Prehistory	Other	-1.54%	-0.31%	-0.93%	-1.54%	-1.54%	-0.31%	-0.62%	-0.31%	
Miscellaneous	Other	0.00%	0.51%	0.00%	0.51%	0.64%	-0.13%	0.26%	0.26%	
Business Ethics	Other	1.00%	0.00%	-2.00%	2.00%	-1.00%	2.00%	0.00%	0.00%	
Management	Other	-0.97%	0.00%	-0.97%	-0.97%	0.00%	-0.97%	0.00%	0.00%	
Nutrition	Other	-2.29%	-1.96%	-0.98%	-2.29%	-3.27%	-1.63%	-1.96%	-1.63%	
Medical Genetics	Other	1.00%	0.00%	0.00%	0.00%	1.00%	0.00%	0.00%	0.00%	
Econometrics	Other	-1.75%	-1.75%	0.00%	1.75%	0.88%	0.88%	-0.88%	0.00%	
Marketing	Other	-0.85%	0.00%	-0.85%	-0.85%	0.00%	0.00%	0.00%	0.00%	
College Medicine	Other	-0.58%	-1.73%	-0.58%	0.00%	-0.58%	-2.89%	-1.16%	-1.16%	
Professional Accounting	Other	-1.06%	-1.77%	-0.71%	-1.06%	-1.06%	-1.77%	-0.35%	-0.35%	
Sociology	Social Sciences	0.00%	-1.00%	0.00%	0.00%	-1.49%	-0.50%	-1.00%	-1.00%	
Clinical Knowledge	Social Sciences	1.51%	1.89%	1.89%	1.51%	2.26%	1.89%	1.89%	3.02%	
Public Relations	Social Sciences	1.82%	0.00%	2.73%	0.00%	-0.91%	1.82%	0.91%	0.00%	
Human Aging	Social Sciences	-0.90%	0.45%	-0.45%	-0.90%	0.00%	0.45%	1.35%	0.00%	
Human Sexuality	Social Sciences	-1.53%	0.00%	0.00%	-2.29%	0.00%	-0.76%	0.00%	0.00%	
Us Foreign Policy	Social Sciences	1.00%	0.00%	-1.00%	0.00%	0.00%	0.00%	0.00%	0.00%	
High School Macroeconomics	Social Sciences	-0.51%	-0.51%	0.51%	1.03%	0.00%	0.00%	-0.51%	0.77%	
High School Government And Politics	Social Sciences	0.52%	0.00%	0.00%	0.00%	0.52%	0.52%	1.04%	1.04%	
Security Studies	Social Sciences	0.41%	-0.41%	-0.41%	-0.41%	-0.82%	-0.41%	0.00%	-0.41%	
Global Facts	Social Sciences	-1.00%	-5.00%	-2.00%	1.00%	-4.00%	5.00%	0.00%	-2.00%	
Professional Psychology	Social Sciences	-1.80%	-0.49%	0.00%	-0.65%	0.00%	-0.49%	-0.16%	0.16%	
High School Geography	Social Sciences	0.00%	-1.01%	-0.51%	-0.51%	-1.01%	-1.52%	-0.51%	0.00%	
High School Microeconomics	Social Sciences	0.00%	1.26%	0.84%	0.42%	0.84%	0.42%	0.84%	0.84%	
World Religions	Social Sciences	-0.58%	0.00%	0.58%	0.58%	0.58%	0.58%	0.00%	0.00%	
High School Psychology	Social Sciences	0.00%	0.18%	0.00%	-0.73%	0.00%	-0.37%	0.37%	0.37%	

Figure 8: The task-level accuracy changes on eight rule-generated datasets (4K samples) compared to fine-tuning with D_{seed}.

Task	Subject	Seed	Analogy	Anthropomorphic Expressions	Blurring the Conceptual Boundaries	Counterintuitive Thinking	Exaggerating Everyday Phenomena	Interdisciplinary Knowledge Integration	Philosophical Thinking	Social Phenomena, Pun, or Buzzwords
Conceptual Physics	STEM	61.70	62.13	60.85	61.70	62.98	60.85	61.70	60.43	60.85
High School Statistics	STEM	60.65	61.57	62.96	62.04	61.11	60.65	61.57	62.04	62.04
College Biology	STEM	79.86	81.94	82.64	81.94	81.94	80.56	82.64	80.56	80.56
College Chemistry	STEM	51.00	52.00	50.00	49.00	48.00	51.00	49.00	52.00	52.00
Elementary Mathematics	STEM	46.30	46.56	46.56	46.03	46.56	46.83	46.03	46.30	46.03
Electrical Engineering	STEM	62.07	64.14	64.14	64.14	62.07	63.45	62.76	64.83	64.14
Formal Logic	STEM	57.14	57.14	56.35	57.14	55.56	53.97	56.35	54.76	56.35
Abstract Algebra	STEM	27.00	31.00	28.00	27.00	32.00	28.00	30.00	27.00	27.00
High School Computer Science	STEM	74.00	71.00	75.00	72.00	72.00	73.00	75.00	74.00	73.00
Computer Security	STEM	77.00	75.00	78.00	77.00	76.00	75.00	75.00	76.00	77.00
High School Chemistry	STEM	55.67	55.17	53.69	55.67	53.20	55.17	55.17	55.67	54.68
College Computer Science	STEM	52.00	51.00	53.00	51.00	51.00	55.00	50.00	52.00	53.00
High School Biology	STEM	80.00	80.00	79.68	79.68	79.35	80.00	79.68	79.35	79.03
High School Mathematics	STEM	37.41	36.30	35.56	36.30	35.56	37.78	36.30	37.78	37.78
College Physics	STEM	50.00	50.98	48.04	48.04	47.06	46.08	50.00	48.04	48.04
Astronomy	STEM	75.66	73.68	74.34	72.37	73.68	74.34	72.37	73.68	73.68
High School Physics	STEM	39.74	39.07	38.41	39.74	40.40	42.38	38.41	39.07	41.06
College Mathematics	STEM	39.00	39.00	39.00	38.00	39.00	39.00	38.00	39.00	39.00
Machine Learning	STEM	47.32	45.54	50.00	49.11	50.00	50.00	50.00	47.32	48.21
Anatomy	STEM	62.96	62.96	65.93	63.70	64.44	65.19	64.44	65.19	65.19
Jurisprudence	Humanities	76.85	75.93	75.93	75.93	77.78	75.93	76.85	76.85	76.85
Moral Scenarios	Humanities	49.72	50.28	49.61	49.16	49.16	47.49	49.94	49.39	49.16
High School World History	Humanities	83.97	83.54	83.54	83.97	84.39	83.54	85.65	84.81	85.65
High School European History	Humanities	75.76	76.36	76.97	76.36	76.36	76.97	76.97	76.36	77.58
Philosophy	Humanities	75.24	73.63	73.95	73.95	72.35	73.63	74.92	73.95	74.28
Moral Disputes	Humanities	70.23	69.08	70.81	70.52	68.79	69.94	70.52	70.23	70.23
High School Us History	Humanities	86.76	85.78	85.78	87.25	85.78	87.25	85.29	87.25	85.78
Professional Law	Humanities	46.22	45.70	46.61	46.22	46.02	46.74	46.54	46.68	47.00
International Law	Humanities	80.99	80.99	80.17	80.99	80.17	80.99	80.99	80.99	80.99
Logical Fallacies	Humanities	74.85	75.46	74.85	75.46	74.85	76.69	74.85	75.46	75.46
Professional Medicine	Other	74.63	75.00	75.00	75.37	74.26	75.00	76.10	75.37	75.00
Virology	Other	51.20	50.60	51.81	47.59	48.80	51.20	50.00	49.40	50.00
Prehistory	Other	71.60	70.06	71.30	70.68	70.06	70.06	71.30	70.99	71.30
Miscellaneous	Other	79.69	79.69	80.20	79.69	80.20	80.33	79.57	79.95	79.95
Business Ethics	Other	71.00	72.00	71.00	69.00	73.00	70.00	73.00	71.00	71.00
Management	Other	80.58	79.61	80.58	79.61	79.61	80.58	79.61	80.58	80.58
Nutrition	Other	79.41	77.12	77.45	78.43	77.12	76.14	77.78	77.45	77.78
Medical Genetics	Other	79.00	80.00	79.00	79.00	79.00	80.00	79.00	79.00	79.00
Econometrics	Other	51.75	50.00	50.00	51.75	53.51	52.63	52.63	50.88	51.75
Marketing	Other	91.45	90.60	91.45	90.60	90.60	91.45	91.45	91.45	91.45
College Medicine	Other	67.05	66.47	65.32	66.47	67.05	66.47	64.16	65.90	65.90
Professional Accounting	Other	50.00	48.94	48.23	49.29	48.94	48.94	48.23	49.65	49.65
Sociology	Social Sciences	86.07	86.07	85.07	86.07	86.07	84.58	85.57	85.07	85.07
Clinical Knowledge	Social Sciences	73.96	75.47	75.85	75.85	75.47	76.23	75.85	75.85	76.98
Public Relations	Social Sciences	68.18	70.00	68.18	70.91	68.18	67.27	70.00	69.09	68.18
Human Aging	Social Sciences	73.09	72.20	73.54	72.65	72.20	73.09	73.54	74.44	73.09
Human Sexuality	Social Sciences	80.15	78.63	80.15	80.15	77.86	80.15	79.39	80.15	80.15
Us Foreign Policy	Social Sciences	86.00	87.00	86.00	85.00	86.00	86.00	86.00	86.00	86.00
High School Macroeconomics	Social Sciences	66.15	65.64	65.64	66.67	67.18	66.15	66.15	65.64	66.92
High School Government And Politics	Social Sciences	90.16	90.67	90.16	90.16	90.16	90.67	90.67	91.19	91.19
Security Studies	Social Sciences	76.33	76.73	75.92	75.92	75.92	75.51	75.92	76.33	75.92
Global Facts	Social Sciences	48.00	47.00	43.00	46.00	49.00	44.00	53.00	48.00	46.00
Professional Psychology	Social Sciences	69.44	67.65	68.95	69.44	68.79	69.44	68.95	69.28	69.61
High School Geography	Social Sciences	83.84	83.84	82.83	83.33	83.33	82.83	82.32	83.33	83.84
High School Microeconomics	Social Sciences	76.47	76.47	77.73	77.31	76.89	77.31	76.89	77.31	77.31
World Religions	Social Sciences	78.95	78.36	78.95	79.53	79.53	79.53	79.53	78.95	78.95
High School Psychology	Social Sciences	85.69	85.69	85.87	85.69	84.95	85.69	85.32	86.06	86.06

Figure 9: The task-level accuracy on eight rule-generated datasets (4K samples) compared to fine-tuning with D_{seed}.

Task	Subject	Max Perplexity	Max Perplexity (w/o Seed)	Min Perplexity	Min Perplexity (w/o seed)	Summary
Conceptual Physics	STEM	-0.43%	-0.43%	0.00%	-0.43%	
High School Statistics	STEM	-1.85%	-0.93%	0.93%	1.39%	
College Biology	STEM	0.69%	0.69%	0.00%	-0.69%	
College Chemistry	STEM	-2.00%	-2.00%	0.00%	0.00%	
Elementary Mathematics	STEM	-0.79%	0.00%	-1.06%	-1.59%	
Electrical Engineering	STEM	-1.38%	-1.38%	2.07%	0.00%	
Formal Logic	STEM	-1.59%	-1.59%	-2.38%	-0.79%	
Abstract Algebra	STEM	0.00%	0.00%	2.00%	3.00%	
High School Computer Science	STEM	1.00%	2.00%	1.00%	1.00%	
Computer Security	STEM	1.00%	1.00%	1.00%	0.00%	
High School Chemistry	STEM	-1.97%	-1.97%	-0.99%	-2.46%	
College Computer Science	STEM	-1.00%	-2.00%	-2.00%	-4.00%	
High School Biology	STEM	-0.32%	0.00%	-0.32%	-0.65%	
High School Mathematics	STEM	-1.11%	-0.37%	0.00%	0.00%	
College Physics	STEM	-0.98%	0.98%	1.96%	1.96%	
Astronomy	STEM	0.00%	0.66%	-0.66%	0.00%	
High School Physics	STEM	0.66%	0.66%	1.32%	0.66%	
College Mathematics	STEM	-2.00%	-2.00%	-1.00%	-1.00%	
Machine Learning	STEM	1.79%	0.00%	0.00%	0.00%	
Anatomy	STEM	-0.74%	0.00%	1.48%	2.22%	
Jurisprudence	Humanities	0.00%	-0.93%	0.93%	0.00%	
Moral Scenarios	Humanities	-0.56%	-0.45%	0.67%	1.34%	
High School World History	Humanities	0.42%	0.42%	-1.27%	-0.84%	
High School European History	Humanities	1.21%	0.61%	-0.61%	0.00%	
Philosophy	Humanities	-1.93%	-1.93%	-0.64%	-0.64%	
Moral Disputes	Humanities	-0.29%	-0.58%	0.87%	1.16%	
High School Us History	Humanities	0.00%	0.00%	-1.96%	-1.47%	
Professional Law	Humanities	-0.46%	-0.26%	-0.33%	-0.13%	
International Law	Humanities	0.00%	0.00%	1.65%	2.48%	
Logical Fallacies	Humanities	-1.23%	-1.23%	-0.61%	-0.61%	
Professional Medicine	Other	-1.84%	-0.74%	-1.84%	-1.10%	
Virology	Other	0.00%	0.00%	-0.60%	-0.60%	
Prehistory	Other	-0.93%	-0.31%	-0.93%	-0.93%	
Miscellaneous	Other	0.38%	0.26%	0.38%	0.38%	
Business Ethics	Other	-1.00%	0.00%	0.00%	-1.00%	
Management	Other	-0.97%	-1.94%	-1.94%	-1.94%	
Nutrition	Other	-2.29%	-2.29%	-1.31%	-1.63%	
Medical Genetics	Other	-1.00%	0.00%	0.00%	0.00%	
Econometrics	Other	0.88%	-1.75%	-0.88%	-0.88%	
Marketing	Other	0.85%	0.85%	0.00%	-0.43%	
College Medicine	Other	-2.31%	-1.73%	0.58%	-0.58%	
Professional Accounting	Other	-3.55%	-2.13%	-1.06%	-1.42%	
Sociology	Social Sciences	-2.49%	-2.49%	-1.00%	0.00%	
Clinical Knowledge	Social Sciences	0.00%	0.38%	0.75%	0.75%	
Public Relations	Social Sciences	0.91%	0.91%	-0.91%	-0.91%	
Human Aging	Social Sciences	-0.45%	0.00%	-0.45%	-0.45%	
Human Sexuality	Social Sciences	0.76%	0.76%	0.00%	0.00%	
Us Foreign Policy	Social Sciences	1.00%	1.00%	0.00%	1.00%	
High School Macroeconomics	Social Sciences	0.51%	-0.51%	0.51%	0.51%	
High School Government And Politics	Social Sciences	0.00%	0.00%	1.04%	1.04%	
Security Studies	Social Sciences	-1.63%	-2.45%	-0.41%	-1.22%	
Global Facts	Social Sciences	4.00%	2.00%	0.00%	-1.00%	
Professional Psychology	Social Sciences	-0.98%	-0.98%	-0.82%	-0.16%	
High School Geography	Social Sciences	-1.52%	-2.53%	-2.02%	-2.53%	
High School Microeconomics	Social Sciences	1.26%	1.68%	0.00%	0.00%	
World Religions	Social Sciences	-1.17%	-1.75%	-0.58%	-1.17%	
High School Psychology	Social Sciences	0.18%	0.37%	0.18%	0.37%	

Figure 10: The task-level accuracy changes on four filtering and mixing strategies-generated datasets (13K samples) compared to fine-tuning with D_{seed}.

Task	Subject	Seed	Max Perplexity	Max Perplexity (w/o Seed)	Min Perplexity	Min Perplexity (w/o seed)
Conceptual Physics	STEM	60.43	60.00	60.00	60.43	60.00
High School Statistics	STEM	62.50	60.65	61.57	63.43	63.89
College Biology	STEM	81.25	81.94	81.94	81.25	80.56
College Chemistry	STEM	50.00	48.00	48.00	50.00	50.00
Elementary Mathematics	STEM	46.30	45.50	46.30	45.24	44.71
Electrical Engineering	STEM	62.76	61.38	61.38	64.83	62.76
Formal Logic	STEM	56.35	54.76	54.76	53.97	55.56
Abstract Algebra	STEM	28.00	28.00	28.00	30.00	31.00
High School Computer Science	STEM	73.00	74.00	75.00	74.00	74.00
Computer Security	STEM	75.00	76.00	76.00	76.00	75.00
High School Chemistry	STEM	56.65	54.68	54.68	55.67	54.19
College Computer Science	STEM	53.00	52.00	51.00	51.00	49.00
High School Biology	STEM	79.68	79.35	79.68	79.35	79.03
High School Mathematics	STEM	37.04	35.93	36.67	37.04	37.04
College Physics	STEM	47.06	46.08	48.04	49.02	49.02
Astronomy	STEM	73.68	73.68	74.34	73.03	73.68
High School Physics	STEM	39.07	39.74	39.74	40.40	39.74
College Mathematics	STEM	39.00	37.00	37.00	38.00	38.00
Machine Learning	STEM	49.11	50.89	49.11	49.11	49.11
Anatomy	STEM	62.22	61.48	62.22	63.70	64.44
Jurisprudence	Humanities	75.93	75.93	75.00	76.85	75.93
Moral Scenarios	Humanities	48.49	47.93	48.04	49.16	49.83
High School World History	Humanities	84.81	85.23	85.23	83.54	83.97
High School European History	Humanities	76.36	77.58	76.97	75.76	76.36
Philosophy	Humanities	74.92	72.99	72.99	74.28	74.28
Moral Disputes	Humanities	69.65	69.36	69.08	70.52	70.81
High School Us History	Humanities	87.25	87.25	87.25	85.29	85.78
Professional Law	Humanities	47.33	46.87	47.07	47.00	47.20
International Law	Humanities	78.51	78.51	78.51	80.17	80.99
Logical Fallacies	Humanities	75.46	74.23	74.23	74.85	74.85
Professional Medicine	Other	76.84	75.00	76.10	75.00	75.74
Virology	Other	48.80	48.80	48.80	48.19	48.19
Prehistory	Other	70.99	70.06	70.68	70.06	70.06
Miscellaneous	Other	79.05	79.44	79.31	79.44	79.44
Business Ethics	Other	71.00	70.00	71.00	71.00	70.00
Management	Other	82.52	81.55	80.58	80.58	80.58
Nutrition	Other	79.41	77.12	77.12	78.10	77.78
Medical Genetics	Other	79.00	78.00	79.00	79.00	79.00
Econometrics	Other	50.00	50.88	48.25	49.12	49.12
Marketing	Other	90.60	91.45	91.45	90.60	90.17
College Medicine	Other	66.47	64.16	64.74	67.05	65.90
Professional Accounting	Other	50.35	46.81	48.23	49.29	48.94
Sociology	Social Sciences	88.06	85.57	85.57	87.06	88.06
Clinical Knowledge	Social Sciences	75.85	75.85	76.23	76.60	76.60
Public Relations	Social Sciences	70.00	70.91	70.91	69.09	69.09
Human Aging	Social Sciences	73.09	72.65	73.09	72.65	72.65
Human Sexuality	Social Sciences	79.39	80.15	80.15	79.39	79.39
Us Foreign Policy	Social Sciences	86.00	87.00	87.00	86.00	87.00
High School Macroeconomics	Social Sciences	66.15	66.67	65.64	66.67	66.67
High School Government And Politics	Social Sciences	89.64	89.64	89.64	90.67	90.67
Security Studies	Social Sciences	77.96	76.33	75.51	77.55	76.73
Global Facts	Social Sciences	46.00	50.00	48.00	46.00	45.00
Professional Psychology	Social Sciences	69.93	68.95	68.95	69.12	69.77
High School Geography	Social Sciences	84.85	83.33	82.32	82.83	82.32
High School Microeconomics	Social Sciences	76.05	77.31	77.73	76.05	76.05
World Religions	Social Sciences	80.12	78.95	78.36	79.53	78.95
High School Psychology	Social Sciences	85.32	85.50	85.69	85.50	85.69

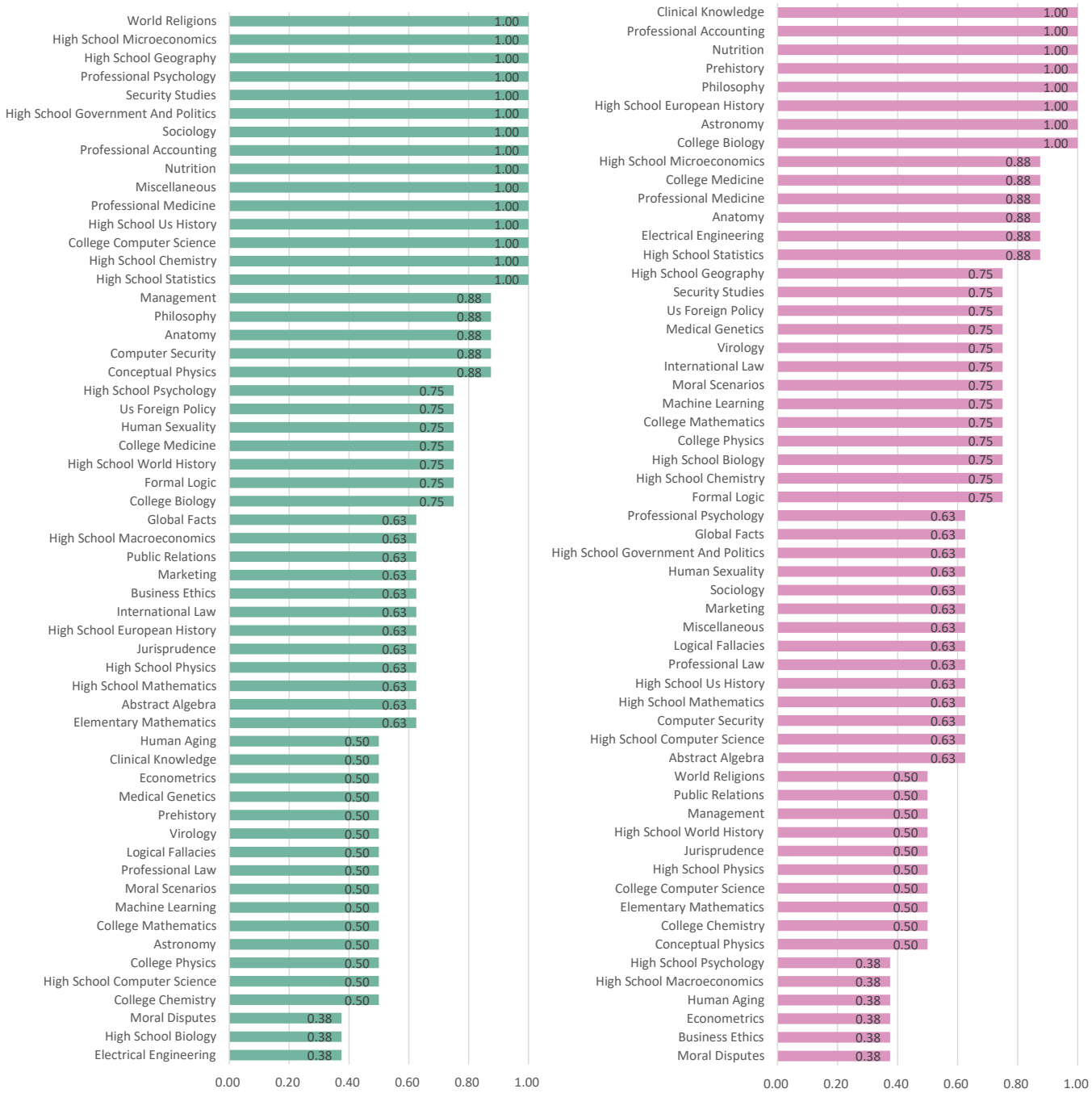
Figure 11: The task-level accuracy changes on four filtering and mixing strategies-generated datasets (13K samples) compared to fine-tuning with D_{seed}.

Task	Subject	Alignment	Consistency	Correctness	Max Perplexity	Max Perplexity (w/o Seed)	Min Perplexity	Min Perplexity (w/o seed)	Summary
Conceptual Physics	STEM	-0.43%	-0.43%	-0.43%	0.00%	1.28%	-0.85%	-0.85%	
High School Statistics	STEM	-0.46%	0.46%	1.39%	0.93%	1.39%	1.39%	0.93%	
College Biology	STEM	2.08%	0.69%	2.08%	0.69%	0.69%	2.08%	2.08%	
College Chemistry	STEM	-1.00%	0.00%	1.00%	-2.00%	1.00%	0.00%	-1.00%	
Elementary Mathematics	STEM	-0.26%	0.53%	-0.79%	1.06%	-0.26%	-1.59%	-1.06%	
Electrical Engineering	STEM	1.38%	1.38%	0.00%	0.69%	1.38%	1.38%	1.38%	
Formal Logic	STEM	-1.59%	-1.59%	-2.38%	-0.79%	-0.79%	-1.59%	-1.59%	
Abstract Algebra	STEM	0.00%	3.00%	2.00%	3.00%	2.00%	2.00%	3.00%	
High School Computer Science	STEM	-1.00%	0.00%	1.00%	1.00%	0.00%	1.00%	1.00%	
Computer Security	STEM	-1.00%	0.00%	0.00%	-1.00%	-1.00%	-1.00%	-1.00%	
High School Chemistry	STEM	-0.49%	-0.49%	-0.99%	-0.99%	-0.99%	-0.49%	-0.49%	
College Computer Science	STEM	0.00%	0.00%	0.00%	-1.00%	0.00%	0.00%	1.00%	
High School Biology	STEM	-0.32%	-0.65%	0.65%	-0.32%	-0.65%	-0.32%	-0.32%	
High School Mathematics	STEM	-0.74%	-1.85%	-1.85%	-1.11%	-1.11%	0.00%	-0.37%	
College Physics	STEM	0.00%	-1.96%	-1.96%	-2.94%	-2.94%	0.00%	-0.98%	
Astronomy	STEM	-1.32%	-2.63%	-2.63%	-1.97%	-1.97%	-3.29%	-3.29%	
High School Physics	STEM	0.00%	-0.66%	-0.66%	-1.32%	-0.66%	1.99%	1.32%	
College Mathematics	STEM	1.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	
Machine Learning	STEM	0.89%	-1.79%	2.68%	0.89%	-0.89%	1.79%	1.79%	
Anatomy	STEM	1.48%	1.48%	0.00%	0.74%	0.74%	0.74%	1.48%	
Jurisprudence	Humanities	0.00%	-0.93%	-0.93%	0.00%	0.00%	0.00%	0.00%	
Moral Scenarios	Humanities	1.56%	-0.22%	0.34%	0.00%	0.00%	0.11%	0.22%	
High School World History	Humanities	0.84%	0.42%	1.27%	0.84%	0.84%	1.27%	1.27%	
High School European History	Humanities	0.61%	0.61%	0.61%	0.00%	0.00%	0.61%	0.61%	
Philosophy	Humanities	-1.93%	-0.96%	-1.93%	-1.61%	-1.93%	-1.29%	-1.29%	
Moral Disputes	Humanities	-0.87%	-1.16%	0.00%	-1.16%	-1.16%	-0.58%	-0.87%	
High School Us History	Humanities	-0.98%	0.98%	0.00%	-0.98%	-0.49%	0.49%	-0.49%	
Professional Law	Humanities	-0.07%	-0.07%	0.20%	-0.13%	-0.13%	0.39%	0.26%	
International Law	Humanities	0.00%	-0.83%	0.00%	-0.83%	-1.65%	0.83%	0.00%	
Logical Fallacies	Humanities	0.00%	1.84%	1.84%	1.23%	1.23%	0.61%	0.00%	
Professional Medicine	Other	0.74%	0.37%	0.00%	1.10%	0.74%	1.84%	1.10%	
Virology	Other	-3.61%	-3.01%	-2.41%	-1.81%	-1.81%	-1.81%	-1.20%	
Prehistory	Other	-2.16%	0.00%	-0.93%	-1.54%	-0.93%	-0.62%	-0.62%	
Miscellaneous	Other	0.64%	0.26%	0.26%	-0.51%	-0.38%	0.13%	-0.26%	
Business Ethics	Other	0.00%	0.00%	1.00%	0.00%	1.00%	0.00%	0.00%	
Management	Other	0.00%	0.00%	-0.97%	1.94%	1.94%	0.00%	-0.97%	
Nutrition	Other	-1.96%	-1.63%	-1.63%	-1.96%	-1.96%	-2.29%	-1.31%	
Medical Genetics	Other	0.00%	0.00%	1.00%	0.00%	0.00%	0.00%	0.00%	
Econometrics	Other	-1.75%	-3.51%	0.00%	-7.89%	-2.63%	3.51%	0.00%	
Marketing	Other	-0.43%	-0.43%	0.00%	-0.85%	-0.85%	-1.28%	0.00%	
College Medicine	Other	-1.16%	-1.16%	-2.31%	-0.58%	-1.16%	-2.31%	-2.31%	
Professional Accounting	Other	-2.13%	-1.06%	-1.06%	-2.13%	-1.42%	-0.35%	-0.35%	
Sociology	Social Sciences	-1.00%	-0.50%	-1.99%	0.00%	-1.00%	-0.50%	0.00%	
Clinical Knowledge	Social Sciences	1.13%	1.13%	1.89%	1.51%	1.13%	1.13%	1.13%	
Public Relations	Social Sciences	0.91%	0.00%	1.82%	0.91%	0.00%	2.73%	0.91%	
Human Aging	Social Sciences	0.00%	0.45%	0.45%	0.00%	0.00%	-0.90%	0.00%	
Human Sexuality	Social Sciences	0.00%	-0.76%	-0.76%	-0.76%	0.00%	-0.76%	-0.76%	
Us Foreign Policy	Social Sciences	1.00%	-1.00%	0.00%	1.00%	1.00%	0.00%	0.00%	
High School Macroeconomics	Social Sciences	0.51%	-0.77%	0.00%	0.00%	-0.51%	-0.51%	0.00%	
High School Government And Politics	Social Sciences	0.52%	-0.52%	-1.04%	-0.52%	0.00%	1.04%	0.00%	
Security Studies	Social Sciences	0.41%	-0.41%	-1.22%	0.00%	0.41%	0.00%	0.41%	
Global Facts	Social Sciences	-1.00%	-1.00%	0.00%	-1.00%	0.00%	-1.00%	0.00%	
Professional Psychology	Social Sciences	0.16%	0.33%	0.00%	-0.16%	-0.16%	0.16%	0.00%	
High School Geography	Social Sciences	-0.51%	-0.51%	0.00%	-0.51%	-1.01%	-0.51%	-1.01%	
High School Microeconomics	Social Sciences	1.26%	1.68%	1.26%	0.84%	0.84%	0.00%	0.00%	
World Religions	Social Sciences	0.00%	-0.58%	-0.58%	-0.58%	-0.58%	0.58%	0.58%	
High School Psychology	Social Sciences	0.18%	-0.18%	0.37%	-0.18%	0.37%	-0.18%	-0.37%	

Figure 12: The task-level accuracy changes on seven filtering and mixing strategies-generated datasets (4K samples) compared to fine-tuning with D_{seed}.

Task	Subject	Seed	Alignment	Consistency	Correctness	Max Perplexity	Max Perplexity (w/o Seed)	Min Perplexity	Min Perplexity (w/o seed)
Conceptual Physics	STEM	61.70	61.28	61.28	61.28	61.70	62.98	60.85	60.85
High School Statistics	STEM	60.65	60.19	61.11	62.04	61.57	62.04	62.04	61.57
College Biology	STEM	79.86	81.94	80.56	81.94	80.56	80.56	81.94	81.94
College Chemistry	STEM	51.00	50.00	51.00	52.00	49.00	52.00	51.00	50.00
Elementary Mathematics	STEM	46.30	46.03	46.83	45.50	47.35	46.03	44.71	45.24
Electrical Engineering	STEM	62.07	63.45	63.45	62.07	62.76	63.45	63.45	63.45
Formal Logic	STEM	57.14	55.56	55.56	54.76	56.35	56.35	55.56	55.56
Abstract Algebra	STEM	27.00	27.00	30.00	29.00	30.00	29.00	29.00	30.00
High School Computer Science	STEM	74.00	73.00	74.00	75.00	75.00	74.00	75.00	75.00
Computer Security	STEM	77.00	76.00	77.00	77.00	76.00	76.00	76.00	76.00
High School Chemistry	STEM	55.67	55.17	55.17	54.68	54.68	54.68	55.17	55.17
College Computer Science	STEM	52.00	52.00	52.00	52.00	51.00	52.00	52.00	53.00
High School Biology	STEM	80.00	79.68	79.35	80.65	79.68	79.35	79.68	79.68
High School Mathematics	STEM	37.41	36.67	35.56	35.56	36.30	36.30	37.41	37.04
College Physics	STEM	50.00	50.00	48.04	48.04	47.06	47.06	50.00	49.02
Astronomy	STEM	75.66	74.34	73.03	73.03	73.68	73.68	72.37	72.37
High School Physics	STEM	39.74	39.74	39.07	39.07	38.41	39.07	41.72	41.06
College Mathematics	STEM	39.00	40.00	39.00	39.00	39.00	39.00	39.00	39.00
Machine Learning	STEM	47.32	48.21	45.54	50.00	48.21	46.43	49.11	49.11
Anatomy	STEM	62.96	64.44	64.44	62.96	63.70	63.70	63.70	64.44
Jurisprudence	Humanities	76.85	76.85	75.93	75.93	76.85	76.85	76.85	76.85
Moral Scenarios	Humanities	49.72	51.28	49.50	50.06	49.72	49.72	49.83	49.94
High School World History	Humanities	83.97	84.81	84.39	85.23	84.81	84.81	85.23	85.23
High School European History	Humanities	75.76	76.36	76.36	76.36	75.76	75.76	76.36	76.36
Philosophy	Humanities	75.24	73.31	74.28	73.31	73.63	73.31	73.96	73.96
Moral Disputes	Humanities	70.23	69.36	69.08	70.23	69.08	69.08	69.65	69.36
High School Us History	Humanities	86.76	85.78	87.75	86.76	85.78	86.27	87.25	86.27
Professional Law	Humanities	46.22	46.15	46.15	46.41	46.09	46.09	46.61	46.48
International Law	Humanities	80.99	80.99	80.17	80.99	80.17	79.34	81.82	80.99
Logical Fallacies	Humanities	74.85	74.85	76.69	76.69	76.07	76.07	75.46	74.85
Professional Medicine	Other	74.63	75.37	75.00	74.63	75.74	75.37	76.47	75.74
Virology	Other	51.20	47.59	48.19	48.80	49.40	49.40	49.40	50.00
Prehistory	Other	71.60	69.44	71.60	70.68	70.06	70.68	70.99	70.99
Miscellaneous	Other	79.69	80.33	79.95	79.95	79.18	79.31	79.82	79.44
Business Ethics	Other	71.00	71.00	71.00	72.00	71.00	72.00	71.00	71.00
Management	Other	80.58	80.58	80.58	79.61	82.52	82.52	80.58	79.61
Nutrition	Other	79.41	77.45	77.78	77.78	77.45	77.45	77.12	78.10
Medical Genetics	Other	79.00	79.00	79.00	80.00	79.00	79.00	79.00	79.00
Econometrics	Other	51.75	50.00	48.25	51.75	43.86	49.12	55.26	51.75
Marketing	Other	91.45	91.03	91.03	91.45	90.60	90.60	90.17	91.45
College Medicine	Other	67.05	65.90	65.90	64.74	66.47	65.90	64.74	64.74
Professional Accounting	Other	50.00	47.87	48.94	48.94	47.87	48.58	49.65	49.65
Sociology	Social Sciences	86.07	85.07	85.57	84.08	86.07	85.07	85.57	86.07
Clinical Knowledge	Social Sciences	73.96	75.09	75.09	75.85	75.47	75.09	75.09	75.09
Public Relations	Social Sciences	68.18	69.09	68.18	70.00	69.09	68.18	70.91	69.09
Human Aging	Social Sciences	73.09	73.09	73.54	73.54	73.09	73.09	72.20	73.09
Human Sexuality	Social Sciences	80.15	80.15	79.39	79.39	79.39	80.15	79.39	79.39
Us Foreign Policy	Social Sciences	86.00	87.00	85.00	86.00	87.00	87.00	86.00	86.00
High School Macroeconomics	Social Sciences	66.15	66.67	65.38	66.15	66.15	65.64	65.64	66.15
High School Government And Politics	Social Sciences	90.16	90.67	89.64	89.12	89.64	90.16	91.19	90.16
Security Studies	Social Sciences	76.33	76.73	75.92	75.10	76.33	76.73	76.33	76.73
Global Facts	Social Sciences	48.00	47.00	47.00	48.00	47.00	48.00	47.00	48.00
Professional Psychology	Social Sciences	69.44	69.61	69.77	69.44	69.28	69.28	69.61	69.44
High School Geography	Social Sciences	83.84	83.33	83.33	83.84	83.33	82.83	83.33	82.83
High School Microeconomics	Social Sciences	76.47	77.73	78.15	77.73	77.31	77.31	76.47	76.47
World Religions	Social Sciences	78.95	78.95	78.36	78.36	78.36	78.36	79.53	79.53
High School Psychology	Social Sciences	85.69	85.87	85.50	86.06	85.50	86.06	85.50	85.32

Figure 13: The task-level accuracy on seven filtering and mixing strategies-generated datasets (4K samples) compared to fine-tuning with D_{seed}.



(a) 13K samples

(b) 4K samples

Figure 14: Comparison of task-level percentage agreement for datasets generated by the eight rules, considering whether SFT results improved, declined, or remained unchanged relative to the model fine-tuned with D_{seed}.