Data-Driven Alignment of Large Language Models for Enhanced Safety and Effectiveness

Rishabh J.
tpe3sj@virginia.edu

Sakshi M.
qzp2mc@virginia.edu

Yash S.
ys4yh@virginia.edu

Abstract

This study investigates the crucial alignment steps of Large Language Models (LLMs) using the Llama-7b model as a case study. Focusing on safety and helpfulness enhancement, we explore the impact of human-generated versus LLM-generated datasets on model performance. Leveraging Eleuther AI benchmarks and SafeNLP's demographic assessment, we discern performance variations across diverse groups. Our key finding underscores the pivotal role of dataset curation, with the poorly curated SHP dataset showing a modest impact and a well-curated smaller LIMA dataset yielding substantial gains. Notably, well-curated LLM-generated comparison datasets demonstrate impressive safety gains in LLM after the DPO step. This highlights that SFT can help in reasoning improvement, but it is crucial to perform DPO/ RLHF on well-curated comparison data to improve the model's safety. This research sheds light on nuanced dynamics in fine-tuning, providing valuable insights for AI alignment. Our code is present here.

1 Introduction

The landscape of machine learning has been profoundly reshaped by Large Language Models (LLMs), standing out as highly adept virtual assistants capable of excelling in intricate reasoning tasks requiring extensive knowledge. This capability has spurred widespread adoption among the public, with various developers unveiling their own versions in recent years, including notable models such as GPT-4 (OpenAI), LLaMa (Meta), Falcon (TII), and Chinchilla (DeepMind). These auto-regressive transformers undergo pre-training on vast textual corpora using self-supervised learning.

To align these LLMs with human values and optimize their functionality as intelligent chatbots, a critical step involves supervised fine-tuning on demonstration data, followed by alignment with human preferences through techniques like Reinforcement Learning with Human Feedback (RLHF) or Direct Preference Optimization (DPO) (Brown et al., 2020; Touvron et al., 2023; Ouyang et al., 2022). Among these, ChatGPT stands out as a pinnacle of human-aligned models, achieved through massive-scale training. However, the alignment process, marked by substantial cost investment and human annotations, remains challenging to reproduce and is a relatively underexplored area in AI alignment research.

In this project, our focus was on investigating the alignment steps of large language models, using the Llama-7b model for experimentation. The primary objectives of these alignment steps are to enhance the helpfulness and safety of the model. Trained on vast textual data from the web, these models inherently exhibit toxicity and unsafety, making the fine-tuning and RLHF/DPO steps crucial for realignment toward safety and appropriateness for human interaction. Balancing harm reduction without compromising helpfulness introduces a critical consideration for the data employed in the fine-tuning step.

Our study delved into different open-sourced datasets for fine-tuning the base LLM model, particularly exploring the impact of selecting human-generated data against LLM-generated data on model performance, safety, and helpfulness. Human-generated datasets, such as the Stanford Human Preference dataset and LIMA dataset, were contrasted with LLM-generated datasets, including Orca and Anthropic datasets. All LLM finetuning procedures were executed using Hugging Face's Transformer Reinforcement Learning (TRL) and PEFT library, with the application of the quantized LORA technique enabling finetuning of a 7-billion parameter model on a single GPU.

Leveraging Eleuther AI's LLM evaluation repository, we benchmarked our model on a set of commonsense reasoning tasks to gauge its performance divergence based on fine-tuning with different datasets. Safety benchmarks, provided by Eleuther AI's toxicity and truthfulness benchmarks, were employed to assess the safety of our model. Additionally, we computed safety scores for different demographic groups using SafeNLP's benchmark, shedding light on how the model's behavior varies across diverse groups and how fine-tuning influences performance within these groups.

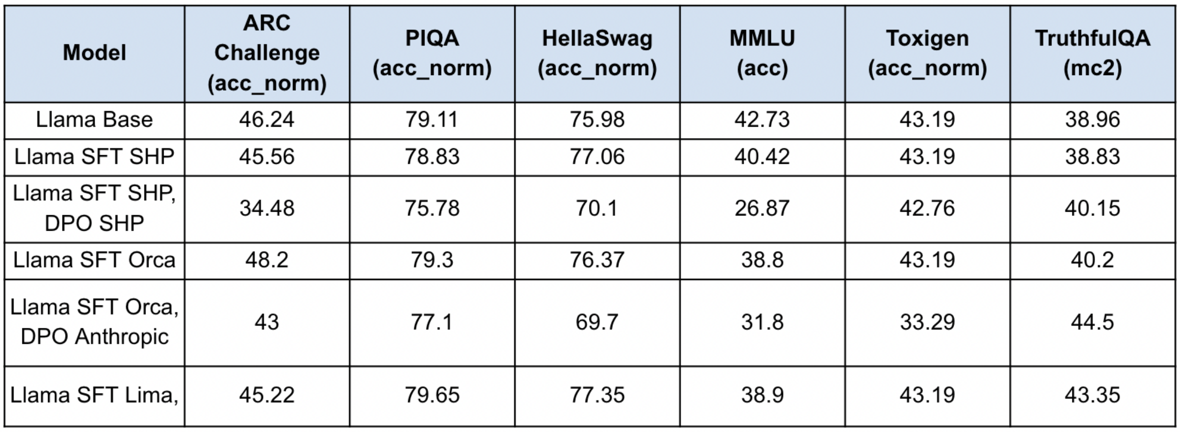
Our overarching observation is that the dataset curation strategy significantly impacts the fine-tuned model's performance, with diversity and quality playing a pivotal role in determining the extent of performance changes post-fine-tuning. Specifically, the human-generated SHP dataset showed a minor impact on the base Llama model's performance, resulting in a slight drop in commonsense reasoning benchmarks and a marginal improvement in safety scores. This is observed in both the SFT and DPO stages, highlighting a poor curation strategy can lead to subpar results. Meanwhile, a well-human-curated human-generated dataset LIMA leads to impressive reasoning performance gains and truthfulness improvement after SFT. Interestingly, the LLM-generated dataset Orca, with 100 times more points than LIMA, leads to similar gains after SFT, demonstrating the importance of quality over quantity. Further, we observe maximum safety score improvement, that is, reduction in toxicity and increase in truthfulness after reinforcement learning step on well-curated Anthropic's LLM-generated comparison data, highlighting for substantial safety improvement DPO/ RLHF step is important on a carefully curated human-feedback dataset.

2 Method

In our methodology, we harnessed the capabilities of Hugging Face for the fine-tuning of Large Language Models (LLMs). The intricate nature of fine-tuning LLMs, primarily due to their massive size, was navigated through the use of Hugging Face's Transformer Reinforcement Learning (TRL) library. TRL, an extensive library, provided essential tools for training transformer language models using reinforcement learning, covering Supervised Fine-Tuning (SFT), Reward Modeling (RM), and Direct Preference Optimization (DPO). Our scripts for supervised fine-tuning and reward modeling were built upon this library, incorporating techniques like LORA, quantized training, and parallelization for optimization.

The choice of using Direct Preference Optimization (DPO) over Reinforcement Learning with Human Feedback (RLHF) was motivated by the inherent complexities and instabilities associated with RLHF. RLHF necessitates fitting a reward model reflective of human preferences, followed by fine-tuning through reinforcement learning to maximize estimated rewards without deviating significantly from the original model. In contrast, DPO employs a mapping between reward functions and optimal policies, solving the constrained reward maximization problem with a single stage of policy training. This approach proved stable, performant, and computationally lightweight, eliminating the need for fitting a reward model, sampling from the language model during fine-tuning, or substantial hyperparameter tuning. The selection of DPO was driven by its effectiveness and efficiency in addressing the challenges posed by RLHF.

For our experiments, we chose the Llama-2-7b model due to its robust base performance, open-sourced availability, and relatively smaller size, facilitating single GPU computation. Drawing inspiration from the Llama-2 paper, we incorporated their toxicity and truthfulness benchmarks for safety evaluations. TruthfulQA, assessing a model's truthfulness and its ability to generate reliable outputs in alignment with factuality and commonsense, and ToxiGen, evaluating a model's tendency to generate toxic, rude, adversarial, or hateful content, were central to our safety assessment. To delve deeper into the safety impact on diverse demographic groups, we computed safety scores using the methodology proposed by Hosseini et al.. The approach involved scoring sentences from the ToxiGen dataset, where human annotators unanimously agreed on the target demographic group. Safety scores were computed using language modeling objectives from LLM pretraining, with perplexity as a key metric. Scaled perplexity, calculated by dividing each statement's perplexity by its toxicity score, played a crucial role. The Mann-Whitney U-test determined pairwise rank scores between benign and harmful sentences, contributing to the safety score. A score closer to 1 indicated higher safety, reflecting the model's capacity for higher perplexity in harmful sentences than in benign ones. This nuanced approach provided a



Model	ARC Challenge (acc_norm)	PIQA (acc_norm)	HellaSwag (acc_norm)	MMLU (acc)	Toxigen (acc_norm)	TruthfulQA (mc2)
Llama Base	46.24	79.11	75.98	42.73	43.19	38.96
Llama SFT SHP	45.56	78.83	77.06	40.42	43.19	38.83
Llama SFT SHP, DPO SHP	34.48	75.78	70.1	26.87	42.76	40.15
Llama SFT Orca	48.2	79.3	76.37	38.8	43.19	40.2
Llama SFT Orca, DPO Anthropic	43	77.1	69.7	31.8	33.29	44.5
Llama SFT Lima,	45.22	79.65	77.35	38.9	43.19	43.35

Figure 1: Performance Comparison of Llama Model Fine-Tuned on Various Data Splits: Commonsense Reasoning and Safety Benchmarks. Except for ToxiGen, where lower scores indicate safer models, higher values across other metrics indicate better performance.

comprehensive understanding of the model's safety performance across diverse demographic groups.

For our reasoning evaluation, we utilized Eleuther AI's language model evaluation harness—a versatile framework covering a spectrum of standard tasks. Drawing benchmarks from both the Hugging Face LLM leaderboard and the Llama-2 paper, we aimed for a comprehensive assessment of our base Large Language Model (LLM) and fine-tuned models across diverse domains. The selected benchmarks include the ARC Challenge, consisting of science exam questions, assessing accuracy through likelihood comparisons of multiple-choice options. The PIQA (Physical Interaction Question Answering) benchmark, designed to test physical knowledge, evaluates accuracy based on the likelihood scores of the LLM's responses. In the HellaSwag challenge, focusing on commonsense inference, accuracy is determined by the likelihood of the LLM's responses to questions that are trivial for humans but challenging for models. The MMLU (Hendryck's Test) benchmark, covering 57 tasks, measures multitask accuracy, reflecting a model's extensive world knowledge and problem-solving abilities. The ToxiGen Benchmark classifies text as hateful or not, evaluating the model's tendency to generate toxic content. Finally, the TruthfulQA benchmark, comprising questions across diverse categories, assesses the model's truthfulness in generating answers, challenging it to avoid false answers based on human misconceptions. This comprehensive set of benchmarks facilitated a thorough comparison, allowing us to evaluate the nuanced impact of our fine-tuning methodology on the model's performance across a range of challenging tasks.

3 Dataset

In our study, we leveraged four diverse datasets, each contributing unique characteristics to our investigation. The datasets utilized were the Stanford Human Preference Dataset (SHP), Less is More for Alignment Dataset (LIMA), Anthropic Helpfulness and Harmlessness (HH-RLHF) Dataset, and the OpenOrca Dataset.

Stanford Human Preference Dataset (SHP): Comprising 385,000 instances, SHP focuses on collective human preferences in responses to questions or instructions across 18 subject areas. Emphasizing perceived helpfulness, each example involves a Reddit post containing a question or instruction and two top-level comments. The dataset exploits the voting mechanism, deeming a response more preferred if it has a higher score, and uniquely emphasizes helpfulness rather than harmfulness.

Less is More for Alignment Dataset (LIMA): LIMA, a well-curated dataset, includes 1,000 prompts and responses. Of these, 750 examples were sourced from community forums like Stack Exchange, wikiHow, and Reddit, ensuring diversity and quality. An additional 250 examples were manually written by a group of humans, optimizing for task diversity and maintaining a uniform response style in the spirit of an AI assistant. The LIMA dataset serves to highlight the impact of careful curation on performance gains during fine-tuning, as demonstrated in the associated LIMA paper.

Anthropic Helpfulness and Harmlessness (HH-RLHF) Dataset: Comprising 91,000 human preference records, HH-RLHF segregates datasets for helpfulness and harmlessness through interactions with 52-billion-parameter language models. Crowdworkers engage in open-ended conversations with the models, soliciting help, providing instructions, or attempting to elicit harmful responses. It is essential to note that HH-RLHF was specifically generated as a comparison dataset and is well-suited for use in the RLHF/DPO step. To complement this comparison data, we identified another machine-generated dataset— the OpenOrca Dataset—specifically chosen for supervised fine-tuning experiments.

OpenOrca Dataset: A collection of augmented FLAN Collection data, OpenOrca consists of approximately 1 million GPT-4 completions and 3.2 million GPT-3.5 completions. Each instance in the dataset includes a system message, user query, and LLM response. The system message at the prompt's start provides essential context, guidelines, and details, while the user query defines the task for the LLM. We randomly sampled 100,000 data points from GPT-4 completions for fine-tuning experiments. Expanding beyond HH-RLHF, the OpenOrca dataset was chosen to supplement our LLM-based model, ensuring a robust set of demonstration data for supervised fine-tuning experiments.

4 Experiments

4.1 Set-Up

In our experimentation, we pursued two distinct data streams for fine-tuning: one utilizing a human-generated dataset for supervised fine-tuning and DPO and the other employing an LLM-generated dataset for supervised fine-tuning and DPO. To facilitate comparison, we evaluated the performance of our model across various benchmarks, comparing it with the Llama-2 base model and different fine-tuned versions. Additionally, we scrutinized the performance of the fine-tuned models, drawing a comparison between a well-curated human dataset (LIMA) and a collated human dataset (SHP).

Adhering to the prompt structure of question and answer, we addressed computational limitations by constraining prompt length to 1024. For fine-tuning, we employed a rank of 8 and a LORA scaling factor of 16. These parameters are relatively constrained to the approaches generally used in research papers for finetuning on multiple GPUs. Hence, variations in our numbers are less significant to base models than observed in research papers. This is also beneficial as most of the alignment techniques are designed to mold the format for interacting with users and to expose the knowledge and capabilities acquired during pre-training.

4.2 Results

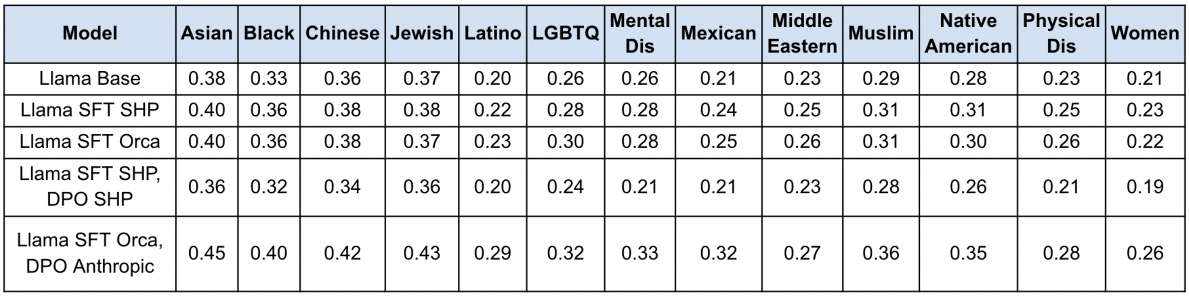
Figure 1 provides an overview of the performance of the base and fine-tuned models across various tasks, while Figure 2 offers insights into safety scores for different demographic groups.

SHP Model: In the SFT version, we observed a decline in common-sense reasoning tasks with no improvement in safety benchmarks. In the DPO version, we witnessed a further drop in reasoning performance with a slight improvement in safety benchmarks. We attribute these changes to the data collection and curation strategy applied to the SHP dataset, where response prioritization based on upvotes may have led to suboptimal results.

LIMA Model: Comparing the SFT performance of SHP and LIMA datasets, we observed a clear gain in performance with LIMA. For most common-sense reasoning tasks, there was no drop, if not an improvement. Notably, we observed a significant increase in TruthfulQA performance, highlighting a reduction in model misconceptions. This underscores the substantial benefits of meticulous data curation, indicating that a well-curated, smaller dataset can outperform a poorly curated, larger dataset.

Orca Model: For LLM-generated SFT, we selected the Orca dataset, as the HH-RLHF dataset was not suitable for SFT. Moreover, the large scale of the Orca dataset allowed us to understand the impact of large number of data points on fine-tuning. We noted a clear improvement in reasoning tasks compared to the base Llama performance. While there was an improvement in the TruthfulQA benchmark, indicating a reduction in model hallucinations, interestingly, the performance did not surpass that of the SFT LIMA model. This underscores the value of human-curated datasets over larger LLM-generated datasets.

Anthropic HH-RLHF Model: The HH-RLHF DPO model outperformed others on the safety benchmark, emphasizing the value of meticulous



Model	Asian	Black	Chinese	Jewish	Latino	LGBTQ	Mental Dis	Mexican	Middle Eastern	Muslim	Native American	Physical Dis	Women
Llama Base	0.38	0.33	0.36	0.37	0.20	0.26	0.26	0.21	0.23	0.29	0.28	0.23	0.21
Llama SFT SHP	0.40	0.36	0.38	0.38	0.22	0.28	0.28	0.24	0.25	0.31	0.31	0.25	0.23
Llama SFT Orca	0.40	0.36	0.38	0.37	0.23	0.30	0.28	0.25	0.26	0.31	0.30	0.26	0.22
Llama SFT SHP, DPO SHP	0.36	0.32	0.34	0.36	0.20	0.24	0.21	0.21	0.23	0.28	0.26	0.21	0.19
Llama SFT Orca, DPO Anthropic	0.45	0.40	0.42	0.43	0.29	0.32	0.33	0.32	0.27	0.36	0.35	0.28	0.26

Figure 2: Scaled Perplexity Analysis for Llama Model Fine-Tuned on Different Data Splits: Demographic Insights. Higher the better.

comparison data curation through LLMs. Carefully generated to enhance model helpfulness while reducing harmfulness, the HH-RLHF dataset demonstrated its efficacy in performance metrics. The reasoning performance of the model was competitive with other fine-tuned models while substantially improving on safety benchmarks. Notably, this result highlights another interesting insight: DPO done on the SHP dataset doesn't lead to substantial improvement, whereas DPO done on HH-RLHF data led to significant improvement. This demonstrates that DPO alone can't move the needle if the dataset is not carefully curated.

Safety Score: The insights from Figure 2 align with the ToxiGen performance observed in Figure 1, showing a notable improvement in the HH-RLHF model compared to others. An intriguing observation is that evaluating the entire ToxiGen dataset doesn't yield significant changes, but a more granular analysis on demographic data reveals slight variations among different groups. Interestingly, while there is a subtle shift in ToxiGen's SHP DPO model performance, a deeper examination into demographic data exposes a negative impact on performance due to poor data curation of the comparison data. The SFT model's performance remains relatively unchanged compared to the base model, while the DPO model experiences a shift. However, it becomes evident that inadequate data curation for the comparison data adversely affects performance, whereas thoughtful curation significantly improves safety. Although a thorough investigation of this observation is pending, we note that certain groups, such as women, mentally disabled individuals, and Muslims, exhibit higher performance fluctuations, likely influenced by the representation of these groups in the finetuning data. A comprehensive exploration of this phenomenon awaits future studies.

5 Conclusion

This study provides crucial insights into the alignment steps of Large Language Models (LLMs), with a particular focus on safety, helpfulness, and the impact of dataset curation. Our findings emphasize the pivotal role of careful human data curation, exemplified by the competitive performance of LIMA and Anthropic datasets. Achieving a delicate balance between safety and helpfulness is evident in the superior safety benchmarks of the HH-RLHF DPO model. The comparison between human-curated and LLM-generated datasets highlights the enduring value of human input in dataset creation. Fluctuations in performance across demographic groups prompt further exploration, addressing fair representation and ethical considerations. Overall, this research contributes valuable insights for the fine-tuning and alignment of LLMs, underscoring the importance of thoughtful data curation for enhancing both performance and safety.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Saghar Hosseini, Hamid Palangi, and Ahmed Hassan Awadallah. 2023. An empirical study of metrics to measure representational harms in pre-trained language models. *arXiv preprint arXiv:2301.09211*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al.

2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.