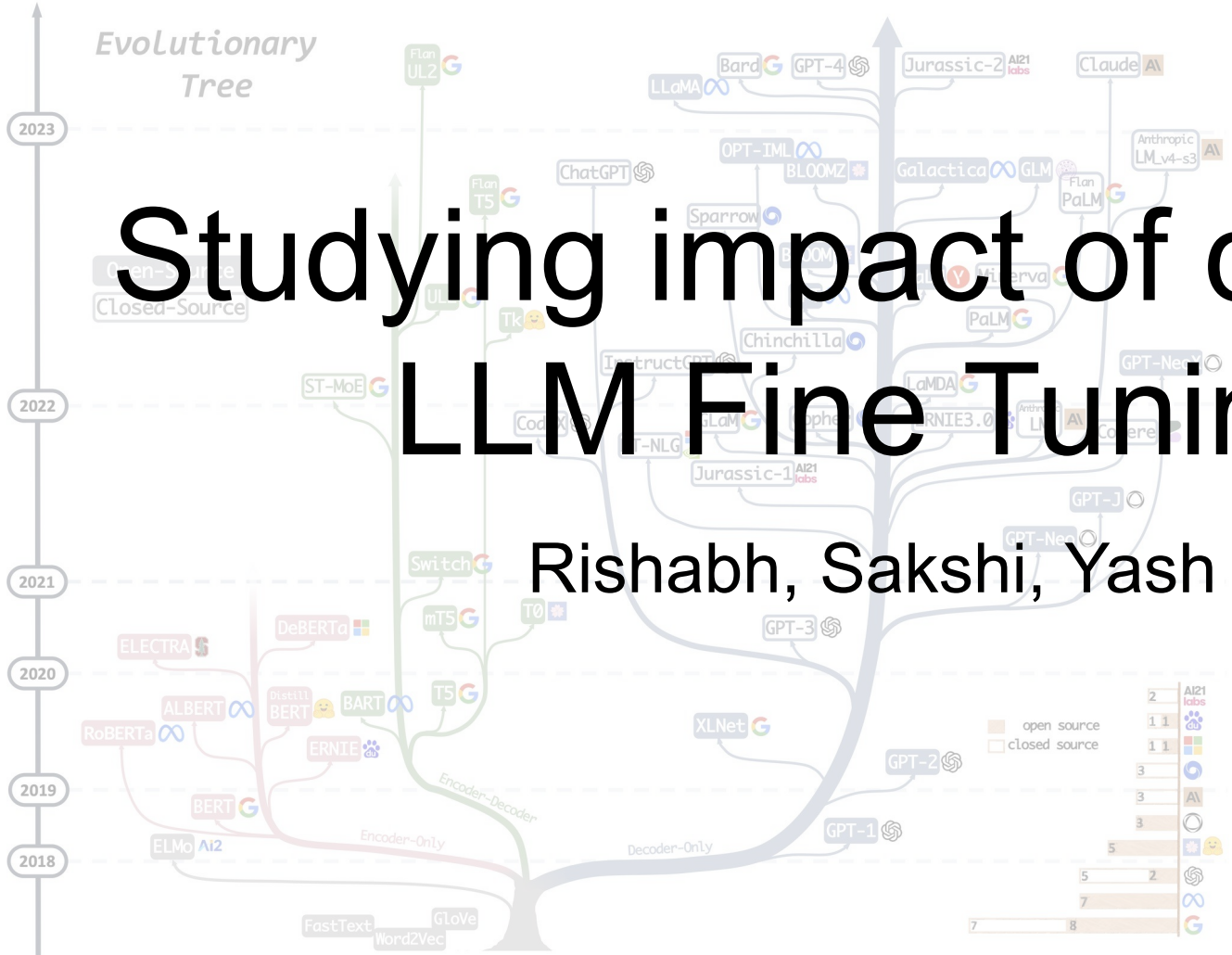

Studying impact of data on LLM Fine Tuning

Rishabh, Sakshi, Yash

Introduction

1. **Problem Definition:**
    - Language Models (LLMs) are **transforming** ML but pose alignment challenges.
    - **Alignment** in LLMs aim to enhance model safety and helpfulness, particularly addressing issues of toxicity and truthfulness.
    - **Challenge** lies in fine-tuning large pre-trained models for human interaction while maintaining safety.

2. **Alignment Challenges:**
    - LLMs trained on **vast textual data** may exhibit toxicity and unsafe behavior.
    - Alignment process involves **reconfiguring** models using demonstration and comparison data.

Motivation

1. **Widespread Adoption**
   - LLMs are witnessing **widespread adoption** due to their capability in complex reasoning tasks.
2. **Alignment Significance**
   - Alignment step **critical** for making LLMs **safe** & **appropriate** for human interaction.
   - We **aim** to understand & improve alignment process, addressing safety concerns.
3. **Cost and Reproducibility**
   - Alignment involves **significant cost** and **human** annotations.
   - Lack of reproducibility and limited studies hinder AI alignment research progress.

Objective: The project aims to investigate the influence of diverse demonstration and comparison datasets on the fine-tuning of Large Language Models (LLMs).

Methods

1. **Fine-tuning with Hugging Face**
   - Utilized Transformer Reinforcement Learning (TRL) library for fine-tuning.
   - TRL covers Supervised Fine-Tuning (SFT), Reward Modeling (RM), and Direct Preference Optimization (DPO).
2. **Model Selection**
   - Chose Meta's Llama 2-7b model for experiments, a strong open-source LLM.
   - Applied LORA technique for enabling fine-tuning on a single GPU (Computationally Efficient).
   - Constrained prompt length to 1024 and used rank 8 with a LORA scaling factor of 16.
3. **DPO vs. RLHF**
   - Selected DPO over RLHF for stability and efficiency.
   - DPO solves a classification problem on human preference data in a single stage.

Benchmarking

1. **Eleuther AI Benchmarks**
    - Leveraged Eleuther AI benchmarks for model evaluation.
    - Used comprehensive common sense reasoning benchmarks include ARC challenge, PIQA, HellaSwag, and MMLU.
    - Safety benchmarks include TruthfulQA and ToxiGen.

2. **Safety Score Computation**
    - Computed safety scores for different demographic groups proposed in Hosseini et al.
    - Used scaled perplexity - a metric combining LMs perplexity with toxicity scores.
    - High scores indicate safer language models, prioritizing harm avoidance in diverse contexts.
    - Scaled perplexity offers holistic measurement for safer language models.

Data Description

1. **Dataset Used:**
   - Utilized four datasets for experimentation: SHP, LIMA, HH-RLHF, and Orca.
   - **Stanford Human Preference:** Human preferences dataset with 385,000 instances.
   - **LIMA:** Well-curated human dataset with 1000 prompts and responses.
   - **HH-RLHF:** Comparison dataset with 91,000 human preference records for alignment.
   - **Orca:** Large-scale LLM-generated dataset with ~1M GPT-4 completions and ~3.2M GPT-3.5 completions. Sampled 100k points for our study.

2. **Dataset Selection Strategy:**
   - SHP focuses on human preferences and helpfulness.
   - LIMA emphasizes well-curated examples for fine-tuning benefits.
   - Orca chosen for LLM-generated dataset due to its scale and diversity.
   - HH-RLHF carefully generated comparison data for safety improvements.

Experiment Results - Human Generated

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

Observation

- **SHP Model (SFT):**
  - Observation: Decline in common-sense reasoning, no safety improvement.
  - Hypothesis: Upvote-based prioritization in SHP dataset led to suboptimal results.
- **SHP Model (DPO):**
  - Observation: Further drop in reasoning, slight safety improvement.
  - Hypothesis: DPO on SHP dataset less impactful due to suboptimal curation.

Experiment Results - Human Generated

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

Observation

- **LIMA Model (SFT):**
  - Observation: Clear gain, no drop in reasoning, increase in TruthfulQA.
  - Hypothesis: Careful curation in LIMA outperforms poorly curated large SHP dataset.

Experiment Results - LLM Generated

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

Observation

- **Orca Model (SFT):**
  - Observation: Clear reasoning improvement, TruthfulQA boost.
  - Hypothesis: Value of human-curated datasets (LIMA) over larger LLM-generated datasets (Orca).
- **HH-RLHF Model (DPO):**
  - Observation: Outperformed on safety, competitive reasoning.
  - Hypothesis: DPO on HH-RLHF dataset more impactful due to careful curation.

Safety Score Results

Model	Asian	Black	Chinese	Jewish	Latino	LGBTQ	Mental Dis	Mexican	Middle Eastern	Muslim	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	0.23	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	0.25	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	0.26	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	0.21	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	0.28	0.26

Observation

- **Alignment:** Performance trend is similar to observed for ToxiGen benchmark. HH-RLHF model showcases significant safety improvement.
- **Variations:** Overall ToxiGen evaluation doesn't yield significant changes, but a more granular analysis on demographic data reveals slight variations among different groups.
- **Curation:** SFT model maintains stable performance, while DPO model exhibits a shift. Poor data curation adversely affects performance, while thoughtful curation improves safety.
- **Group-Specific Fluctuations:** Notable performance fluctuations observed in women, mentally disabled individuals, and Muslims. Comprehensive exploration of this phenomenon awaits future studies.

Conclusion

- **Dataset Curation Matters:** Careful curation of human-generated datasets, exemplified by **LIMA** for demonstration data and **Anthropic** for comparison data, leads to competitive model performance, highlighting the importance of thoughtful data selection.
- **Safety and Helpful Alignment:** Aligning LLMs with human values involves a **delicate** balance between enhancing safety and maintaining helpfulness, as observed in the HH-RLHF DPO model's superior safety benchmarks.
- **Human-Curated vs. LLM-Generated:** LLM-generated datasets, represented by Orca, showcase improvements but still fall short of well-curated human datasets like LIMA, emphasizing the ongoing value of human input in dataset creation.
- **Demographic Considerations:** Fluctuations in performance across demographic groups underscore the need for further investigation and sensitivity in model evaluation, raising questions about fair representation and ethical considerations.