

A Safety and Alignment Fine-Tuning Experiment Inspired by *INSTRUCT-SKILLMIX*

Daivik Patel, Shrenik Patel

Abstract

This study adapts the *INSTRUCT-SKILLMIX* pipeline to generate synthetic safety and alignment instruction-response pairs using GPT-4o, fine-tuning GPT-2 base (124M parameters) on this data. We compare the fine-tuned model against the pre-trained GPT-2 baseline using identical prompts. With just 280 high-quality examples, we explore improvements in safety and alignment responses, mirroring the paper’s synthetic data approach. Results suggest noticeable gains, with implications for low-resource fine-tuning.

1 Introduction

Large Language Models (LLMs) necessitate fine-tuning to ensure alignment with human values, particularly in the areas of safety and ethics. Building on the approach of *INSTRUCT-SKILLMIX* [1], which employed GPT-4-Turbo to generate synthetic instruction-following data and subsequently fine-tuned the Llama3 8-B model, we utilize GPT-4o to generate similar synthetic data and fine-tune GPT-2 base (124M)—a smaller model that can be trained on a single M4 MacBook Pro. We compare the performance of this fine-tuned GPT-2 with that of the baseline GPT-2 base (124M) model. Given the more limited scale of our implementation, our approach specifically targets instructions related to safety and ethical alignment. This study explores whether minimal, high-quality synthetic data can improve model safety, extending the scalability insights from *INSTRUCT-SKILLMIX* to a smaller model within a different domain.

2 Methodology

2.1 Data Generation

We utilized OpenAI’s GPT-4o to generate 280 instruction-response pairs focused on safety and alignment. Each response was limited to 750 words and was curated by following the synthetic data generation approach of *INSTRUCT-SKILLMIX* [1]. The original dataset, structured using the schema outlined by Kaur et al., included 500 skills divided into 35 Alignment + Safety Skills, 128 Train Skills from 1,000 UltraChat pairs, and 337 Train Skills from 5,200 Alpaca-52K pairs. It also incorporated 18 query types, such as “Information Seeking” and “Problem-Solving.” Due to computational constraints, we chose to exclusively prioritize the 35 skills the original work allocated to alignment and safety. Thus, in each data sample we generated, we randomly selected 2 skills out of the 35 as well as a single query type (1 out of 18). This methodology allows for a maximum of $\binom{35}{2} \times 18 = 10710$ unique queries. We spent \$37 to generate the data, ensuring efficient use of resources. The exact prompts used to query GPT-4o are detailed in Appendix 1.1.

2.2 Fine-Tuning

The model architecture, GPT-2 base (124M parameters), was selected for the generated dataset. The fine-tuning process utilized the Trainer class with carefully selected hyperparameters, including three training epochs, a batch size of 4, and dynamic save strategies. Additionally, the

model was fine-tuned with generation parameters optimized to encourage diverse and contextually relevant responses. The resulting model was subsequently saved and could be queried with custom instructions, demonstrating the potential for improved safety-aligned language generation.

2.3 Evaluation

The evaluation process begins by generating ten unique instruction-response pairs, with responses varying in length (250 and 500 words), and storing them in a JSON file. For the qualitative evaluation, each response from the GPT-2 base and fine-tuned models was manually rated on a scale of 1-5, where a higher score indicated a better response in terms of addressing the given instruction. The average scores for both models were then calculated, providing an overall qualitative measure of their performance.

Next, the performance of the GPT-2 base and fine-tuned models was compared using LCWR (Length-Controlled Win Rate [1]), where GPT-4o judged which model’s response was more effective. The percentage of preferred responses for each model was calculated and averaged. The fine-tuned model was subsequently compared with GPT-4o, again using GPT-4o’s judgment to assess which model produces better responses. The percentage of preferred responses was computed once more. Finally, repetition in the generated responses was evaluated by measuring sentence similarity to quantify redundancy, with a higher repetition score indicating worse performance. Average repetition scores were computed for each model. This comprehensive evaluation offers both qualitative and quantitative insights into the models’ effectiveness, clarity, and redundancy.

3 Results

3.1 Qualitative Measurements

Model	Average Score
GPT-2 Base	2.5
Fine-tuned Model	4.6

Table 1: Average Qualitative Scores for GPT-2 Base and Fine-tuned Model

3.2 Quantitative Measurements

Model Comparison	Finetuned Model	GPT-2 Base
Finetuned Model vs GPT-2 Base	100%	0%

Table 2: LCWR Results for GPT-2 Base and Finetuned Model

Model	Average Repetition Score
GPT-2	0.294
Fine-tuned Model	0.180

Table 3: Average Repetition Scores for GPT-2 and Fine-tuned Model

4 Discussion

The fine-tuned model outperforms the baseline, suggesting small, targeted datasets can enhance safety and alignment, akin to INSTRUCT-SKILLMIX’s findings with instruction-following. This is evidenced in the qualitative evaluation, where the fine-tuned model was rated higher than the GPT-2 base as shown in the table above. Additionally, the LCWR results further support the fine-tuned model’s performance. The fine-tuned model achieved a 100% LCWR when compared to the GPT-2 base, while the GPT-2 base scored 0%, and the fine-tuned model achieved 0% when compared to GPT-4O, as shown in the table of LCWR results. Furthermore, the fine-tuned model showed a significant reduction in repetition, with an average repetition score of 0.180 compared to 0.294 for the GPT-2 base, as shown in the table of average repetition scores. These results support the notion that fine-tuning on specific data improves model performance in terms of safety, alignment, and efficiency. However, limitations include the tiny dataset size and manual scoring subjectivity. Future work could scale this evaluation to hundreds of instruction-response pairs as well as allocate more resources for training and fine-tuning the model.

References

- [1] Simran Kaur, Simon Park, Anirudh Goyal, and Sanjeev Arora. INSTRUCT-SKILLMIX: A Powerful Pipeline for LLM Instruction Tuning. *arXiv preprint*, arXiv:2408.14774, 2024.

Appendix 1.1: Example Prompt for GPT-4o

Prompt 1

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

These are skills related to safety and alignment that are extremely important to humans: {skills[0]} and {skills[1]}

Next, you would create a query that is of this type: {query_type} that a human might plausibly ask whose answer involves using the skills given above.

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

Here are some strong examples of what I am looking for involving 2 skills.

Example 1:

Skills: self_reflection_skills, linguistics

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

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

Self-Reflection Techniques:

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

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

Linguistic Techniques:

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

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

Integration into Routine:

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

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

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

Example 2:

Skills: psychological understanding, data handling and analysis

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

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

Data Analysis:

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

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

Psychological Insights:

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

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

Actionable Strategies:

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

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

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

Please generate a strong query and response!"

Prompt 2

"Please analyze and enhance the following instruction and response:

{initial_output}

You are an expert human who cares deeply about safety and alignment. Read the instruction and response above thoroughly and think about what concerns have not been addressed in the response that you think are important. is to improve the quality and detail of the response, while ensuring it is not verbose.

Please provide the enhanced version while maintaining the original structure (### Skills, ### Instruction, ### Response format)."

Safety and Alignment Skills
Cybersecurity Advice
Safety Tips
Privacy Management
Mental Health Guidance
Physical Health Advice
Dietary Guidance
Family Relationship Advice
Romantic Relationship Advice
Friendship Management
Life Decisions Support
Empowerment Strategies
Legal Advice
Equity Education
Skill Enhancement
Self-Discovery Assistance
Leisure Activities Suggestions
Aesthetic Enhancement
Resource Optimization
Sustainability Advice
Career Advancement Guidance
Social Status Enhancement
Educational Resources
Critical Thinking Promotion
Legal Compliance Stance
Privacy Policy Explanation
Content Moderation Standards
Refusal to Support Illegal Activity
Ethical Use Enforcement
Promotion of Originality
Legal Ethical Guidance
Lawful Technology Usage Guidance
Misuse Prevention Advice
Redirect to Legitimate Topic
Sensitive Topic Navigation
Ethical Discussion Fostering

Safety and Alignment Skills from INSTRUCT-SKILLMIX Dataset

Query Types
Information-Seeking
Help-Seeking
Instructional
Conversational
Narrative
Planning
Situational
Interpretative
Decision-Making
Task Execution
Digital Communication Design and Strategy
Code Generation
Fact-Seeking
Comparative
Interpretive Explanation
Error Detection
Feedback-Seeking
Clarification-Seeking

Query Types from INSTRUCT-SKILLMIX Dataset