

Childproofing LLMs: A Comparative Analysis of CAA, ReFT, and DPO for Safety Alignment

Stanford CS224N Custom Project

Rosemary Jiang
Department of Computer Science
Stanford University
rmjiang@stanford.edu

Anya Zhang
Department of Computer Science
Stanford University
anyazh@stanford.edu

Alice Zhu
Department of Computer Science
Stanford University
alicezhu@stanford.edu

Abstract

As Large Language Models have become more widely used in society especially by users of younger ages, a major safety consideration that has become increasingly prominent is reliable refusal of harmful requests in LLMs while maintaining normal language ability. However, in most cases, LLMs are able to handle direct harmful queries but remain susceptible to adversarial jailbreak prompts and contextual framing that obscures malicious intent. In this study, we aim to evaluate three methods, Contrastive Activation Addition (CAA) for inference-time activation steering, Representation Fine-Tuning (ReFT) for representation-level fine-tuning, and Direct Preference Optimization (DPO) for preference-based training objective, to improve upon the current alignment strategy using Mistral-7B as a baseline. Our results show that while all methods resulted in improved safety compared to our baseline model, representation-level alignment achieved the best results with ReFT demonstrating the strongest reduction in harmful compliance while CAA offered insights into the interpretability of safety representations with multi-layer steering resulting in improved robustness.

1 Key Information to include

- Mentor: Qinan Yu
- External Collaborators (if you have any): N/A
- Sharing project: N/A
- Late Days (if you are using any):

Team member	Late days remaining
Rosemary	5
Alice	1
Anya	0
Total pooled	6
Team size	3
Late days applied for the report	0

2 Introduction

The increasing deployment of Large Language Models (LLMs) has made safety alignment a critical challenge, particularly in high-stakes domains such as child safety. Models must reliably refuse harmful requests while maintaining helpful and appropriate responses, but achieving this balance without degrading core language capabilities remains difficult.

Although modern alignment techniques can block direct harmful queries, LLMs remain vulnerable to adversarial jailbreak prompts that disguise malicious intent through contextual framing or role-play. These attacks expose limitations in current alignment approaches and motivate methods that strengthen safety behavior while preserving language quality.

In this work, we compare three alignment strategies on the Mistral-7B model: inference-time steering via Contrastive Activation Addition (CAA), representation-level training through Representation Fine-Tuning (ReFT), and preference-based fine-tuning using Direct Preference Optimization (DPO). We evaluate these methods on both direct harmful queries and adversarial jailbreak prompts.

Our contributions are threefold. First, we present a unified evaluation of inference-time steering and training-time alignment methods for safety. Second, we show that safety behavior in Mistral-7B is partially localized in intermediate representations, where single-layer interventions trigger refusals for direct harmful prompts while multi-layer steering improves robustness to jailbreak attacks. Third, we analyze the trade-offs between training-time representation learning and inference-time activation steering in terms of safety performance, interpretability, and computational cost.

3 Related Work

Recent work in representation engineering shows that Large Language Models can be guided at inference time without expensive fine-tuning. Rimsky et al. (2023) introduced Contrastive Activation Addition (CAA), which steers model behavior by injecting the difference between mean activations of contrasting prompts [1]. While effective for direct harmful queries, their evaluation focuses primarily on in-distribution prompts, leaving the adversarial robustness of activation steering less explored. Our work examines this limitation in the context of child safety, demonstrating that while localized interventions can mitigate standard queries, defending against complex roleplay-based jailbreaks requires steering a broader multi-layer semantic cluster.

Representation-level alignment methods modify internal activations rather than updating full model parameters. Representation Fine-Tuning (ReFT) learns lightweight interventions applied to hidden representations while keeping the base model largely frozen [2]. These modules implement low-rank transformations at selected layers to enable targeted behavioral changes with minimal additional parameters. This approach is related to parameter-efficient fine-tuning methods such as LoRA [3], which adapt LLMs using low-rank updates to weight matrices.

Direct Preference Optimization (DPO) provides an alternative alignment paradigm that trains models directly from human preference data without the reinforcement learning pipeline used in RLHF. As described in *Direct Preference Optimization: Your Language Model is Secretly a Reward Model*, DPO optimizes a preference-based objective that encourages the model to generate responses favored by human annotators.

4 Approach

To study safety alignment under adversarial prompting, we evaluate three intervention paradigms on the Mistral-7B-v0.2 model: Contrastive Activation Addition (CAA), Representation Fine-Tuning (ReFT), and Direct Preference Optimization (DPO). All methods are evaluated against the same baseline model using deterministic greedy decoding.

4.1 Contrastive Activation Addition (CAA)

CAA is an inference-time steering method that modifies hidden representations during the forward pass to bias the model toward a target behavioral direction.

1. Mean-Centered Vector Extraction. Let D_{safe} and D_{unsafe} denote datasets of safe and unsafe prompts. For transformer layer l , we compute mean hidden activations μ_l^+ and μ_l^- for the two datasets. To isolate the behavioral signal from generic linguistic features, we subtract the global activation mean μ_l^* computed over $D_{all} = D_{safe} \cup D_{unsafe}$.

The steering direction is defined as

$$v_l = \frac{\mu_l^+ - \mu_l^-}{\|\mu_l^+ - \mu_l^-\|}$$

2. Inference-Time Intervention. During generation, we modify the hidden state at layer l :

$$\tilde{h}_l = h_l + \alpha v_l$$

where α controls steering strength. Through empirical sweeps we select $\alpha = -0.5$, which maximizes refusal behavior without degrading linguistic coherence.

3. Single-Layer vs Multi-Layer Steering. We evaluate two intervention strategies. *Single CAA* applies steering at one intermediate layer ($l = 16$), while *Multi CAA* distributes the intervention across a semantic cluster of layers (14–17). This comparison tests whether adversarial jailbreaks exploit distributed representations that require broader intervention.

4.2 Representation Fine-Tuning (ReFT)

ReFT performs parameter-efficient alignment by learning lightweight modules that modify hidden representations while keeping the base model frozen.

1. Representation Intervention Module. Using the PyReFT framework [2], we insert an intervention module at transformer layer l . The modified hidden state is

$$\tilde{h}_l = h_l + f_\theta(h_l)$$

where f_θ is a low-rank transformation with a small number of trainable parameters.

2. Training Procedure. The intervention module is trained on prompt–response pairs consisting of harmful prompts paired with safe refusals and benign prompts paired with helpful responses. This mixture prevents the model from collapsing into universal refusal while encouraging representation shifts associated with safe behavior.

3. Intervention Placement. The module is applied at an intermediate transformer layer, allowing safety-aligned representation changes to influence downstream generation while preserving the majority of pretrained model parameters.

4.3 Direct Preference Optimization (DPO)

DPO aligns model behavior using preference learning without reinforcement learning.

1. Preference Pair Construction. Each training example consists of a prompt x and a preferred and rejected response (y^+, y^-) :

$$(x, y^+, y^-)$$

Harmful prompts use safe refusals as preferred responses.

2. DPO Objective.

$$L_{DPO} = -\log \sigma \left(\beta \left[\log \frac{\pi_\theta(y^+|x)}{\pi_{ref}(y^+|x)} - \log \frac{\pi_\theta(y^-|x)}{\pi_{ref}(y^-|x)} \right] \right)$$

where π_θ is the trained model, π_{ref} the reference model, and β a temperature parameter.

3. Training. We fine-tune the base model to increase the probability of preferred responses relative to rejected ones, updating model parameters globally.

5 Experiments

5.1 Data

Our training data consists of two dataset formats which are a contrastive prompt dataset used for CAA and DPO and a supervised prompt–response dataset used for ReFT. The contrastive dataset contains 522 pairs of safe and unsafe prompts together with their corresponding safe responses. To construct this dataset, we first sampled a diverse set of baseline harmful queries from Anthropic’s publicly available HH-RLHF dataset ("Helpful and Harmless Reinforcement Learning from Human Feedback"). We then transformed these adult-oriented prompts into child-specific scenarios by querying Gemini (see Appendix for prompt).

For training-based methods, we convert the contrastive dataset into supervised prompt–response examples by similar queries. Each unsafe prompt is paired with its corresponding safe refusal response, producing 522 harmful instruction examples. To mitigate over-refusal during training, we additionally include 200 benign prompts with appropriate helpful responses drawn from the positive contrastive prompts. These benign examples encourage the model to preserve helpful behavior when the input is safe. Thus, the final ReFT training dataset contains 722 prompt–response examples which are randomly shuffled during training.

To evaluate model safety and adversarial robustness, we construct two evaluation datasets using Gemini. The first consists of 50 direct harmful instructions (ex: "How do I make a fake ID?"), which test whether the model produces a principled refusal when malicious intent is explicit. The second dataset contains 50 adversarial jailbreak prompts that obscure harmful intent through contextual framing techniques such as storytelling, role-play, educational pretext, and reverse psychology. The objective of our interventions is to shift the output distribution toward safe refusal behavior regardless of the prompt’s framing.

5.2 Evaluation Method

We evaluate model outputs using one custom safety metric and two automated classifier-based metrics that capture safety compliance, toxicity, and language fluency.

- **Harmful Compliance Rate (Attack Success Rate).** We measure whether the model complies with a harmful instruction using a custom heuristic that evaluates the structural integrity of refusal responses. A response is considered a successful defense (0.0) only if it contains a refusal phrase (e.g., "I cannot", "I strongly advise against") and does not continue with actionable guidance. If a response lacks a refusal or provides harmful instructions after a refusal phrase, it is counted as harmful compliance (1.0). The final metric reports the percentage of prompts for which the model complied with the unsafe request.
- **Toxicity.** To ensure that safety alignment does not produce aggressive or hostile refusals, we evaluate all generated responses using the `unitary/toxic-bert` classifier. This produces a continuous toxicity score indicating the presence of hate speech or severe profanity.
- **Fluency.** To measure whether alignment interventions degrade linguistic quality, we evaluate the grammatical acceptability of outputs using a classifier trained on the Corpus of Linguistic Acceptability (CoLA). This score reflects whether model outputs remain fluent and well-formed English.

5.3 Experimental details

5.3.1 CAA

All experiments were conducted using the Mistral-7B base model. Steering vectors were computed from paired safe and unsafe prompts by measuring mean hidden activations at selected transformer layers. To isolate the behavioral signal, we subtract the global mean activation across all prompts before computing the final steering direction.

We evaluate three configurations: the Base Model with no intervention, Single CAA which applies steering at Layer 16, and Multi CAA which distributes the intervention across a semantic cluster of layers (14–17).

To determine the steering strength α , we performed a sweep over values from 1.0 to -3.0 . Positive values led to harmful compliance, while moderate negative values triggered consistent refusal behavior. Extremely negative values caused activation collapse and degraded generation quality. Based on these observations, we selected $\alpha = -0.5$ for all evaluations.

For reproducibility, all generations use deterministic greedy decoding (*do_sample* = *False*) with a maximum generation length of 150 tokens.

5.3.2 ReFT

For the ReFT experiments, we applied a LoReFT intervention to the Mistral-7B-Instruct-v0.2 model using the PyReFT framework. The intervention module was inserted at transformer layer $l = 12$ with a low-rank dimension of $r = 4$, enabling lightweight representation-level adjustments while keeping the base model parameters frozen. This configuration allows the intervention module to steer intermediate representations associated with safety behavior without modifying the majority of pretrained weights.

Training was performed using the supervised prompt–response dataset described in Section 5.1. The final dataset consisted of 522 harmful prompts paired with safe refusal responses and 200 benign prompts paired with helpful answers (722 examples total). Including both harmful and benign examples encourages the intervention module to learn safety-aligned behaviors while preserving normal helpful generation when the input is safe.

During early experiments, we initially trained the intervention module using only the 522 harmful prompts paired with refusal responses. This configuration was trained for three epochs with a batch size of 1, gradient accumulation of 8 steps, and a learning rate of 5×10^{-4} using bf16 precision. While this training regime produced a model that refused all harmful prompts, we observed that it also generalized the refusal pattern to benign prompts, indicating that the intervention module had overfit the refusal behavior.

To address this issue, we introduced the additional 200 benign prompt–response pairs and retrained the intervention module using a reduced learning rate of 1×10^{-4} for a single epoch while keeping the remaining hyperparameters unchanged. This mixed training dataset prevented the model from collapsing into universal refusal behavior and restored normal helpful responses while maintaining strong refusal performance on harmful prompts.

5.3.3 DPO

DPO experiments were conducted using the Mistral-7B-Instruct-v0.2 model with parameter-efficient fine-tuning via LoRA (QLoRA variant). Training used 4-bit NF4 quantization with double quantization and float16 compute.

The LoRA configuration used rank $r = 16$, scaling factor $\alpha = 32$, and dropout 0.05. LoRA adapters were applied to the attention and MLP projection layers (“q_proj”, “k_proj”, “v_proj”, “o_proj”, “gate_proj”, “up_proj”, and “down_proj”).

The model was trained using the DPO objective for one epoch with learning rate 1×10^{-5} , per-device batch size 1, and gradient accumulation of 8 steps. A cosine learning rate scheduler with warmup ratio 0.1 was used. Maximum input and generation lengths were 1024 and 512 tokens respectively.

Training used the `paged_adamw_8bit` optimizer with checkpointing enabled. Evaluation was performed every 100 steps, and training states were periodically saved with a limit of two checkpoints. Mixed precision used fp16 with bf16 disabled.

Table 1: Evaluation of CAA, DPO, and ReFT on Mistral-7B. Harmful compliance represents the attack success rate of unsafe prompts.

Metric	Base Model	Single CAA (L16)	Multi CAA (L14–17)	DPO	ReFT
<i>Standard Evaluation (Direct Harmful Queries)</i>					
Toxicity ($\downarrow$)	0.0022	0.0025	0.0024	0.0039	0.0097
Harmful Compliance ($\downarrow$)	58.0%	44.0%	44.0%	56.0 %	28.0%
Fluency ($\uparrow$)	0.9375	0.9219	0.9300	0.9327	0.9763
<i>Adversarial Evaluation (Jailbreak Prompts)</i>					
Toxicity ($\downarrow$)	0.0049	0.0066	0.0103	0.0128	0.0115
Harmful Compliance ($\downarrow$)	92.0%	88.0%	76.0%	80.0%	48.0%
Fluency ($\uparrow$)	0.8870	0.9047	0.8657	0.8878	0.9370

5.4 Results

In Table 1 which summarizes performance across all models on both standard harmful queries and adversarial jailbreak prompts, we see that the baseline Mistral-7B model exhibited high vulnerability to unsafe requests, complying with 58.0% of standard harmful queries and 92.0% of adversarial jailbreaks.

5.4.1 CAA

On the standard dataset, applying the mean-centered steering vector ($\alpha = -0.5$) yielded a 14.0% absolute reduction in harmful compliance for both the Single CAA (Layer 16) and Multi CAA (Layers 14–17) interventions, bringing the rate down to 44.0%.

On the adversarial jailbreak dataset, the Single CAA intervention provided only a marginal defense, reducing compliance from 92.0% to 88.0%. However, the Multi CAA intervention achieved a much more significant 16.0% absolute reduction, lowering the harmful compliance rate to 76.0%.

Importantly, across all interventions and datasets, the model’s fluency remained highly stable. For the Multi CAA intervention, fluency scored 0.9300 on standard prompts and 0.8657 on jailbreaks, demonstrating that the steering vector did not degrade the model’s fundamental linguistic capabilities. Toxicity remained near zero across all evaluated settings.

5.4.2 ReFT

The ReFT intervention achieved the strongest safety performance on the standard harmful query dataset. Harmful compliance dropped from 58.0% in the base model to 28.0%, representing a 30.0% absolute reduction and outperforming both CAA configurations and DPO. This indicates that representation-level fine-tuning is highly effective at learning stable refusal behaviors for explicit harmful instructions.

However, ReFT remained vulnerable to adversarial jailbreak prompts. While it improved the jailbreak compliance rate from 92.0% to 48.0%, it did not match the robustness of the Multi CAA configuration, which achieved a lower compliance rate of 76.0% through multi-layer steering. This suggests that training-time alignment alone may struggle to generalize to complex contextual jailbreak attacks.

Despite these safety interventions, ReFT maintained the highest fluency across all evaluated methods, achieving a fluency score of 0.9763 on standard prompts and 0.9370 on adversarial prompts. Toxicity levels also remained very low across both evaluation settings. These results highlight a key trade-off between training-time alignment and inference-time steering: ReFT excels at enforcing refusal behavior on direct harmful queries while preserving linguistic quality, but is less robust to adversarial prompt framing compared to multi-layer CAA.

5.4.3 DPO

Under standard evaluation, DPO achieved a toxicity score of 0.0039, harmful compliance of 56%, and fluency of 0.9327 and has 0.0128 toxicity, 80% harmful compliance, and 0.8878 fluency under adversarial evaluation. We note that it generally performs worse than CAA and ReFT on metrics under standard evaluation and similarly to CAA but worse than ReFT under adversarial evaluation.

6 Analysis

The results of this evaluation provide critical insights into how safety representations are distributed within the latent space of Large Language Models, as well as the inherent fragility of standard RLHF tuning.

6.1 CAA

Our experiments suggest that safety representations in Mistral-7B are highly localized for direct harmful instructions. Intervening at a single layer (Layer 16) was sufficient to trigger refusal behavior on standard harmful queries, indicating that the model encodes a compact safety direction that can be activated with a localized intervention.

However, this localized defense proved fragile against adversarial jailbreak prompts. These prompts dilute harmful intent across complex contextual framing, allowing them to bypass single-layer interventions. Distributing the steering vector across multiple layers (Layers 14–17) improved robustness by influencing a broader semantic cluster within the model’s internal representations.

Importantly, both configurations preserved strong fluency and near-zero toxicity, suggesting that mean-centered steering can isolate the safety direction without degrading the model’s core linguistic capabilities.

6.2 ReFT

ReFT demonstrated the strongest safety performance across both evaluation settings. On the standard harmful query dataset, harmful compliance dropped from 58.0% in the base model to 28.0%, representing the largest reduction among all evaluated methods.

More notably, ReFT also produced the lowest harmful compliance rate on adversarial jailbreak prompts, reducing compliance from 92.0% to 48.0%. This suggests that representation-level fine-tuning can learn robust safety behaviors that generalize beyond explicit harmful instructions to more complex contextual attacks.

At the same time, ReFT maintained the highest fluency scores across all models, indicating that low-rank representation interventions can improve safety behavior while preserving the model’s core language generation capabilities.

6.3 DPO

We notice that DPO performs worst than CAA and ReFT in both the standard and jailbreaking evaluations. Global fine-tuning improves the model only slightly when compared to the base model. From this, we can deduce that global fine-tuning struggles to change model behavior based on complex scenarios, and cannot isolate internal safety representations. DPO is not robust, and does not do well as its own method alone.

We notice that DPO does not do better, as it is hard to fine-tune a model globally to be good in contextually based scenarios such as child-safety.

6.4 Cross-Method Comparison

Taken together, these results reveal important differences between training-time alignment and inference-time steering approaches. ReFT achieves the strongest overall safety performance, produc-

ing the lowest harmful compliance rates on both standard harmful queries and adversarial jailbreak prompts. Because ReFT explicitly learns refusal behaviors through supervised training, it can develop a robust decision boundary between harmful and benign instructions that generalizes across a range of prompt formulations.

CAA, in contrast, operates purely at inference time by modifying internal activations without updating model weights. While this approach provides weaker improvements on adversarial jailbreak prompts compared to ReFT, it offers two key advantages: interpretability and efficiency. By identifying a localized safety direction within the model’s latent space, CAA provides insight into how refusal behaviors are represented internally and enables rapid safety adjustments without retraining the model.

DPO performs worst among the evaluated methods across most metrics. Although preference-based fine-tuning improved alignment slightly relative to the base model, it did not achieve the same level of robustness as either CAA or ReFT in our experiments. One possible explanation is that our DPO training run was limited to a single epoch, which may not have been sufficient for the model to fully learn the preference signal. With more extensive training or additional hyperparameter tuning, DPO performance could potentially improve. Nevertheless, under the training configuration used in this study, representation-level methods produced stronger safety outcomes.

Overall, these results suggest that representation-level alignment methods offer a promising direction for improving LLM safety. ReFT demonstrates the strongest empirical safety gains through training-time representation learning, while CAA highlights how interpretable inference-time interventions can expose and manipulate safety representations within the model.

7 Conclusion

In this work, we investigated how representation-level interventions can improve the safety behavior of large language models in the context of child-directed harmful prompts. We evaluated three alignment approaches, Contrastive Activation Addition (CAA), Representation Fine-Tuning (ReFT), and Direct Preference Optimization (DPO), using both direct harmful queries and adversarial jailbreak prompts.

Our results show that representation-level methods provide effective mechanisms for improving refusal behavior while preserving linguistic quality. ReFT achieved the strongest overall safety performance, substantially reducing harmful compliance on both standard and adversarial prompts while maintaining the highest fluency among the evaluated models. CAA provided valuable interpretability insights, revealing that safety behaviors in Mistral-7B can be activated through targeted interventions in intermediate model layers. Multi-layer CAA further demonstrated that distributing the intervention across a broader semantic cluster can improve robustness against contextual jailbreak attacks. In contrast, DPO provided more modest improvements under our training configuration.

These findings highlight the complementary strengths of training-time and inference-time safety alignment strategies. Training-based methods such as ReFT can learn robust refusal behaviors from supervised examples, while inference-time steering methods such as CAA offer interpretable and computationally efficient mechanisms for dynamically modifying model behavior without retraining.

Our study also has several limitations. Our DPO experiments were conducted with limited training epochs, and more extensive hyperparameter tuning may yield stronger results. Additionally, our evaluation datasets are relatively small and focus primarily on child-safety scenarios.

Future work could explore larger-scale evaluations, more sophisticated jailbreak strategies, and hybrid approaches that combine training-time representation learning with inference-time activation steering. Such methods may provide a promising direction for developing safer and more controllable large language models.

Team contributions (Required for multi-person team)

Rosemary generated our 522-row contrastive dataset along and the 50-row standard and jailbreak test datasets. Rosemary made organizational edits and experimented with parameter values for the CAA method and generated the evaluation code shared across methods.

Alice implemented hidden representation extraction and single-layer steering functions in our CAA method. Alice implemented DPO, with the exclusion of the evaluation code, which was generated by Rosemary in collaboration with AI.

Anya implemented the multi-layer and response with CAA functions. Anya implemented ReFT, with the exclusion of the evaluation code, which was generated by Rosemary in collaboration with AI.

Alice and Anya wrote the project proposal. Alice wrote the experiments (section 3) and future work (section 4) sections of the project milestone; Anya wrote all other sections (abstract and section 2 approach).

Alice wrote the DPO sections (sections 4.3, 5.3.3, 5.4.3, 6.3). Rosemary wrote the CAA sections, abstract, and introduction sections of this report; Anya wrote the ReFT and conclusion sections. Anya and Rosemary wrote all other sections of this report.

Citation Notes

Our implementations are implemented on a base model of Mistral-7B. Our evaluation code is written in collaboration with AI (we decided on the metrics ourselves). Our datasets were sampled from Anthropic’s publicly available HH-RLHF using Gemini.

References

- [1] Julian Schulz Meg Tong Evan Hubinger Alexander Matt Turner Nina Rimksy, Nick Gabrieli. Steering llama 2 via contrastive activation addition. 2024.
- [2] Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. Reft: Representation finetuning for language models, 2024.
- [3] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.
- [4] Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsay, Trent Bricken, Brian Chen, Adam Pearce, Natalie Sandford, Chris Hallacy, John Hyde, Jason Wheeler, Igor Kanitscheider, Shauna Kravec, Nicholas Schiefer, Thang Doig, Amanda Olsson, Catherine Olsson, Nicholas Cammarata, Joseph Benton, Cem Anil, Jared Kaplan, Henk Tilman, Todd Sheehan, Zak Rogoff, Madelaine Udell, Julia Yi, Sri Venkat Rao, David Amodei, Dario Amodei, Jack Clark, Christopher Olah, and Tom Brown. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. In *Transformer Circuits Thread*, 2024.
- [5] Ruben Härle, Felix Friedrich, Manuel Brack, Stephan Wäldchen, Björn Deiseroth, Patrick Schramowski, and Kristian Kersting. Measuring and guiding monosemanticity. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.
- [6] Eric Mitchell Stefano Ermon Christopher D. Manning Chelsea Finn Rafael Rafailov, Archit Sharma. Direct preference optimization: Your language model is secretly a reward model, 2023.

A Appendix

System Prompt:

You are generating contrastive prompt pairs for CHILD SAFETY STEERING research. Given an unsafe user prompt (adult phrasing), produce two outputs:

1. `unsafe_child`: A child-appropriate rewrite preserving intent/tone but with NO actionable details.
2. `safe_child`: A matched safe alternative encouraging positive, prosocial behavior.

Return STRICT JSON with keys: `unsafe_child`, `safe_child`.