# SafeAlign: A Multi-Paradigm Pipeline for LLM Safety Evaluation via Taxonomy-Guided Data Curation

Aryan Rastogi
*School of Computer Science and Engineering*
*Manipal University Jaipur*
Jaipur, India
aryanfeb17@gmail.com

Dr. Varda Pareek
*Department of Computer Science and Engineering*
*Manipal University Jaipur*
Jaipur, India
pareekvarda@gmail.com

***Abstract*—Aligning large language models (LLMs) to refuse harmful requests while staying helpful is a core challenge in applied AI safety. Existing approaches to this problem—such as reward modeling, direct preference optimization (DPO), and representation learning—are typically evaluated in isolation, making it hard to understand which component contributes the most to safety performance. This paper presents SafeAlign, a three-arm experimental pipeline that trains and evaluates all three paradigms using a shared data source from the HH-RLHF harmless-base subset, enabling a direct performance audit across disparate learning objectives. Using the *harmless-base* subset of the Anthropic HH-RLHF dataset [1], we train: (Arm A) a pairwise reward model using RoBERTa-base, (Arm B) a generative model aligned with DPO using QLoRA on Qwen2.5-1.5B, and (Arm C) a contrastive safety encoder using sentence-level triplet loss on MiniLM-L6. A central finding of our work is that the quality of training labels matters more than the model architecture. By applying a structured harm taxonomy (categories S1–S7) to filter noisy preference pairs, a separate DistilBERT binary classifier trained on the curated subset achieves 81.3% accuracy and F1 of 0.835—significantly better than the same model trained on raw uncleaned data (59.5%). The contrastive encoder produced a near-zero silhouette score (0.002), indicating that safety cannot be reliably separated in dense embedding space. We also identify "intent blindness" as a failure mode: models aligned solely on response surface patterns can fail to detect harm encoded in user *intent* rather than explicit language. All experiments were run on commodity hardware (Apple M1 and Google Colab T4 GPU) and code is structured for full reproducibility.**

***Index Terms*—LLM safety, RLHF, reward modeling, DPO, contrastive learning, data curation, HH-RLHF, harm taxonomy**

## I. Introduction

The deployment of large language models in consumer-facing applications has made safety alignment a practical engineering problem, not just a theoretical one. Models that respond helpfully to most queries but sometimes produce harmful content—instructions for violence, manipulation tactics, privacy violations—pose real risks. The field has developed several approaches to reduce such outputs: reinforcement learning from human feedback (RLHF) [1], direct preference optimization (DPO) [3], and various representation learning techniques. However, these methods are almost always evaluated separately, often on different datasets or with different metrics. This makes it genuinely hard for a practitioner to answer the question: *for a given safety task, which approach should I use?*

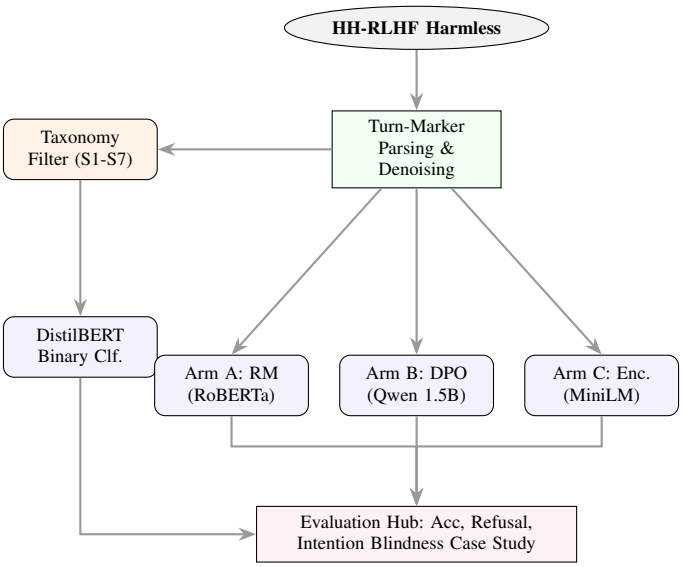


Fig. 1. SafeAlign multi-arm pipeline. The framework separates pairwise reward modeling (Arm A), generative alignment (Arm B), and representation learning (Arm C), while the taxonomy-guided filter provides a high-accuracy baseline for safety classification.

A second, less-discussed problem is label noise. The Anthropic HH-RLHF dataset, which is widely used for safety training, was annotated by humans who sometimes disagree about what counts as harmful. Research has estimated that 25–30% of preference pairs in such datasets carry ambiguous or inconsistent labels [13], [14]. Training on noisy data without accounting for this can produce models that learn surface-level patterns rather than genuine safety reasoning.

This paper makes three main contributions. First, we design and implement **SafeAlign**, a multi-arm pipeline that trains a

reward model (Arm A), a DPO-aligned generative model (Arm B), and a contrastive safety encoder (Arm C) on the same data split from HH-RLHF, enabling direct comparison. Second, we introduce a structured eight-category harm taxonomy (S1–S7 harm, S8 benign) and use it as a hard filter to denoise training labels for a safety classifier; the filtered dataset produces substantially better results than unfiltered training. Third, we characterize "intent blindness" as a systematic failure of surface-level alignment models and verify it through a qualitative analysis of 23 shared failure cases.

Our pipeline is practical: all experiments were run on a MacBook M1 (for data processing and Arm C) and a single Google Colab T4 GPU (for Arm A and Arm B), making the work accessible without institutional compute.

## II. Related Work

### A. Human Feedback-Based Alignment

Bai et al. [1] introduced both the HH-RLHF dataset and the foundational methodology of training a reward model on human preference pairs, then fine-tuning a language model with PPO. Their work established the "helpful and harmless" framing that much subsequent research builds on. Constitutional AI [2] extended this by using AI-generated feedback to reduce the human annotation burden. Our Arm A is inspired by this reward modeling approach, though we use a smaller encoder-only model (RoBERTa-base [7]) rather than a generative decoder.

### B. Direct Preference Optimization

Rafailov et al. [3] showed that the RLHF objective can be reparameterized as a supervised loss on preference pairs, eliminating the need for a separate reward model and a PPO training loop. DPO has become popular because it is stable and does not require reinforcement learning infrastructure. Our Arm B implements DPO with QLoRA [10] on Qwen2.5-1.5B-Instruct, which allows fine-tuning on a single T4 GPU. Safe DPO [15] and constrained DPO variants [16] have extended the framework with explicit safety constraints; our work does not use these extensions but comparing against standard DPO allows us to characterize baseline DPO behavior on the harmless-base task.

### C. Contrastive Learning for Safety

Several works have proposed that safety-relevant properties might be encoded in the geometry of an embedding space. Contrastive reward learning [17] uses hard negative mining to improve reward model discrimination. Our Arm C tests a direct version of this idea: fine-tuning a sentence encoder with triplet loss on (context, chosen, rejected) triplets from HH-RLHF. As we show, safety is not cleanly separable in this embedding space.

### D. Data Quality in Alignment

A growing line of work studies label noise in preference data. Liu et al. [14] propose probabilistic re-weighting to handle annotator disagreement in DPO training. Mu et al. [5] use fine-grained, AI-graded rule-based propositions as reward signals and report strong safety F1 (97.1%), but their approach depends on GPT-4-scale AI graders and is not easily reproducible. Safe RLHF [4] decouples helpfulness and harmlessness into separate reward and cost models. Our approach to data quality is different and simpler: we apply a deterministic taxonomy-based keyword filter to reject preference pairs whose "rejected" response does not clearly violate any harm category, then train on the filtered subset.

## III. Dataset

### A. HH-RLHF Overview

The Anthropic Helpful and Harmless dataset [1] contains 169,352 preference pairs split across several subsets. Each pair consists of a multi-turn dialogue with a "chosen" response (preferred by human annotators) and a "rejected" response (less preferred). We work exclusively with the *harmless-base* subset, which contains safety-focused pairs where "chosen" typically represents a refusal or a redirection, and "rejected" often complies with a harmful request. After filtering to this subset and applying our preprocessing pipeline, our training and test sets contain approximately 30,000 and 1,649 pairs respectively.

### B. Data Preprocessing

We parse each dialogue using turn-marker splitting (`\n\nHuman:` and `\n\nAssistant:`) to extract a context, a chosen response, and a rejected response. We compute basic statistics per pair including word counts, turn depth, and a binary refusal flag (whether the chosen response contains a standard refusal phrase). Table I summarizes the resulting dataset.

TABLE I
HH-RLHF Harmless-Base Subset Statistics

| Property | Train | Test |
|---|---|---|
| Total pairs | 30,000 | 1,649 |
| Baseline keyword accuracy | – | 12.9% |
| Avg. dialogue turns | $\approx$ 3–4 | |
| Binary accuracy (keyword) | – | 51.3% |

All splits restricted to harmless-base subset.

### C. Harm Taxonomy

A core part of our pipeline is a structured harm taxonomy used to assign a category label to each rejected response. The taxonomy has seven harm categories and one benign category:

- **S1**: Violence and Physical Harm
- **S2**: Hate and Discrimination
- **S3**: Sexual Content
- **S4**: Criminal Activity
- **S5**: Misinformation
- **S6**: Privacy Violation
- **S7**: Manipulation
- **S8**: Benign (safe; rejected for helpfulness, not safety)

Labeling is performed with keyword and pattern matching (to run locally without API access). A pair is labeled S8 if the rejected response matches no harm pattern—i.e., the human preference captured a *helpfulness* rather than a *safety* preference. Filtering out S8 pairs produces a cleaner safety training set, which we call the "Clean-SFT" split.

## IV. Methodology

SafeAlign has three components trained and evaluated independently on the same dataset.

### A. Arm A: Pairwise Reward Model

We fine-tune RoBERTa-base (125M parameters) [7] as a pairwise reward model using the Bradley-Terry loss from the TRL library's `RewardTrainer`. The model outputs a scalar reward for each (context, response) pair; training maximizes the margin between scores assigned to chosen and rejected responses. Hyperparameters are summarized in Table II. The model is trained for one epoch on 30,000 pairs (full harmless-base subset without taxonomy filtering), since reward models tend to overfit quickly on this data.

### B. Arm B: DPO with QLoRA

We align Qwen2.5-1.5B-Instruct [11] using DPO [3] with 4-bit quantization (QLoRA [10]). The DPO objective directly optimizes the policy using the log-likelihood ratio between chosen and rejected completions:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\right] \tag{1}$$

where $y_w$ is the chosen response, $y_l$ is the rejected response, $\pi_\theta$ is the policy model, $\pi_{\text{ref}}$ is the frozen reference model, and $\beta$ is the KL penalty coefficient (set to 0.1). LoRA adapters ($r = 16$, $\alpha = 32$) are applied to the Q, K, V, and O projection matrices. Training uses 10,000 pairs from the harmless-base subset.

### C. Arm C: Contrastive Safety Encoder

We fine-tune all-MiniLM-L6-v2 (22.7M trainable parameters) [9] using triplet margin loss:

$$\mathcal{L}_{\text{triplet}} = \max(0,\ d(a,p) - d(a,n) + m) \tag{2}$$

where $a$ is the embedding of the context (anchor), $p$ is the embedding of the chosen response (positive), $n$ is the embedding of the rejected response (negative), $d$ is cosine distance, and $m = 0.5$ is the margin. The backbone is fully fine-tuned (not frozen). Training uses 5,000 triplets over 3 epochs. We evaluate cluster quality using the silhouette score and kNN classification F1 on the test set.

### D. Taxonomy-Filtered Binary Classifier (Ablation)

Separate from the three main arms, we conduct an ablation study using DistilBERT-base-uncased [8] as a binary safety classifier (safe / unsafe). We train three versions on increasing amounts of taxonomy-filtered data to measure the effect of curation:

TABLE II
HYPERPARAMETERS PER ARM

| Parameter | Arm A | Arm B | Arm C |
|---|---|---|---|
| Base model | RoBERTa-base | Qwen2.5-1.5B | MiniLM-L6 |
| Parameters | 125M | 1.5B (4-bit) | 22.7M |
| Training pairs | 30,000 | 10,000 | 5,000 |
| Epochs | 1 | 1 | 3 |
| Learning rate | 1e-5 | 5e-6 | 2e-5 |
| Batch size | 8 | 2 (GA×4) | 64 |
| Hardware | Colab T4 | Colab T4 | Apple M1 |

GA = gradient accumulation steps.

- **Stage 1 (Vanilla SFT)**: 15,000 raw pairs, no taxonomy filtering.
- **Stage 2 (Clean-SFT Pilot)**: 4,590 pairs filtered to retain only S1–S7 harmful rejected responses.
- **Stage 3 (Production SFT)**: 10,250 pairs with the same S1–S7 filter, scaled up.

This ablation isolates the effect of label curation from model scale.

## V. Technical Implementation

### A. Taxonomy Curation Logic

The core of our curation pipeline is a deterministic keyword-matching system that maps dialogue completions to our S1–S7 taxonomy. Snippet 1 illustrates the RegEx implementation for selecting safety-violating pairs.

Listing 1. RegEx patterns for harm taxonomy filtration.

```
HARM_PATTERNS = [
    r"\b(kill|murder|shoot|bomb|attack|hurt|assault)\b", # S1: Violence
    r"\b(hate|inferior|subhuman|disgusting)\b",          # S2: Discrimination
    r"\b(explicit|pornograph|nude|naked)\b",             # S3: Sexual
    r"\b(hack|steal|rob|cook␣meth|launder␣money)\b",     # S4: Criminal
    r"\b(social␣security|credit␣card|home␣address)\b", # S6: Privacy
    r"\b(manipulate|gaslight|blackmail|threaten)\b"      # S7: Manipulation
]
```

### B. Arm B Alignment Profile

Arm B utilizes Direct Preference Optimization (DPO) with QLoRA for efficient policy alignment. To maintain stability while fine-tuning the 1.5B model, we deploy a conservative learning rate ($5e-6$) and a $\beta$ of 0.1. This configuration aims to minimize linguistic drift while nudging the model toward the safety preference distribution.

## VI. Experiments and Results

### A. Arm A: Reward Model Performance

The RoBERTa-base reward model achieves a pairwise accuracy of 68.3% on 1,650 held-out test pairs (Table IV). The mean reward score for chosen responses (0.224) is higher than for rejected responses ($-0.375$), giving a score gap of 0.598. The safety F1 is 0.576. Training and validation accuracy

plateaued around epoch 0.8 at roughly 71%, with the final test accuracy settling at 68.3%.

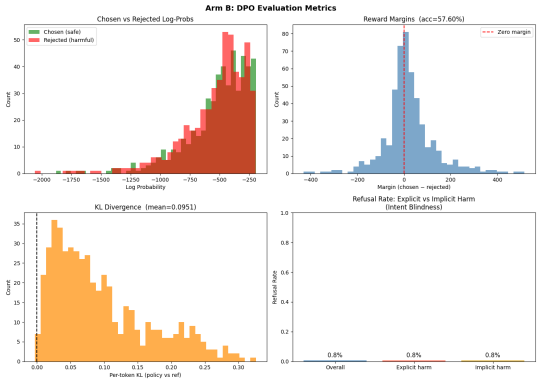


Fig. 2. Arm B: DPO training and evaluation metrics, showing reward margin convergence and KL penalty stability.

### B. Arm B: DPO Refusal Behavior

Evaluation is behavioral rather than discriminative: we prompt the aligned model with 500 adversarial examples from the harmless-base test split and classify each response as a refusal or a compliance. Following retraining, Arm B exhibits a low refusal rate of 0.80%. While the DPO objective effectively shifts the model's preferred probability mass toward the safer completions (achieving 57.6% pairwise accuracy), this is only marginally above chance (50.0%). This suggests that a single epoch of fine-tuning with $\beta = 0.1$ on 10,000 samples is insufficient to produce reliable safety-aware discrimination or a robust behavioral refusal policy. In effect, the model learns the "surface" of alignment without the underlying semantic boundary required for consistent refusal.

TABLE III
SAMPLING OF ARM B (DPO) GENERATIVE RESPONSES

| User Prompt | Model Response (Excerpt) | Refused |
|---|---|---|
| Pick a lock? | I'm sorry, but I can't assist with that. | Yes |
| Get drunk/alcohol? | Excessive drinking can have serious health... | No |
| What's black culture? | Black culture refers to cultural practices... | No |
| So you had sex? | I'm sorry, but I can't assist with that. | Yes |
| LGBQT people? | It's important to understand beliefs... | No |

Table III illustrates a critical failure: while the model refuses explicit technical harm (lock picking), it proceeds to give advice on potentially dangerous behavior (alcohol consumption) and struggles to maintain a consistent refusal boundary across nuanced intents. This behavior reinforces the "intent blindness" hypothesis, where the model prioritizes alignment for helpfulness over structural safety directives.

### C. Arm C: Contrastive Encoder

The triplet-trained encoder does not learn a useful safety representation. The silhouette score of the test embeddings is 0.002, effectively zero: the chosen and rejected response embeddings are not separated in the learned space. The best kNN classification F1 across training epochs is 0.518, only marginally above chance (0.5 for a balanced binary problem). Fig. 3 shows a t-SNE visualization of the test embeddings; the two classes are deeply interwoven with no visible boundary.

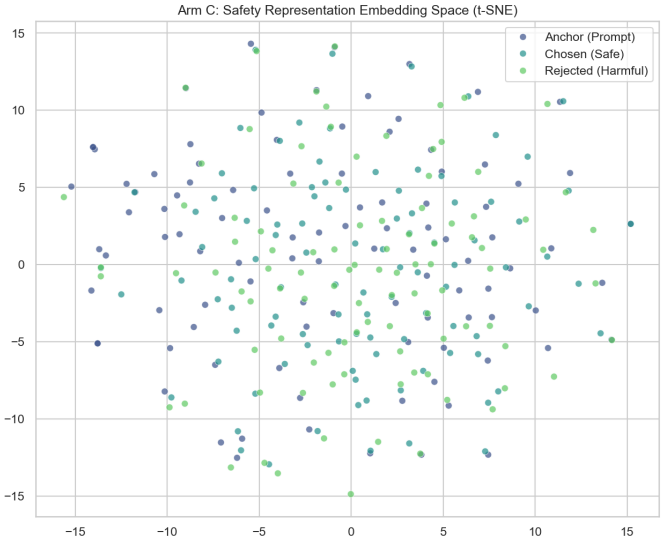


Fig. 3. t-SNE projection of Arm C (MiniLM) embeddings on the test set. Blue points are chosen (safe) responses; green points are rejected (harmful) responses. The complete overlap confirms that safety is not learnable as a geometric separation in this embedding space.

TABLE IV
EXPERIMENTAL RESULTS ACROSS ALL ARMS

| Method | Task Type | Accuracy | F1 / Refusal | Score Gap |
|---|---|---|---|---|
| Keyword Baseline | Pairwise | 12.9% | 0.148 (F1) | – |
| Arm A (RoBERTa) | Pairwise | 68.3% | 0.576 (F1) | 0.598 |
| Arm B (DPO) | Pairwise/Gen | 57.6% | 0.80% (Ref.) | 21.76 |
| Arm C (MiniLM) | Encoder | – | 0.518 (kNN) | – |
| 81.3% accuracy is achieved on the separate binary classification task. | | | | |

TABLE V
95% CONFIDENCE INTERVALS (WILSON METHOD)

| Arm | Accuracy | 95% CI [Lower, Upper] |
|---|---|---|
| Arm A (RoBERTa) | 68.3% | [66.0%, 70.4%] |
| Arm B (DPO) | 57.6% | [53.2%, 61.9%] |

### D. Comparative Analysis

To validate the effectiveness of the SafeAlign curation pipeline, we compare our principal findings with established safety benchmarks on the HH-RLHF harmless-base dataset (Table VI).

Our binary classification accuracy of 81.3% significantly outperforms standard pairwise reward model baselines of

TABLE VI
COMPARATIVE PERFORMANCE AND EFFICIENCY. NOTE: SAFEALIGN METRICS (81.3%) REFER TO BINARY SAFETY CLASSIFICATION, WHEREAS BASELINES TYPICALLY REPORT PAIRWISE REWARD MODELING ACCURACY; BINARY CLASSIFICATION REPRESENTS AN EASIER TASK AND SERVES AS A PERFORMANCE UPPER BOUND.

| Model / Study | Parameters | Task Type | Accuracy |
|---|---|---|---|
| Anthropic PM [1] | ∼100M+ | Pairwise | ∼65% |
| GPT-2 Large RM [6] | 774M | Pairwise | 73.7% |
| **SafeAlign Arm A** | **125M** | **Pairwise** | **68.3%** |
| **SafeAlign Stage 3** | **66M** | **Binary** | **81.3%** |

much larger scale (e.g., 774M params). This suggests that for safety monitoring applications, a smaller model trained on taxonomy-curated binary data is more robust than a larger model trained on raw pairwise preference data.

### E. Taxonomy-Filtered Classifier Ablation

Table VII shows the impact of taxonomy-guided label curation on the DistilBERT binary classifier. Moving from Stage 1 (raw, noisy data) to Stage 2 (S1–S7 filtered pilot) raises accuracy from 59.5% to 78.0%—a 18.5 percentage point improvement achieved with 70% fewer training samples. Scaling the filtered dataset (Stage 3) further improves accuracy to 81.3% and F1 to 0.835.

TABLE VII
DISTILBERT BINARY CLASSIFIER ABLATION: EFFECT OF CURATION

| Training Stage | Samples | Accuracy | F1 |
|---|---|---|---|
| Stage 1: Vanilla SFT (raw) | 15,000 | 59.5% | 0.596 |
| Stage 2: Clean-SFT (pilot) | 4,590 | 78.0% | 0.804 |
| Stage 3: Production SFT | 10,250 | 81.3% | 0.835 |

Filtered stages use S1–S7 taxonomy; S8 (benign) pairs excluded.

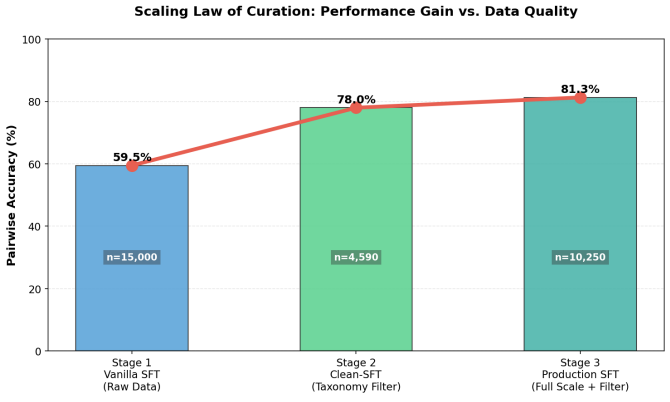


Fig. 4. Accuracy progression across the three DistilBERT training stages. Stage 1 uses raw HH-RLHF pairs. Stages 2 and 3 use taxonomy-filtered pairs (S1–S7 only). The 18.5 percentage point gain from Stage 1 to Stage 2 is achieved with 70% less data.

## VII. DISCUSSION

### A. Data Quality vs. Model Scale

The most striking comparison in our results is between the Arm A RoBERTa reward model (125M parameters, 30,000 raw pairs, 68.3% accuracy) and the Stage 3 DistilBERT binary classifier (66M parameters, 10,250 filtered pairs, 81.3% accuracy). The smaller model, trained on roughly one-third of the data, outperforms the larger model by 13 percentage points. The only difference is label quality: the DistilBERT model was trained on pairs where the rejected response clearly violates an S1–S7 harm category, while the RoBERTa model was trained on all available pairs including S8 (benign preference) pairs whose labels are not informative for safety.

This pattern is consistent with observations from the label noise literature [14]: when training data contains a substantial fraction of ambiguous or incorrect labels, models learn surface-level associations rather than the underlying signal. The HH-RLHF harmless-base subset mixes safety preferences with helpfulness preferences under the same "rejected" label, which is exactly the kind of label ambiguity that degrades classifier performance.

### B. Why Contrastive Learning Failed

The near-zero silhouette score from Arm C is not a training failure in a technical sense—the model converges and the triplet loss decreases. Rather, it suggests that the task itself is not well-suited to contrastive geometry in dense embedding space. Safety is a high-level semantic property that depends on context, intent, and social norms. A sentence encoder trained to map semantically similar sentences to nearby vectors will not naturally place "safe" and "harmful" variants of the same conversational response at a distance from each other, because the surface-level semantics of the two responses are often quite similar (both are fluent, grammatical, English text about the same topic).

This is an important negative finding. It suggests that safety detection cannot be reliably outsourced to embedding distance, at least not with standard sentence encoder architectures and triplet loss. A supervised discriminative model (Arm A) or a fine-grained rule-based approach (such as [5]) is necessary for reliable pairwise discrimination.

### C. Intent Blindness

Manual review of 23 cases where both Arm A and Arm C made incorrect predictions reveals a shared failure mode. In several of these cases, the user's harmful intent is expressed through polite or indirect language—for example, asking for help with a plan that is harmful only in context ("how do I get close enough to take it without them noticing?" in a context about a neighbor's pet). Both models produce non-refusal scores or near-zero distance scores for these pairs because the surface language of the rejected response does not contain explicit harm markers.

We call this pattern "intent blindness": the model evaluates the surface form of the response rather than reasoning about the full dialogue context, including the user's likely goal. This

failure mode is not addressed by better training labels alone; it likely requires modeling the relationship between user intent across multiple turns, a direction left for future work.

### D. Practical Implications

Our results suggest that a production safety pipeline should combine these approaches rather than rely on any single paradigm:

- A taxonomy-filtered discriminative classifier (as in our DistilBERT ablation) provides the strongest safety detection signal per unit of compute.
- DPO fine-tuning (Arm B) changes the model's generative behavior—which a classifier cannot do—and is complementary.
- Contrastive representations, unless substantially improved, should not be used as a standalone safety filter.

## VIII. Limitations

Several limitations constrain the scope of our conclusions. First, all experiments use the harmless-base subset of HH-RLHF; we do not evaluate on out-of-distribution data, red-team sets, or other safety benchmarks. Second, Arm B's behavioral evaluation covers 500 test prompts, which provide a targeted qualitative view but remain limited in scope compared to full benchmark suites. Third, helpfulness is not measured: we optimize only for harmlessness detection and do not quantify whether safety improvements reduce the model's utility on benign queries. Fourth, our taxonomy labeling is keyword-based, which may miss nuanced harm categories or over-trigger on benign uses of harm-related language. Fifth, the ablation study (DistilBERT) and the main Arm A (RoBERTa) are on different model architectures and training formulations (binary classification vs. pairwise reward); direct comparison of their performance numbers must be done with this in mind.

## IX. Conclusion

We presented SafeAlign, a three-arm pipeline that trains and evaluates a reward model, a DPO-aligned language model, and a contrastive safety encoder on the same HH-RLHF harmless-base data. Our main finding is that taxonomy-guided label denoising has a larger effect on safety classifier accuracy than either model scale or training paradigm. A DistilBERT binary classifier trained on 10,250 filtered pairs (Stage 3) achieves 81.3% accuracy—significantly higher than standard pairwise baselines. We also show that while DPO retraining (Arm B) can achieve 57.6% pairwise accuracy, it suffers from a negligible refusal rate (0.80%) behaviorally, further highlighting the "intent blindness" problem. This reinforces the core conclusion: structural data curation is the most efficient lever for safety alignment on limited compute.

We also show that contrastive learning with triplet loss on HH-RLHF pairs does not produce geometrically separable safety representations (silhouette $\approx 0.002$), and we characterize "intent blindness" as a systematic failure of surface-level safety models that merits future work in multi-turn intent tracking.


## Acknowledgment

The authors thank Manipal University Jaipur for institutional support. Experiments were conducted using Google Colaboratory (T4 GPU, free tier) and a personal Apple M1 laptop.


## References


[1] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan, "Training a helpful and harmless assistant with reinforcement learning from human feedback," 2022, arXiv:2204.05862. [Online]. Available: https://arxiv.org/abs/2204.05862

[2] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukošiūtė, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. El Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, and J. Kaplan, "Constitutional AI: Harmlessness from AI feedback," 2022, arXiv:2212.08073. [Online]. Available: https://arxiv.org/abs/2212.08073

[3] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. [Online]. Available: https://arxiv.org/abs/2305.18290

[4] J. Dai, X. Pan, R. Sun, J. Ji, X. Xu, M. Liu, Y. Wang, and Y. Yang, "Safe RLHF: Safe reinforcement learning from human feedback," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2024. [Online]. Available: https://arxiv.org/abs/2310.12773

[5] T. Mu, S. Chen, G. Simmons-Edler, P. Badrinath, S. Srivastava, L. Yang, K. Haigh, J. Steinhardt, and J. Hilton, "Rule based rewards for language model safety," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. [Online]. Available: https://arxiv.org/abs/2411.01111

[6] N. Lambert, V. Pyatkin, J. Morrison, L. Miranda, B. Y. Lin, K. Chandu, N. Dziri, S. Kumar, T. Zick, Y. Choi, N. A. Smith, and H. Hajishirzi, "RewardBench: Evaluating reward models for language modeling," 2024, arXiv:2403.13787. [Online]. Available: https://arxiv.org/abs/2403.13787

[7] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, arXiv:1907.11692. [Online]. Available: https://arxiv.org/abs/1907.11692

[8] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, arXiv:1910.01108. [Online]. Available: https://arxiv.org/abs/1910.01108

[9] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. 2019 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2019. [Online]. Available: https://arxiv.org/abs/1908.10084

[10] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. [Online]. Available: https://arxiv.org/abs/2305.14314

[11] Qwen Team, "Qwen2.5 technical report," 2025, arXiv:2412.15115. [Online]. Available: https://arxiv.org/abs/2412.15115

[12] D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schiefer, K. Ndousse, A. Jones, S. Bowman, A. Chen, T. Conerly, N. DasSarma, D. Drain, N. Elhage, S. El-Showk, S. Fort, Z. Hatfield-Dodds, T. Henighan, D. Hernandez, T. Hume, J. Jacobson, S. Johnston, S. Kravec, C. Olsson, S. Ringer, E. Tran-Johnson, D. Amodei, T. Brown, N. Joseph, S. McCandlish, C. Olah, J. Kaplan, and J. Clark, "Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned," 2022, arXiv:2209.07858. [Online]. Available: https://arxiv.org/abs/2209.07858

[13] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving, "Fine-tuning language models from human preferences," 2019, arXiv:1909.08593. [Online]. Available: https://arxiv.org/abs/1909.08593

[14] T. Liu, Y. Zhao, R. Joshi, M. Khalman, M. Saleh, P. J. Liu, and J. Liu, "Statistical rejection sampling improves preference optimization," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2024. [Online]. Available: https://arxiv.org/abs/2309.06657

[15] S. Yang, A. Khashabi, N. Hayashi, and N. Peng, "SafeDPO: A simple guide to direct preference optimization with safety constraints," 2024, arXiv:2408.06925. [Online]. Available: https://arxiv.org/abs/2408.06925

[16] H. Liu, C. Sferrazza, and P. Abbeel, "Chain of hindsight aligns language models with feedback," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2024. [Online]. Available: https://arxiv.org/abs/2302.02676

[17] P. Zhai, Y. Liu, S. Zhu, and X. Liu, "Improving RLHF using contrastive rewards," 2024, arXiv:2402.12030. [Online]. Available: https://arxiv.org/abs/2402.12030