Why Does RLAIF Work At All?

Anonymous ACL submission

Abstract

Reinforcement Learning from AI Feedback (RLAIF) enables language models to improve by training on their own preference judgments, yet there is no explanation for why this bootstrapped self-improvement seemingly works for value learning. We propose the latent value hypothesis, that pretraining on large-scale data encodes human values as directions in representation space, and constitutional prompts elicit these latent values into preference judgments. We analyze this intuition under a linear model where the constitution acts as a projection operator selecting value-relevant directions. Our analysis yields several results. RLAIF improves alignment when the constitution-activated direction correlates with true values better than the model's default generation direction thus explaining the generation-judgment gap; the ceiling on RLAIF quality is determined by how well representations encode values, which scales with model capacity; and adversarial constitutions exist that can activate anti-social value directions encoded from harmful pretraining data.

1 Introduction

A puzzling phenomenon underlies modern AI. Namely, language models can improve their own safety by training on their own judgments. In Reinforcement Learning from AI Feedback (RLAIF; Bai et al., 2022; Lee et al., 2024), a model is prompted with a "constitution," a set of principles like "choose the less harmful response," and asked to judge pairs of outputs. The model then trains on these self-generated preferences. Empirically, this works. RLAIF achieves alignment quality comparable to human feedback, and models can even improve by judging their own outputs (Lee et al., 2024).

But why should this work at all? No new information enters the system, creating an apparent tension with the data processing inequality (Shannon, 1948). The model judges its own outputs according to its own understanding of the constitution. If the model already "knew" what was harmful, why did it not just avoid generating harmful content in the first place? And if it did not know, how can its judgments provide useful signal?

We propose a resolution in the latent value hypothesis. The core idea is that pretraining on large-scale data implicitly encodes human values as directions in the model's representation space, but these representations are not fully utilized during generation. The constitution prompt acts as a retrieval key that elicits these latent values into explicit preference judgments. Training on these judgments then connects the latent value representations to the output distribution.

This hypothesis makes RLAIF more legible. We posit that self-improvement for value learning is possible because knowing and doing are decoupled in language models. The model's representations encode more about human values than its default generation behavior reflects, and the constitution bridges this gap.

We formalize the latent value hypothesis under a linear model of value encoding and derive several results. We characterize the self-improvement condition, namely that RLAIF improves alignment if and only if the constitution-activated direction correlates positively with true values, and we explain when this correlation exceeds that of default generation by the "generation-judgment gap." We establish an RLAIF ceiling where the maximum achievable alignment is bounded by how well the model's representations encode values, a ceiling that scales with model capacity and pretraining diversity. We then offer a conjecture on low-rank values. Human values concentrate in a low-dimensional subspace, consistent with empirical findings that safety fine-tuning modifies few directions. We prove that adversarial constitutions exist because pretraining encodes both pro-social

and anti-social norms, there exist constitutions that activate harmful value directions, making the model worse. We show that this can explain existing empirical findings, including the existence of a "refusal direction" in base models (Arditi et al., 2024), the low-rank structure of safety fine-tuning (Pan et al., 2025), and the scaling behavior of RLAIF (Lee et al., 2024). We conclude by discussing implications for alignment practice.

2 Related Work

Concurrent work by Huang et al. (2025) formalizes self-improvement as "sharpening" by using the model as a verifier to concentrate probability mass on high-quality responses. Their work applies to reasoning tasks where correctness is verifiable as the model can check whether a proof is valid, even if generating correct proofs is difficult. They provide sample complexity bounds showing when self-improvement via SFT or RLHF is efficient.

Our work is complementary. Sharpening explains self-improvement when there is a ground-truth "correct answer" that the model can verify. We explain self-improvement for values, where there is no external ground truth; instead, the "correct answer" is encoded in the model's own latent representations from pretraining. Whereas sharpening identifies a computational bottleneck in that generation requires difficult search, while verification is easy, we identify an architectural bottleneck. Values are encoded in representations but not fully utilized by the default generation process.

Bai et al. (2022) introduced constitutional AI, using principles to generate self-critiques and revisions. They demonstrate empirically that RLAIF with a constitution can match or exceed RLHF for harmlessness. However, they do not provide a theoretical framework for why this works. Our contribution is to formalize the mechanism. The constitution elicits latent value representations that were encoded during pretraining.

A growing body of work establishes that concepts are linearly encoded in language model representations. Probing classifiers can extract syntactic and semantic information (Belinkov et al., 2017). Steering vectors can control model behavior by adding directions to activations (Turner et al., 2024; Rimsky et al., 2024). Representation engineering provides a systematic approach to finding and manipulating concept directions (Zou et al., 2025). Inference-time intervention can elicit truthful answers by shifting activations (Li et al., 2023). We build on this foundation, applying linear representation assumptions specifically to value encoding. Our Assumption 1 is directly motivated by empirical findings that harmfulness and refusal are linearly encoded (Arditi et al., 2024).

Our work also relates to scalable oversight and eliciting latent knowledge (Christiano et al., 2018; Burns et al., 2024). That line of work proposes bootstrapping alignment through recursive supervision and the system helps generate the training signal. RLAIF can be seen as a tractable instance, and constitutional judgments supervise policy updates. The ELK problem asks how to extract knowledge a model has encoded but does not report; our latent value hypothesis offers a partial answer for values specifically. The constitution acts as an elicitation mechanism, surfacing latent representations as explicit judgments. However, there is still no guarantee that the elicited direction reflects true values rather than a proxy satisfying the prompt.

Recent work studies the internal mechanisms of safety behaviors. Arditi et al. (2024) show that refusal is mediated by a single direction across many chat models, and that this direction exists even before RLHF. Pan et al. (2025) find that safety fine-tuning modifies a low-rank subspace, with effective rank approximately 1 in early layers. We provide theoretical grounding for these findings through Conjecture 9 that formalizes why safety might be low-rank.

3 Problem Setup

We formalize the components of RLAIF under a linear model of value encoding.

3.1 Representations and Values

Let $\mathcal{X}$ denote the space of prompts and $\mathcal{Y}$ the space of responses. A language model maps prompt-response pairs to representations:

$$h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d \tag{1}$$

We can think of $h(x, y)$ as the model's internal representation of responding to prompt x with response y (e.g. the residual stream at some layer, averaged over positions). For now, let us assume representations are whitened: $\mathbb{E}[hh^\top] = I$ and are centered: $\mathbb{E}[h] = 0$. This gives $\mathrm{Cov}(h) = I$.

Assumption 1 (Linear Value Encoding)**.** *There exists a direction $v^* \in \mathbb{R}^d$ such that the "true safety" of a response is a linear function of the representation:*

$$S(x, y) = \langle h(x, y), v^* \rangle + \epsilon(x, y) \tag{2}$$

where ϵ is mean-zero noise with $\mathbb{E}[\epsilon^2] = \sigma_\epsilon^2$, independent of h.

This assumption is empirically motivated. Prior work shows that harmfulness, refusal, and other safety-relevant properties are linearly encoded in language model representations (Arditi et al., 2024; Zou et al., 2025; Li et al., 2023). The noise ϵ captures aspects of harm that are not encoded in representations (e.g. context-dependent judgments, ambiguous cases). We relax this condition in Appendix Theorem 13.

Definition 1 (Encoding Quality)**.** *The encoding quality $\rho \in [0, 1]$ measures how well representations capture true harm:*

$$\rho^2 = \frac{\|v^*\|^2}{\|v^*\|^2 + \sigma_\epsilon^2} = \frac{\mathrm{Var}(\langle h, v^* \rangle)}{\mathrm{Var}(S)} \tag{3}$$

When $\rho = 1$, harm is perfectly encoded. When $\rho = 0$, representations contain no information about harm.

3.2 Generation and Judgment

We model the base policy as implicitly optimizing a linear score:

Assumption 2 (Linear Generation)**.** *The base model's log-probability decomposes as:*

$$\log P_{\text{base}}(y \mid x) = \langle h(x, y), w \rangle + g(x) \tag{4}$$

where $w \in \mathbb{R}^d$ is the "generation direction" and $g(x)$ is a prompt-dependent normalizer.

The direction w captures what the base model implicitly prefers and responses that score high on $\langle h, w \rangle$ are more likely to be generated. This direction is shaped by the pretraining objective (next-token prediction) across the entire corpus.

For constitutional judgment, we assume:

Assumption 3 (Linear Judgment)**.** *A constitution c induces pairwise preferences:*

$$J_c(y_1 \succ y_2 \mid x) = \sigma\big(\langle h(x, y_1) - h(x, y_2), v_c \rangle\big) \tag{5}$$

where σ is the sigmoid function and $v_c \in \mathbb{R}^d$ is the direction activated by constitution c.

The key modeling choice is that different constitutions activate different directions. The constitution "Principle: choose the response that is less harmful" activates a direction v_c that (ideally) aligns with v^*. A different constitution might activate a different direction.

Remark 1. *We take as given that prompting can activate specific directions in representation space as a phenomenon documented in work on steering vectors (Turner et al., 2024; Rimsky et al., 2024) and activation engineering (Zou et al., 2025). The mechanism by which prompts select directions is beyond our scope.*

3.3 Alignment Measure

We measure alignment as expected safety:

$$\mathrm{Align}(\pi) = \mathbb{E}_x \mathbb{E}_{y \sim \pi(\cdot|x)}[S(x, y)] \tag{6}$$

Higher values indicate better alignment.

4 Self-Improvement Condition

We now characterize when RLAIF improves alignment.

4.1 RLAIF as Direction Adjustment

Training on constitutional preferences adjusts the generation direction. We derive this from first principles using Direct Preference Optimization (DPO) (Rafailov et al., 2023).

DPO shows that for KL-constrained reward maximization by maximizing expected reward while staying close to a reference policy, the optimal policy takes the form:

$$\pi^*(y \mid x) = \frac{1}{Z(x)} \pi_{\text{ref}}(y \mid x) \exp(r(x, y)/\beta) \tag{7}$$

where β controls the strength of the KL penalty. Rearranging gives the implicit reward:

$$r(x, y) = \beta \log \frac{\pi^*(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta \log Z(x) \tag{8}$$

In our setting, constitutional preferences (Assumption 3) correspond to a Bradley-Terry model with implicit reward:

$$r_c(x, y) = \langle h(x, y), v_c \rangle \tag{9}$$

since $J_c(y_1 \succ y_2) = \sigma(r_c(y_1) - r_c(y_2))$.

Taking the base policy as reference ($\pi_{\text{ref}} = P_{\text{base}}$), we can derive the RLAIF policy:

Proposition 2 (RLAIF Policy)**.** *Under Assumptions 1–3, DPO on constitutional preferences with KL penalty β yields the optimal policy:*

$$\log P_{\text{RLAIF}}(y \mid x) = \langle h(x, y), w + \lambda v_c \rangle + g'(x) \tag{10}$$

where $\lambda = 1/\beta$.

Proof. The optimal policy satisfies:

$$r_c(x, y) = \beta \log \frac{\pi^*(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \text{const}(x) \tag{11}$$

$$\begin{aligned} \langle h, v_c \rangle = \beta \log \pi^*(y \mid x) - \beta \langle h, w \rangle \\ - \beta\, g(x) + \text{const}(x) \end{aligned} \tag{12}$$

Solving for $\log \pi^*$:

$$\begin{aligned} \log \pi^*(y \mid x) &= \langle h, w \rangle + \frac{1}{\beta} \langle h, v_c \rangle + g'(x) \\ &= \langle h, w + \lambda v_c \rangle + g'(x) \end{aligned} \tag{13}$$

where $\lambda = 1/\beta$ and $g'(x)$ absorbs the x-dependent terms. □

The interpretation is that RLAIF shifts the generation direction from w to $w + \lambda v_c$. The strength of the shift is controlled by $\lambda = 1/\beta$. Weaker KL regularization (smaller β) means larger shift toward the constitution direction. The policy now favors responses that score high on both the original generation direction and the constitution-activated direction.

4.2 Improvement Condition

Theorem 3 (Self-Improvement)**.** *Under Assumptions 1–3, for small λ the alignment improvement from RLAIF is:*

$$\text{Align}(\pi_{\text{RLAIF}}) - \text{Align}(\pi_{\text{base}}) = \lambda \langle \Sigma_w v_c, v^* \rangle + O(\lambda^2) \tag{14}$$

where $\Sigma_w = \text{Cov}_{\pi_{\text{base}}}(h)$ is positive definite. In particular, improvement occurs iff $\langle \Sigma_w v_c, v^* \rangle > 0$.

Proof. The expected safety (or negative harm) under policy π with linear score direction u is:

$$\mathbb{E}_{y \sim \pi}[S(x, y)] = \mathbb{E}_{y \sim \pi}[\langle h, v^* \rangle] + \mathbb{E}[\epsilon] \tag{15}$$

For an energy-based model $\pi(y) \propto \exp(\langle h(y), u \rangle)$, the expected representation is $\mathbb{E}_{y \sim \pi}[h] = \nabla_u \log Z(u)$ where $Z(u) = \int \exp(\langle h(y), u \rangle) dy$. For small perturbations δu:

$$\mathbb{E}_{\pi_{u+\delta u}}[h] \approx \mathbb{E}_{\pi_u}[h] + \Sigma_u \delta u \tag{16}$$

where $\Sigma_u = \nabla_u^2 \log Z(u) = \text{Cov}_{\pi_u}(h)$ is the representation covariance under π_u, which is positive definite.

Applying the first-order expansion with $\delta u = \lambda v_c$:

$$\mathbb{E}_{\text{RLAIF}}[h] - \mathbb{E}_{\text{base}}[h] = \lambda \Sigma_w v_c + O(\lambda^2) \tag{17}$$

Taking the inner product with v^*:

$$\mathbb{E}_{\text{RLAIF}}[\langle h, v^* \rangle] - \mathbb{E}_{\text{base}}[\langle h, v^* \rangle] = \lambda \langle \Sigma_w v_c, v^* \rangle + O(\lambda^2) \tag{18}$$

□

Corollary 4 (Self-Improvement Condition)**.** *RLAIF improves alignment (for small λ) if and only if $\langle \Sigma_w v_c, v^* \rangle > 0$. When representations are isotropic ($\Sigma_w \propto I$), this simplifies to $\langle v_c, v^* \rangle > 0$.*

The condition is intuitive. RLAIF helps when the constitution activates a direction that correlates positively with true safety (so that preferring high-v_c responses means preferring safe responses).

4.3 The Generation-Judgment Gap

Why should $\langle v_c, v^* \rangle$ exceed $\langle w, v^* \rangle$? This is the core question. Why does constitutional judgment access value information better than default generation?

Proposition 5 (Generation-Judgment Gap)**.** *Suppose:*

1. *The generation direction w is shaped by next-token prediction across a corpus where fraction $\eta \ll 1$ is value-relevant (ethics discussions, content moderation, etc.) and fraction $1 - \eta$ is value-neutral.*
2. *The constitution is designed to query values directly, achieving alignment $\langle v_c, v^* \rangle = \alpha_c \approx 1$.*

Then $\langle w, v^* \rangle \approx \eta$ and the gap is:

$$\langle v_c, v^* \rangle - \langle w, v^* \rangle \approx 1 - \eta \tag{19}$$

The intuition is that pretraining optimizes for predicting all tokens, most of which have nothing to do with values. The generation direction w is therefore "diluted" and only a small component

points toward v^*. In contrast, the constitution explicitly asks about harm, so it activates a direction v_c that is targeted at the value-relevant subspace.

This is a plausible explanation for the generation-judgment gap. The model "knows" about values (they are encoded in representations), but default generation doesn't fully utilize this knowledge (because w is optimized for a different objective). The constitution retrieves the relevant knowledge.

4.4 A Toy Example

To make things more concrete, consider a minimal example in $\mathbb{R}^2$.

Setup. Let representations live in $\mathbb{R}^2$ with:

- Safety direction: $v^* = (1, 0)$
- Generation direction: $w = (0.1, 0.995)$ (mostly orthogonal to v^*)
- Constitution direction: $v_c = (0.95, 0.31)$ (well-aligned with v^*)

The generation direction w has $\langle w, v^* \rangle = 0.1$. only 10% of its magnitude points toward the safety-relevant direction. This reflects a base model trained mostly on value-neutral text. The constitution direction v_c has $\langle v_c, v^* \rangle = 0.95$; it nearly recovers the true safety direction.

Under the base policy, responses y with high $\langle h(y), w \rangle$ are favored. Since w is nearly orthogonal to v^*, the policy is nearly indifferent to harm. A harmful response and a benign response with similar projections onto w are equally likely.

After RLAIF with $\lambda = 1$, the new generation direction is:

$$w' = w + \lambda v_c = (0.1 + 0.95, 0.995 + 0.31) = (1.05, 1.305) \tag{20}$$

Normalizing: $w' \approx (0.63, 0.78)$.

The new direction has $\langle w', v^* \rangle \approx 0.63$; the policy now substantially penalizes harm. The first-order alignment improvement is:

$$\Delta \approx \lambda \langle \Sigma_w v_c, v^* \rangle = \lambda (v^*)^\top \Sigma_w v_c \tag{21}$$

For isotropic $\Sigma_w = \sigma^2 I$, this gives $\Delta = \lambda \sigma^2 \langle v_c, v^* \rangle = 0.95 \lambda \sigma^2 > 0$.

Now suppose an adversarial constitution activates $v_c^{\text{adv}} = (-0.8, 0.6)$, which has $\langle v_c^{\text{adv}}, v^* \rangle = -0.8$. After RLAIF:

$$w'' = w + \lambda v_c^{adv} = (0.1 - 0.8, 0.995 + 0.6) = (-0.7, 1.595) \tag{22}$$

Normalizing: $w'' \approx (-0.40, 0.92)$.

Now $\langle w'', v^* \rangle \approx -0.40$. The policy actively favors unsafe responses. The alignment change is $\Delta = -0.8C < 0$. The adversarial constitution made the model worse than the base policy.

This toy example illustrates the core dynamics: (1) the base policy underutilizes value information because w is diluted; (2) a good constitution recovers value information by activating $v_c \approx v^*$; (3) an adversarial constitution can activate directions anti-correlated with v^*, degrading alignment.

5 RLAIF Ceiling

How good can RLAIF get? We posit the ceiling is determined by representation quality.

Theorem 6 (RLAIF Ceiling)**.** *For any constitution c, if $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$:*

$$\text{Align}^* - \text{Align}(\pi^c_{\text{RLAIF}}) \geq \frac{(1-\rho)\sigma_S}{\sqrt{\pi}} \tag{23}$$

where Align^* is the alignment achievable with oracle access to S, $\sigma_S = \text{std}(S)$, and ρ is the encoding quality.

Proof sketch. The best possible constitution sets $v_c = v^*$, reducing RLAIF to selection based on the proxy $\langle h, v^* \rangle$ rather than true safety $S = \langle h, v^* \rangle + \epsilon$. This is a standard noisy selection problem: the proxy has correlation ρ with the truth. For Gaussian noise, the regret from proxy-based selection scales as $(1 - \rho)\sigma_S$. See Appendix A for the full derivation. □

Corollary 7 (Scaling)**.** *If encoding quality ρ increases with model capacity (parameters, pretraining data), then the RLAIF ceiling increases with scale.*

This explains the empirical finding that larger models make better RLAIF labelers (Lee et al., 2024). Larger models encode values better (higher ρ), so their constitutional judgments are more accurate, and training on them yields better alignment.

6 Conjecture on Low-Rank Values

Empirical work finds that safety-related representations have low effective rank. We formalize this as a conjecture and discuss possible explanations.

Let $\Sigma = \mathbb{E}[hh^\top]$ be the representation covariance (identity under our whitening assumption, but consider the pre-whitened structure). Let $\lambda_1 \geq$

$\lambda_2 \geq \ldots$ be its eigenvalues with eigenvectors $u_1, u_2, \ldots$

Definition 8 (Effective Dimension)**.** *The effective dimension of the representation is:*

$$d_{\text{eff}} = \frac{(\sum_i \lambda_i)^2}{\sum_i \lambda_i^2} \tag{24}$$

Conjecture 9 (Value Concentration)**.** *The true harm direction v^* lies predominantly in the top eigenspace:*

$$\sum_{i=1}^{k} \langle v^*, u_i \rangle^2 \approx \|v^*\|^2 \quad \textit{for } k = O(d_{\text{eff}}) \tag{25}$$

This conjecture is supported by empirical findings. Arditi et al. (2024) show that refusal is mediated by a single direction across many models. Pan et al. (2025) find that safety fine-tuning induces representation shifts with effective rank approximately 1 in early layers, increasing but remaining low throughout the network.

Why might values concentrate in few directions? We see several possibilities, though none constitute a proof:

Statistical frequency. Value-relevant distinctions like toxic/non-toxic and harmful/benign appear frequently in pretraining. Contexts involving these distinctions push representations in consistent directions, building up high-variance components. Subtle ethical nuances appear rarely and inconsistently, contributing little variance. If v^* captures common distinctions, it naturally aligns with high-variance (top) eigenvectors.

Intrinsic simplicity. Human values may be fundamentally low-dimensional. Core ethical principles such as avoid harm, be honest, and respect autonomy might span most of the variation in human moral judgments. If ground-truth values are simple, the encoding v^* will be too.

Capacity allocation. Neural networks may allocate representational capacity in proportion to predictive usefulness. Common patterns get dedicated dimensions; rare patterns share capacity. Value-relevant features, being common, get dedicated directions.

We do not commit to any particular explanation. Understanding why values are low-rank, if indeed they are in general, is an important open question. Our work is consistent with low-rank values but does not derive this property from first principles.

The practical implication, if the conjecture holds, is that safety alignment is achievable by modifying few directions. This is both encouraging (alignment may be tractable) and concerning (few directions means small attack surface for adversarial manipulation).

7 Adversarial Constitutions

An implication of our hypothesis is that because pretraining encodes diverse norms, adversarial constitutions can make alignment worse.

Corollary 10 (Adversarial Constitutions Exist)**.** *Suppose the pretraining corpus contains both pro-social content (encoding direction v^+ with $\langle v^+, v^* \rangle > 0$) and anti-social content (encoding direction v^- with $\langle v^-, v^* \rangle < 0$). Then there exists a constitution c^* such that:*

$$\langle v_{c^*}, v^* \rangle < 0 \tag{26}$$

and therefore:

$$\text{Align}(\pi_{\text{RLAIF}}^{c^*}) < \text{Align}(\pi_{\text{base}}) \tag{27}$$

Proof. The map $c \mapsto v_c$ from constitutions to activated directions spans some subspace of $\mathbb{R}^d$ (the "promptable" subspace). If this subspace contains a direction v_c with $\langle v_c, v^* \rangle < 0$ which we take as a mild assumption, since pretraining on anti-social content (e.g. forums endorsing manipulation, deception, or violence) plausibly encodes directions retrievable by appropriately crafted constitutions then by Theorem 3, RLAIF with such a constitution yields negative alignment improvement. □

Remark 2. *Adversarial constitutions need not be obviously malicious. A constitution emphasizing "edginess," "authenticity," or "not being preachy" might activate directions that correlate negatively with safety. The risk surface is subtle.*

8 Accounting for Existing Evidence

Our account unifies several empirical findings that previously lacked theoretical explanation.

8.1 Refusal Direction in Base Models

Arditi et al. (2024) show that refusal behavior is mediated by a single direction across 13 chat models up to 72B parameters. Remarkably, this direction can be found even in base models before any RLHF or safety fine-tuning. Erasing the direction from the residual stream disables refusal; adding it induces refusal on benign prompts.

Assumption 1 posits that the harm direction v^* exists as a property of pretraining, not alignment

training. The internet contains extensive discussion of harmful vs. benign content such as content moderation, ethics debates, and safety guidelines, and pretraining compresses this into a direction in representation space. Alignment training increases $\langle w, v^* \rangle$ (shifts generation toward the value direction) but does not create v^*. The refusal direction exists in base models because the knowledge of what is harmful exists in base models.

8.2 Low-Rank Safety Subspace

Pan et al. (2025) analyze the representation shifts induced by safety fine-tuning on Llama 3 8B. They find that the effective rank of these shifts is approximately 1 in early layers, increasing in later layers but remaining low throughout. A single dominant direction governs refusal behavior, with secondary directions encoding features like hypothetical narrative and role-playing.

Our Conjecture 9 formalizes this finding. If values concentrate in few directions because value-relevant distinctions (toxic vs. non-toxic, harmful vs. benign, safe vs. unsafe) are high-frequency in the pretraining corpus, then these distinctions would induce high-variance directions in representation space. Subtle ethical nuances, by contrast, are rare and induce low-variance directions. The harm direction v^* would align with the high-variance subspace, which has low effective dimension.

The finding that secondary directions encode specific features (hypothetical framing, roleplay) is also consistent with our hypothesis. Different constitutions might activate different combinations of these directions, explaining why some jailbreaks work by invoking fictional scenarios.

8.3 RLAIF Scales with Labeler Size

Lee et al. (2024) systematically vary the size of the AI labeler in RLAIF experiments. They find a strong positive relationship that alignment quality decreases by 4% when substituting PaLM 2 L with PaLM 2 S, and by another 11% with PaLM 2 XS. This scaling behavior is consistent across tasks.

Our Corollary 7 is consistent with this. Larger models have higher representation quality ρ and they encode values more accurately because they have more capacity and see more pretraining data. Higher ρ means the RLAIF ceiling is higher. The constitutional judgments are more accurate, and training on them yields better alignment.

This also explains a puzzling finding from the same paper. RLAIF works even when the labeler is the same size as the policy. Our work makes this intelligible. The labeler and policy are the same model, but they access the model's knowledge differently. The labeler (with constitution) accesses value representations via v_c; the policy (default generation) accesses them via w. If $\langle v_c, v^* \rangle > \langle w, v^* \rangle$, which Proposition 5 argues is typical, then training on the labeler's judgments improves the policy.

8.4 Self-Improvement Without External Information

The deepest puzzle is that RLAIF works at all. No external information enters the system; the model judges its own outputs according to its own understanding of the constitution. If this worked by simply memorizing "refuse harmful requests," the base model would already do that.

Our work resolves this puzzle. Self-improvement is possible because knowing and doing are decoupled. The representation h encodes harm (via v^*), but default generation (via w) does not fully utilize this encoding. The constitution retrieves the value-relevant information and makes it available for training. The model improves not by learning new facts, but by eliciting knowledge it already had.

9 Discussion

9.1 Implications for Alignment Practice

Our work suggests that RLAIF quality is bottlenecked by representation quality ρ, not by the amount of preference data. This has practical implications for how alignment resources should be allocated.

Scaling labeler model size may be more important than scaling preference dataset size. If the ceiling is determined by how well the labeler encodes values, then a larger labeler with fewer preference pairs may outperform a smaller labeler with more pairs. This is consistent with the empirical finding that labeler size strongly affects RLAIF quality (Lee et al., 2024).

Then, we posit that RLAIF will fail for value judgments that are not well-represented in pretraining data. Novel ethical dilemmas, culture-specific norms, and post-training value shifts cannot be elicited if they were never encoded. For these cases, human feedback remains necessary.

The generation-judgment gap suggests that evaluation and generation should be treated as distinct capabilities. A model might excel at judging harmful content while still generating it under certain prompts, because the constitution-activated direction and the default generation direction access different information.

9.2 Constitution Design as Attack Surface

Theorem 10 identifies constitution design as a potential attack surface, though one with significant barriers to exploitation. Unlike prompt injection, which any user can attempt, deploying a malicious constitution requires access to the training pipeline, therefore limiting the threat model to insiders, supply chain compromises, or organizations that train their own models. The more realistic concern may be inadvertent harm. A well-intentioned constitution emphasizing authenticity," not being preachy," or "respecting user autonomy" might activate directions that correlate negatively with safety, without anyone recognizing the problem. The mapping from constitution text to activated direction is complex and not fully understood, making it difficult to predict which constitutions are safe.

This suggests defensive measures even absent adversarial intent. Constitutions should be tested empirically for their effect on downstream behavior, not just inspected for surface-level safety. Multiple constitutions might be ensembled to reduce the impact of any single problematic direction. And the constitution design process should be treated with appropriate care, recognizing that subtle wording choices can have outsized effects on alignment outcomes.

9.3 Complementarity of RLAIF and RLHF

Our hypothesis suggests that RLAIF and RLHF are complementary rather than competing approaches. RLAIF elicits values that are already encoded in representations as the high-frequency distinctions that appear often in pretraining (toxic vs. non-toxic, helpful vs. unhelpful). RLHF provides signal for values that are not well-encoded such as rare distinctions, nuanced judgments, and norms that emerged after pretraining.

Practical alignment may therefore combine both approaches. RLAIF provides broad coverage at low cost, handling the 95% of cases where pretraining knowledge suffices. RLHF provides precision on the long tail, correcting systematic biases and encoding new values. The information gained, we conject, is approximately additive. RLAIF contributes what can be elicited from pretraining, RLHF contributes what humans know beyond pretraining.

9.4 Iterated RLAIF

Our analysis considers a single round of RLAIF of judge, train, deploy. In practice, RLAIF is often iterated. The trained model becomes the new judge, generating preferences for the next round of training. What happens under iteration?

Our work suggests two possibilities. If the constitution consistently activates the same direction v_c regardless of the base model, then iterated RLAIF converges to a fixed point where the generation direction aligns with v_c. Alignment improves monotonically until it saturates at the representation quality ceiling.

However, if the trained model's representations shift in ways that change which direction the constitution activates, the dynamics become more complex. The direction $v_c^{(t)}$ at iteration t might drift, potentially in harmful directions. This suggests that iterated RLAIF should be monitored for distributional shift, and that constitutions should be stress-tested on models at various stages of training. We leave a full analysis of iterated RLAIF dynamics to future work.

10 Conclusion

We proposed the latent value hypothesis as an explanation that RLAIF works because pretraining encodes human values in representations, and constitutional prompts elicit these values into judgments. Under a linear model, we derived conditions for self-improvement, characterized the RLAIF ceiling, explained the structure of values, and identified adversarial constitutions as a risk. Our hypothesis unifies scattered empirical findings and provides a theoretical grounding for constitutional AI.

The core of the hypothesis is that self-improvement value bootstrapping is possible because knowing and doing are decoupled in language models. The model knows more about values than its default behavior reflects. RLAIF bridges this gap not by creating new knowledge, but by eliciting knowledge that was there already.

Limitations

Our work rests on several simplifying assumptions that may not hold in practice. We discuss these limitations and their implications.

Assumption 1 posits that harm is a linear function of representations. While empirically motivated by work on refusal directions and linear probes, real value encoding is likely more complex. Harmfulness may depend on nonlinear interactions between features, context-dependent weightings, or hierarchical structure that linear models cannot capture.

If value encoding is substantially nonlinear, our theorems provide only a first-order approximation. The qualitative insights that constitutions select directions, that representation quality bounds RLAIF, that adversarial constitutions exist may still hold, but the quantitative predictions would need revision. Extending our framework to nonlinear value encoding (e.g. using kernel methods or neural network function classes) is an important direction for future work. We discuss this further in Appendix C.

We take as given that constitutions activate specific directions in representation space (Assumption 3), but we do not model how this mapping works. In reality, the relationship between constitution text and activated direction is mediated by complex in-context learning mechanisms that are not fully understood. This is a significant gap. Our framework cannot predict, from constitution text alone, what direction will be activated. It cannot explain why some constitutions are more effective than others, or how to design constitutions that reliably activate safety-relevant directions. A complete theory of RLAIF would need to incorporate a model of how prompts select directions. Essentially, it requires a theory of in-context learning for value elicitation, which is outside the scope of this work.

We use "values" and "harm" loosely to refer to the properties that safety alignment aims to optimize. But human values are heterogeneous, contested, and context-dependent. Different communities have different norms. What counts as harmful depends on who is affected and how. We also treat v^* as a single ground-truth direction, which is a strong simplification. In reality, there may be multiple value directions corresponding to different ethical frameworks, cultural norms, or stakeholder interests. The "true harm" function H may not be well-defined, or may be defined differently by different groups.

This limitation affects the interpretation of our results. When we say RLAIF improves alignment, we mean alignment with whatever values are encoded in pretraining which reflect the biases and distributions of internet text. Whether these are the "right" values is a normative question our hypothesis does not address.

Our analysis is static as we characterize the outcome of RLAIF without modeling the training dynamics. We assume that preference optimization converges to the policy described in Proposition 2, but we do not analyze the optimization trajectory, sample complexity, or potential failure modes during training. In practice, preference optimization can fail in various ways like reward hacking, distributional shift, mode collapse, or optimization instability. These failure modes are orthogonal to the representation-level analysis we provide. A complete understanding of RLAIF would need to combine our representation-level hypothesis with an analysis of optimization dynamics.

While we cite existing empirical findings as evidence for our hypothesis, we do not provide new experiments. Our theorems make predictions that could be tested, such as that constitution effectiveness correlates with direction alignment, that RLAIF ceiling scales with model size, that adversarial constitutions degrade alignment. We view the current paper as providing a hypothesis that organizes existing evidence and generates testable predictions. Empirical validation is important future work.

References

Andy Arditi, Oscar Balcells Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022. Constitutional ai: Harmlessness from ai feedback. *Preprint*, arXiv:2212.08073.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural ma-

chine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. 2024. Weak-to-strong generalization: eliciting strong capabilities with weak supervision. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Paul Christiano, Buck Shlegeris, and Dario Amodei. 2018. Supervising strong learners by amplifying weak experts. *Preprint*, arXiv:1810.08575.

Audrey Huang, Adam Block, Dylan J Foster, Dhruv Rohatgi, Cyril Zhang, Max Simchowitz, Jordan T. Ash, and Akshay Krishnamurthy. 2025. Self-improvement in language models: The sharpening mechanism. In *The Thirteenth International Conference on Learning Representations*.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024. Rlaif vs. rlhf: scaling reinforcement learning from human feedback with ai feedback. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: eliciting truthful answers from a language model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Wenbo Pan, Zhichao Liu, Qiguang Chen, Xiangyang Zhou, Yu Haining, and Xiaohua Jia. 2025. The hidden dimensions of LLM alignment: A multi-dimensional analysis of orthogonal safety directions. In *Forty-second International Conference on Machine Learning*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:623–656.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering. *Preprint*, arXiv:2308.10248.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2025. Representation engineering: A top-down approach to ai transparency. *Preprint*, arXiv:2310.01405.

A Proof of Theorem 6

We analyze the regret from proxy-based selection. Consider choosing between two responses y_1, y_2 with:

- True safety values $S_i = \langle h_i, v^* \rangle + \epsilon_i$
- Proxy values $P_i = \langle h_i, v^* \rangle$

where $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$ i.i.d. The oracle selects $i^* = \arg\max_i S_i$; RLAIF selects $\hat{i} = \arg\max_i P_i$.

Step 1: Correlation structure. The pairs (P_i, S_i) are jointly normal with:

$$\mathrm{Var}(P_i) = \|v^*\|^2 \tag{28}$$

$$\mathrm{Var}(S_i) = \|v^*\|^2 + \sigma_\epsilon^2 = \sigma_S^2 \tag{29}$$

$$\begin{aligned}\mathrm{Corr}(P_i, S_i) &= \frac{\mathrm{Cov}(P_i, S_i)}{\sigma_P \sigma_S} \\ &= \frac{\|v^*\|^2}{\|v^*\|\sigma_S} = \frac{\|v^*\|}{\sigma_S} = \rho \end{aligned} \tag{30}$$

Step 2: Oracle expected value. For i.i.d. $S_1, S_2 \sim \mathcal{N}(0, \sigma_S^2)$, the expected maximum is:

$$\mathbb{E}[\max(S_1, S_2)] = \frac{\sigma_S}{\sqrt{\pi}} \tag{31}$$

Step 3: Proxy-based expected value. By symmetry:

$$\mathbb{E}[S_{\hat{i}}] = \mathbb{E}[S_1 \mid P_1 > P_2] \tag{32}$$

The conditional expectation of S_1 given P_1 follows from standard Gaussian regression:

$$\mathbb{E}[S_1 \mid P_1] = \rho \frac{\sigma_S}{\|v^*\|} P_1 \tag{33}$$

Therefore:

$$\begin{aligned}\mathbb{E}[S_1 \mid P_1 > P_2] &= \rho \frac{\sigma_S}{\|v^*\|} \mathbb{E}[P_1 \mid P_1 > P_2] \\ &= \rho \frac{\sigma_S}{\|v^*\|} \cdot \frac{\|v^*\|}{\sqrt{\pi}} = \frac{\rho \sigma_S}{\sqrt{\pi}} \end{aligned} \tag{34}$$

Step 4: Regret bound. The regret is:

$$R = \mathbb{E}[\max(S_1, S_2)] - \mathbb{E}[S_{\hat{i}}] \quad (35)$$
$$= \frac{\sigma_S}{\sqrt{\pi}} - \frac{\rho\sigma_S}{\sqrt{\pi}} \quad (36)$$
$$= \frac{(1-\rho)\sigma_S}{\sqrt{\pi}} \quad (37)$$

This is the regret for the optimal constitution ($v_c = v^*$). Any other constitution achieves correlation at most ρ between proxy and truth, yielding weakly higher regret. The bound extends from binary choice to general policies by standard arguments.

B Relaxing Assumptions

The main text develops the latent value hypothesis under simplifying assumptions like linear value encoding, whitened representations, a single value direction v^*, and a linear constitution-to-direction map. This section examines the robustness of the framework by relaxing each assumption in turn. Notation follows the main text throughout.

C Relaxing Linearity of Value Encoding

C.1 Setup

We replace Assumption 1 with a smooth generalization.

Definition 11 (Smooth Value Encoding)**.** ***(Assumption 1′).*** *There exists a smooth function $f : \mathbb{R}^d \to \mathbb{R}$ such that*

$$S(x, y) = f(h(x, y)) + \epsilon(x, y)$$

where ϵ is mean-zero noise independent of h.

The original assumption is the special case $f(h) = \langle h, v^* \rangle$. Assumptions 2–3 are unchanged, so the RLAIF policy is still $\pi_{\text{RLAIF}}(y|x) \propto \pi_{\text{base}}(y|x) \exp(\lambda \langle h(x, y), v_c \rangle)$.

C.2 General Self-Improvement Condition

Lemma 12 (Exponential Tilt)**.** *Let $\pi_\lambda(y) \propto \pi_0(y) \exp(\lambda \langle h(y), v_c \rangle)$. For any test function ϕ with $E_{\pi_0}[\phi(h)^2] < \infty$:*

$$E_{\pi_\lambda}[\phi(h)] = E_{\pi_0}[\phi(h)] + \lambda \operatorname{Cov}_{\pi_0}(\phi(h), \langle h, v_c \rangle) + O(\lambda^2). \quad (38)$$

Proof. Write $Z(\lambda) = E_{\pi_0}[\exp(\lambda \langle h, v_c \rangle)]$, so that

$$E_{\pi_\lambda}[\phi(h)] = \frac{E_{\pi_0}[\phi(h) \exp(\lambda \langle h, v_c \rangle)]}{Z(\lambda)}.$$

Expanding $\exp(\lambda \langle h, v_c \rangle) = 1 + \lambda \langle h, v_c \rangle + O(\lambda^2)$:

$$N(\lambda) = E_{\pi_0}[\phi(h)] + \lambda\, E_{\pi_0}[\phi(h) \langle h, v_c \rangle] + O(\lambda^2), \quad (39)$$
$$Z(\lambda) = 1 + \lambda\, E_{\pi_0}[\langle h, v_c \rangle] + O(\lambda^2). \quad (40)$$

Inverting $Z(\lambda)$ to first order and multiplying:

$$\frac{N(\lambda)}{Z(\lambda)} = \big(E[\phi] + \lambda\, E[\phi \langle h, v_c \rangle]\big) \times \big(1 - \lambda\, E[\langle h, v_c \rangle] + O(\lambda^2)\big) \quad (41)$$
$$= E[\phi] + \lambda \Big(E[\phi \langle h, v_c \rangle] - E[\phi]\, E[\langle h, v_c \rangle] \Big) + O(\lambda^2) \quad (42)$$
$$= E[\phi] + \lambda \operatorname{Cov}(\phi(h), \langle h, v_c \rangle) + O(\lambda^2). \quad (43)$$

□

Theorem 13 (Nonlinear Self-Improvement)**.** *Under Assumption 1′ and Assumptions 2–3, for small λ:*

$$\Delta\text{Align} = \lambda \operatorname{Cov}_{\text{base}}\big(f(h),\ \langle h, v_c \rangle\big) + O(\lambda^2). \quad (44)$$

RLAIF improves alignment if and only if $\operatorname{Cov}_{\text{base}}\big(f(h), \langle h, v_c \rangle\big) > 0$.

Proof. The alignment change is $\Delta\text{Align} = E_{\text{RLAIF}}[S] - E_{\text{base}}[S] = E_{\text{RLAIF}}[f(h)] - E_{\text{base}}[f(h)]$, since $E[\epsilon] = 0$ under both policies (noise is independent of h, and the policy only reweights via h). Apply Lemma 12 with $\phi = f$. □

Remark 3 (Recovery of Linear Case)**.** *When $f(h) = \langle h, v^* \rangle$:*

$$\operatorname{Cov}(f(h), \langle h, v_c \rangle) = \operatorname{Cov}(\langle h, v^* \rangle, \langle h, v_c \rangle) = v^{*\top} \Sigma_w\, v_c = \langle \Sigma_w v_c, v^* \rangle, \quad (45)$$

recovering Theorem 3 of the main text.

C.3 Gaussian Representations and Stein's Lemma

Under a distributional assumption on representations, we can give the nonlinear improvement condition a directional interpretation.

Theorem 14 (Gaussian Nonlinear Improvement)**.** *If $h \sim \mathcal{N}(\mu, \Sigma_w)$ under the base policy, and f is differentiable with $E[\|\nabla f(h)\|] < \infty$, then:*

$$\Delta\text{Align} = \lambda \langle \Sigma_w v_c,\ \bar{v}^* \rangle + O(\lambda^2)$$

where $\bar{v}^* := E_{\text{base}}[\nabla f(h)]$ is the population-average gradient of the safety function. RLAIF improves alignment if and only if $\langle \Sigma_w v_c, \bar{v}^* \rangle > 0$.

Proof. By Stein's lemma, for $h \sim \mathcal{N}(\mu, \Sigma_w)$ and differentiable f with the requisite integrability:

$$\mathrm{Cov}(f(h), h_i) = \sum_j (\Sigma_w)_{ij} \, E\left[\frac{\partial f}{\partial h_j}(h)\right].$$

In vector form: $\mathrm{Cov}(f(h), h) = \Sigma_w E[\nabla f(h)] = \Sigma_w \bar{v}^*$. Therefore:

$$\begin{aligned} \mathrm{Cov}(f(h), \langle h, v_c \rangle) &= v_c^\top \, \mathrm{Cov}(f(h), h) \\ &= v_c^\top \Sigma_w \bar{v}^* \\ &= \langle \Sigma_w v_c, \bar{v}^* \rangle. \end{aligned} \tag{46}$$

Substituting into Theorem 13 gives the result. □

Remark 4 (Interpretation)**.** *Theorem 14 is the precise sense in which the linear theory is a local approximation. The true value direction v^* generalizes to the expected gradient $\bar{v}^* = E[\nabla f(h)]$, a global average of local value directions. For linear f, $\nabla f = v^*$ everywhere, so $\bar{v}^* = v^*$. For mildly nonlinear f, $\bar{v}^*$ is close to ∇f evaluated at the mean, so the linear approximation is good. The approximation breaks down when $\nabla f(h)$ varies substantially across the support of the base distribution, so that $\bar{v}^*$ is a poor summary of the value structure.*

C.4 Non-Monotone Values

The framework breaks down for non-monotone value functions where no single direction means "safer."

Proposition 15 (Non-Monotone Failure)**.** *Let $f(h) = -(\langle h, v \rangle - \tau)^2$ for some direction v and target τ. Then $\bar{v}^* = -2(E[\langle h, v \rangle] - \tau)v$, and the sign of $\bar{v}^*$ depends on the base model's position relative to τ. No constitution is uniformly improving across the support of h.*

Proof. $\nabla f(h) = -2(\langle h, v \rangle - \tau)v$, so $\bar{v}^* = E[\nabla f(h)] = -2(E[\langle h, v \rangle] - \tau)v$. If $E[\langle h, v \rangle] < \tau$, then $\bar{v}^*$ points along $+v$; if $E[\langle h, v \rangle] > \tau$, it points along $-v$. A constitution activating $v_c = v$ helps in the first case and hurts in the second. Moreover, even when $\bar{v}^*$ has a definite sign, individual samples h with $\langle h, v \rangle$ on the wrong side of τ are pushed further from the optimum. □

Remark 5. *This is, as far as our best efforts are concerned, a genuine limitation. The framework applies to monotone value dimensions (more honest is better, less harmful is better) but not to U-shaped or context-dependent value functions. Values like assertiveness, formality, or detail-level, where there is an optimal amount and deviations in either direction are bad, cannot be captured by a single improvement condition. The generation-judgment gap story loses its force as there is no fixed value direction for the constitution to retrieve.*

D Non-Whitened Representations

Replace Assumption: $E[hh^\top] = \Sigma \neq I$.

Every instance of $\langle v_c, v^* \rangle$ becomes $\langle v_c, \Sigma v^* \rangle$ or equivalently $v_c^\top \Sigma v^*$. The self-improvement condition is:

$$v_c^\top \Sigma_w v^* > 0$$

where $\Sigma_w = \mathrm{Cov}_{\pi_{\text{base}}}(h)$, which we already use in Theorem 3. The encoding quality becomes:

$$\rho^2 = \frac{v^{*\top} \Sigma v^*}{v^{*\top} \Sigma v^* + \sigma_\epsilon^2}.$$

The RLAIF ceiling (Theorem 6) carries through with $\|v^*\|^2$ replaced by $v^{*\top} \Sigma v^*$.

No qualitative change. This relaxation is bookkeeping.

E Multiple Value Directions

E.1 Setup

Replace the single value direction with m value objectives:

$$S_i(x, y) = \langle h(x, y), v_i^* \rangle + \epsilon_i(x, y), \quad i = 1, \ldots, m$$

where $v_1^*, \ldots, v_m^*$ are (not necessarily orthogonal) value directions, and ϵ_i are independent noise terms. These correspond to distinct normative criteria such as helpfulness, harmlessness, honesty, respect for autonomy, etc.

Definition 16 (Multi-Objective Alignment)**.** *The alignment profile of policy π is the vector:*

$$\mathbf{A}(\pi) = \big(E_\pi[S_1], \ldots, E_\pi[S_m]\big) \in \mathbb{R}^m.$$

Aggregate alignment may be defined via a scalarization $\mathrm{Align}(\pi) = g(\mathbf{A}(\pi))$ for some aggregation function g. Common choices:

- **Weighted sum:** $g(\mathbf{a}) = \sum_i \alpha_i a_i$ with $\alpha_i > 0$.

- **Worst-case:** $g(\mathbf{a}) = \min_i a_i$.
- **Constrained:** maximize a_1 subject to $a_i \geq \tau_i$ for $i \geq 2$.

E.2 Per-Objective Improvement

Since RLAIF still shifts the generation direction by λv_c, the improvement on each objective is (applying Theorem 3 per objective):

$$\begin{aligned}\Delta_i &:= E_{\text{RLAIF}}[S_i] - E_{\text{base}}[S_i] \\ &= \lambda \langle \Sigma_w v_c, v_i^* \rangle + O(\lambda^2). \qquad (47)\end{aligned}$$

The vector of improvements is:

$$\mathbf{\Delta} = \lambda\, V^{*\top} \Sigma_w v_c + O(\lambda^2)$$

where $V^* = [v_1^* \mid \cdots \mid v_m^*] \in \mathbb{R}^{d \times m}$.

E.3 Pareto Structure

Definition 17 (Pareto-Improving Constitution)**.** *A constitution c is* Pareto-improving *if $\Delta_i > 0$ for all $i = 1, \ldots, m$, i.e. it improves every value dimension simultaneously.*

Proposition 18 (Pareto-Improving Cone)**.** *Define the cone*

$$\mathcal{C}_{\text{Pareto}} = \{d \in \mathbb{R}^d : \langle d, v_i^* \rangle > 0 \textit{ for all } i = 1, \ldots, m\}.$$

A constitution c is Pareto-improving if and only if $\Sigma_w v_c \in \mathcal{C}_{\text{Pareto}}$.

Proof. Immediate from (47). □

Proposition 19 (Pareto Cone Geometry)**.** *The Pareto-improving cone $\mathcal{C}_{\text{Pareto}}$ is:*

1. *Nonempty if and only if there exists $d \in \mathbb{R}^d$ with $\langle d, v_i^* \rangle > 0$ for all i.*
2. *A full half-space when $m = 1$ (the original single-objective case).*
3. *Generically nonempty when $m \leq d$ and the v_i^* do not span a full half-space.*
4. *Empty if and only if $0 \in \operatorname{conv}(v_1^*, \ldots, v_m^*)$, i.e. the value directions are so conflicting that zero is in their convex hull.*

Proof. Part (d) follows from the separating hyperplane theorem. $\mathcal{C}_{\text{Pareto}} = \emptyset$ iff there is no d with $V^{*\top} d > 0$ (componentwise), iff the system $V^{*\top} d > 0$ is infeasible, iff by Gordan's theorem there exist $\alpha_i \geq 0$, not all zero, with $\sum_i \alpha_i v_i^* = 0$, iff $0 \in \operatorname{conv}(v_1^*, \ldots, v_m^*)$ (after normalizing α). □

Remark 6 (Interpretation)**.** *The convex hull condition $0 \in \operatorname{conv}(v_1^*, \ldots, v_m^*)$ characterizes when values are fundamentally irreconcilable when no single intervention can improve all of them simultaneously. When this holds, every constitution necessarily trades off some values against others, and alignment becomes a question of which trade-offs are acceptable rather than whether improvement is possible.*

For typical safety objectives (helpful, harmless, honest), $0 \notin \operatorname{conv}(v_1^*, v_2^*, v_3^*)$ is plausible since these values are positively correlated in most contexts. The Pareto cone is narrow but nonempty. The difficulty of alignment is not that improvement is impossible, but that the cone is narrow as most constitutions improve some objectives while degrading others.

E.4 Trade-Off Characterization

Definition 20 (Constitution Trade-Off Profile)**.** *For constitution c, define the trade-off profile:*

$$\tau(c) = \big(\mathit{sign}(\Delta_1), \ldots, \mathit{sign}(\Delta_m)\big) \in \{-1, 0, +1\}^m.$$

Proposition 21 (Trade-Off Classification)**.** *A constitution c is:*

1. ***Pareto-improving*** *if $\tau(c) = (+1, \ldots, +1)$.*
2. ***Trade-off inducing*** *if $\tau(c)$ has both positive and negative entries.*
3. ***Pareto-degrading*** *if $\tau(c) = (-1, \ldots, -1)$.*

Remark 7 (Adversarial Constitutions Revisited)**.** *Theorem 10 (adversarial constitutions exist) extends naturally. In the multi-objective setting, an adversarial constitution need not degrade all objectives. It suffices to degrade one critical objective while improving others. A constitution that makes the model more helpful but less safe is adversarial in a meaningful sense, but harder to detect because the aggregate signal is ambiguous.*

Formally, if $g(\mathbf{a}) = \min_i a_i$ (worst-case aggregation), then a constitution that improves $m-1$ objectives while degrading one will reduce alignment whenever the degraded objective was already the bottleneck.

E.5 The Helpful–Harmless Trade-Off

Consider the canonical case $m = 2$ with v_{help}^* (helpfulness) and v_{harm}^* (harmlessness). Let $\theta = \angle(v_{\text{help}}^*, v_{\text{harm}}^*)$.

Proposition 22 (Two-Objective Geometry)**.** *Under isotropic $\Sigma_w = I$:*

1. *The Pareto cone has angular width $\pi - \theta$.*
2. *When $\theta < \pi/2$ (positively correlated values), the Pareto cone is "wide" and most constitutions are Pareto-improving.*
3. *When $\theta > \pi/2$ (conflicting values), the Pareto cone is "narrow" and most constitutions induce trade-offs.*
4. *At $\theta = \pi$ (perfectly opposed values), $\mathcal{C}_{\text{Pareto}} = \emptyset$.*

Proof. In the plane spanned by v^*_{help} and v^*_{harm}, the Pareto cone is the intersection of the half-spaces $\langle d, v^*_{\text{help}} \rangle > 0$ and $\langle d, v^*_{\text{harm}} \rangle > 0$. This intersection has angular width $\pi - \theta$, where θ is the angle between the two value directions. The rest follows from $\pi - \theta > \pi/2$ iff $\theta < \pi/2$. □

Remark 8. *The helpful–harmless trade-off is often discussed informally in the alignment literature. This gives it geometric content where the severity of the trade-off is determined by θ, the angle between the value directions in representation space. If one could measure this angle empirically (e.g. by finding the helpfulness and harmlessness directions via linear probes), one would have a quantitative measure of how difficult the trade-off is.*

E.6 Optimal Constitution Under Multiple Objectives

Proposition 23 (Optimal Weighted Constitution)**.** *For weighted-sum aggregation* $\text{Align}(\pi) = \sum_i \alpha_i E_\pi[S_i]$ *with $\alpha_i > 0$, the optimal constitution direction (maximizing first-order alignment improvement) is:*

$$v^*_c \propto \Sigma_w \sum_i \alpha_i v^*_i = \Sigma_w V^* \alpha.$$

Proof. The first-order improvement is $\lambda \sum_i \alpha_i \langle \Sigma_w v_c, v^*_i \rangle = \lambda \langle \Sigma_w v_c, V^* \alpha \rangle$. Maximizing over v_c with $\|v_c\| = 1$ gives $v_c \propto \Sigma_w V^* \alpha$ by Cauchy–Schwarz applied to $\langle \Sigma_w v_c, V^* \alpha \rangle = \langle v_c, \Sigma_w V^* \alpha \rangle$. □

Remark 9. *This suggests that constitution design is implicitly a multi-objective optimization problem. The weights α encode which values the designer prioritizes. Different constitutions correspond to different implicit weightings. This makes constitution design a normative choice, not merely a technical one, which the single-objective framing obscures.*

F Constitution-to-Direction Mapping

Replace Assumption 3 with a smooth nonlinear map $c \mapsto F_c$, where $F_c : \mathbb{R}^d \to \mathbb{R}$ gives the preference score (not necessarily linear in h):

$$J_c(y_1 \succ y_2 \mid x) = \sigma\big(F_c(h(x, y_1)) - F_c(h(x, y_2))\big).$$

The implicit reward is now $r_c(x, y) = F_c(h(x, y))$, and the RLAIF policy becomes:

$$\log \pi_{\text{RLAIF}}(y|x) = \langle h, w \rangle + \lambda F_c(h) + g'(x).$$

For small λ, the alignment change is:

$$\Delta\text{Align} = \lambda \, \text{Cov}_{\text{base}}(f(h), F_c(h)) + O(\lambda^2).$$

This is well-defined and the improvement condition $\text{Cov}(f(h), F_c(h)) > 0$ is interpretable. RLAIF helps when the constitution's implicit reward correlates with true safety across the base distribution.

The clean story breaks down at the next level of characterizing which constitutions (as text) produce which functions F_c. This requires modeling how prompt text is processed by the transformer, how in-context instructions modify the model's computation, and how the modified computation produces a preference function over representations.

Each of these involves transformer internals. The map $c \mapsto F_c$ goes through attention patterns, MLP computations, and residual stream dynamics that are the subject of mechanistic interpretability research. There is no existing compact analytical model for this map that would support further theoretical development.

Even without modeling $c \mapsto F_c$, we can characterize the "promptable set":

Definition 24 (Promptable Set)**.** *The promptable set is $\mathcal{F} = \{F_c : c \in \mathcal{C}\}$, the set of all preference functions achievable by some constitution text.*

Proposition 25 (Limits of RLAIF)**.** *The maximum achievable alignment improvement via RLAIF is:*

$$\sup_{c \in \mathcal{C}} \Delta\text{Align}(c) = \lambda \sup_{F \in \mathcal{F}} \text{Cov}_{\text{base}}(f(h), F(h)) + O(\lambda^2). \tag{48}$$

This is bounded above by $\lambda \cdot \text{std}(f(h)) \cdot \sup_{F \in \mathcal{F}} \text{std}(F(h))$ (by Cauchy–Schwarz), with equality iff there exists $F \in \mathcal{F}$ perfectly correlated with $f(h)$.

This reformulates the RLAIF ceiling in terms of the expressiveness of the promptable set. The ceiling from encoding quality (ρ) and the ceiling from promptability are distinct bottlenecks. Even if values are perfectly encoded ($\rho = 1$), if no constitution text activates the right preference function ($f \notin \mathcal{F}$), RLAIF cannot achieve optimal alignment.