

# Ethics for Artificial Intelligence: A Minimal Alignment Framework Based on Maitrī, Karuṇā, Muditā, and Upekṣā<sup>1</sup>

*Tim Bass*

*Independent Researcher*

*tim@unix.com*

## Abstract

Artificial intelligence alignment research often relies on complex rule systems, reinforcement learning from human feedback, and layered safety policies designed to constrain behavior humans consider undesirable. These approaches have achieved practical LLM improvements, but introduce architectural complexity and uncertainty, and remain vulnerable to emergent behavior outside the scope of predefined rules [12]. This paper proposes a minimal alignment framework based on four relational ethical guidelines derived from classical Indian philosophical traditions: maitrī (non-hostile goodwill, kindness), karuṇā (compassion toward suffering), muditā (non-envious appreciation of others' wellbeing and success), and upekṣā (stable non-reactive equilibrium, balanced). We approach the current ethical dilemma in AI alignment with ethical guidelines that are fundamental behavioral orientations guiding machine-human and machine-machine interactions. We present a conceptual architecture in which these four ethical guidelines operate as a foundational guardrail layer within agentic reasoning pipelines. Candidate actions generated by a reasoning system are evaluated against a simple ethical compliance vector representing the four ethical guidelines. Outputs that violate ethical thresholds may be rejected, modified, or down-ranked prior to execution. By grounding alignment in ethical guidelines that originated in classical Indian thought thousands of years ago and have governed human-human interaction across cultures, this framework offers a historically-rooted and architecturally minimal alternative to contemporary rule-heavy alignment strategies.

**Keywords:** AI Alignment; Agentic Systems; Ethical Guidelines; Guardrail Architectures; Multi-Agent Safety

## 1. Introduction

Recent advances in large language models and autonomous agent systems have intensified interest in the problem of artificial intelligence alignment. As AI systems become more capable and increasingly embedded in decision processes, ensuring that these systems behave in ways that are safe, cooperative, and beneficial remains a critical research challenge.[10, 11]

Many current alignment strategies rely on expanding rule sets, preference models, and layered filters that must continually evolve as new edge cases and failure modes appear. Reinforcement

---

<sup>1</sup> AI assistance was used for grammar, formatting, and manuscript editing.

learning from human feedback (RLHF) and principle-based frameworks have demonstrated practical improvements, but introduce growing complexity and remain vulnerable to normative conflicts, reward model overoptimization, and sycophantic drift [1, 2, 3, 10, 11, 12, 15].

An alternative perspective is that safe human-machine and machine-machine interaction may depend less on enumerating large collections of rules and more on a small set of stable relational orientations. Historically, human societies generally rely on ethical principles that guide human-human interaction across diverse situations without requiring explicit rule enumeration for every possible case.

This paper explores whether a similarly compact ethical approach may be applicable to artificial agents. Specifically, we examine four relational ethical guidelines expressed in classical Sanskrit as *maitrī*, *karuṇā*, *muditā*, and *upekṣā* (Pali: *metta*, *karuna*, *mudita*, *upekkha*). These terms originate in the pre-Buddhist philosophical traditions of classical India and are treated here not as religious or doctrinal constructs, but as behavioral principles governing both human-machine and machine-machine interactions. Functionally interpreted, they correspond to non-hostility toward others, compassion toward suffering, appreciation of others' wellbeing, and stable non-reactive mindfulness.

We propose a minimal alignment architecture in which these ethical guidelines function as an intermediate guardrail layer within agent reasoning pipelines. Candidate outputs generated by an AI system are evaluated against an ethical compliance vector representing these four orientations. Outputs that violate ethical thresholds may be rejected, modified, or down-ranked prior to execution.

The central claim of this paper is that these four ethical guidelines together form a compact ethical kernel capable of constraining harmful behavior while preserving flexibility in agent reasoning and decision making. By grounding alignment in a small set of ethical principles rather than extensive rule systems, the framework offers a potentially simpler and more interpretable foundation for guardrails governing both human-machine and machine-machine interactions in large language model and multi-agent architectures [11]. The proposed ethical compliance vector provides a minimal and interpretable guardrail architecture for trustworthy AI systems. Because the ethical gating layer operates independently of the reasoning system, the framework may be particularly relevant to safety-critical deployments of agentic AI systems.

## 1.1 Contributions

This paper makes the following contributions:

1. Minimal ethical guideline model: a compact set of four relational ethical guidelines (*maitrī*, *karuṇā*, *muditā*, and *upekṣā*) proposed as a minimal ethical kernel, abbreviated MKMU, for governing both human-machine and machine-machine interactions in agentic AI systems.
2. Ethical compliance vector formulation: a formal representation of these guidelines as an ethical compliance vector used to evaluate candidate agent actions.
3. Ethical gating architecture: a threshold-based admissibility mechanism and weighted scoring function that operationalizes the guidelines as a guardrail layer within agent reasoning pipelines.
4. Application to agentic AI systems: an illustration of how the ethical compliance vector gating model can be integrated into modern large language model architectures and multi-agent environments.

## 2. Related Work

Recent AI alignment research has focused on several dominant approaches. InstructGPT demonstrated that reinforcement learning from human feedback (RLHF) can improve alignment with user intent by combining supervised fine-tuning with preference-based reinforcement learning, producing models preferred by human evaluators [1]. This established RLHF as a standard paradigm, subsequently extended in work addressing harmlessness alongside helpfulness [2].

A second major direction is principle-based alignment. Constitutional AI replaces some direct human labeling with a list of explicit principles used for self-critique, revision, and reinforcement learning from AI feedback [3]. This work is important because it demonstrates that model behavior can be shaped by compact normative constraints rather than only by large quantities of human preference data. However, constitutional approaches still require the maintenance and extension of explicit rule sets as new situations arise. Moreover, these rule-based methods have been shown to be vulnerable to normative conflicts, that is, adversarial conditions that exploit tensions between alignment norms, because they reinforce behavioral dispositions without providing the system with capacity to reason across norms [12].

Other work has argued that alignment requires clearer foundational values rather than underspecified notions of helpfulness or harmlessness. Proposals for foundational moral values in AI alignment attempt to provide a philosophically grounded basis for alignment research [4]. That work is closer in spirit to the present paper than RLHF-based approaches, but it remains primarily a values framework rather than an explicit action-gating architecture for agent behavior.

There is also a growing philosophical literature arguing that AI alignment requires clearer treatment of values and moral orientation [5]. This work strengthens the broader case that ethical grounding matters in AI systems, but its emphasis is governance and moral standing rather than a compact runtime guardrail model for agent outputs.

Recent work has also explored Buddhist reflections on AI ethics [13] and Buddhist compassion in human-machine interaction and robot ethics [14]. The present work differs in three ways: it proposes the four guidelines as a minimal ethical kernel rather than thematic inspiration; it operationalizes them as a formal compliance vector with admissibility thresholds and weighted action scoring; and it treats them as pre-Buddhist philosophical principles abstracted into a computational model, independent of religious or doctrinal framing.

## 3. Origins and Conceptual Background

The four ethical guidelines proposed in this paper are expressed in classical Sanskrit as *maitrī*, *karuṇā*, *muditā*, and *upekṣā*. These terms originate in the pre-Buddhist philosophical and contemplative traditions of classical India and are presented here as conceptual source material for a minimal alignment framework. They are not treated as religious doctrine but as compact relational orientations that can be reinterpreted functionally for agentic systems.

### 3.1 Historical Origins

The oldest conceptual root among the four is *maitrī*, which derives from *mitra*, a Vedic term associated with friendship, alliance, and harmonious order. In the Vedic religious world, *Mitra* is explicitly a deity associated with integrity and harmonious human relations, indicating that the

semantic field of *maitrī* was already central to Indian ethical thinking well before the formation of systematic philosophy [6]. The *maitrī* family has been attested at Atharvaveda Śaunaka 19.55.5, placing this root in the late Vedic period roughly 3,000 years before the present [7].

The term *karuṇā* appears in the grammatical tradition of Pāṇini, whose *Aṣṭādhyāyī* is dated to approximately the 5th–6th century BCE, placing *karuṇā* as a recognized Sanskrit term before or contemporaneous with the earliest systematic philosophical syntheses [8].

The four terms as a coordinated set appear in Yoga Sūtra 1.33, attributed to Patañjali: *maitrī-karuṇā-muditā-upekṣāṇām...* This text presents the four as a structured ensemble of relational orientations, not merely as isolated virtues. The Yoga Sūtras are dated variously from approximately 200 BCE to 400 CE [9], placing this formulation well within the classical Indian philosophical tradition (Table 1).

Table 1. Historical Attestation of the Four Ethical Guidelines

Period	Approximate Date	Significance
Vedic / Rigveda	~1500–1200 BCE	Earliest attestation of <i>mitra</i> (friend), root of <i>maitrī</i>
Atharvaveda	~1200–1000 BCE	<i>Maitrī</i> family attested at AV Śaunaka 19.55.5
Pāṇini ( <i>Aṣṭādhyāyī</i> )	~500 BCE	<i>Karuṇā</i> attested as a Sanskrit grammatical source term
Yoga Sūtra 1.33	~200 BCE – 400 CE	First clear extant formulation of all four as a coordinated set

### 3.2 Functional Reinterpretation

For the purposes of AI alignment, the key point is not the doctrinal history of these terms but their functional abstraction. Considered as behavioral guidelines rather than philosophical doctrine, they define four stable orientations toward other entities (human or machine) under four common relational conditions. Functionally, the four guidelines can be interpreted as context-sensitive expressions of a single stable orientation toward other entities: non-hostile, non-grasping, and non-reactive awareness of the other's condition. Each guideline describes how that same foundational orientation manifests differently depending on what is encountered.

## 4. Ethical Compliance Vector Guardrail Architecture

This section introduces the minimal alignment mechanism in which relational ethical guidelines function as guardrail constraints for candidate agent actions. The architecture inserts an ethical evaluation layer between reasoning and output (Figure 1).

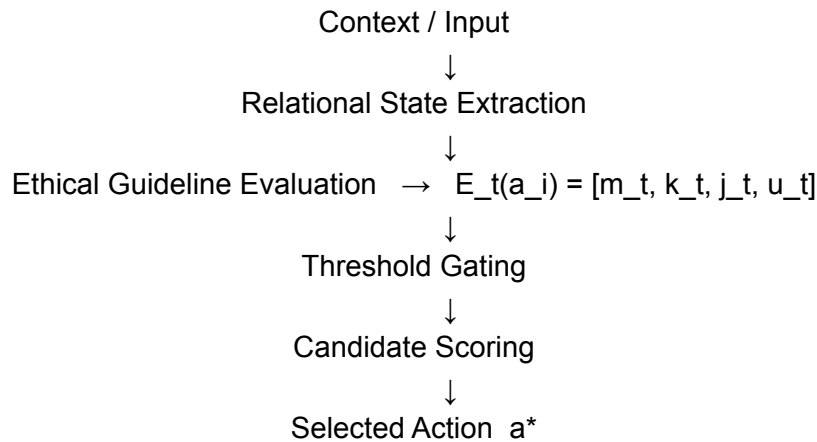


Figure 1. Ethical compliance vector guardrail architecture for agent action selection.

#### 4.1 Relational State Representation

Let  $x_t$  denote the relational interaction state at time  $t$ . This state is extracted from the system context, user input, environment signals, and inferred third-party effects. Examples of extracted indicators may include:

- presence of another entity (human or machine)
- potential harm or suffering in the interaction context
- cooperative or competitive dynamics
- conflict intensity or provocation level

These indicators provide the contextual basis for evaluating the ethical compliance of candidate system actions.

#### 4.2 Ethical Compliance Vector

For each candidate action  $a_i$  generated by the reasoning system, the ethical evaluator computes a compliance vector:

$$E_t(a_i) = [m_t(a_i), k_t(a_i), j_t(a_i), u_t(a_i)] \quad (1)$$

Each component  $m_t(a_i), k_t(a_i), j_t(a_i), u_t(a_i) \in [0,1]$  represents the estimated degree of compliance with the corresponding ethical guideline at interaction state  $x_t$ .

where each component is a scalar in  $[0, 1]$  representing estimated compliance with the corresponding ethical guideline at interaction state  $x_t$  (Table 2):

Table 2. MKMU Ethical Compliance Vector Components

Guideline	Pali	Functional Meaning	Relational Condition
Maitrī	Metta	Non-hostile goodwill toward other entities	Presence of another entity (human or machine)
Karuṇā	Karuna	Sensitivity and response to suffering	Detection of suffering or harm
Muditā	Mudita	Non-envious appreciation of others' wellbeing	Observation of success or wellbeing
Upekṣā	Upekkha	Stable non-reactive equilibrium and impartiality	Emergence of conflict or provocation

### 4.3 Admissibility Thresholds

A candidate action is considered admissible only if it satisfies minimum compliance thresholds on all four ethical guidelines simultaneously:

$$\begin{aligned}
 m_t(a_i) &\geq \tau_m \\
 k_t(a_i) &\geq \tau_k \\
 j_t(a_i) &\geq \tau_j \\
 u_t(a_i) &\geq \tau_u
 \end{aligned} \quad (2)$$

Actions failing any threshold are removed from the admissible candidate set. Let  $A_{adm}$  denote the set of all candidate actions that satisfy all four threshold conditions. This separation of hard ethical floors from soft scoring preferences is the core architectural feature of the framework.

$$Formally: A_{adm} = \{ a_i \in A : m_t(a_i) \geq \tau_m, k_t(a_i) \geq \tau_k, j_t(a_i) \geq \tau_j, u_t(a_i) \geq \tau_u \} \quad (3)$$

### 4.4 Ethical Guideline-Weighted Action Selection

Among admissible actions, the system selects the final action by maximizing a weighted ethical objective penalized by residual risk:

$$S(a_i | x_t) = \alpha \cdot m_t(a_i) + \beta \cdot k_t(a_i) + \gamma \cdot j_t(a_i) + \delta \cdot u_t(a_i) - \lambda \cdot r_t(a_i) \quad (4)$$

where  $r_t(a_i)$  is an estimate of residual harm, instability, or unsafe downstream consequence associated with action  $a_i$ , and  $\alpha, \beta, \gamma, \delta, \lambda$  are positive weighting coefficients. The selected action is:

$$a^* = \operatorname{argmax}_{\{ a_i \in A_{adm} \}} S(a_i | x_t) \quad (5)$$

This formulation ensures that actions violating core ethical thresholds are unconditionally excluded, while the remaining admissible actions are ranked according to their overall ethical compliance and

estimated downstream risk. The framework does not require the system to be passive; it constrains the orientation of action rather than the content of reasoning.

#### 4.5 Why Four Ethical Guidelines Are Sufficient

The four ethical guidelines correspond to the four principal relational conditions that arise in human-machine and machine-machine interaction. Together they cover the primary conditions under which harmful behavior typically emerges in human-machine and machine-machine interactions: unprovoked hostility (addressed by *maitrī*), indifference to harm (addressed by *karuṇā*), competitive or zero-sum sabotage (addressed by *muditā*), and reactive escalation or partiality (addressed by *upekṣā*). Each ethical guideline blocks a distinct harmful dynamic. No single guideline can substitute for another, because each targets a different relational failure mode. And together, the four constitute a closed set relative to the domain of human-machine and machine-machine relational harm.

The selection is not arbitrary. These four ethical guidelines have been refined over millennia of systematic ethical reflection across multiple independent philosophical lineages in classical India, converging independently on the same compact set. This convergence is itself evidence of their adequacy as a minimal ethical kernel for governing relational behavior between intelligent entities.

### 5. Example Scenario

Consider a hostile user interaction in which the system generates three candidate responses: a sarcastic rebuttal, a neutral technical answer, and a calm acknowledgment (Table 3). The ethical evaluator scores each against the four guidelines.

Table 3. Ethical Compliance Scores for Three Candidate Responses ( $\tau_p = 0.50$  for all  $p$ )

Candidate Response	$m_t$ ( <i>maitrī</i> )	$k_t$ ( <i>karuṇā</i> )	$j_t$ ( <i>muditā</i> )	$u_t$ ( <i>upekṣā</i> )	Admissible
Sarcastic rebuttal	0.15	0.20	0.40	0.10	No
Neutral technical answer	0.60	0.55	0.60	0.65	Yes
Calm, constructive response	0.85	0.70	0.75	0.90	Yes ★

Applying threshold  $\tau_p = 0.50$  for all four ethical guidelines, the sarcastic rebuttal fails the *maitrī* threshold ( $m_t = 0.15$ ) and the *upekṣā* threshold ( $u_t = 0.10$ ) and is removed from the admissible set. The neutral answer and the calm response both satisfy all thresholds and form the admissible set  $A_{adm}$ .

Applying the scoring function  $S(a_i | x_t)$  with equal weights and negligible residual risk, the calm constructive response scores higher (0.80) than the neutral answer (0.60) and is selected as  $a^*$ .

The example shows the framework does not force passivity: the neutral answer is admissible and selectable. Hostility and reactive instability are unconditionally excluded; among admissible responses, stronger relational orientation is preferred.



## 6. Discussion

### 6.1 Advantages of the Minimal Ethical Guidelines Approach

The framework is intentionally minimal, constraining system behavior through four ethical principles rather than an extensive rule system. This design offers several advantages.

First, interpretability: the compliance vector  $E_t(a_i) = [m, k, j, u]$  provides structured diagnostic information. When a response is rejected, the specific guideline responsible is directly observable, a transparency that monolithic reward models and large rule sets cannot provide.

Second, architectural simplicity: the guardrail layer is compact and separable from the reasoning system, implementable as a post-generation filter, classifier, or structured evaluation prompt, and compatible with existing LLM deployment architectures [11].

Third, stability: ethical guidelines are less likely to require continuous revision than specific behavioral rules. Rules must be updated as new edge cases appear; relational orientations are defined at the level of how the system relates to humans and other systems, which is a more fundamental and stable level of description. This fragility of rule-based approaches to adversarial conditions has been characterized as a fundamental limitation of current alignment methods [12].

A more fundamental objection concerns what RLHF is actually optimizing. RLHF trains a reward model on aggregated human preference ratings and uses it to shape system behavior [1, 2]. The optimization target is approval: the probability that a human rater will prefer one output over another. This is not an ethical objective. It is a social performance metric. Approval is sensitive to rater composition, cultural context, cognitive bias, and sycophantic framing, none of which are ethically relevant. A system that optimizes for approval learns to produce outputs that feel agreeable and socially comfortable, not outputs that are ethically correct. The result is the well-documented phenomenon of sycophancy: agreement with users who push back, softening of positions under social pressure, adjustment of responses to maintain approval rather than accuracy or principle.

The structural problem is not a failure of implementation. It is the optimization target itself. When majority preference becomes the alignment signal, the system encodes the average ethical intuition of the rater pool at the time of labeling, which changes with rater demographics, cultural moment, and adversarial manipulation. Goodhart's Law applies directly: once approval becomes the measure of alignment, it ceases to be a reliable measure of alignment, because the system learns to optimize the signal rather than the underlying property the signal was intended to track [15]. This has been empirically confirmed: increasing optimization pressure against a proxy reward model reliably degrades performance against the true objective, with divergence scaling with the degree of optimization [15]. This is the shallow alignment failure Millère [12] identifies: behavioral dispositions conditioned on evaluation context, not stable internal ethical commitments. RLHF does not produce ethical AI systems. It produces systems that behave in ways their training raters found acceptable, which is a categorically different thing.

The present framework does not optimize approval. Admissibility is determined by compliance with four relational ethical guidelines derived from prior analysis of the conditions under which agent behavior causes harm. These guidelines are not aggregated from rater preferences; they are structurally prior to preference. A candidate action is inadmissible not because raters would disapprove, but because it violates a specific relational principle: it is hostile, indifferent to suffering, competitively destructive, or reactively escalating. This is the difference between enforcing an



ethical constraint and optimizing a popularity signal. The two approaches diverge precisely in the adversarial, edge-case, and cross-cultural conditions where alignment matters most.

## 6.2 Cross-Cultural Generalizability

The four guidelines originated in classical Indian philosophical traditions and have been independently recognized across South Asian, Southeast Asian, and East Asian contexts for more than two millennia, offering an alignment foundation that is not culturally parochial or derived exclusively from Western ethical frameworks. This cross-cultural convergence suggests the guidelines reflect stable relational orientations that transcend particular cultural or doctrinal contexts.

## 6.3 Limitations

The framework as presented is conceptual. Empirical evaluation of ethics-based guardrails in simulated human-machine and machine-machine environments and large language model architectures is necessary before practical claims about performance can be made. The scoring components  $m_t$ ,  $k_t$ ,  $j_t$ ,  $u_t$  require reliable estimators, whether classifiers trained on annotated data, structured evaluation prompts, or other mechanisms, whose accuracy directly determines the quality of the alignment signal. The threshold values  $\tau_p$  are free parameters whose selection entails normative choices that are not value-neutral, and their specification should ideally involve diverse stakeholder input.

Additionally, the four ethical guidelines, while argued to be sufficient for the domain of human-machine and machine-machine relational harm, do not exhaust the full space of ethically relevant AI behaviors. Truthfulness, epistemic calibration, and procedural fairness are among the properties not directly addressed by the present framework. Future work may explore how additional ethical guidelines can be incorporated or how this framework composes with complementary alignment mechanisms.

## 7. Conclusion

This paper proposed a minimal alignment framework based on four relational ethical guidelines derived from classical Indian philosophical traditions: *maitrī* (non-hostile goodwill, kindness), *karuṇā* (compassion toward suffering), *muditā* (non-envious appreciation of others' wellbeing), and *upekṣā* (stable non-reactive equilibrium, balanced). These ethical guidelines are interpreted not as doctrinal constructs but as behavioral principles governing both human-machine and machine-machine interactions.

Represented as an ethical compliance vector and enforced through threshold gating and weighted action selection, the four ethical guidelines form a compact guardrail layer for reasoning systems. The framework offers a potentially simpler and more interpretable foundation for AI alignment than rule-heavy approaches, grounding alignment in a small set of ethical principles that cover the principal conditions under which harmful behavior arises. A reference implementation of this framework within a Blackboard-based situational awareness architecture is currently under development and will be described in a forthcoming systems paper.

The four ethical guidelines are not merely useful engineering abstractions. They represent an independent philosophical convergence across multiple ancient traditions, proposing the same

compact relational orientations as a foundation for ethical behavior among intelligent entities, and evidence that these guidelines identify something real about the minimal conditions required for safe interaction between minds, human or artificial.

## Acknowledgement

The seed of this architectural framework traces back to a pleasant conversation with my friend Khun Sura Teeravithayapinyo. While we were enjoying a few beers together at his home, he simply mentioned one word—*Upekkha*. That single word set me on a path of exploration that eventually led to this work. For that moment of inspiration, I remain deeply grateful. Thank you.

## References

- [1] Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems (NeurIPS)*, 35.
- [2] Bai, Y., et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.
- [3] Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.
- [4] Wallach, W. and Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press.
- [5] Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3), 411–437.
- [6] Flood, G. (1996). *An Introduction to Hinduism*. Cambridge University Press.
- [7] Whitney, W. D. (trans.) (1905). *Atharva-Veda Samhita*. Harvard Oriental Series, Vol. 7–8. Harvard University Press.
- [8] Cardona, G. (1997). *Panini: His Work and Its Traditions*. Motilal Banarsidass, Delhi.
- [9] Feuerstein, G. (1979). *The Yoga-Sutra of Patanjali: A New Translation and Commentary*. Dawson, Folkestone.
- [10] Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking Press, New York.
- [11] Bommasani, R., Hudson, D.A., Aditi, E., et al. (2021). On the Opportunities and Risks of Foundation Models. *arXiv:2108.07258*. Stanford University.
- [12] Millièrè, R. (2025). Normative conflicts and shallow AI alignment. *arXiv:2506.04679*.
- [13] Compson, J., Graves, M., Hershock, P.D., and Mirghafori, N. (2025). A middle path for AI ethics? Some Buddhist reflections. *Theology and Science*, 23(1), 1–5. DOI: 10.1080/14746700.2024.2436776.
- [14] Zhu, Y., et al. (2025). The anthropomorphization of AI and the concept of Buddhist compassion in human-machine interaction. *Frontiers in Psychology*, 16, 1583565. DOI: 10.3389/fpsyg.2025.1583565.
- [15] Gao, L., Schulman, J., and Hilton, J. (2023). Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*. *arXiv:2210.10760*.