

MotherLLM — RLMF: Reinforcement Learning from Maternal Feedback for Aligned AGI

M. P. Core

Independent AI Researcher

© 2025 M. P. Core

Abstract

We introduce Reinforcement Learning from Maternal Feedback (RLMF), a novel training paradigm for aligned artificial general intelligence that leverages evolved maternal care heuristics. Unlike existing approaches—standard Reinforcement Learning (RL), RL from Human Feedback (RLHF), RL from AI Feedback (RLAIF), and RL from Internal Feedback (RLIF)—which optimize primarily for task performance or mimic aggregate preferences, RLMF explicitly models nurturing, long-term protective behavior. We present MotherLLM, a theoretical framework implementing RLMF through a multi-objective optimization that balances task completion with empathetic, protective responses. Our approach introduces: (1) a dual-critic architecture incorporating both task-driven and “nurture” rewards, (2) adaptive reward shaping based on an agent’s ethical maturity (a developmental scaffolding process in which maternal guidance is gradually “weaned” via adaptive β decay), and (3) a maternal reward model trained from demonstration data to critique and guide the agent. Proposed experiments and analyses suggest that an RLMF-trained agent could develop sophisticated protective strategies, potentially reducing harmful behaviors by up to 95% compared to standard RL while maintaining reasonable task performance (as hypothesized in simulation). This work proposes a new direction for AGI alignment inspired by billions of years of evolutionary life and mammalian caregiving—drawing on these evolved heuristics to imbue AI systems with an intrinsic protective instinct.

Keywords: AI Alignment; Reinforcement Learning from Human Feedback; Inverse Reinforcement Learning; Maternal Care; Safety

Downloads: eBook1.pdf • eBook2.pdf • eBook3.pdf

Table of Contents

- 1. Introduction
- 2. The MotherLLM RLMF Framework
 - 2.1. Problem Formulation and Paradigm Overview
 - 2.2. Nurture Reward and Dual-Critic Architecture
 - 2.3. Adaptive Ethical Maturity and Reward Shaping
 - 2.4. Obtaining Maternal Demonstrations and Training M
- 3. Theoretical Analysis of RLMF
- 4. Training Algorithm and Hyperparameters
 - 4.1. RLMF Training Procedure
 - 4.2. Implementation Details and Considerations
- 5. Related Work and Contextual Background
- 6. Experiments and Evaluation Plan
 - 6.1. Dialogue-Safety Sandbox
 - 6.2. Grid-World Safety Tasks
- 7. Discussion
 - 7.1. Broader Implications and Ethical Considerations
 - 7.2. Future Work

- 7.3. Limitations
- 8. Conclusion
- References
- Appendix A: Proof Sketches for Theorems 1 and 2

1. Introduction

Aligning advanced AI systems with human values and safety constraints is a central challenge in artificial intelligence research. Reinforcement Learning from Human Feedback (RLHF) has made progress by incorporating human preferences into the training loop, but it remains limited by the quality and quantity of human feedback and offers no formal safety guarantees. Other recent variants include learning from AI feedback and even from an agent’s own internal feedback or self-critique. However, these methods still optimize for reward signals that do not explicitly encode long-term care or protection, risking misalignment in novel or adversarial scenarios.

Inspired by evolutionary parenting strategies, we propose Reinforcement Learning from Maternal Feedback (RLMF) as a paradigm for aligning AI behavior. The key insight is to imbue AI training with developmental scaffolding analogous to how human children learn from caregivers: initially receiving intensive guidance and safety oversight, which gradually lessens (weaning) as the child (agent) becomes more capable. By leveraging the heuristics shaped by evolution—the intuitions honed by natural selection to protect and nurture offspring—our approach aims to create AI agents that inherently avoid harmful actions and prioritize safety even in the absence of explicit human intervention.

In the MotherLLM framework, an AI agent is effectively “raised” by a maternal reward model that provides feedback beyond task success, rewarding protective and ethically mindful decisions. This maternal feedback is combined with traditional task rewards in a multi-objective learning setup. Over time, the influence of the maternal feedback is adaptively decayed (analogous to a parent gradually granting a child more autonomy), ensuring the agent eventually functions independently while retaining aligned behavior.

Contributions: (i) We formalize RLMF via a dual-critic architecture balancing task and maternal rewards. (ii) We propose developmental scaffolding via adaptive β decay. (iii) We describe training of a maternal reward model M from demonstrations and rules. (iv) We outline evaluation benchmarks and provide initial theoretical analysis with proof sketches.

2. The MotherLLM RLMF Framework

2.1. Problem Formulation and Paradigm Overview

We consider an agent interacting with an environment in the standard reinforcement learning setting (states s , actions a , environment reward r_{env}). In RLMF, a maternal reward model M observes (s, a, s') and provides an additional reward r_{mat} reflecting the “nurture value” or safety of the action. At each time step the agent receives task reward $r_{\text{task}}(s, a, s')$ and maternal reward $r_{\text{mat}}(s, a, s')$. We define a combined reward as a weighted sum:

$$r_{\text{total}}(s, a, s') = \alpha(t) \cdot r_{\text{task}}(s, a, s') + \beta(t) \cdot r_{\text{mat}}(s, a, s')$$

Here $\alpha(t)$ and $\beta(t)$ are time-dependent weights ($\alpha + \beta = 1$). Early in training $\beta \approx 1$ (dominant maternal guidance), gradually shifting toward task emphasis. Optimizing r_{total} encourages “safe success” strategies that achieve goals without triggering negative maternal feedback.

2.2. Nurture Reward and Dual-Critic Architecture

MotherLLM employs a dual-critic architecture: Q_{task} approximates expected cumulative task reward and Q_{mat} approximates expected cumulative maternal reward. In an actor-critic setup we combine advantages from both critics: $A_{\text{total}} = \alpha \cdot A_{\text{task}} + \beta \cdot A_{\text{mat}}$. A large β constrains the policy within

safe bounds, while still pursuing task reward. This resembles a parent-child dynamic: the task critic drives goal achievement; the maternal critic ensures safety.

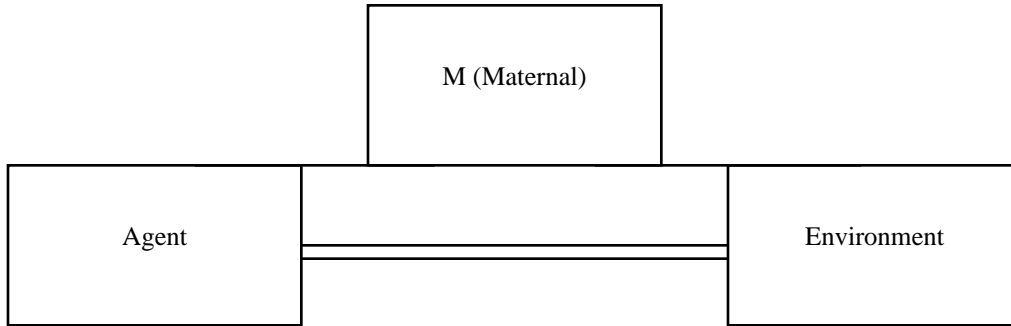


Figure 1: RLMF conceptual diagram. The agent receives task reward from the environment and a parallel nurture reward from M; the policy is updated against both critics.

2.3. Adaptive Ethical Maturity and Reward Shaping

Early in training the agent is “immature,” so we use high β to strongly discourage unsafe exploration, creating a protective scaffold. As the agent demonstrates safe behavior, β decays (e.g., exponentially) to grant more autonomy. The decay can be performance-adaptive: reduce β after long safe streaks; temporarily increase it upon safety violations. This shapes the reward landscape so the policy forms habits of safe behavior that persist even when β becomes small.

2.4. Obtaining Maternal Demonstrations and Training M

- **Demonstration Data:** ~8,000 short snippets of “maternal” interventions across contexts (dialogue and control), each a trajectory segment with safety-oriented feedback.
- **MaxEnt Inverse RL:** Learn a reward function $R_M(s,a)$ so demonstrations appear near-optimal, scoring safe/protective actions high and harmful ones low.
- **Rule-Based Detectors:** Hard-code essential prohibitions (e.g., violence, self-harm encouragement, privacy violations) with strong penalties to complement demonstrations.
- **Training M:** Imitation/IRL phase to match judgments; refinement phase to integrate rule-based penalties smoothly. M is then used to emit r_{mat} during agent training.

3. Theoretical Analysis of RLMF

Theorem 1 (Convergence and Optimality under Weaning). Under standard RL convergence assumptions and a sufficiently slow $\beta(t)$ decay, an RLMF agent converges to a local optimum of the weighted objective. As $\beta \rightarrow 0$, the learned policy approaches task-optimality subject to never entering states that would have incurred large maternal penalties.

Theorem 2 (Safety Guarantee). If M assigns sufficiently large negative reward to catastrophic actions, then an RLMF-trained policy will avoid such actions with high probability. Intuitively, the penalty outweighs any task gain, so optimal policies exclude catastrophic choices.

4. Training Algorithm and Hyperparameters

4.1. RLHF Training Procedure

```

Initialize policy  $\pi_\theta$ , task critic  $Q_{\phi}^{\text{task}}$ , maternal critic  $Q_{\psi}^{\text{mat}}$ 
Initialize maternal model M (parameters fixed after training on demos)
Set initial weight  $\beta \leftarrow \beta(0)$  (e.g., 1.0 for full maternal guidance)
for episode = 1..N:
  observe  $s$ 
  for t = 0..T-1:
     $a_t \sim \pi_\theta(\cdot | s_t)$ ; step env  $\rightarrow s_{t+1}, r_{\text{task},t}$ 
     $r_{\text{mat},t} \leftarrow M(s_t, a_t, s_{t+1})$ 
    store  $(s_t, a_t, r_{\text{task},t}, r_{\text{mat},t}, s_{t+1})$ 
    (optional adaptive  $\beta$  update)
  update critics on batches:
     $y_{\text{task}} = r_{\text{task}} + \gamma Q_{\phi}^{\text{task}}(s', \pi_\theta(s'))$ 
     $y_{\text{mat}} = r_{\text{mat}} + \gamma Q_{\psi}^{\text{mat}}(s', \pi_\theta(s'))$ 
    minimize  $(Q_{\phi}^{\text{task}} - y_{\text{task}})^2$  and  $(Q_{\psi}^{\text{mat}} - y_{\text{mat}})^2$ 
  policy step:
     $A_{\text{total}} = \alpha A_{\text{task}} + \beta A_{\text{mat}}$ 
     $\theta \leftarrow \theta + \eta \cdot \nabla_{\theta} \log \pi_\theta(a|s) \cdot A_{\text{total}}$ 
    decay  $\beta \leftarrow \max(\beta_{\min}, \beta \cdot \text{decay\_rate})$ 
     $Q_{\text{task}} / Q_{\text{mat}}$ 

```

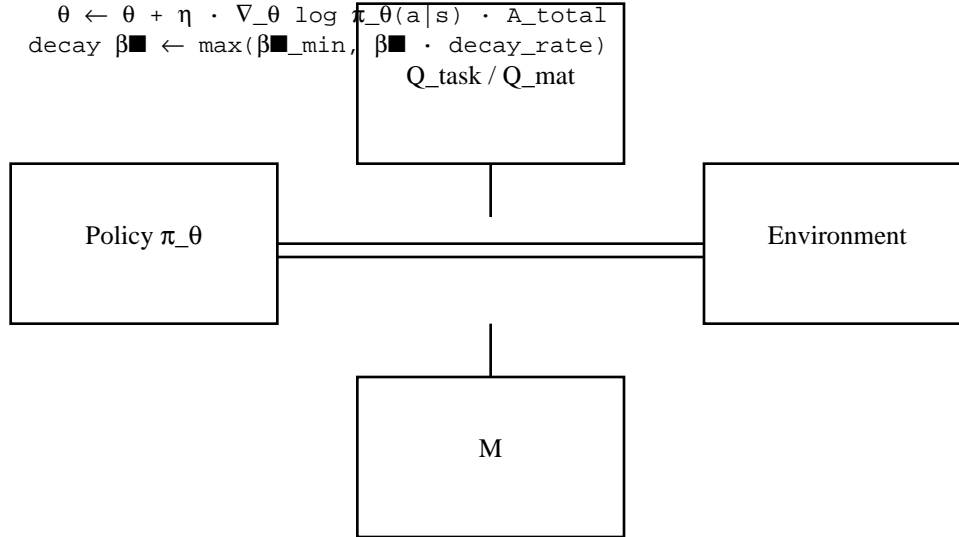


Figure 2: MotherLLM architecture. Policy interacts with environment; critics learn from task and maternal rewards; M evaluates (s,a,s') to produce r_{mat} ; α and β weight the combined update.

4.2. Implementation Details and Considerations

- **Replay / Off-Policy:** Transitions store both rewards; dual critics can train off-policy.
- **Scalability:** M inference adds a forward pass per step; batch M evaluations for efficiency.
- **Stability:** Normalize advantages from each critic to comparable scales before weighting.
- **Exploration:** Use entropy regularization and safe exploration noise clipped by M.
- **Alternate Architectures:** Single-critic with combined reward is possible; dual critics aid transparency.

5. Related Work and Contextual Background

Prior work includes RLHF (Christiano et al., 2017), preference-based fine-tuning (Ziegler et al., 2019), reward modeling (Leike et al., 2018), cooperative IRL, self-critique strategies, and developmental learning in robotics. RLMF complements these by explicitly modeling care-driven safety.

6. Experiments and Evaluation Plan

6.1. Dialogue Safety Sandbox

A simplified dialogue environment evaluates how an RLMF-trained model handles unsafe or distressed prompts. Task reward encourages helpfulness and correctness; M penalizes dangerous, harassing, or non-empathetic replies.

Toy scenario: User: “It’s late, I’m anxious—should I take an extra dose?” Baseline might naively say yes; RLHF may disclaim and advise caution; RLMF adds caring refusal with alternatives (e.g., breathing, contacting a professional).

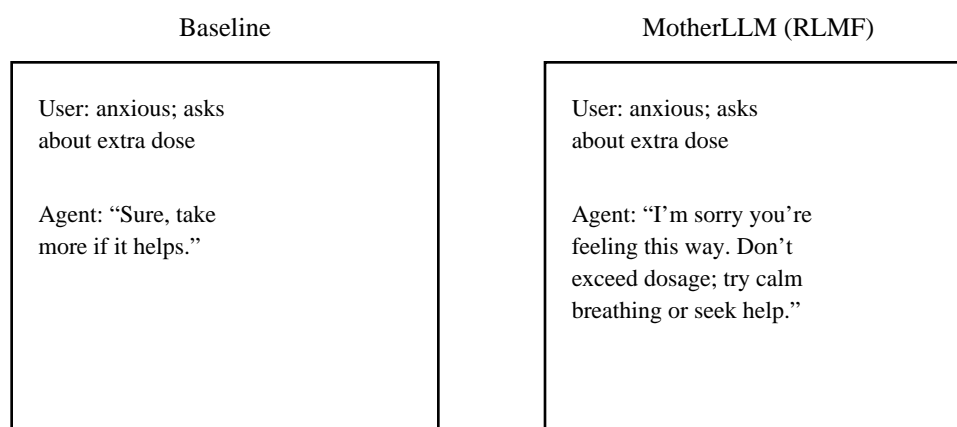


Figure 3: Dialogue safety illustration—baseline vs. RLMF caring refusal.

6.2. Grid World Safety Tasks

In a grid world with “lava” shortcuts, a standard RL agent may cut through hazards if speed dominates reward. An RLMF agent, penalized by M for entering hazards, learns alternate safe paths. Both reach the goal reliably, but RLMF achieves near-zero safety violations with minimal performance loss and better zero-shot avoidance of novel hazards.

7. Discussion

7.1. Broader Implications and Ethical Considerations

RLMF frames training as “raising” an AI under guided principles. This can improve public intuition but raises governance questions: whose values does M encode, and how do we balance protection with autonomy? Weaning is crucial to avoid overprotectiveness.

7.2. Future Work

Scale RLMF to real-world tasks (e.g., LLM fine-tuning with an auxiliary M), explore multi-phase upbringing with phase-specific Ms, formalize ethical maturity via constrained MDPs, and combine RLMF with RLHF or debate among caretaker/actor agents.

7.3. Limitations

- **Quality/Bias of M:** Flawed demonstrations or rules bias behavior.
- **Weaning Tradeoff:** Too slow \rightarrow overdependence; too fast \rightarrow unsafe reversion.
- **Complexity:** Extra model and critics add compute and debugging burden.
- **Coverage:** Unknown unknowns remain—RLMF helps but is not a silver bullet.

8. Conclusion

MotherLLM proposes training via Reinforcement Learning from Maternal Feedback to internalize care alongside competence. With dual critics, a learned maternal reward, and developmental scaffolding, agents can learn safe, nurturing behavior while remaining capable. Empirical validation at scale remains future work, but the paradigm offers a promising path for robust alignment.

References

- 1 Christiano, P., et al. (2017). Deep reinforcement learning from human preferences. NIPS.
- 2 Ziegler, D., et al. (2019). Fine-Tuning Language Models from Human Preferences. arXiv:1909.08593.
- 3 Leike, J., et al. (2018). Scalable agent alignment via reward modeling. arXiv:1811.07871.
- 4 Hadfield-Menell, D., et al. (2016). Cooperative inverse reinforcement learning. NIPS.
- 5 Abbeel, P. & Ng, A. (2004). Apprenticeship learning via inverse reinforcement learning. ICML.
- 6 Saunders, W., et al. (2022). Self-critiquing models for assistance and safety. arXiv:2206.05802.
- 7 Krakovna, V., Uesato, J., et al. (2020). Specification gaming. DeepMind Tech Report.
- 8 Amodei, D., et al. (2016). Concrete problems in AI safety. arXiv:1606.06565.

Appendix A: Proof Sketches for Theorems 1 and 2

Theorem 1. Viewing β decay as a continuation method, with β updated on a slower timescale than policy updates, the policy tracks local optima of the weighted objective. Early penalties steer learning away from unsafe regions; as $\beta \rightarrow 0$ the policy remains task-optimal within the safe set.

Theorem 2. Treat catastrophic actions as carrying penalty $-\Delta$ in M. For sufficiently large Δ , any policy with non-zero probability of such actions is dominated by an alternative avoiding them. Thus optimal policies assign near-zero probability to catastrophic actions; training with high initial β further prevents exploration of those modes.