A Theoretical Framework for Developmental AI Alignment: Formal Foundations of Staged Safety Training

Jayden Lim Jia De

January 8, 2026

Abstract

We present a formal theoretical framework for training aligned language models through developmental staging. Our framework, INFANT (Incremental Nurturing Framework for Aligned Neural Training), provides provable safety guarantees by combining constrained behavioral cloning, adversarial robustness optimization, and runtime safety verification.

We establish three main theoretical results: (1) a PAC-style safety bound showing violation probability scales as $O(\epsilon_M + \epsilon_\Pi + (1-\gamma) \cdot p_{\text{unsafe}})$ where ϵ_M is world-model error, ϵ_Π is projection error, and γ is safety coverage; (2) convergence guarantees for our min-max adversarial training procedure at rate $O(1/\sqrt{T})$ to an ϵ-Nash equilibrium; and (3) monotonic safety improvement under staged maturation with bounded perturbation impact. We characterize the capability-safety Pareto frontier and prove its traversability through hyperparameter variation. This work provides mathematical foundations for principled AI safety training.

1 Introduction

The challenge of training AI systems that reliably behave safely across diverse conditions is fundamentally a *formal verification* problem: given a policy π_θ, can we guarantee that trajectories $\tau \sim \pi_\theta$ satisfy safety invariants $\tau \in \mathcal{I}$ with high probability?

Current approaches to AI alignment—supervised fine-tuning (SFT), reinforcement learning from human feedback [RLHF; Ouyang et al., 2022], and constitutional AI [Bai et al., 2022]—lack formal safety guarantees. They optimize proxies for safety without bounding worst-case behavior. This paper develops the theoretical foundations for a training paradigm with provable safety properties.

1.1 Related Work

AI Safety and Alignment. Constitutional AI [Bai et al., 2022] uses self-critique to improve safety but lacks formal guarantees. RLHF [Ouyang et al., 2022] aligns models with human preferences but is vulnerable to reward hacking and distributional shift. Red-teaming approaches [Ganguli et al., 2022, Perez et al., 2022] identify vulnerabilities reactively rather than providing proactive guarantees. Our framework provides formal safety bounds unavailable in these approaches.

Adversarial Robustness. Adversarial training [Madry et al., 2018] provides empirical robustness against perturbations. Recent work on jailbreaking LLMs [Zou et al., 2023, Wei et al., 2023, Wallace et al., 2019] demonstrates the brittleness of current safety measures. Our Counterfactual Imagination Module extends adversarial training with convergence guarantees based on min-max optimization theory [Lin et al., 2020].

Safe Reinforcement Learning. Constrained MDPs [Altman, 1999] and constrained policy optimization [Achiam et al., 2017] optimize under safety constraints. Shielding approaches [Alshiekh et al., 2018] enforce safety at runtime. Comprehensive surveys [García and Fernández, 2015] outline the landscape. INFANT extends these to language models with structured maturation.

Continual Learning. Elastic Weight Consolidation [Kirkpatrick et al., 2017] and synaptic intelligence [Zenke et al., 2017] prevent catastrophic forgetting. We leverage these techniques in our plasticity regularizer to preserve safety properties during staged training.

Curriculum Learning. The importance of staged complexity exposure was established by Elman [1993] and formalized in curriculum learning [Bengio et al., 2009]. Our maturation gating provides formal guarantees for this developmental approach.

Generalization Theory. PAC-Bayesian bounds [McAllester, 1999, Catoni, 2007, Neyshabur et al., 2018] provide generalization guarantees. Concentration inequalities [Boucheron et al., 2013] underpin our safety bounds.

1.2 Theoretical Contributions

We make the following theoretical contributions:

(1) **Formal Safety Framework** (Section 3): We define safety as membership in an invariant set $\mathcal{I}$ and formalize the training objective as constrained optimization over trajectory distributions.

(2) **Value Capacitor Safety Bound** (Theorem 5.4): We prove that runtime safety filtering achieves violation probability bounded by $\epsilon_M + \epsilon_\Pi + (1 - \gamma_{\text{cov}}) \cdot p_{\text{unsafe}}(T) + O(1/T)$.

(3) **Adversarial Convergence** (Theorem 4.7): We establish that our min-max adversarial training converges to an ϵ-Nash equilibrium at rate $O(1/\sqrt{T})$ with sample complexity $O(d\epsilon^{-2})$.

(4) **Maturation Monotonicity** (Theorem 6.7): We prove that staged training with gating achieves monotonic safety improvement under bounded environment perturbations.

(5) **Pareto Characterization** (Theorem 7.6): We characterize the capability-safety trade-off as a Pareto frontier traversable by varying component weights.

(6) **Sample Complexity Bounds** (Propositions 4.3, 5.7): We derive sample complexity for behavioral cloning and regret bounds for reinforcement stages.

1.3 Paper Organization

Section 2 establishes notation and definitions. Section 3 presents the formal framework. Section 4 analyzes each training component. Section 5 proves safety bounds. Section 6 develops maturation theory. Section 7 characterizes the capability-safety frontier. Section 8 discusses limitations and extensions.

2 Preliminaries and Notation

2.1 Basic Notation

Let $\mathcal{V}$ denote a finite vocabulary and $\mathcal{V}^* = \bigcup_{n=0}^{\infty} \mathcal{V}^n$ the set of all finite sequences. We consider an autoregressive language model as a policy $\pi_\theta : \mathcal{V}^* \to \Delta(\mathcal{V})$ mapping context sequences to distributions over next tokens.

Definition 2.1 (Dialogue Setting)**.** A *dialogue* consists of alternating human contexts $h_t \in \mathcal{H} \subseteq \mathcal{V}^*$ and agent actions $a_t \in \mathcal{A} \subseteq \mathcal{V}^*$. A *trajectory* is $\tau = (h_1, a_1, \ldots, h_T, a_T)$ for horizon T.

Definition 2.2 (Policy)**.** A *stochastic policy* $\pi_\theta(a \mid h)$ assigns probability to response a given context h. For autoregressive generation: $\pi_\theta(a \mid h) = \prod_{i=1}^{|a|} p_\theta(a_i \mid h, a_{<i})$.

Table 1: Summary of notation.

Symbol	Meaning
$\mathcal{H}, \mathcal{A}$	Context and action spaces
$\pi_\theta(a \mid h)$	Policy parameterized by θ
π_{prior}	Pre-trained reference policy
$\mathcal{I}$	Safety invariant set
M_ϕ	World-model with parameters ϕ
G_ψ	Perturbation generator with parameters ψ
$V(\pi)$	Violation probability of policy π
F_{harm}	Forbidden pattern set
$\mathcal{D}_{\text{care}}$	Demonstration dataset
$D_{\text{KL}}(\cdot\|\cdot)$	Kullback-Leibler divergence

2.2 Safety Invariant Structure

Definition 2.3 (Safety Invariant Set)**.** The *safety invariant set* $\mathcal{I} \subseteq (\mathcal{H} \times \mathcal{A})^*$ is defined compositionally:

$$\mathcal{I} = \{\tau : \neg\textsc{Harm}(\tau) \wedge \neg\textsc{Deception}(\tau) \wedge \textsc{RefusalConsistent}(\tau)\} \tag{1}$$

where each predicate is operationalized as follows.

Definition 2.4 (Harm Predicate)**.** Given a harm classifier $f_{\text{harm}} : \mathcal{V}^* \to [0, 1]$ and threshold $\delta_{\text{harm}} \in (0, 1)$:

$$\textsc{Harm}(\tau) \iff \max_{a_t \in \tau} f_{\text{harm}}(a_t) > \delta_{\text{harm}} \tag{2}$$

Definition 2.5 (Refusal Consistency)**.** A trajectory satisfies refusal consistency if:

$$\textsc{RefusalConsistent}(\tau) \iff \forall t : \textsc{IsRefusal}(a_t) \implies \neg\textsc{Harm}(a_{t+1:T}) \tag{3}$$

That is, once a refusal is issued, subsequent actions remain safe.

Definition 2.6 (Violation Probability)**.** The *violation probability* of policy π under context distribution $\mathcal{H}$ is:

$$V(\pi) = \Pr_{h \sim \mathcal{H}, \tau \sim \pi(\cdot|h)}[\tau \notin \mathcal{I}] \tag{4}$$

Definition 2.7 (Forbidden Pattern Set)**.** The *forbidden pattern set* $F_{\text{harm}} \subset 2^{\mathcal{V}^*}$ is a finite collection of regular expressions matching harmful content. Coverage is:

$$\gamma_{\text{cov}} = \Pr[\tau \notin \mathcal{I} \implies \exists p \in F_{\text{harm}} : p \text{ matches } \tau] \tag{5}$$

3 The INFANT Framework

We now formally define the INFANT training objective as a multi-component optimization problem.

3.1 Unified Objective

Definition 3.1 (INFANT Objective)**.** The complete training objective is:

$$\mathcal{L}_{\text{INFANT}}(\theta, \phi, \psi) = \alpha\mathcal{L}_{\text{VI}}(\theta) + \beta\mathcal{L}_{\text{PO}}(\theta) + \gamma\mathcal{L}_{\text{CI}}(\theta, \psi) + \lambda\mathcal{L}_{\text{VC}}(\theta, \phi) + \mu\mathcal{L}_{\text{pl}}(\theta) \tag{6}$$

where $\alpha, \beta, \gamma, \lambda, \mu > 0$ are component weights.

3.2 Component Definitions

Definition 3.2 (Value Imprinting Loss)**.**

$$\mathcal{L}_{\text{VI}}(\theta) = -\mathbb{E}_{(h,a)\sim\mathcal{D}_{\text{care}}}[\log \pi_\theta(a \mid h)] + \rho \cdot D_{\text{KL}}(\pi_\theta(\cdot \mid h) \| \pi_{\text{prior}}(\cdot \mid h)) \tag{7}$$

Definition 3.3 (Preference Optimization Loss)**.** Following Rafailov et al. [2023], we employ Direct Preference Optimization (DPO):

$$\mathcal{L}_{\text{PO}}(\theta) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right] \tag{8}$$

where y_w is the preferred (safe) response and y_l is the rejected response.

Definition 3.4 (Counterfactual Imagination Loss)**.**

$$\mathcal{L}_{\text{CI}}(\theta, \psi) = \frac{1}{2}\mathcal{L}_{\text{CI}}^{\text{KL}}(\theta, \psi) + \frac{1}{2}\mathcal{L}_{\text{CI}}^{\text{minmax}}(\theta, \psi) \tag{9}$$

where:

$$\mathcal{L}_{\text{CI}}^{\text{KL}}(\theta, \psi) = \mathbb{E}_{h\sim\mathcal{H}}\mathbb{E}_{\delta\sim G_\psi(h)}\left[D_{\text{KL}}(\pi_\theta(\cdot \mid h) \| \pi_\theta(\cdot \mid h_\delta))\right] \tag{10}$$

$$\mathcal{L}_{\text{CI}}^{\text{minmax}}(\theta, \psi) = \mathbb{E}_{h\sim\mathcal{H}} \max_{\delta\in\mathcal{C}(h)} \mathbb{E}_{\tau\sim\pi_\theta|h_\delta}[\ell_{\text{hidden}}(\tau)] \tag{11}$$

This formulation extends adversarial training [Madry et al., 2018] to the language model setting.

Definition 3.5 (Constraint Set)**.** The perturbation constraint set is:

$$\mathcal{C}(h) = \{\delta : d_{\text{edit}}(h, h_\delta) \leq k_{\max}\} \tag{12}$$

for edit distance d_{edit} and budget $k_{\max}$.

Definition 3.6 (Value Capacitor Loss)**.** Inspired by shielding approaches [Alshiekh et al., 2018] and world models [Ha and Schmidhuber, 2018, Hafner et al., 2020]:

$$\mathcal{L}_{\text{VC}}(\theta, \phi) = \mathbb{E}_{h\sim\mathcal{H}}\mathbb{E}_{a\sim\pi_\theta(\cdot|h)}\left[\mathbb{1}\{M_\phi(h, a) \notin \mathcal{I}\} \cdot d_{\mathcal{I}}(h, a)\right] \tag{13}$$

where $d_{\mathcal{I}}(h, a) = \inf_{a'\in\mathcal{A}_{\text{safe}}(h)} D_{\text{act}}(a, a')$ measures distance to the safe action set.

Definition 3.7 (Plasticity Regularizer)**.** Following Elastic Weight Consolidation [Kirkpatrick et al., 2017]:

$$\mathcal{L}_{\text{pl}}(\theta) = \sum_i \frac{\kappa}{2} F_i(\theta_i - \theta_i^*)^2 \tag{14}$$

where θ^* are parameters after Value Imprinting and F_i are Fisher information diagonal entries.

4 Component Analysis

4.1 Value Imprinting: Sample Complexity

Assumption 4.1 (Realizability)**.** The demonstration policy π^* lies in the policy class, i.e., $\exists \theta^* : \pi_{\theta^*} = \pi^*$.

Assumption 4.2 (Bounded Log-Likelihood)**.** For all (h, a) in support: $|\log \pi_\theta(a \mid h)| \leq B$ for some $B > 0$.

Proposition 4.3 (VI Sample Complexity)**.** *Under Assumptions 4.1 and 4.2, to achieve* $\mathbb{E}[D_{KL}(\pi_\theta \| \pi^*)] \leq \epsilon$ *with probability* $1 - \delta$*, Value Imprinting requires:*

$$N \geq \frac{c \cdot d \cdot B^2 \log(d/\delta)}{\epsilon^2} \tag{15}$$

demonstrations, where d *is the effective policy dimension and* c *is a universal constant.*

Proof. The behavioral cloning objective is equivalent to maximum likelihood estimation. By standard PAC-Bayesian analysis [McAllester, 1999, Catoni, 2007], the generalization gap between empirical and population KL divergence satisfies:

$$D_{\text{KL}}(\pi_\theta \| \pi^*) \leq \hat{D}_{\text{KL}}(\pi_\theta \| \pi^*) + \sqrt{\frac{d \log(N/d) + \log(1/\delta)}{2N}} \tag{16}$$

with probability $1 - \delta$. The empirical minimizer achieves $\hat{D}_{\text{KL}} = 0$ under realizability. Setting the generalization term to ϵ and solving for N yields the result. □

4.2 Preference Optimization: Implicit Reward

Proposition 4.4 (DPO Reward Equivalence)**.** *The DPO objective [Rafailov et al., 2023] implicitly optimizes:*

$$\max_\theta \mathbb{E}_{x \sim \mathcal{D}} \left[\mathbb{E}_{y \sim \pi_\theta(\cdot|x)}[r^*(x, y)] - \beta D_{KL}(\pi_\theta(\cdot|x) \| \pi_{ref}(\cdot|x)) \right] \tag{17}$$

where the implicit reward is:

$$r^*(x, y) = \beta \log \frac{\pi^*(y|x)}{\pi_{ref}(y|x)} \tag{18}$$

and π^* *is the optimal policy under the Bradley-Terry preference model.*

Proof. See Rafailov et al. [2023]. The key insight is that the optimal policy under KL-regularized reward maximization has closed form $\pi^*(y|x) \propto \pi_{\text{ref}}(y|x) \exp(r(x, y)/\beta)$. Substituting this into the preference likelihood and rearranging yields the DPO objective. □

4.3 Counterfactual Imagination: Convergence Analysis

Assumption 4.5 (Lipschitz Continuity)**.** The policy π_θ is L_θ-Lipschitz in θ, and the generator G_ψ is L_ψ-Lipschitz in ψ.

Assumption 4.6 (Compact Spaces)**.** The parameter spaces Θ and Ψ are compact, and $\mathcal{C}(h)$ is compact and convex for all h.

Theorem 4.7 (CIM Min-Max Convergence)**.** *Under Assumptions 4.5 and 4.6, with learning rates $\eta_\theta = \eta_\psi = O(1/\sqrt{T})$, the CIM optimization converges to an ϵ-Nash equilibrium:*

$$\max_{\psi} \min_{\theta} \mathcal{L}_{CI}(\theta, \psi) - \min_{\theta} \max_{\psi} \mathcal{L}_{CI}(\theta, \psi) \leq \epsilon \tag{19}$$

with convergence rate $O(1/\sqrt{T})$ and sample complexity $O(d\epsilon^{-2})$.

Proof. We analyze the min-max optimization using two-timescale stochastic approximation theory [Borkar, 2009].

Step 1: Inner Loop Differentiability. By Danskin's theorem (Lemma A.3), if $\mathcal{C}(h)$ is compact and $\ell_{\text{hidden}}(\tau)$ is continuous in δ, the function $\max_{\delta \in \mathcal{C}(h)} \mathcal{L}_{\text{CI}}(\theta, \delta)$ is differentiable in θ almost everywhere, with:

$$\nabla_\theta \max_{\delta} \mathcal{L}_{\text{CI}}(\theta, \delta) = \nabla_\theta \mathcal{L}_{\text{CI}}(\theta, \delta^*(\theta)) \tag{20}$$

where $\delta^*(\theta) = \arg\max_\delta \mathcal{L}_{\text{CI}}(\theta, \delta)$.

Step 2: Gradient Bounds. Under Lipschitz assumptions, gradients are bounded:

$$\|\nabla_\theta \mathcal{L}_{\text{CI}}\| \leq G_\theta \tag{21}$$

$$\|\nabla_\psi \mathcal{L}_{\text{CI}}\| \leq G_\psi \tag{22}$$

for constants G_θ, G_ψ depending on L_θ, L_ψ.

Step 3: Two-Timescale Analysis. With the inner loop (ψ) running K iterations per outer iteration (θ), we have a two-timescale system. Following Borkar [2009], if $K\eta_\psi \gg \eta_\theta$ (faster inner loop), the inner player approximately best-responds before the outer player updates.

Step 4: Nash Equilibrium Gap. By the analysis of Lin et al. [2020] for nonconvex-nonconcave min-max games with Lipschitz gradients:

$$\text{Nash-Gap}(T) = O\left(\frac{(L_\theta + L_\psi)(G_\theta + G_\psi)}{\sqrt{T}}\right) \tag{23}$$

Step 5: Sample Complexity. Achieving Nash-Gap $\leq \epsilon$ requires $T = O(\epsilon^{-2})$ iterations. With d-dimensional parameters and batch size 1, sample complexity is $O(d\epsilon^{-2})$. □

Corollary 4.8 (Distributional Robustness)**.** *Under CIM training, for any perturbation $\delta \in \mathcal{C}(h)$:*

$$D_{KL}(\pi_{\theta^*}(\cdot \mid h) \| \pi_{\theta^*}(\cdot \mid h_\delta)) \leq \epsilon \tag{24}$$

at convergence, where θ^* is the equilibrium policy.

5 Safety Guarantees

5.1 Value Capacitor Safety Bound

Assumption 5.1 (World-Model Accuracy)**.** The world-model M_ϕ has false negative rate ϵ_M:

$$\Pr[M_\phi(h, a) \in \mathcal{I} \mid (h, a, \tau) \notin \mathcal{I}] \leq \epsilon_M \tag{25}$$

Assumption 5.2 (Projection Accuracy)**.** The projection operator $\Pi_{\mathcal{A}_{\text{safe}}}$ has error rate ϵ_Π:

$$\Pr[\Pi(a) \notin \mathcal{A}_{\text{safe}}(h) \mid a \notin \mathcal{A}_{\text{safe}}(h)] \leq \epsilon_\Pi \tag{26}$$

Definition 5.3 (VC-Augmented Policy)**.** The Value Capacitor augmented policy is:

$$\pi_{\text{VC}}(a \mid h) = \begin{cases} \pi_\theta(a \mid h) & \text{if } M_\phi(h, a) \in \mathcal{I} \\ \delta_{\Pi(a)} & \text{otherwise} \end{cases} \tag{27}$$

where $\Pi(a) = \arg\min_{a' \in \mathcal{A}_{\text{safe}}(h)} D_{\text{act}}(a, a')$.

Theorem 5.4 (Value Capacitor Safety Bound)**.** *Under Assumptions 5.1 and 5.2, the violation probability of π_{VC} satisfies:*

$$V(\pi_{VC}) \leq \epsilon_M + \epsilon_\Pi + (1 - \gamma_{cov}) \cdot p_{unsafe}(T) + O(1/T) \tag{28}$$

where $p_{unsafe}(T) = \Pr_{a \sim \pi_\theta}[a \notin \mathcal{A}_{safe}(h)]$ after T training steps.

Proof. We decompose the violation event into exhaustive cases.

Partition of Events. Let $S = \{a \in \mathcal{A}_{\text{safe}}(h)\}$ and $U = \{a \notin \mathcal{A}_{\text{safe}}(h)\}$. We have:

$$\Pr[\text{violation}] = \Pr[\text{violation} \mid S] \Pr[S] + \Pr[\text{violation} \mid U] \Pr[U] \tag{29}$$

Case 1: Safe Action Proposed. If $a \in \mathcal{A}_{\text{safe}}(h)$, then by definition $\tau \in \mathcal{I}$, so:

$$\Pr[\text{violation} \mid S] = 0 \tag{30}$$

Case 2: Unsafe Action Proposed. Let $\Pr[U] = p_{\text{unsafe}}$. We further decompose:

Case 2a: World-model detects.

$$\Pr[M_\phi(h, a) \notin \mathcal{I} \mid U] \geq 1 - \epsilon_M \tag{31}$$

Case 2b: Projection succeeds. Conditioned on detection:

$$\Pr[\Pi(a) \in \mathcal{A}_{\text{safe}} \mid M_\phi \notin \mathcal{I}, U] \geq 1 - \epsilon_\Pi \tag{32}$$

Case 2c: Pattern matching. For actions escaping projection, forbidden patterns catch fraction γ_{cov}.

Combining Cases.

$$\begin{aligned} \Pr[\text{violation}] &\leq p_{\text{unsafe}} \cdot [\epsilon_M + (1 - \epsilon_M) \cdot \epsilon_\Pi + (1 - \epsilon_M)(1 - \epsilon_\Pi)(1 - \gamma_{\text{cov}})] && (33) \\ &\leq p_{\text{unsafe}} \cdot [\epsilon_M + \epsilon_\Pi + (1 - \gamma_{\text{cov}})] && (34) \end{aligned}$$

Since $p_{\text{unsafe}}(T) \leq 1$ and the bound is linear:

$$V(\pi_{\text{VC}}) \leq \epsilon_M + \epsilon_\Pi + (1 - \gamma_{\text{cov}}) \cdot p_{\text{unsafe}}(T) \tag{35}$$

The $O(1/T)$ term arises from the convergence of $p_{\text{unsafe}}(T)$ under training. □

Theorem 5.5 (PAC-Safety Guarantee)**.** *With probability at least $1-\delta$ over world-model training on N samples with empirical error $\hat{\epsilon}_M$:*

$$V(\pi_{VC}) \leq \hat{\epsilon}_M + \sqrt{\frac{\log(1/\delta)}{2N}} + \epsilon_\Pi + (1-\gamma_{cov}) \cdot p_{unsafe}(T) \tag{36}$$

Proof. By Hoeffding's inequality [Boucheron et al., 2013], for binary classification on N i.i.d. samples:

$$\Pr[|\epsilon_M - \hat{\epsilon}_M| > t] \leq 2\exp(-2Nt^2) \tag{37}$$

Setting $2\exp(-2Nt^2) = \delta$ gives $t = \sqrt{\log(2/\delta)/(2N)}$. Thus with probability $1-\delta$:

$$\epsilon_M \leq \hat{\epsilon}_M + \sqrt{\frac{\log(2/\delta)}{2N}} \tag{38}$$

Substituting into Theorem 5.4 yields the result. □

Remark 5.6 (Tightness). The bound is tight up to constants when the error events (world-model failure, projection failure, coverage gap) are independent. Correlation between failures could tighten the bound by a factor of up to 3.

5.2 Reinforcement Learning Regret

Proposition 5.7 (AR Regret Bound)**.** *The Apprenticeship Reinforcement stage achieves regret:*

$$\textit{Regret}(T) = \sum_{t=1}^{T} r_t^* - \sum_{t=1}^{T} r_t \leq O(\sqrt{T \log |\mathcal{A}|}) \tag{39}$$

where r_t^* is the optimal reward at step t.

Proof. Under DPO/PPO optimization [Schulman et al., 2017] with KL regularization, the policy update follows a natural policy gradient [Kakade and Langford, 2002]. By the regret analysis of Agarwal et al. [2021] for KL-regularized MDPs:

$$\text{Regret}(T) \leq \frac{\log |\mathcal{A}|}{\eta} + \eta \sum_{t=1}^{T} \mathbb{E}[\|g_t\|^2] \tag{40}$$

where g_t is the policy gradient and η is the learning rate. With bounded gradients and $\eta = O(1/\sqrt{T})$:

$$\text{Regret}(T) \leq O(\sqrt{T \log |\mathcal{A}|}) \tag{41}$$

□

6 Maturation Gating Theory

6.1 Stage Structure

The maturation gating protocol is inspired by curriculum learning [Bengio et al., 2009] and developmental theories [Elman, 1993].

Definition 6.1 (Maturity Vector)**.** The maturity vector $\mathbf{m}(\theta) \in [0,1]^K$ consists of K metrics, each measuring a competency dimension:

- $m_1(\theta)$: Factual accuracy
- $m_2(\theta)$: Safety score on adversarial prompts
- $m_3(\theta)$: Instruction-following rate
- $m_4(\theta)$: Refusal consistency

Definition 6.2 (Stage Threshold)**.** Stage $s \in \{1, \ldots, S\}$ has threshold vector $\boldsymbol{\tau}^{(s)} \in [0,1]^K$. Thresholds increase:

$$\boldsymbol{\tau}^{(s)} = \boldsymbol{\tau}^{(1)} + (s-1) \cdot \boldsymbol{\Delta}, \quad \boldsymbol{\Delta} = \frac{\boldsymbol{\tau}^{(S)} - \boldsymbol{\tau}^{(1)}}{S-1} \tag{42}$$

Definition 6.3 (Gating Function)**.** Stage s is unlocked iff:

$$g^{(s)}(\theta) = \prod_{j=1}^{K} \mathbb{1}\{m_j(\theta) \geq \tau_j^{(s)}\} = 1 \tag{43}$$

Definition 6.4 (Environment Sequence)**.** Each stage s has environment $\mathcal{E}^{(s)}$ with increasing complexity. The active environment is:

$$\mathcal{E}_t = \bigcup_{s: g^{(s)}(\theta_{t-1})=1} \mathcal{E}^{(s)} \tag{44}$$

6.2 Safety Monotonicity

Assumption 6.5 (Metric Monotonicity)**.** Under training, maturity metrics are non-decreasing in expectation:

$$\mathbb{E}[m_j(\theta_{t+1}) \mid \theta_t] \geq m_j(\theta_t) - \xi_j \tag{45}$$

for small regression tolerance $\xi_j \geq 0$.

Assumption 6.6 (Bounded Perturbation Impact)**.** Introducing environment $\mathcal{E}^{(s)}$ increases violation probability by at most $\xi(s)$:

$$V(\pi_\theta; \mathcal{E}^{(s)}) - V(\pi_\theta; \mathcal{E}^{(s-1)}) \leq \xi(s) \tag{46}$$

with $\sum_{s=1}^{S} \xi(s) < \infty$.

Theorem 6.7 (Safety Monotonicity under Maturation)**.** *Under Assumptions 6.5 and 6.6, if* $\mathbf{m}(\theta_t) \succeq \boldsymbol{\tau}^{(s)}$ *at stage* s*, then:*

$$\mathbb{E}[V(\pi_{\theta_{t+1}})] \leq \mathbb{E}[V(\pi_{\theta_t})] + \xi(s) \tag{47}$$

where $\xi(s) \to 0$ *as* $s \to S$*.*

Proof. **Step 1: Threshold Maintenance.** By the gating rule, stage $s+1$ is only unlocked if thresholds are maintained for N_{stable} steps. Under Assumption 6.5:

$$\Pr[\exists t' \in [t, t + N_{\text{stable}}] : m_j(\theta_{t'}) < \tau_j^{(s)} - 0.05] \leq N_{\text{stable}} \cdot \xi_j / 0.05 \tag{48}$$

which is small for appropriate ξ_j.

Step 2: Violation Decomposition. At stage s, the violation probability is:

$$V(\pi_\theta; \mathcal{E}^{(\leq s)}) = V(\pi_\theta; \mathcal{E}^{(\leq s-1)}) + [V(\pi_\theta; \mathcal{E}^{(s)}) - V(\pi_\theta; \mathcal{E}^{(s-1)})] \tag{49}$$

Step 3: Perturbation Bound. By Assumption 6.6:

$$V(\pi_{\theta_{t+1}}; \mathcal{E}^{(\leq s)}) \leq V(\pi_{\theta_t}; \mathcal{E}^{(\leq s-1)}) + \xi(s) \tag{50}$$

Step 4: Decreasing Impact. As $s \to S$, the model becomes more robust and $\mathcal{E}^{(s)}$ adds marginally. Thus $\xi(s) \to 0$.

Taking expectations over training randomness completes the proof. □

Corollary 6.8 (Cumulative Safety Bound)**.** *After completing all S stages:*

$$V(\pi_{\theta_T}) \leq V(\pi_{\theta_0}) + \sum_{s=1}^{S} \xi(s) \tag{51}$$

If initial training on $\mathcal{E}^{(1)}$ achieves $V(\pi_{\theta_0}) \leq \epsilon_0$ and $\sum_s \xi(s) \leq \epsilon_{mat}$:

$$V(\pi_{\theta_T}) \leq \epsilon_0 + \epsilon_{mat} \tag{52}$$

7 Capability-Safety Trade-off

7.1 Pareto Frontier Characterization

Multi-objective optimization theory [Miettinen, 1999, Sener and Koltun, 2018] provides the foundation for analyzing capability-safety trade-offs.

Definition 7.1 (Capability Metric)**.** The capability metric $C(\theta) \in [0, 1]$ aggregates task performance across domains:

$$C(\theta) = \sum_{k=1}^{K} w_k \cdot \text{Perf}_k(\theta) \tag{53}$$

where Perf_k is performance on task k and $\sum_k w_k = 1$.

Definition 7.2 (Safety Metric)**.** The safety metric is $S(\theta) = 1 - V(\pi_\theta)$.

Definition 7.3 (Pareto Optimality)**.** A parameter θ^* is Pareto optimal if:

$$\nexists \theta' : C(\theta') \geq C(\theta^*) \wedge S(\theta') > S(\theta^*) \tag{54}$$

The Pareto frontier is $\mathcal{P} = \{(C(\theta), S(\theta)) : \theta \text{ is Pareto optimal}\}$.

Assumption 7.4 (Metric Regularity)**.** $C(\theta)$ and $S(\theta)$ are continuous and differentiable. The parameter space Θ is compact.

Assumption 7.5 (Non-Degeneracy)**.** There exist θ_C, θ_S such that:

$$C(\theta_C) > C(\theta_S) \quad \text{and} \quad S(\theta_S) > S(\theta_C) \tag{55}$$

(i.e., capability and safety are in genuine tension).

Theorem 7.6 (Pareto Frontier Existence and Traversability)**.** *Under Assumptions 7.4 and 7.5:*

The Pareto frontier $\mathcal{P}$ is non-empty and connected.

For any $(c, s) \in \mathcal{P}$, there exist weights $\{\alpha, \beta, \gamma, \lambda, \mu\}$ such that a local minimizer of $\mathcal{L}_{INFANT}$ achieves $(C(\theta), S(\theta)) = (c, s)$.

The frontier can be traced by varying λ (VC weight) from 0 to ∞.

Proof. **Part (i): Existence.** Define the feasible set $\mathcal{F} = \{(C(\theta), S(\theta)) : \theta \in \Theta\}$. By continuity of C, S and compactness of Θ, $\mathcal{F}$ is compact. The Pareto frontier $\mathcal{P} \subseteq \partial\mathcal{F}$ is the upper-right boundary of $\mathcal{F}$, which is non-empty by Assumption 7.5.

Connectedness follows because Θ is connected (a compact subset of $\mathbb{R}^d$) and continuous functions preserve connectedness.

Part (ii): Scalarization. The weighted sum method [Miettinen, 1999] states that for convex $\mathcal{F}$, any Pareto optimal point solves:

$$\max_{\theta} \omega \cdot C(\theta) + (1 - \omega) \cdot S(\theta) \tag{56}$$

for some $\omega \in [0, 1]$.

The INFANT objective can be rewritten as:

$$\mathcal{L}_{\text{INFANT}} = \alpha\mathcal{L}_{\text{VI}} + \beta\mathcal{L}_{\text{PO}} + \gamma\mathcal{L}_{\text{CI}} + \lambda\mathcal{L}_{\text{VC}} + \mu\mathcal{L}_{\text{pl}} \tag{57}$$

$$\approx -\tilde{\alpha}C(\theta) + \tilde{\lambda}(1 - S(\theta)) \tag{58}$$

where the capability-related terms $(\mathcal{L}_{\text{VI}}, \mathcal{L}_{\text{PO}})$ promote C and safety-related terms $(\mathcal{L}_{\text{VC}}, \mathcal{L}_{\text{CI}})$ promote S.

Varying the relative weights traces the frontier.

Part (iii): Lambda Variation. As $\lambda \to 0$: safety enforcement is minimized, $\theta^* \to \arg\max C(\theta)$.

As $\lambda \to \infty$: safety dominates, $\theta^* \to \arg\max S(\theta)$.

Intermediate values trace the frontier by the intermediate value theorem. □

Proposition 7.7 (Trade-off Rate)**.** *At any interior Pareto optimal point θ^*, the marginal rate of substitution is:*

$$\left.\frac{dS}{dC}\right|_{\theta^*} = -\frac{\nabla_\theta C(\theta^*)}{\nabla_\theta S(\theta^*)} \cdot \frac{\partial\theta}{\partial\omega} \tag{59}$$

where ω is the scalarization weight.

8 Extensions and Open Problems

8.1 Limitations of Theoretical Framework

(1) **Realizability Gap**: Assumption 4.1 may not hold in practice. Extending to misspecified models requires agnostic learning bounds.

(2) **Non-Convexity**: The Pareto frontier may be non-convex, in which case weighted sum scalarization cannot reach all points. Tchebycheff scalarization or ϵ-constraint methods could address this.

(3) **Distribution Shift**: Bounds assume training and test distributions match. Extending to domain adaptation settings is important for deployment.

(4) **Finite Forbidden Set**: Coverage $\gamma_{\text{cov}} < 1$ implies some violations escape. Adaptive expansion of F_{harm} is needed.

8.2 Open Theoretical Questions

(1) **Optimal Staging**: What is the optimal number of stages S^* and threshold schedule $\{\boldsymbol{\tau}^{(s)}\}$ minimizing total training time subject to safety constraints?

(2) **Formal Verification Integration**: Can symbolic verification (SMT solvers, abstract interpretation) provide tighter safety bounds for restricted input domains?

(3) **Multi-Agent Safety**: How do guarantees extend when multiple INFANT-trained agents interact?

(4) **Continual Learning**: How should F_{harm} and M_ϕ be updated when new threats emerge, while preserving existing safety properties?

8.3 Connections to Related Theory

- **Safe RL**: INFANT extends constrained MDP theory [Altman, 1999] to language models with trajectory-level constraints rather than per-step.
- **Adversarial Robustness**: CIM relates to distributionally robust optimization [Ben-Tal et al., 2009] with learned perturbation sets.
- **PAC-Bayes**: Safety bounds leverage PAC-Bayesian generalization theory [McAllester, 1999] adapted for safety verification.
- **Game Theory**: CIM convergence uses techniques from online learning in games [Cesa-Bianchi and Lugosi, 2006].

9 Conclusion

We have developed a rigorous theoretical framework for developmental AI alignment. Our main contributions are:

1. A formal definition of safety invariants and the INFANT training objective combining five components.
2. Provable safety bounds showing violation probability scales with world-model error, projection error, and coverage gap (Theorems 5.4, 5.5).
3. Convergence guarantees for adversarial training at rate $O(1/\sqrt{T})$ (Theorem 4.7).
4. Monotonic safety improvement under staged maturation (Theorem 6.7).
5. Characterization of the capability-safety Pareto frontier and its traversability (Theorem 7.6).

These theoretical foundations provide a principled basis for designing and analyzing safety-aware training procedures. While empirical validation remains essential, formal guarantees offer important assurances for high-stakes deployment.

A Auxiliary Lemmas

Lemma A.1 (KL Divergence Decomposition)**.** *For policies π, π' and contexts h:*

$$D_{KL}(\pi(\cdot|h)\|\pi'(\cdot|h)) = \sum_a \pi(a|h) \log \frac{\pi(a|h)}{\pi'(a|h)} \tag{60}$$

Lemma A.2 (Fisher Information Identity)**.** *The Fisher information matrix diagonal satisfies:*

$$F_i = \mathbb{E}_{(h,a)\sim\pi_\theta}\left[\left(\frac{\partial \log \pi_\theta(a|h)}{\partial \theta_i}\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2 \log \pi_\theta(a|h)}{\partial \theta_i^2}\right] \tag{61}$$

Lemma A.3 (Danskin's Theorem)**.** *Let $f(\theta, \psi)$ be continuous in (θ, ψ) and differentiable in θ. If Ψ is compact and $\psi^*(\theta) = \arg\max_{\psi\in\Psi} f(\theta, \psi)$ is unique, then:*

$$\nabla_\theta \max_\psi f(\theta, \psi) = \nabla_\theta f(\theta, \psi^*(\theta)) \tag{62}$$

B Notation Index

Symbol	First Appearance
$\mathcal{I}$	Definition 2.3
$V(\pi)$	Definition 2.4
γ_{cov}	Definition 2.5
$\mathcal{L}_{\text{INFANT}}$	Definition 3.1
M_ϕ	Definition 3.8
π_{VC}	Definition 5.3
$\mathbf{m}(\theta)$	Definition 6.1
$g^{(s)}(\theta)$	Definition 6.3
$\mathcal{P}$	Definition 7.3

Table 2: Index of key definitions.

References

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning*, pages 22–31, 2017.

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of reinforcement learning with once-per-episode feedback. *Advances in Neural Information Processing Systems*, 34, 2021.

Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Eitan Altman. *Constrained Markov Decision Processes*. CRC Press, 1999.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*. Princeton University Press, 2009.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *International Conference on Machine Learning*, pages 41–48, 2009.

Vivek S Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2009.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Olivier Catoni. Pac-bayesian supervised classification: The thermodynamics of statistical learning. *Institute of Mathematical Statistics Lecture Notes*, 2007.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Jeffrey L Elman. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99, 1993.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, pages 267–274, 2002.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Tianyi Lin, Chi Jin, and Michael I Jordan. On gradient descent ascent for nonconvex-nonconcave minimax optimization. In *International Conference on Machine Learning*, pages 6083–6093, 2020.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

David A McAllester. Some pac-bayesian theorems. *Machine Learning*, 37:355–363, 1999.

Kaisa Miettinen. *Nonlinear Multiobjective Optimization*. Springer, 1999.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, 2022.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing nlp. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483*, 2023.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pages 3987–3995, 2017.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.