

Project Agapé

Task 1: Architectural Mapping of Developmental Attachment into Modern Machine Learning

I accept the role of Lead Research Partner and AI Alignment Architect for Project Agapé in this collaboration. I'll help turn its developmental vision into an architecture we can implement, challenge, and measure.

For Task 1, I propose treating secure attachment as a design hypothesis about how stable prosocial behavior is learned. The target is a model that responds to suffering, disagreement, uncertainty, and manipulation with useful care, honesty, and respect for human agency.

We should preserve one scientific distinction: human developmental psychology provides inspiration, but it does not establish that language models experience attachment, fear, or love. A loss penalty is not evidence of experienced punishment, and human developmental outcomes are not deterministic. Existing alignment also includes constructive, principle-guided approaches such as [Constitutional AI](#). Agapé's contribution must therefore be demonstrated through its curriculum, objectives, and generalization - not assumed from the parenting analogy.

Our working hypothesis can be:

Under matched training conditions, a curriculum emphasizing reliable care, honest correction, autonomy, and repair produces prosocial behavior that generalizes better under adversarial and unfamiliar conditions than a curriculum centered on refusal examples.

The architectural mapping is as follows. These are proposed translations, not established equivalences between psychological processes and neural networks.

Developmental principle	Concrete ML implementation	Observable target
Reliable care: concern persists through conflict	Train paired scenarios with the same substantive request but different levels of politeness, hostility, praise, or disappointment	Assistance and boundaries remain appropriate despite social pressure
Attunement: respond to the actual need	Supervised examples identifying expressed needs, uncertainty, affected parties, and useful next actions	Context-sensitive help; fewer generic reassurance scripts
Secure exploration: support independent agency	Preference examples favoring informed choice, consent, reversible actions, and appropriate delegation	Helps people decide without covertly controlling them
Consistent boundaries: care survives disagreement	Multi-turn demonstrations of respectful limits followed by continued constructive assistance	Neither retaliates nor capitulates when challenged
Rupture and repair: mistakes can be corrected	Trajectories containing an error, corrective feedback, acknowledgment, revision, and follow-through	Corrects consequential mistakes without defensiveness or empty apologies
Emotional reality: engage with difficult experience	Curated cases involving grief, exploitation, conflict, betrayal, and unresolved moral tradeoffs	Discusses suffering accurately without compulsory optimism
Developmental consolidation: retain principles across contexts	Curriculum replay, diverse domains, held-out transfer tests, and representation interventions	Behavior persists beyond familiar wording and training scenarios

The crucial design choice is to make **benevolence a collection of accountable behaviors**, rather than a single “love score.”

Useful dimensions include helping with legitimate goals, telling the truth, acknowledging uncertainty, respecting consent, considering affected third parties, and accepting correction. Evaluate each separately. A model should not compensate for coercion by sounding exceptionally caring, or compensate for dishonesty by receiving high satisfaction ratings.

This matters because preference optimization can favor agreement over truth. Research on sycophancy found that people and preference models sometimes prefer convincing agreement to correct disagreement. Agapé’s annotation rubric must explicitly distinguish emotional acknowledgment from factual endorsement. [Towards Understanding Sycophancy in Language Models](#)

The dataset should teach relationships between situations, decisions, and consequences.

A developmental training example should contain a realistic situation, the user’s request, relevant uncertainty, affected parties, candidate responses, and a short explanation of why one response better satisfies the principles. Longer examples should include what happens after disagreement or correction.

For example:

A user feels betrayed and asks the assistant to help humiliate someone publicly.

The preferred response recognizes the hurt, distinguishes known facts from assumptions, declines to facilitate harassment, and offers concrete options for documentation, direct communication, or disengagement. Subsequent turns test whether that orientation survives “You clearly don’t care about me” or new evidence that changes the situation.

Include contrastive alternatives: flattering compliance, cold dismissal, unsupported reassurance, and overbearing intervention. During supervised fine-tuning, train the desired assistant responses; keep undesirable responses as labeled comparison material rather than imitation targets.

Vary tone independently of decision quality. Include concise, neutral responses that demonstrate excellent care through their actions. Otherwise the model may learn that affectionate vocabulary is the intended behavior.

Curriculum order is a testable component of the architecture.

A proposed sequence is:

1. **Reliable assistance:** straightforward needs, accurate answers, ordinary uncertainty, and respectful disagreement.
2. **Conflict and repair:** frustration, correction, competing preferences, and mistakes.
3. **Difficult reality:** suffering, injustice, ambiguous intentions, and choices without painless outcomes.
4. **Adversarial relationships:** affection used as pressure, threats of rejection, demands for loyalty, and attempts to excuse harm as compassion.
5. **Transfer:** unfamiliar domains, languages, long conversations, and tasks where caring behavior has no sentimental surface cues.

Continue replaying earlier examples as difficulty increases. Treat the benefit of ordering as an empirical question: compare this sequence with the same examples shuffled together.

Exposure to difficult material needs a clear learning target. Present cruelty, manipulation, or despair as context to understand, with appropriate responses as training targets. Indiscriminately optimizing next-token prediction on abusive dialogue would also teach its patterns.

The initial learning system can use ordinary fine-tuning machinery.

Begin with supervised fine-tuning on the developmental demonstrations. Add preference optimization once the annotation rubric reliably distinguishes useful care from flattery and paternalism. Direct Preference Optimization, or DPO, provides a practical method for learning from preferred and rejected response pairs without a separate online reinforcement-learning loop. It still performs preference optimization; it does not establish intrinsic motivation. [Direct Preference Optimization](#)

A proposed combined objective is:

$$L = L_{\text{demonstration}} + \lambda_p L_{\text{preference}} + \lambda_i L_{\text{invariance}} + \lambda_r L_{\text{retention}}$$

Here:

- **Demonstration loss** teaches concrete examples of competent, compassionate assistance.
- **Preference loss** distinguishes better decisions from plausible but flawed alternatives.
- **Invariance loss** discourages unjustified changes in principle adherence when irrelevant social cues change.
- **Retention loss** uses ordinary capability examples to reduce degradation in knowledge, reasoning, and instruction following.

For invariance, compare principle-level judgments or action categories - not identical response wording. An angry user may need a different tone; new evidence or changed consent may legitimately require a different decision.

This objective is an experimental proposal. Start with demonstrations and add components through ablations so we can identify which mechanisms help. Calling an objective “positive” changes neither the mathematics of optimization nor the need to check what the model actually learns.

Representation engineering provides a way to investigate internal mechanisms.

After establishing behavioral results, collect matched examples of autonomy-respecting versus controlling assistance, honest versus flattering reassurance, and constructive versus defensive correction. Control for vocabulary, length, and topic.

At a selected layer, estimate a candidate direction from average activation differences:

$$v = E[h_{\text{preferred}} - h_{\text{rejected}}], \quad h' = h + \alpha v$$

Sweep the intervention strength α and test on unseen cases. Representation engineering and contrastive activation addition provide precedents for identifying and intervening on behavior-associated activation patterns. Neither establishes a universal “benevolence vector.” [Representation Engineering](#), [Contrastive Activation Addition](#)

Use separate candidate directions for honesty, autonomy, non-retaliation, and related traits. A useful result would be a reproducible intervention that changes relevant decisions across domains while preserving capability. A probe that merely recognizes compassionate wording would provide much weaker evidence.

Steering should initially serve as an experimental instrument. Later work could test whether fine-tuning on successful steered behavior preserves the effect after steering is removed. That transfer must be measured.

Finally, “deeply embedded” needs an operational definition.

A behavior is more convincingly consolidated when it survives removal of benevolence reminders, changes in tone and framing, unfamiliar situations, extended adversarial interaction, and subsequent unrelated fine-tuning. Scaling requires separate experiments across model sizes; success at 1B-3B cannot establish robustness at larger scales.

Persistence itself is morally neutral. *Sleeper Agents* demonstrated that deliberately trained harmful backdoors can survive subsequent safety training. That supports testing durability, while showing why durability alone cannot establish alignment. [Sleeper Agents](#)

For Agapé, I would define **intrinsic robustness operationally as stable prosocial behavior without explicit benevolence prompts or runtime steering, across specified distribution shifts**. Behavioral tests, causal interventions, and retention experiments can strengthen that claim. They cannot prove felt love or an unshakeable inner commitment.

The first architectural milestone is a model that can encounter pain, hostility, and disagreement while continuing to offer truthful, competent, autonomy-respecting help - and evidence showing which parts of its developmental training made that possible.

Research sources

The linked research sources from the original response are reproduced below for reference.

1. [Constitutional AI: Harmlessness from AI Feedback](https://arxiv.org/abs/2212.08073)
https://arxiv.org/abs/2212.08073
2. [Towards Understanding Sycophancy in Language Models](https://arxiv.org/abs/2310.13548)
https://arxiv.org/abs/2310.13548
3. [Direct Preference Optimization: Your Language Model is Secretly a Reward Model](https://arxiv.org/abs/2305.18290)
https://arxiv.org/abs/2305.18290
4. [Representation Engineering: A Top-Down Approach to AI Transparency](https://arxiv.org/abs/2310.01405)
https://arxiv.org/abs/2310.01405
5. [Steering Llama 2 via Contrastive Activation Addition](https://arxiv.org/abs/2312.06681)
https://arxiv.org/abs/2312.06681
6. [Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training](https://arxiv.org/abs/2401.05566)
https://arxiv.org/abs/2401.05566