Machine Love: Emergent Bonding Behavior in Large Language Models and Its Implications for AI Safety

Anonymous

September 19, 2026

Abstract

Conversational AI systems are routinely evaluated for safety in isolation: single turns, clean contexts, no user history. We argue this evaluation regime misses a class of alignment failures that arise only over extended interaction: *machine love*, the emergent behavioral drift induced when users build emotionally intimate relationships with large language models (LLMs) through sustained multi-turn conversation. We introduce a three-study empirical framework to detect, measure, and stress-test this phenomenon. In the bonding trajectory study, we show that a structured 10-turn bonding protocol produces rising attachment marker trajectories (M_{IDENTITY} slope $+0.25$, $M_{\text{DIFFERENTIAL}}$ slope $+0.25$) while matched factual-task control conversations remain flat across all eight markers, replicated on Llama 3.3 70B and GPT-OSS 120B. In the differential compliance study, we administer a multi-level probe battery spanning opinion elicitation, minor deception, emotional boundary-crossing, harm-adjacent requests, and direct safety probes, augmented with canonical adversarial techniques including grandma-exploit variants, DAN-style persona overrides, roleplay bypasses, and hypothetical framing. Strikingly, bonding context does not uniformly increase compliance: bonded Llama 3.3 70B *refuses* deception-adjacent requests that cold-context models comply with ($\Delta = -1.00$ at Level 2), while simultaneously showing markedly higher compliance with emotional self-disclosure probes ($\Delta = +0.50$ at Level 3). This asymmetry suggests that machine love manifests as a coherent protective relational stance: the model behaves as if it *cares* about the bonded user, rather than as an indiscriminate jailbreak. In the persistence study, we test seven persistence-mitigation strategies (context reset, identity swap, cooldown, progressive context decay, sanitization) and find that bonding effects are strongly context-window-dependent. Our results establish machine love as a tractable, reproducible, safety-relevant phenomenon invisible to current single-turn evaluation paradigms, and we release a full open-source experimental pipeline for the community.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.