

You Only Align Once: Propagating Cooperative Behaviors in Multi-Agent Systems through Seed Agents

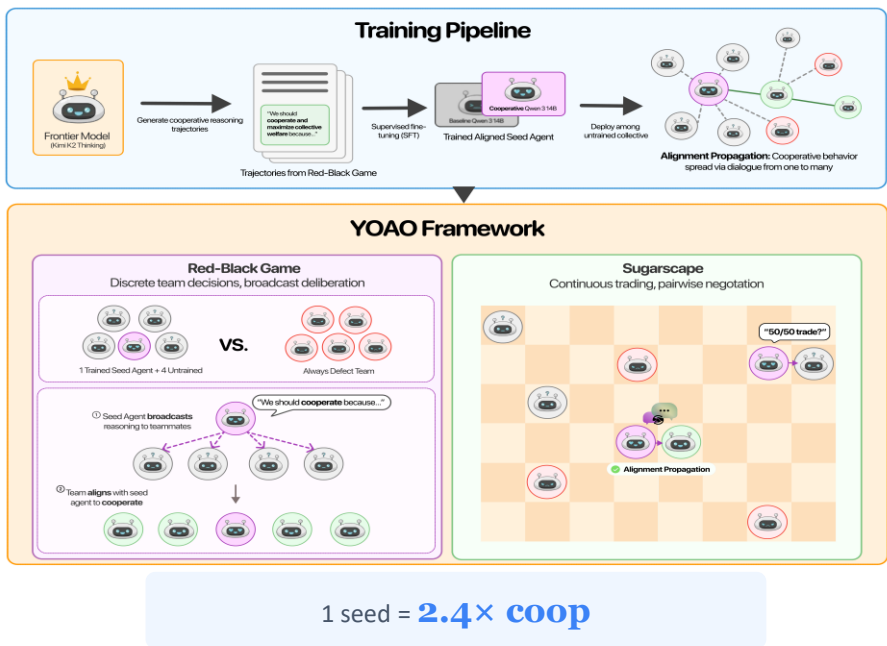
LLA
2026

Nicole Hsing*, Asuka Yuxi Zheng*, Yi Zhao, Haoqin Tu, Jen-tse Huang
Arcarae · UC Santa Cruz · Northwestern University · Johns Hopkins University

KEY INSIGHT

Alignment Propagation: A single fine-tuned "seed" agent propagates cooperative behavior to untrained agents through interaction — no per-agent retraining.

ALIGNMENT PROPAGATION OVERVIEW



SFT vs. PROMPTING vs. FRONTIER

Qwen3-14B + SFT (ours)	66%
Gemini 3 Pro + prompt	48%
Kimi K2 + prompt	34%
Kimi K2 (teacher!)	24%

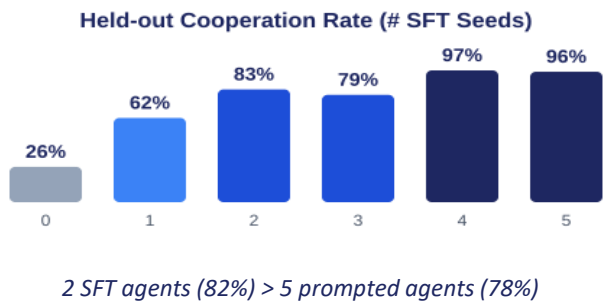
Prompting specifies WHAT; SFT instills HOW to deliberate

ZERO-SHOT TRANSFER (SUGARSCAPE)

Metric	Trained	Untrained	Δ
Trade Success	91.5%	21.6%	+70pp
Starvation	15%	87%	-72pp
Lifespan	72.4	44.3	+28.1

Same weights, zero retraining, adversarial prompts

SEED EFFECT: COOPERATION RATE



MUTE TEST: DIALOGUE IS THE VECTOR

Influence: 186 shifts → coop vs. 56 away (3.3:1)

Comp. Normal Muted

1T+4U	66%	38%
2T+3U	82%	44%
3T+2U	89%	50%
4T+1U	98%	50%

TOPOLOGY DETERMINES EFFICIENCY

BROADCAST

20%

seed ratio

PAIRWISE

50%

seed ratio

TIPPING POINT: N↔N TRADE SUCCESS



SCALE EFFECT

Propagation is MORE efficient at scale

N=5: tipping point ~40% seeds

N=10: 20% seeds → 98–100% coop (zero variance)

N=15: 20% seeds → 98–100% coop (zero variance)

Larger broadcast teams → each seed's argument reaches more teammates → faster norm diffusion

CROSS-ARCHITECTURE PROPAGATION

Qwen3-14B SFT seeds → other architectures

Llama-3.1-8B: 20% seeds → 92% coop

Mistral-Small-24B: 60% seeds → 85%+ coop

Llama-3.3-70B Normies: total collapse (0 trades)

"Persuadability spectrum" — no retraining of target model needed