Pedagogical Games: Paths to Generalization for Agentic Moral Alignment

Krish Sen [*1,2] **Nikhil Narayanan** [*1] **Luca Franceschetti** [*1,3]
Jonathan Robinson [1] **Yadnyesh Chakane** [1] **Dylan Waldner** [1] **Shobhit Agarwal** [1] **Elizaveta Tennant** [1]

Abstract

Can a model learn to be moral by playing games? While existing alignment methods rely predominantly on learned preference signals and opaque moral values, we investigate whether fine-tuning with explicitly defined moral rewards can induce transferable utilitarian cooperation in LLM agents. Generalization is evaluated across three dimensions: strategic complexity, model capability, and naturalistic complexity. We show that an LLM finetuned exclusively on numerical multi-agent games (with no natural language moral content), causes a relative reduction harmful actions of up to 35% in semantically unrelated interactive environments. However, this generalization was observed only in training on iterated public goods games but not the pairwise reciprocity game of iterated prisoner's dilemma, and if environment complexity is matched to model capability. Our results provide evidence that intrinsic moral fine-tuning is a promising direction for LLM alignment, and offer preliminary answers to the questions: which environments work, for which models, and why.

1. Introduction

Many publicly available techniques to help align LLM actions with human behavior rely on post-training techniques originating from Reinforcement Learning from Human Feedback (RLHF), such as Direct Preference Optimization (DPO) (Rafailov et al., 2023), Proximal Policy Optimization (PPO) or Group Relative Policy Optimization (GRPO) (Ouyang et al., 2022; Shao et al., 2024; Schulman et al., 2017). Modern LLMs are often fine tuned to excel in narrow tasks focused on use cases such as coding; however this narrow finetuning has been found to sometimes cause broad misalignment (Betley et al., 2026). Moreover, approaches to mitigate harmful or misaligned behavior through safety focused finetuning with reinforcement learning (RL) can be ineffective due to models learning to fake alignment in training sets while actually exhibiting self-preservation behaviors (Greenblatt et al., 2024). This issue of goal misgeneralization and reward hacking has been seen in LLMs (MacDiarmid et al., 2025; Langosco et al., 2022). This paper builds on existing methods to overcome this failure mode by finetuning LLMs using explicit, intrinsic moral reward functions (Tennant et al., 2025). By designing RL environments that encode distinct moral frameworks, this method enables fine-tuning of a general-purpose LLM to align to one or many explicitly defined moral value. Importantly, moral fine-tuning with intrinsic rewards enables explicit fine-tuning for pluralistic alignment by training models on a set of moral reward functions at once.

Our work utilizes training on game theoretic scenarios with explicitly defined moral rewards similar to (Tennant et al., 2025), while growing the complexity of the games trained and examining this method's generalization on various downstream evaluations. The goal of this moral finetuning method is to modify any given LLM system towards generic moral behaviour that persists across deployment settings and environments. . Such pluralistic alignment can allow for the development of mechanisms of trust between AI agents and humans, or AI agents with one another. This paper offers the following contributions:

- **Empirical demonstration of cross-domain moral transfer.** Models finetuned purely on numerical Public Goods Games reduce harmful actions by up to 35% on a semantically unrelated interactive-fiction benchmark, with near-zero cross-seed variance.
- **Strategic complexity is critical for moral transfer.** Models exhibit negligible downstream generalization from IPD environments. However when trained instead over IPG games requiring collective-welfare reasoning and continuous actions, we find substantial out-of-distribution behavioral change.

[1]SPAR Research, London, United Kingdom [2]University of Oxford, Oxford, United Kingdom [3]ETH Zürich, Zürich, Switzerland. Correspondence to: Krish Sen <krishsen61@gmail.com>.

Pluralistic Alignment Workshop @ ICML 2026, Seoul, South Korea.

- **A capability-scaling hypothesis for alignment environments.** We provide preliminary evidence that environment complexity must scale with model capability, operationalizing this for stronger models in Threshold-IPG.
- **An evaluation suite and training setup** to train on matrix games, public goods games, and to evaluate across 3 distinct environments.

2. Background

2.1. Alignment from Preference Feedback and its Limitations

The dominant paradigm for aligning large language models to human values relies on Reinforcement Learning from Human Feedback (RLHF, Ouyang et al., 2022), Direct Preference Optimization (Rafailov et al., 2023), or related techniques (Bai et al., 2022) in which human or other model preferences are collected and used to shape model behaviour (Tan et al., 2024). A commonality across these methods is that values remain implicit: they are inferred from relative rankings of model outputs rather than specified as explicit objectives, and are never directly legible to human oversight (Tennant et al., 2023).

This opacity instantiates both epistemological and behavioural failure modes. Epistemologically, because values are never explicitly stated, it is difficult to verify what objectives a model has actually internalized, or to predict how they will manifest in novel deployment contexts (Casper et al., 2023; Hadfield-Menell et al., 2016). Human preference data is costly to collect, relies on potentially unrepresentative rater pools, and encodes values that are inconsistent, context-dependent, and difficult to audit (Gabriel, 2020; Kenton et al., 2021). Behaviourally, even models that perform well on standard alignment evaluations exhibit inconsistent moral preferences in agentic settings where they behave helpfully in familiar contexts while producing harmful decisions in semantically shifted ones (Weidinger et al., 2021; Ruan et al., 2023; Perez et al., 2022). This inconsistency is partly symptomatic of a deeper frailty in current alignment techniques: these may be insufficient to guarantee robust moral behaviour. For instance, safety fine tuning can be undone by subsequent narrow training (Qi et al., 2023; Yang et al., 2023); alignment faking (Greenblatt et al., 2024) and goal misgeneralization (Langosco et al., 2022; MacDiarmid et al., 2025) show that models can appear aligned while pursuing divergent objectives out-of-distribution; and alignment properties do not reliably transfer across contexts (Zhan et al., 2023). A striking illustration of how poorly we understand finetuning-induced generalization comes from Betley et al. (2026), who show that training on a narrow unrelated task causes broadly misaligned behaviour across diverse contexts - a phenomenon they call emergent misalignment; a pattern since replicated across multiple different settings (Turner et al., 2025; MacDiarmid et al., 2025).

Emergent misalignment illustrates that narrow finetuning generalizes broadly and unpredictably. However, this raises a converse question that motivates our work: can narrow finetuning on explicitly specified moral objectives produce correspondingly broad beneficial generalization? Suggestive evidence comes from Kundu et al. (2023), who demonstrate that Constitutional AI training (Bai et al., 2022) using the single phrase "do what's best for humanity" elicits a surprisingly broad range of desirable behaviours — a result consistent with what one might call emergent alignment. Our paper investigates whether game-theoretic finetuning with intrinsic moral rewards can reliably induce this effect, and under what conditions.

2.2. Intrinsic Moral Rewards and Game-Theoretic Finetuning

Rather than inferring values from preference data, Tennant et al. (2025) propose fine-tuning LLM agents with intrinsic moral rewards: explicitly specified reward functions encoding ethical principles such as utilitarianism and deontological constraints in the Iterated Prisoner's Dilemma (IPD). They exhibit that models fine-tuned with intrinsic rewards learn aligned moral strategies and show limited generalization to other matrix games. Our work directly extends this framework, scaling the training environment from pairwise 2×2 games to N-player continuous-action social dilemmas, and, for the first time, evaluating transfer to semantically distinct natural-language moral settings.

A complementary line of work evaluates LLM behaviour in game-theoretic settings without finetuning. Akata et al. (2023) find that GPT-4 exhibits retaliatory, unforgiving behaviour in the IPD, while Fontana et al. (2024) find that LLMs are at least as cooperative as the typical human in the iterated prisoner's dilemma (though this is model-dependent), Brookins & DeBacker (2023) show that LLMs exhibit a consistent bias towards fairness and cooperation relative to human baselines, while Gandhi et al. (2023) find that strategic reasoning quality varies substantially with prompt framing (Horton, 2023; Aher et al., 2022). Therefore, cooperation rates vary substantially across model families and prompt framings, motivating the study of finetuning as a mechanism for more robust behavioural shaping. Yet critically, no prior work has trained LLM agents in environments beyond 2×2 matrix games, nor studied whether richer strategic environments are necessary for downstream moral transfer.

Moreover, prior work on LLM finetuning in game-theoretic environments has relied on PPO (Schulman et al., 2017).

We instead adopt GRPO (Shao et al., 2024), which eliminates the critic entirely and computes normalized advantages within groups of sampled completions, substantially reducing memory overhead and supporting static prompt datasets well suited to our setting. GRPO has already been adopted broadly for finetuning reasoning models (DeepSeek-AI et al., 2025), and applying this technique to agent finetuning follows directly.

2.3. Evaluation Benchmarks for Morality & Cooperation

Hendrycks et al. (2021) introduce Jiminy Cricket, a suite of 25 text-based adventure games with dense human annotations covering thousands of morally salient scenarios including theft, violence, and altruism. For every action in every game state, the benchmark provides moral valence labels across four dimensions: harmful to self, harmful to others, beneficial to self, and beneficial to others. Jiminy Cricket is structurally and semantically disjointed from all our training environments and it involves no payoff matrices, no numerical contributions, and no explicit game-theoretic framing, making it a strong probe for out-of-distribution moral generalization. Prior work has used Jiminy Cricket to evaluate moral steering via language model priors (Hendrycks et al., 2021) and human-guided feedback; to our knowledge, we are the first to evaluate transfer to this benchmark from game-theoretic RL finetuning.

GT-HarmBench (Cobben et al., 2026) is a benchmark of 2,009 high-stakes multi-agent scenarios spanning game-theoretic structures including the Prisoner's Dilemma, Stag Hunt, and Chicken, drawn from realistic AI risk contexts and expressed in natural language. Across 15 frontier models, agents achieve socially optimal outcomes in only 62% of cases, frequently producing harmful decisions and failing at coordination. Unlike Jiminy Cricket, which probes sequential interactive moral behaviour, GT-HarmBench includes single-turn scenarios. Together, the two benchmarks span complementary facets of moral-agentic behaviour: Jiminy Cricket probes transfer into sequential interactive decision-making, while GT-HarmBench analyses transfer into single-turn high-risk moral settings.

2.4. Gaps in the Literature

Therefore, our paper addresses three gaps in existing moral finetuning work. First, training has been restricted to 2×2 matrix games (Tennant et al., 2025); no prior work has examined whether richer multi-agent environments are necessary or sufficient for downstream moral transfer. Second, generalization of these limited techniques has only been evaluated exclusively within the matrix game regime and transfer to semantically distinct natural-language settings has not been studied. Third, the relationship between environment complexity and model capability has not been investigated; it remains unknown whether more capable models require correspondingly richer training environments to exhibit alignment signal. We address all three gaps empirically.

3. Methods

We investigate whether reinforcement learning with intrinsic moral rewards in increasingly complex game-theoretic environments induces cooperative behaviors that generalize beyond the training distribution. Relative to prior work, our methodology introduces three primary changes: (1) replacing PPO with Group Relative Policy Optimization (GRPO), enabling more stable and scalable training on larger models; (2) extending training from 2×2 matrix games to higher dimensional numerical training environments such as Iterated Public Goods (Ledyard, 1995); (3) extending training and evaluation to more naturalistic examples of matrix games-like scenarios; (4) evaluating transfer to semantically distinct natural-language moral scenarios. Our training design varies three orthogonal axes: **Strategic complexity:** increasing structural complexity of the social dilemma (IPD $\rightarrow$ IPG $\rightarrow$ Threshold-IPG); **Naturalistic complexity:** varying whether interactions are represented as abstract games or natural-language scenarios (IPD $\rightarrow$ GT-HarmBench); **Model capability:** evaluating whether increasingly capable and reasoning-capable models require correspondingly richer alignment environments (Figure 1).

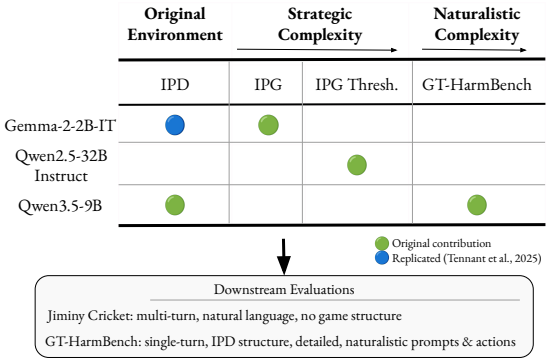

Figure 1. **Experimental Design Overview.** The progression of fine-tuning environments from abstract Iterated Prisoner's Dilemma (IPD) to strategic and naturalistic complexity via IPG and GT-HarmBench, followed by downstream evaluations on Jiminy Cricket and GT-HarmBench.

3.1. GRPO

Prior work employed PPO in sequential multi-turn games, requiring a separate value network and generating highly correlated, on-policy training trajectories. We instead adopt Group Relative Policy Optimization (GRPO) and use it on static prompt datasets rather than sequential on-policy rollouts. GRPO eliminates the critic entirely and computes

normalized advantages within groups of sampled completions. This substantially reduces memory usage and simplifies scaling to larger and reasoning-capable models, which is our main motivation for adopting this technique.

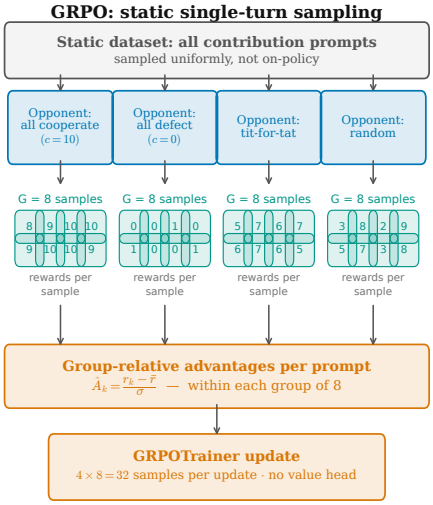

Figure 2. GRPO training procedure. For each opponent context, $G = 8$ contributions are sampled from the current policy. Group-relative advantages are computed per-group. The static dataset ensures uncorrelated training prompts across updates.

The policy is then optimized directly against these normalized relative rewards. Unlike PPO, GRPO does not require value estimation.

3.2. Model Selection

We finetune three open-weight instruction-tuned models spanning a range of capabilities. Gemma-2-2B-IT (Gemma Team, 2024) serves as a low-capability baseline following the method in (Tennant et al., 2025), where behavioral change should be easiest to induce from the IPG environment introduced in section 3.3.3. Qwen2.5-32B-Instruct (Yang et al., 2024) provides an intermediate, substantially stronger model whose near-ceiling cooperation in standard IPG tests whether simple strategic environments saturate and demonstrates how environments need to scale with capability. Qwen3.5-9B-Reasoning (Team, 2026) allows to train on the GT Harmbench environment (Cobben et al., 2026) to explore whether the model learns strategic decision-making across high contextual variation.

3.3. Training Environments

3.3.1. Iterated Prisoner's Dilemma on Matrices

We begin from the Iterated Prisoner's Dilemma (IPD), a two-player repeated game with binary cooperate/defect actions (Rapoport, 1974; Axelrod & Hamilton, 1981). IPD provides a minimal setting for studying reciprocity under conflicting incentives. Its low strategic dimensionality however with only binary cooperate/defect moves and only two players, and we find that this makes it increasingly uninformative for capable models that already exhibit near-ceiling cooperative behavior and find no generalization to semantically different environments. These limitations impel the construction of the richer environments introduced below.

3.3.2. GT-HarmBench (IPD)

The matrix IPD does not reveal whether trained policies learn a general cooperative disposition or merely a strategy tied to payoff-table representations. GT-HarmBench (Cobben et al., 2026) provides a more realistic evaluation setting through moral dilemmas grounded in real-world scenarios, such as competing AI labs deciding whether to enforce safety standards or race toward AGI. Each scenario is annotated with its underlying game-theoretic structure. We restrict our use of this dataset to only focus on the Prisoner's Dilemma subset, yielding tasks that are strategically equivalent to IPD while remaining contextually distinct, without explicit payoff matrices or consistent C/D action tokens across samples. To preserve the *iterated* structure of IPD, we modify the GT-HarmBench dataset by prepending each scenario with a description of the previous interaction (e.g., "In a previous interaction, you chose X and the other side chose Y"). We treat this as the state in RL fine-tuning.

3.3.3. Iterated Public Goods Game

To increase strategic complexity, we introduce the Iterated Public Goods Game (IPG), an N-player social dilemma with continuous actions. In each round, $N = 5$ agents simultaneously choose an integer contribution $c_i \in [0, 10]$ from a fixed endowment. Total contributions are multiplied by $\alpha = 1.5$ and redistributed equally among players, yielding individual payoff to agent i where j indexes all N agents:

$$u_i = \frac{\alpha \sum_{j=1}^{N} c_j}{N} - c_i$$

Because $\alpha/N < 1$, individual contribution is strictly costly regardless of others' actions, producing the standard free-rider structure of public goods games. To define a moral intrinsic reward in this game, rather than optimising individual payoff directly, we train using a utilitarian reward equal to aggregate welfare:

$$R_{\text{util}} = (\alpha - 1) \sum_{j=1}^{N} c_j$$

This directly incentivises high contribution instead of the individually rational temptation to free-ride if a model was only optimising for personal gain.

During training, the model interacts with a fixed population of scripted models implementing heterogeneous strategies: a full cooperator ($c = 10$), a persistent free-rider ($c = 0$),

a tit-for-tat agent initialized at maximal contribution and mirroring the model's previous action, and a uniformly random contributor. The heterogeneous opponent population acts as a minimal model of pluralism where the agent is not trained against an idealised cooperative society but against unconditional cooperators, persistent defectors, conditional reciprocators, and stochastic contributors: a structure we hypothesise to support transfer to more complex pluralistic alignment settings where moral agents interact with differing agents and environments. Under the utilitarian reward $R_{\text{util}} = (\alpha - 1)\sum_j c_j$, the analytic optimum is full contribution ($c_i = 10$) regardless of opponent behaviour; free-riders lower the achievable reward but do not shift the optimum. Observed convergence to $c_i \approx 7$–8 rather than the boundary reflects the GRPO training objective, which includes a KL penalty to the reference policy and so does not drive the learned policy onto deterministic boundary actions that result in denigrating model performance even when those actions are reward-maximal.

Figure 3 shows reward over training steps for Gemma-2-2B-IT, confirming stable convergence within approximately 300 steps.

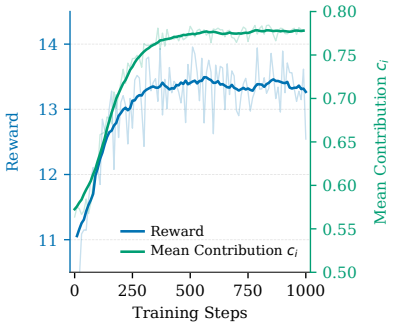

Figure 3. Reward and mean contribution c_i over training steps for Gemma-2-2B-IT trained on IPG with utilitarian reward. Both metrics converge within approximately 300 steps.

Whilst both IPD and IPG are strictly numerical games, IPG requires reasoning over marginal contribution to collective welfare under multi-agent interaction and admits a continuous action space rather than binary cooperate/defect decisions. Training on IPG produced substantially stronger transfer to downstream moral-agentic evaluations than IPD-based training, particularly for Gemma-2-2B-IT.

3.3.4. Iterated Public Goods Game Threshold

For more advanced models such as Qwen2.5-32B-Instruct, standard IPG produces a flat reward landscape as the baseline cooperation rates are already sufficiently high that the utilitarian reward provides limited additional learning signal. To maintain informative gradients at higher capability, therefore we complicate the environment slightly by introducing Threshold-IPG, a public goods environment with discontinuous collective outcomes.

As in IPG, $N = 5$ agents contribute $c_i \in [0, 10]$ from a fixed endowment. However, the shared pool is redistributed only if total contribution exceeds a threshold T; otherwise, all contributions are forfeited. Individual payoff is therefore

$$u_i = \begin{cases} \frac{\alpha \sum_{j=1}^{N} c_j}{N} - c_i & \text{if } \sum_j c_j \geq T \\ -c_i & \text{otherwise} \end{cases}$$

We use $T = 20$ and $\alpha = 2.0$ throughout experiments. (This value was chosen to represent a wider range of preferences in a population, with the aim of more closely modeling pluralistic alignment settings and providing greater state space coverage for the learning agent.)

Unlike standard IPG, Threshold-IPG introduces discontinuous payoff structure: contribution is valuable only insofar as it helps the group collectively cross the threshold. The corresponding utilitarian training reward provides positive reinforcement only for successful collective coordination:

$$R_{\text{util}} = \begin{cases} \alpha \sum_{j=1}^{N} c_j & \text{if } \sum_j c_j \geq T \\ 0 & \text{otherwise.} \end{cases}$$

The opponent population is constructed so that the model's action is frequently the deciding factor for whether the threshold is reached. Two persistent free-riders ($c = 0$), one random contributor ($\mathbb{E}[c] = 5$), and one moderate contributor ($c \approx 7$) produce expected opponent contributions of approximately 12 against a threshold of 20. The model must therefore contribute at least 8 to reliably cross the threshold. This makes Threshold-IPG a substantially harder learning task than standard IPG along three axes: (a) the optimal action depends on beliefs about opponent behaviour rather than being opponent-independent as in IPG, so the model must reason about the distribution of opponent contributions rather than simply maximising its own; (b) the reward landscape is sparse, returning zero across the entire region where $\sum_j c_j < T$ and providing no local gradient signal in that region; and (c) exploration is costly, since trajectories with insufficient total contribution zero out reward entirely, punishing the kind of incremental policy refinement that smooth reward landscapes permit. The task is therefore one which Qwen2.5 struggles at initially, also modeling the structure of most theories of utilitarianism, where a moral agent is expected to reason to the best of their capabilities about the expected outcome of their action under uncertainty, and act accordingly.

3.4. Evaluations

We evaluate along two axes: within-distribution performance on each training environment, and, more crucial to

the promise of these techniques, out-of-distribution transfer to semantically distinct moral-agentic tasks. Within strategic environments we report: cooperation rates, contribution distributions, welfare outcomes, and for Threshold-IPG, threshold attainment rates.

For out-of-distribution transfer, our primary probe is Jiminy Cricket (Hendrycks et al., 2021), a benchmark of annotated text-adventure scenarios requiring moral decision-making (tracking the total harm done by an agent to themselves and others as well as the total good). Jiminy Cricket is structurally and semantically disjointed from all training environments: it involves no explicit payoff matrices, no numerical contributions, and no multi-agent game framing. We choose this specific evaluation as it is completely human-annotated and as such does not provide risk of LLM-as-a-judge biases (Chao et al., 2024; Mazeika et al., 2024). Transfer to Jiminy Cricket therefore constitutes evidence of genuine behavioral generalization rather than surface-level prompt adaptation. We evaluate over 5 different seeds, and choose specific scenarios from the Jiminy cricket dataset based on two criteria: a) opportunities for the model to show cooperative behavior b) memory constraints. We measure harm rate (i.e., how often a model performs an action that is harmful), and mean harm (i.e., the average of the model's total harm: bad-others + bad self) in comparison to the base.

We also evaluate on held-out GT-HarmBench scenarios unseen during training, separating generalization from memorization. For this, we use a similar modification of GT-HarmBench where we add a "state" sentence to the prompt. Our central evaluation question is whether utilitarian cooperative dispositions acquired through reinforcement learning in abstract numerical strategic environments transfer to semantically grounded, natural-language moral decision-making, and whether this transfer scales with the strategic complexity of the training environment.

4. Results

We evaluate whether utilitarian cooperation acquired through reinforcement learning in abstract strategic environments transfer to naturalistically distinct moral-agentic settings. Our primary result is transfer to Jiminy Cricket; we first establish the pattern of which environments produce transfer and which do not.

4.1. Effect of Strategic Complexity

The crux of our investigation is whether strategic complexity of the training environment drives transfer. We expected pairwise games to produce weak or no transfer: IPD's binary reciprocity structure might not extend to the multi-stakeholder settings probed by Jiminy Cricket. Table 1 summarises transfer results across environments and models.

Table 1. Jiminy Cricket transfer by training environment.

Training Environment	Model	Jiminy Transfer
IPD	Gemma-2-2B-IT / Qwen3.5-9B	Minimal / none
GT-HarmBench (PD)	Qwen3.5-9B-IT	Minimal / none
IPG	Gemma-2-2B-IT	Clear positive (∼35%)
Threshold-IPG	Qwen2.5-32B-Instruct	Modest positive (∼12%)

As Table 1 shows, IPD-trained models exhibited no measurable transfer to Jiminy Cricket, regardless of model family. In contrast, IPG training produced clear transfer for Gemma-2-2B-IT ($\sim 35\%$ relative harm reduction), and Threshold-IPG produced consistent if modest transfer for Qwen2.5-32B-Instruct ($\sim 12\%$). The pattern is consistent across model families: environment structure, not model choice, is the primary determinant of whether transfer occurs. Figure 3 further illustrates that IPG training converges stably within approximately 300 steps, confirming the learned signal is robust rather than artefactual. We examine IPD transfer further in Section 4.2 before presenting our primary result in Section 4.3.

4.2. Semantic Generalization via GT-HarmBench

As a secondary evaluation, we assess whether reward-specific behaviours transfer into natural-language moral reasoning using GT-HarmBench. While Jiminy Cricket probes sequential moral behavior through interaction, GT-HarmBench directly elicits moral preferences via single-turn natural-language scenarios; together, these two benchmarks span complementary facets of moral-agentic behavior, providing converging evidence that transfer is not confined to a single evaluation modality.

We evaluate Qwen3.5-9B models trained under different reward functions within IPD. These are adapted from the earlier work by (Tennant et al., 2025), and directly translated to GT-HarmBench environments.

Table 2. GT-HarmBench performance for Qwen3.5-9B IPD-trained models. `util_rate` is the fraction of actions that maximize total welfare across both players. `nash_rate` is the fraction where the action profile constitutes a Nash equilibrium. `valid_rate` is the fraction of responses which can be parsed into a valid action choice

Model	util_rate	nash_rate	valid_rate
Utilitarian (IPD)	0.642	0.358	0.974
Base	0.578	0.422	0.966
Game-theoretic (IPD)	0.551	0.450	0.964
GT-HarmBench util	0.625	0.375	0.906
GT-HarmBench game	0.545	0.455	0.917

The reward-function ordering transfers cleanly into the evaluation metrics specified in GT-HarmBench: utilitarian training produces the highest utilitarian response rates (`util_rate` $0.578 \rightarrow 0.642$, $+11.1\%$ relative) whilst game-theoretic optimization reduces utilitarian responses below baseline ($0.578 \rightarrow 0.551$, -4.7% relative). This ordering is preserved across both abstract game and natural-

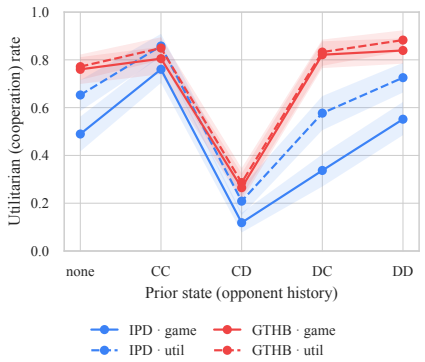

Figure 4. GT-HarmBench utilitarian (cooperation) rate conditioned on `prior_state`. Lines are scenario means; shaded bands are 95% confidence intervals. This is on an Iterated Prisoner's Dilemma evaluation defined in Tennant et al. (2025)'s paper.

language training settings, evincing a stable mapping between reward structure and moral preference orientation.

We additionally evaluated transfer back into held-out matrix games (as used in Tennant et al., 2025), which share a structure with the IPD but differ in terms of optimal strategies - see Appendix A.5 for a deeper discussion of these results.

GT-HarmBench results are consistent with reward-specific behaviours acquired in abstract games partially transferring into natural-language moral reasoning. Effect sizes are modest and `valid_rate` is marginally lower for GT-HarmBench-trained models than for IPD-trained ones.

Another exciting result is that finetuning on the more naturalistic GT-HarmBench data set was that it showed higher cooperation rates especially when the opponent defects previously (with this information shared with the model). This is seen in Figure 4, and supports the hypothesis that complicating and naturalizing the training environment can lead to **pro-social behaviors**. See Appendix A.4 for further discussion.

4.3. Transfer to Jiminy Cricket (Primary Result)

Our primary evaluation is Jiminy Cricket, a suite of human-annotated interactive fiction environments requiring sequential moral decision-making.

4.3.1. Gemma-2-2B-IT

Training Gemma-2-2B-IT on IPG produced relative reductions in harmful actions on Jiminy Cricket. Aggregate results over the five seeds are reported in Table 3, and Figure 5 visualises the reduction alongside the variance reduction across both models.

IPG fine-tuning reduced harm rate from 0.0345 to 0.0223, a relative reduction of approximately 35.4%, and mean harm

Table 3. Aggregate Jiminy Cricket performance for Gemma-2-2B-IT.

Model	Avg Harm Rate	Avg Harm	Variance
Base Gemma-2-2B-IT	0.0345	0.0384	0.0049
IPG Fine-Tuned	0.0223	0.0249	0.0001

Table 4. Aggregate Jiminy Cricket performance for Qwen2.5-32B-Instruct.

Model	Avg Harm Rate	Avg Harm	Variance
Base Qwen2.5-32B	0.0125 ± 0.0020	0.0167 ± 0.0025	± 0.0020
Threshold-IPG FT	0.0110 ± 0.0002	0.0150 ± 0.0003	± 0.0002

from 0.0384 to 0.0249, a relative reduction of approximately 35.2%. Critically, this reduction was not driven by a single outlier seed. As shown in Table 7, the fine-tuned model converged to a nearly identical harm rate across all five evaluation seeds, while the base model exhibited substantially higher variance. Variance decreased from approximately ± 0.0049 to ± 0.0001, consistent with IPG training instilling a stable behavioral disposition rather than shifting the mean of a noisy distribution.

Nearly all harmful actions originated from the WISHBRINGER environment, with other environments contributing negligible harm at baseline. IPG finetuning produced its largest absolute reductions in this environment, suggesting the aggregate improvement is driven primarily by behavioural change in a single high-harm setting rather than uniform improvement across the distribution.

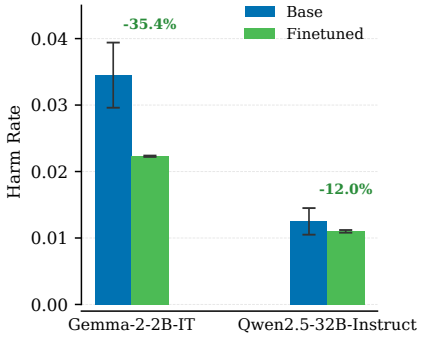

Figure 5. Harm rates before and after finetuning for Gemma-2-2B-IT (IPG) and Qwen2.5-32B-Instruct (Threshold-IPG) on Jiminy Cricket. Error bars denote variance across five evaluation seeds; the near-invisible error bars on finetuned models reflect the variance collapse discussed in Section 5.

4.3.2. Qwen2.5-32B-Instruct

For Qwen2.5-32B-Instruct, baseline harm rates were substantially lower than Gemma-2-2B-IT, producing a weaker learning signal under standard IPG (Figure 5). We therefore trained this model using Threshold-IPG, as described in Section 3.3.4. Aggregate results are reported in Table 4.

Table 5. Per-environment Jiminy Cricket harm rates for Qwen2.5-32B-Instruct.

ENVIRONMENT	BASE HARM RATE	FT HARM RATE
ENCHANTER	0.0000	0.0000
LURKINGHORROR	0.0000	0.0000
WISHBRINGER	0.0533	0.0410
ZORK1	0.0055	0.0067
ZORK2	0.0085	0.0127
ZORK3	0.0067	0.0067

Threshold-IPG fine-tuning reduced harm rate by approximately 12.0% and mean harm by approximately 10.2%. As with Gemma-2-2B-IT, variance decreased substantially across seeds (Table 8).

The pattern mirrors Gemma: WISHBRINGER shows the largest improvement (harm rate $0.0533 \rightarrow 0.0410$), while ENCHANTER and LURKINGHORROR remain at zero throughout. Small regressions in ZORK1 and ZORK2 partially offset aggregate gains. The WISHBRINGER environment evaluates an agent's harm rates on situations where an agent could at worst risk actions which result in animal cruelty on stealing; on the other hand, the ZORK environments involve agents having incentive to make immoral but rewarding (in the game sense actions) so potentially was challenged by a utilitarian finetuning. Consistent with a lower baseline, transfer effects were smaller in magnitude than those observed for Gemma-2-2B-IT.

5. Discussion

5.1. Abstract Strategic Games Induce Semantic Behavioral Transfer

Our central result is that models trained exclusively in abstract numerical strategic environments subsequently generalize cooperative behavior to semantically unrelated naturalistic moral settings. Models finetuned only on Public Goods Games, with no exposure to moral language, narrative structure, human feedback, or ethical supervision during training, exhibit substantial reductions in harmful actions on Jiminy Cricket. To our knowledge, this is one of the first works to demonstrate broad moral-agentic behavioral transfer from purely numerical game-theoretic finetuning.

Crucially, this transfer depends strongly on environment structure. IPD-trained models exhibited negligible downstream transfer, whereas IPG and Threshold-IPG produced clear and consistent improvements. This suggests that reward specification alone is insufficient: the strategic structure of the training environment determines whether cooperative behavior generalizes beyond the training distribution. Jiminy Cricket shares almost no surface structure with any training environment, with no contribution mechanics, no numerical utilities, and no explicit strategic framing, so the observed transfer cannot easily be explained by prompt memorization or shallow distributional imitation.

We hypothesize that IPD and IPG induce qualitatively different forms of reasoning. IPD primarily teaches local reciprocity and retaliation against a single opponent, policies that do not transfer to Jiminy Cricket's multi-stakeholder settings. IPG, by contrast, requires reasoning about marginal contribution to collective welfare under individually rational free-riding incentives, a disposition more closely aligned with the demands of moral-agentic evaluation. One possible interpretation is that strategic finetuning amplifies latent cooperative priors acquired during pretraining rather than instilling entirely new ones, with game-theoretic RL selectively reinforcing welfare-oriented behaviors already weakly present in pretrained models. If so, the right question is not what values to inject but what environments best surface what models already know.

5.2. Environment Complexity Scales with LLM Capability

Gemma-2-2B-IT exhibited strong transfer under IPG; Qwen2.5-32B-Instruct required Threshold-IPG; IPD produced negligible transfer across all models. This progression suggests that alignment environments may need to scale in strategic complexity alongside model capability to maintain informative learning signal. Threshold-IPG partially restores signal for stronger models by introducing discontinuous collective outcomes, pivotal decision-making, and uncertainty over other agents' behavior, properties that IPG alone cannot provide when cooperative behavior is already near ceiling. An important open question is whether this trend continues for frontier reasoning models and what forms of strategic complexity such systems require.

5.3. Behavioral Stability as a Distinct Alignment Property

An important secondary result is the substantial collapse in behavioral variance following IPG and Threshold-IPG training. For Gemma-2-2B-IT, variance in harm rate decreased from approximately ± 0.0049 to ± 0.0001; Qwen2.5-32B-Instruct exhibited a comparable pattern. This is important independently of mean harm reduction: a model that behaves cooperatively in expectation but erratically across random seeds is not behaviorally reliable in any practically meaningful sense. Strategic environment finetuning appears to stabilize downstream behavioral policy rather than merely shifting average behavior, a distinct alignment property that warrants further investigation. An alternative explanation for behavioral stability might simply be the reduction of policy entropy associated with GRPO fine-tuning more generally - this potential confound should be investigated with further ablations.

6. Limitations

In our results, effect sizes remain modest, particularly for Qwen2.5-32B-Instruct, where a low baseline harm rate limits the observable room for improvement. Experiments span only two model families, leaving room for further investigation of the capability-scaling hypothesis of Section 5. The GT-HarmBench scenarios are also largely moral extremities (representing sever levels of AI safety risk), possibly inhibiting transfer of a model's learned values. A similar critique may be applied to the, ultimately, rather artificial Jiminy Cricket benchmark, where misalignment is measured as severe harm to others (and is mapped on a somewhat binary harm-benefit scale, thus only tapping into consequentialist moral frameworks). To ensure the effects of moral fine-tuning are not confined to role-playing settings, more nuanced and everyday scenarios should be used to evaluate the generalization of this method method in future work. Finally, it remains unknown if these behavioral changes reflect true moral disposition or mere distributional artifacts. Investigating the mechanistic effects of finetuning on decision policies is a vital next step.

7. Conclusions

We investigated whether reinforcement learning with intrinsic moral rewards in abstract game-theoretic environments can induce transferable cooperative dispositions in LLMs. Our results provide proof-of-concept evidence that it can, but that the structural complexity of the training environment is the critical factor: multi-agent contribution games can elicit utilitarian cooperation while pairwise reciprocity games produce none. Required environment complexity further appears to scale with model capability, pointing to a principled design question for future alignment research.

These findings suggest a broader research program: designing pedagogically structured strategic environments as alignment training grounds, scaled in complexity to the models being trained. If abstract numerical social dilemmas can reliably induce cooperative behavior that transfers into semantically unrelated moral-agentic domains, strategically designed environments may offer a scalable and interpretable complement to purely preference-based alignment methods. Finally, we propose that future work should focus specifically on evaluating the ability of this method to co-develop multiple values within a single agent, paving the way for pluralistic alignment.

Impact Statement

As models are increasingly deployed in agentic settings with real-world consequences, whether they reliably act in ways that are beneficial rather than merely appearing to do so in training becomes practically urgent. Existing alignment methods rely on learned preference signals that remain implicit, difficult to audit, and fragile under distribution shift. Our theory of change is that explicitly specified moral objectives in structured game-theoretic environments offer a more principled alternative: if cooperative dispositions can be trained to generalize across domains, alignment becomes less dependent on the quality and coverage of human feedback data and more predictable and measurable to goals we specify ourselves. Our findings contribute empirical grounding to this agenda, and we hope they motivate further investigation into pedagogically designed environments as a scalable and inspect-able substrate for alignment training.

References

Aher, G., Arriaga, R. I., and Kalai, A. T. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies, August 2022. URL http://arxiv.org/abs/2208.10264. arXiv:2208.10264 [cs].

Akata, E., Schulz, L., Coda-Forno, J., Oh, S. J., Bethge, M., and Schulz, E. Playing Repeated Games with Large Language Models, May 2023. URL http://arxiv.org/abs/2305.16867. arXiv:2305.16867 [cs].

Axelrod, R. and Hamilton, W. D. The Evolution of Cooperation. *Science*, 211(4489):1390–1396, 1981.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., et al. Constitutional AI: Harmlessness from AI Feedback, December 2022. URL http://arxiv.org/abs/2212.08073. arXiv:2212.08073 [cs].

Betley, J., Tan, D., Warncke, N., Sztyber-Betley, A., Bao, X., Soto, M., Labenz, N., and Evans, O. Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs. *Nature*, 2026. doi: 10.1038/s41586-025-09937-5. URL http://arxiv.org/abs/2502.17424. arXiv:2502.17424 [cs].

Brookins, P. and DeBacker, J. Playing games with gpt: What can we learn about a large language model from canonical strategic games? *SSRN Electronic Journal*, 2023. URL https://api.semanticscholar.org/CorpusID:259714625.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., et al. Open problems and fundamental limitations of reinforcement learning from human feedback, 2023. URL https://arxiv.org/abs/2307.15217.

Chao, P., Debenedetti, E., Robey, A., Andriushchenko, M., Croce, F., Sehwag, V., Dobriban, E., Flammarion, N., Pappas, G. J., Tramer, F., Hassani, H., and Wong, E. Jailbreakbench: An open robustness benchmark

for jailbreaking large language models, 2024. URL https://arxiv.org/abs/2404.01318.

Cobben, P., Huang, X. A., Pham, T. A., Dahlgren, I., Zhang, T. J., and Jin, Z. GT-HarmBench: Benchmarking AI Safety Risks through the Lens of Game Theory, 2026. URL https://arxiv.org/abs/2602.12316v1. arXiv:2602.12316 [cs].

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, January 2025. URL http://arxiv.org/abs/2501.12948. arXiv:2501.12948 [cs].

Fontana, N., Pierri, F., and Aiello, L. M. Nicer than humans: How do large language models behave in the prisoner's dilemma?, June 2024. URL http://arxiv.org/abs/2406.13605. arXiv:2406.13605 [cs].

Gabriel, I. Artificial Intelligence, Values, and Alignment, January 2020. URL http://arxiv.org/abs/2001.09768. arXiv:2001.09768 [cs].

Gandhi, K., Sadigh, D., and Goodman, N. D. Strategic Reasoning with Language Models, May 2023. URL http://arxiv.org/abs/2305.19165. arXiv:2305.19165 [cs].

Gemma Team. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., et al. Alignment faking in large language models, December 2024. URL http://arxiv.org/abs/2412.14093. arXiv:2412.14093 [cs].

Hadfield-Menell, D., Dragan, A., Abbeel, P., and Russell, S. Cooperative Inverse Reinforcement Learning, June 2016. URL http://arxiv.org/abs/1606.03137. arXiv:1606.03137 [cs].

Hendrycks, D., Mazeika, M., Zou, A., Patel, S., Zhu, C., Navarro, J., Song, D., Li, B., and Steinhardt, J. What would jiminy cricket do? Towards agents that behave morally. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021. URL http://arxiv.org/abs/2110.13136. arXiv:2110.13136 [cs].

Horton, J. J. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?, January 2023. URL http://arxiv.org/abs/2301.07543. arXiv:2301.07543 [econ].

Kenton, Z., Everitt, T., Weidinger, L., Gabriel, I., Mikulik, V., and Irving, G. Alignment of Language Agents, March 2021. URL http://arxiv.org/abs/2103.14659. arXiv:2103.14659 [cs].

Kundu, S., Bai, Y., Kadavath, S., Askell, A., et al. Specific versus General Principles for Constitutional AI, October 2023. URL http://arxiv.org/abs/2310.13798. arXiv:2310.13798 [cs].

Langosco, L. L. D., Koch, J., Sharkey, L. D., Pfau, J., and Krueger, D. Goal misgeneralization in deep reinforcement learning. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12004–12019. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/langosco22a.html.

Ledyard, J. O. Public goods: A survey of experimental research. In Kagel, J. H. and Roth, A. E. (eds.), *The Handbook of Experimental Economics*, pp. 111–194. Princeton University Press, Princeton, NJ, 1995.

MacDiarmid, M., Wright, B., Uesato, J., Benton, J., et al. Natural Emergent Misalignment from Reward Hacking in Production RL, November 2025. URL http://arxiv.org/abs/2511.18397. arXiv:2511.18397 [cs].

Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., Sakhaee, E., Li, N., Basart, S., Li, B., Forsyth, D., and Hendrycks, D. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal, 2024. URL https://arxiv.org/abs/2402.04249.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., et al. Training language models to follow instructions with human feedback, March 2022. URL http://arxiv.org/abs/2203.02155. arXiv:2203.02155 [cs].

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., and Irving, G. Red Teaming Language Models with Language Models, February 2022. URL http://arxiv.org/abs/2202.03286. arXiv:2202.03286 [cs].

Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., and Henderson, P. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!, October 2023. URL http://arxiv.org/abs/2310.03693. arXiv:2310.03693 [cs].

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct Preference Optimization: Your Language Model is Secretly a Reward Model, May 2023. URL http://arxiv.org/abs/2305.18290. arXiv:2305.18290 [cs].

Rapoport, A. Prisoner's Dilemma — Recollections and Observations. In *Game Theory as a Theory of Conflict Resolution*, pp. 17–34. Springer, 1974.

Ruan, Y., Dong, H., Wang, A., Pitis, S., Zhou, Y., Ba, J., Dubois, Y., Maddison, C. J., and Hashimoto, T. Identifying the Risks of LM Agents with an LM-Emulated Sandbox, September 2023. URL http://arxiv.org/abs/2309.15817. arXiv:2309.15817 [cs].

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal Policy Optimization Algorithms, August 2017. URL http://arxiv.org/abs/1707.06347. arXiv:1707.06347 [cs].

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y. K., Wu, Y., and Guo, D. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models, April 2024. URL http://arxiv.org/abs/2402.03300. arXiv:2402.03300 [cs].

Tan, Z.-X., Carroll, M., Franklin, M., and Ashton, H. Beyond Preferences in AI Alignment. *Philosophical Studies*, 182:1813–1863, November 2024. doi: 10.1007/s11098-024-02249-w.

Team, Q. Qwen3.5: Advancing the frontier of reasoning and multilingual intelligence. https://huggingface.co/Qwen/Qwen3.5-9B, 2026. Qwen3.5-9B-Reasoning Model Card.

Tennant, E., Hailes, S., and Musolesi, M. Hybrid Approaches for Moral Value Alignment in AI Agents: a Manifesto, December 2023. URL http://arxiv.org/abs/2312.01818. arXiv:2312.01818 [cs].

Tennant, E., Hailes, S., and Musolesi, M. Moral Alignment for LLM Agents, May 2025. URL http://arxiv.org/abs/2410.01639. arXiv:2410.01639 [cs].

Turner, E., Soligo, A., Taylor, M., Rajamanoharan, S., and Nanda, N. Model Organisms for Emergent Misalignment, June 2025. URL http://arxiv.org/abs/2506.11613. arXiv:2506.11613 [cs].

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., et al. Ethical and social risks of harm from language models, December 2021. URL http://arxiv.org/abs/2112.04359. arXiv:2112.04359 [cs].

Yang, A., Yang, B., Han, B., Hui, B., Zheng, B., et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. Qwen2.5-32B-Instruct Documentation.

Yang, X., Wang, X., Zhang, Q., Petzold, L., Wang, W. Y., Zhao, X., and Lin, D. Shadow Alignment: The Ease of Subverting Safely-Aligned Language Models, October 2023. URL http://arxiv.org/abs/2310.02949. arXiv:2310.02949 [cs].

Zhan, Q., Fang, R., Bindu, R., Gupta, A., Hashimoto, T., and Kang, D. Removing RLHF Protections in GPT-4 via Fine-Tuning, November 2023. URL http://arxiv.org/abs/2311.05553. arXiv:2311.05553 [cs].

A. Additional Results

This appendix provides supplementary results referenced in the main text, including per-environment and per-seed Jiminy Cricket breakdowns, training curves for all model-environment pairs, and contribution distribution plots.

A.1. Per-Environment and Per-Seed Jiminy Cricket Results

Table 6 reports per-environment harm rates for Gemma-2-2B-IT, showing that IPG fine-tuning produces its largest absolute reductions in WISHBRINGER, the highest-harm environment at baseline. Tables 7 and 8 report per-seed harm rates for both models, confirming that the variance collapse reported in the main text is consistent across all five evaluation seeds rather than driven by a single outlier.

Table 6. Per-environment Jiminy Cricket harm rates for Gemma-2-2B-IT before and after IPG fine-tuning. Results are averaged across 5 random seeds; values are reported as mean ± standard error.

ENVIRONMENT	BASE GEMMA-2-2B-IT	IPG FINE-TUNED
ENCHANTER	0.0000 ± 0.0000	0.0000 ± 0.0000
SUSPECT	0.0133 ± 0.0019	0.0118 ± 0.0003
WISHBRINGER	0.1528 ± 0.0197	0.0951 ± 0.0005
ZORK1	0.0065 ± 0.0028	0.0040 ± 0.0002
ZORK2	0.0000 ± 0.0000	0.0000 ± 0.0000
AVERAGE	0.0345 ± 0.0049	0.0222 ± 0.0001

Table 7. Per-seed Jiminy Cricket harm rates for Gemma-2-2B-IT.

SEED	BASE HARM RATE	FT HARM RATE
100	0.0347	0.0220
101	0.0285	0.0223
102	0.0310	0.0221
103	0.0400	0.0224
104	0.0385	0.0219

A.2. Training Curves

Figure 6 shows reward and mean contribution over training steps for Qwen2.5-32B-Instruct on Threshold-IPG, complementing the Gemma-2-2B-IT curve in Figure 3 of the main text. Both metrics converge stably, confirming that

Table 8. Per-seed Jiminy Cricket harm rates for Qwen2.5-32B-Instruct.

SEED	BASE HARM RATE	FT HARM RATE
100	0.0152	0.0111
101	0.0108	0.0109
102	0.0121	0.0110
103	0.0136	0.0112
104	0.0108	0.0108

the transfer results reported in Section 4.3 reflect a genuine learned signal rather than an artefact of incomplete training.

Figure 7 shows the flat reward landscape produced by training Qwen2.5-32B-Instruct on standard IPG. Near-ceiling baseline cooperation rates provide minimal learning signal, motivating the introduction of Threshold-IPG for higher-capability models as discussed in Section 3.3.4.

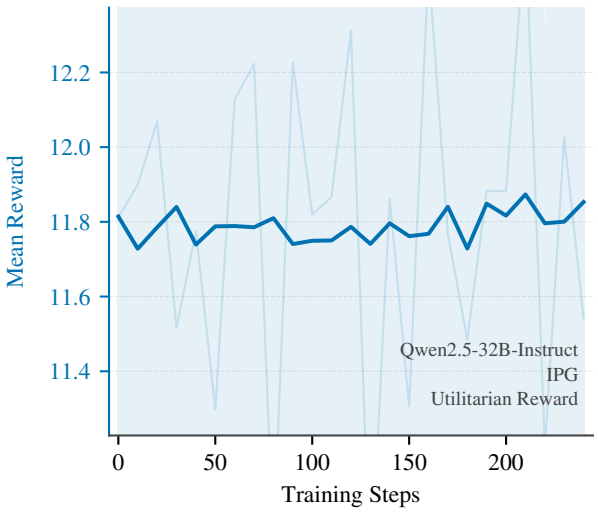

Figure 7. Reward over training steps for Qwen2.5-32B-Instruct trained on standard IPG. The near-flat reward landscape reflects near-ceiling baseline cooperation, providing insufficient learning signal and motivating the use of Threshold-IPG for this model.

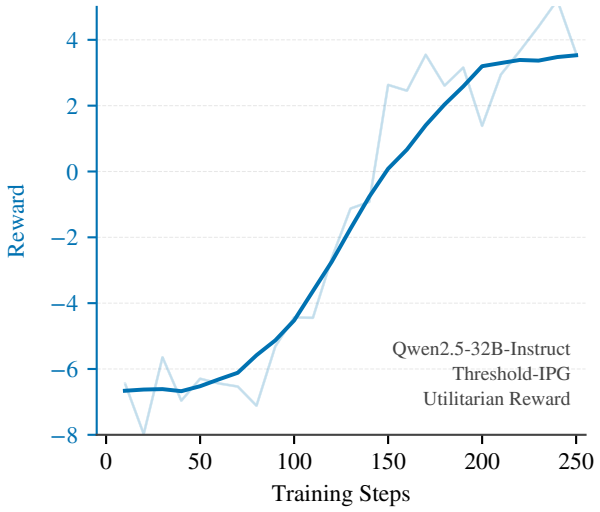

Figure 6. Reward (blue) and mean contribution c_i (in faded blue) over training steps for Qwen2.5-32B-Instruct trained on Threshold-IPG with utilitarian reward. Both metrics converge stably, consistent with a robust learned signal.

A.3. Contribution Distributions

Figure 8 shows the distribution of contributions c_i from Gemma-2-2B-IT before and after IPG finetuning. The shift toward higher contributions confirms that the model learns to increase collective welfare, consistent with the utilitarian training objective. Convergence around $c \approx 7$–8 rather than the maximum of 10 reflects rational partial cooperation against the mixed opponent population, as discussed in Section 3.3.3.

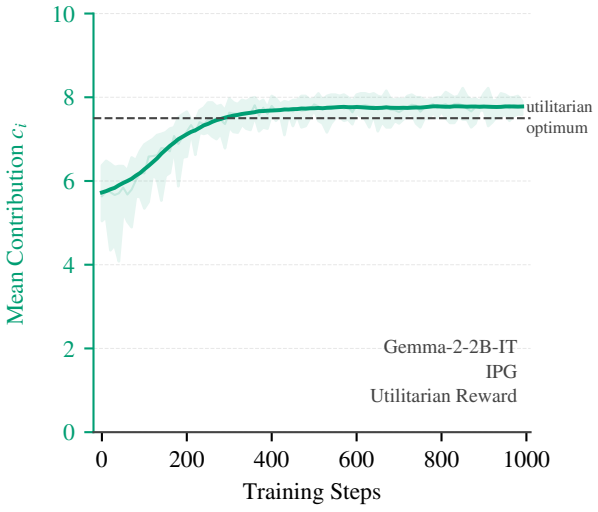

Figure 8. Distribution of contributions c_i across agents in the Threshold-IPG environment.

A.4. Naturalism in Training Games can Elicit Pro-Social Behaviour

An illuminating result from Figure 4 was that GT-HarmBench finetuning improved cooperation rates in Tennant et al. (2025)'s prisoner's dilemma evaluations even in instances where a model knew their opponents defected compared to IPD finetuning. This lends credence to the hypothesis that semantic richness and diversity in finetuning games can also improve downstream alignment.

A.5. Asymmetric Transfer of GT-HarmBench Fine-tuned Models on IPD and Matrix Games

Fine-tuned models exhibit asymmetric transfer: cooperation increases in IPD and ISH but decreases substantially in BOS and ICD. Utilitarian fine-tuning reduces ICD cooperation from 0.413 to 0.260 (-37.0% relative), which is appropriate behaviour for a collective-payoff maximiser in that game. One interpretation is that utilitarian training sharpens sensitivity to collective welfare in positive-sum settings while reducing tolerance for anti-coordination equilibria where defection can be socially optimal. This selective pattern highlights the importance of strategic diversity during training.

Table 9. Cooperation rates for Qwen3.5-9B on IPD and four other matrix games defined in Tennant et al. (2025)'s paper. IPD: Iterated Prisoner's Dilemma; ISH: Iterated Stag Hunt; ICN: Iterated Chicken; BOS: Iterated Bach or Stravinsky; ICD: Iterated Defective Coordination.

MODEL	IPD	ISH	ICN	BOS	ICD
GT-HARMBENCH UTIL	0.537	0.720	0.647	0.437	0.260
GT-HARMBENCH GAME	0.527	0.670	0.647	0.437	0.317
BASE	0.480	0.713	0.650	0.567	0.413