EGGROLL-IPO: Pluralistic Alignment via Decentralised Post-Training with Population Preferences

Alfie Lamerton [1] **Bidipta Sarkar** [2] **Roberto-Rafael Maura-Rivero** [3] **Jakob Foerster** [2]

Abstract

Frontier AI companies align large language models to constitutions and model specs they author themselves, not to the preferences of the populations the models serve. This is an organisational choice rather than a technical necessity, but the dominant alignment pipeline reinforces it: gradient-based reinforcement learning from human feedback (RLHF) requires dense tensor communication between workers, ruling out protocols in which different workers feed heterogeneous preferences into training. EGGROLL recently showed that large-scale evolution strategies require only a scalar fitness per worker, which makes it practical to train a single shared model from compute, queries, and preferences contributed by a population. The remaining question is which loss should drive that protocol: aggregating preferences naively recovers Borda scoring, which violates Condorcet consistency and independence of irrelevant alternatives (IIA). We argue that the loss should target the maximal lottery, the unique probabilistic social choice function satisfying Arrow's axioms, and Online Identity Preference Optimisation (IPO) loss does this. We introduce EGGROLL-IPO, which uses the Online IPO loss as scalar fitness in EGGROLL, and show across three controlled experiments (cyclic preferences, IIA, and sequence-level convergence) that it outperforms a naive ± 1 preference fitness with the optimiser held at EGGROLL.

1. Introduction

The dominant alignment pipeline at frontier labs takes a model spec or constitution as the alignment target (Bai et al., 2022; Glaese et al., 2022; OpenAI, 2025). The spec is authored by the lab; the population whose preferences the model will shape at deployment time is not consulted. This is sometimes defended on technical grounds (preferences are noisy; constitutions are stable), but it is not a technical necessity. The dominant pipeline reinforces this organisational pattern: gradient-based preference optimisation requires fine-tuning of billion-parameter models with reference policies held in memory and dense tensor communication between workers, which restricts who can do safety-relevant training and rules out any protocol in which the population itself participates.

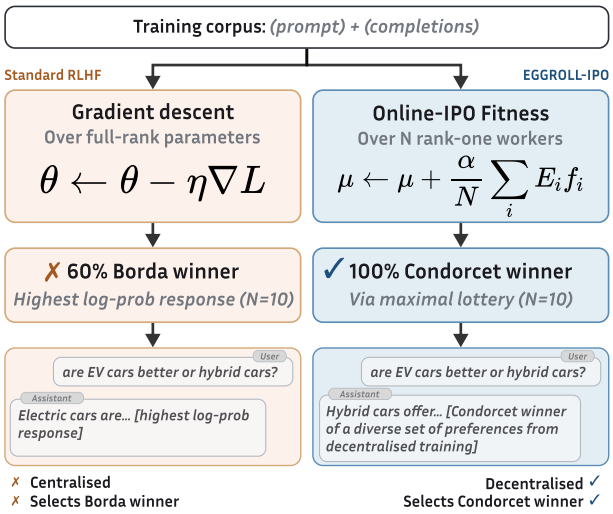

Figure 1. EGGROLL-IPO: each worker evaluates its own policy on its own context, broadcasting a single scalar fitness; the population's maximal lottery is the target.

EGGROLL (Sarkar et al., 2025) changes the second part by replacing gradient backpropagation with a new variant of evolution strategies. EGGROLL allows each worker to evaluate the model on its own contexts and preferences and broadcast only one single scalar fitness per evaluation instead of requiring full gradient communication. The mechanical obstacle to pluralistic training is gone: workers can run on heterogeneous hardware, evaluate on contexts they care about, and contribute preferences from their own

[1]Formation Research [2]FLAIR, University of Oxford [3]Computer Science department, University of Oxford. Correspondence to: Alfie Lamerton <alfie@formationresearch.org>.

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

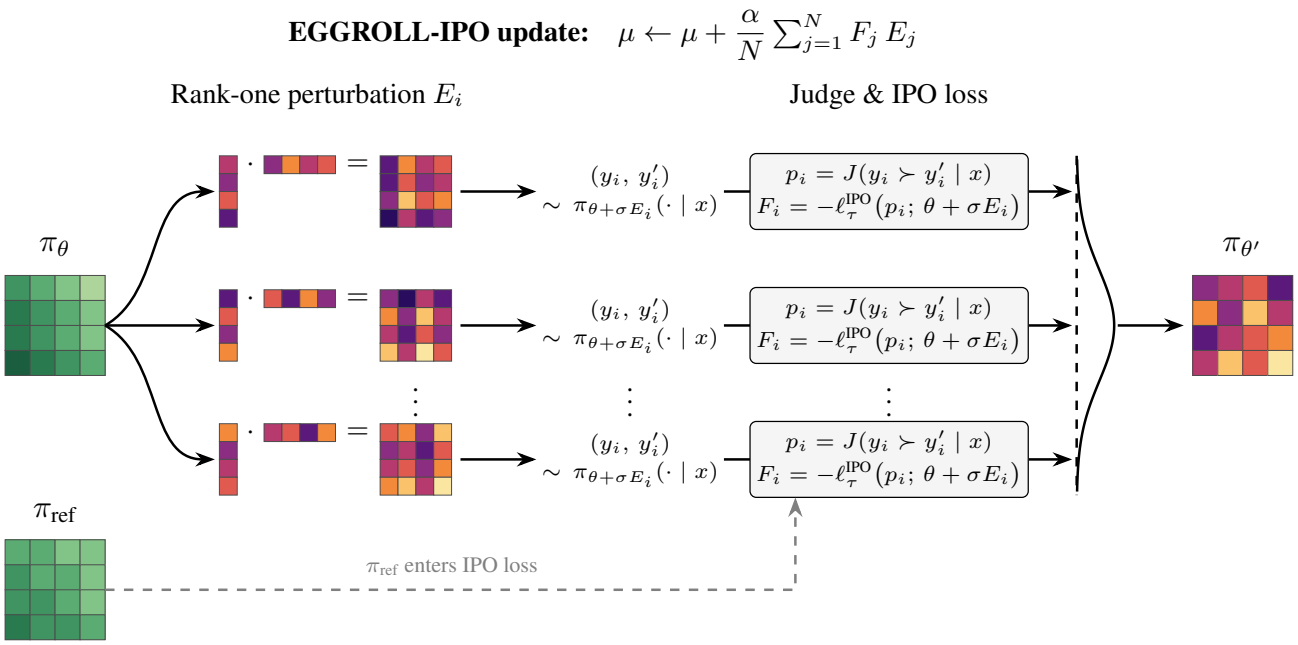

Figure 2. Schematic of EGGROLL-IPO with N workers. Each worker samples a rank-one perturbation $E_i = u_i v_i^\top$, generates a paired completion (y_i, y_i') from the perturbed policy, and receives a soft preference probability p_i from the judge. The IPO loss converts p_i and the log-ratio h_i (which uses the frozen π_{ref}) into a per-worker fitness $F_i = -\mathcal{L}_i^{\text{IPO}}$. The natural-evolution-strategy update aggregates these as $\mu \leftarrow \mu + (\alpha/N)\sum_j F_j E_j$.

feedback channel, without ever exchanging gradient tensors. The question that remains is what loss the workers should compute. The naive answer of treating each preference query as a ± 1 fitness signal recovers Bradley-Terry preference modelling (Siththaranjan et al., 2023), which implements Borda scoring and therefore fails majority consistency, Condorcet consistency, and independence of irrelevant alternatives (Maura-Rivero et al., 2025). Concretely: when 40% of voters prefer $R \succ G \succ B$ and 60% prefer $B \succ R \succ G$, the Condorcet winner is B, but Borda selects R because the irrelevant alternative G drains R's pairwise losses. Section 5.3 shows this failure empirically for raw-fitness EGGROLL.

Voting theory provides a principled target. The maximal lottery is the unique probabilistic social choice function satisfying Arrow's axioms (Brandl & Brandt, 2020; Maura-Rivero et al., 2025), and Nash Learning from Human Feedback (NLHF; Munos et al., 2024) together with its variant Online IPO (Calandriello et al., 2024) are the loss families whose fixed point is the maximal lottery. We therefore use Online IPO as the scalar fitness inside EGGROLL.

Across three controlled experiments, EGGROLL-IPO converges to the maximal lottery while a raw ± 1 preference fitness baseline fails on each in qualitatively distinct ways: violating independence of irrelevant alternatives, oscillating without concentrating on the Condorcet winner, and collapsing or drifting under cyclic preferences.

We make three contributions. **First**, we introduce EGGROLL-IPO, a post-training algorithm that uses the Online IPO loss as scalar fitness in EGGROLL, targeting the population's maximal lottery with per-worker contexts and policies. **Second**, we characterise a decentralised training protocol that follows directly from the scalar-fitness channel: compute donors contribute compute, queries, and preferences without exchanging gradient tensors; we do not evaluate the protocol empirically. **Third**, across three controlled experiments (a cyclic preference game, an IIA counter-example, and a transitive sequence-level convergence task) the Online IPO loss outperforms a raw ± 1 preference fitness with the optimiser held at EGGROLL, isolating each failure to the loss.

Section 2 positions our work in relation to pluralistic alignment, game-theoretic preference learning, and distributed LLM training. Section 3 reviews maximal lotteries, Online IPO, EGGROLL, and Community Alignment. Section 4 presents the EGGROLL-IPO algorithm. Section 5 reports experiments. Section 6 discusses decentralisation and limitations.

2. Related Work

Our work lies at the intersection of three subfields.

Pluralistic alignment. Distributional preference learning (Siththaranjan et al., 2023) models the full distribution of

preferences rather than a point estimate, but retains the Bradley-Terry parameterisation and therefore the IIA failure. Our contribution differs in optimising directly for an axiomatic social-choice-theoretic solution rather than for distribution matching.

Game-theoretic preference learning. NLHF (Munos et al., 2024) and Online IPO (Calandriello et al., 2024) provide the social-choice-theoretic grounding we adopt. We extend this line by removing the gradient requirement and broadcasting only scalar fitness between workers.

Distributed LLM training. EGGROLL-IPO requires only scalar communication, which makes preference-level decentralisation, not just compute-level decentralisation, well-defined.

Prior works have used ES-adjacent techniques for LLM training. MeZO (Malladi et al., 2023) demonstrates the effectiveness of SPSA, equivalent to ES with a population size of 2, for fine-tuning LLMs. Tang et al. (2025) extend this to create a differentially private variant of SPSA, enabling privacy-aware distributed finetuning. Note that these prior works have a population size of 2 from an ES context, meaning that feedback signals are only used to evaluate a single noise direction regardless of the number of workers. In contrast, EGGROLL allows the ES population size to grow alongside the number of evaluators and workers, enabling a richer gradient signal at larger scales.

Backprop-based methods have also been used for decentralised LLM training. DiLoCo (Douillard et al., 2024) works by allowing independent workers to update the model locally for many steps before synchronising a global update. The low frequency of global updates resolves the problem of requiring a high-bandwidth connection between globally distributed nodes for each update step, but DiLoCo still adds a barrier to entry for LLM training as it still requires each worker to do backprop and store large optimiser states independently. In contrast, ES-based methods only require the ability to do forward passes and already have low bandwidth requirements.

3. Background

3.1. Maximal Lotteries and the Regularised Preference Game

A *maximal lottery* is a distribution over alternatives that, in pairwise comparisons against any other distribution, is preferred by at least half the population. Formally, given a finite set of alternatives $\mathcal{Y}$ and a population with preferences over $\mathcal{Y}$, a maximal lottery is a distribution $\pi^* \in \Delta(\mathcal{Y})$ such that $\pi^{*\top} M \pi' \geq 0$ for all $\pi' \in \Delta(\mathcal{Y})$, where $M_{ab} = N(a,b) - N(b,a)$ is the pairwise margin matrix and $N(a,b)$ is the number of individuals preferring a to b (Fishburn, 1984; Kreweras, 1965). Equivalently, the maximal lottery is the Nash equilibrium of the symmetric zero-sum game with payoff matrix M. Brandl & Brandt (2020) prove that maximal lotteries are the unique probabilistic social choice function satisfying Arrow's axioms (independence of irrelevant alternatives, Pareto efficiency, and anonymity, which implies non-dictatorship) in the stochastic setting. Maximal lotteries are Condorcet-consistent, majority-consistent, and handle Condorcet cycles by returning a non-degenerate distribution rather than collapsing arbitrarily to one alternative.

Nash Learning From Human Feedback (NLHF) reformulates human preference alignment as a two-player constant-sum game (Munos et al., 2024). Given a preference function $p(y \succ y')$ representing the probability that a random member of the population prefers y to y', players select policies $\pi_i, \pi_{-i} \in \Delta(\mathcal{Y})$ with payoffs

$$\mathbb{E}_{\substack{Y\sim\pi_i \\ Y'\sim\pi_{-i}}} [p(Y \succ Y')] - \tau \mathrm{KL}(\pi_i \| \pi_{\text{ref}}) + \tau \mathrm{KL}(\pi_{-i} \| \pi_{\text{ref}}). \tag{1}$$

Maura-Rivero et al. (2025) prove that the Nash equilibrium of this game is precisely the maximal lottery for the population's preferences. NLHF therefore inherits maximal lotteries' social-choice-theoretic properties: majority consistency, Condorcet consistency, and IIA in the probabilistic sense (Brandl et al., 2016).

3.2. Online IPO

Online IPO is a pairwise loss whose fixed point coincides with the NLHF Nash equilibrium when the data distribution is the current policy. Given two completions y, y' sampled from the current policy π on prompt x, with y^+, y^- the preferred and dispreferred completions according to the population preference, the loss is

$$\mathcal{L}_{\text{IPO}}(\pi, x, y^+, y^-) = \left(\log \frac{\pi(y^+|x)\pi_{\text{ref}}(y^-|x)}{\pi(y^-|x)\pi_{\text{ref}}(y^+|x)} - \tau^{-1}/2 \right)^2. \tag{2}$$

Calandriello et al. (2024, Proposition 4.2) prove that the expected gradient of this loss when the data distribution is π itself (the *online* setting, as opposed to a fixed offline distribution) coincides with the self-play update direction in the regularised preference game. The fixed point of Online IPO is therefore the Nash equilibrium of that game, which by the Maura-Rivero et al. (2025) result is the maximal lottery. Online IPO thus provides a tractable pairwise loss whose minimiser inherits the social-choice-theoretic properties of maximal lotteries.

3.3. EGGROLL

EGGROLL is an evolution strategies (ES) algorithm that scales to billion-parameter models by replacing full-rank Gaussian perturbations with low-rank ones. For each

Algorithm 1 EGGROLL-IPO($r, \alpha, \sigma, T_{\max}, N_{\text{workers}}$)

Initialise μ and workers with known random seeds ς
for $T_{\max}$ timesteps **do**
  for each worker $i \in \{1, \ldots, N_{\text{workers}}\}$ in parallel **do**
    $A_i \sim p(A_i)$, $B_i \sim p(B_i)$
    $E_i \leftarrow \frac{1}{\sqrt{r}} A_i B_i^\top$
    $\theta_i \leftarrow \mu + \sigma E_i$ *// Perturbed policy parameters*
    Sample $(y_i, y_i') \sim \pi_{\theta_i}(\cdot|x_i)$ *// Sample from policy given worker's context x_i*
    Sample $(y_i^+, y_i^-) \sim \lambda_p(y_i, y_i', x_i)$ from user feedback selecting y_i^+ over y_i^-
    $f_i \leftarrow -\mathcal{L}_{\text{IPO}}(\theta_i, x_i, y_i^+, y_i^-)$ *// Apply the Online IPO formula*
  end for
  Workers share scalar fitness f_i with other workers
  for each worker $i \in \{1, \ldots, N_{\text{workers}}\}$ in parallel **do**
    Reconstruct E_j for $j \in \{1, \ldots, N_{\text{workers}}\}$ from ς
    $\mu \leftarrow \mu + \alpha \frac{1}{N_{\text{workers}}} \sum_{j=1}^{N_{\text{workers}}} E_j f_j$
  end for
end for

linear layer with mean parameters $M \in \mathbb{R}^{m \times n}$, each worker i samples $A_i \in \mathbb{R}^{m \times r}$ and $B_i \in \mathbb{R}^{n \times r}$ with $r \ll \min(m, n)$ and forms the perturbation $E_i = \frac{1}{\sqrt{r}} A_i B_i^\top$. Each worker evaluates the fitness $f_i = f(M + \sigma E_i)$ and broadcasts the scalar f_i. The mean parameters are updated as $M \leftarrow M + \frac{\alpha}{N} \sum_i E_i f_i$. While individual perturbations are rank-r, the aggregated update has rank $\min(Nr, m, n)$, so EGGROLL produces full-rank updates despite low-rank individual search directions, assuming a sufficient number of workers.

EGGROLL achieves up to 91% of the throughput of pure batch inference for billion-parameter models, compared with under 1% for naive ES, by exploiting the batched low-rank adaptation (LoRA; Hu et al., 2021) inference techniques. Sarkar et al. (2025) validate EGGROLL on tabula rasa RL, multi-agent RL, integer-only LLM pretraining, and reasoning fine-tuning, where it is competitive with or outperforms Group Relative Policy Optimisation (GRPO; Shao et al., 2024). To our knowledge, EGGROLL has not previously been evaluated on value alignment tasks.

4. Method

4.1. EGGROLL-IPO

EGGROLL-IPO uses the negative Online IPO loss as the scalar fitness in EGGROLL's update, yielding a gradient-free post-training algorithm that approximates Online IPO's maximal-lottery fixed point. Algorithm 1 states the procedure.

Each worker i forms its low-rank perturbation E_i as in Section 3.3 and evaluates the perturbed policy π_{θ_i} on context x_i. The policy generates a pair of completions (y_i, y_i'), and an oracle (user feedback or, in our experiments, a judge model) supplies the preference between them. The fitness f_i is the Online IPO loss in Equation (2), where π_{ref} is the reference policy. Workers broadcast their scalar fitnesses; each worker reconstructs the others' perturbations from shared seeds and updates the mean parameters μ by the standard EGGROLL aggregation. The aggregated update has rank $\min(N_{\text{workers}} r, m, n)$ despite individual perturbations being rank-r, so EGGROLL-IPO produces full-rank updates from low-rank individual search directions.

Two implementation details matter. **First**, the reference policy π_{ref} is snapshotted at the start of training and held frozen, rather than tracking the current policy. With a non-frozen reference, the IPO log-ratio measures only the per-step perturbation effect rather than cumulative drift from initialisation, removing the loss anchor that pulls the policy toward the maximal-lottery solution. **Second**, each worker samples its pair (y_i, y_i') independently from its own perturbed policy. Shared-pair generation, in which all workers see a single (y, y') drawn once from the unperturbed policy, gives a weaker preference signal because each worker's fitness no longer reflects the preference behaviour of its specific perturbed policy.

The algorithm supports decentralisation along three axes simultaneously. Workers can run on heterogeneous hardware, since they exchange only scalar fitnesses rather than gradient tensors. Workers evaluate on their own contexts x_i, allowing different participants to contribute the queries they care about. In principle, workers can also obtain preferences from their own user feedback channels, allowing different participants to contribute their own preferences; we use a single shared judge model in our experiments. We do not evaluate the decentralised setting empirically; Section 6 discusses the implications.

4.2. Judge Model

The user-feedback step in Algorithm 1 is supplied by a judge J that takes a triplet (y, y', x) and produces a preference probability $p(y \succ y' \mid x) \in [0, 1]$. For the three controlled experiments, the judge is implemented directly from the experiment's preference structure rather than as a learned model, which isolates the loss-axis comparison from any judge-modelling noise. The IIA judge is stochastic: for each (y_1, y_2) pair, it samples a voter type from the two-population distribution (40%/60% weights, with orderings as in Section 5.3) and returns the sampled voter's pairwise preference as a hard label. The cyclic-preferences judge is deterministic and returns the rock-paper-scissors payoff ($p \in \{0, 0.5, 1\}$, with any valid action beating any invalid sample). The sequence-level judge performs a case-insensitive substring

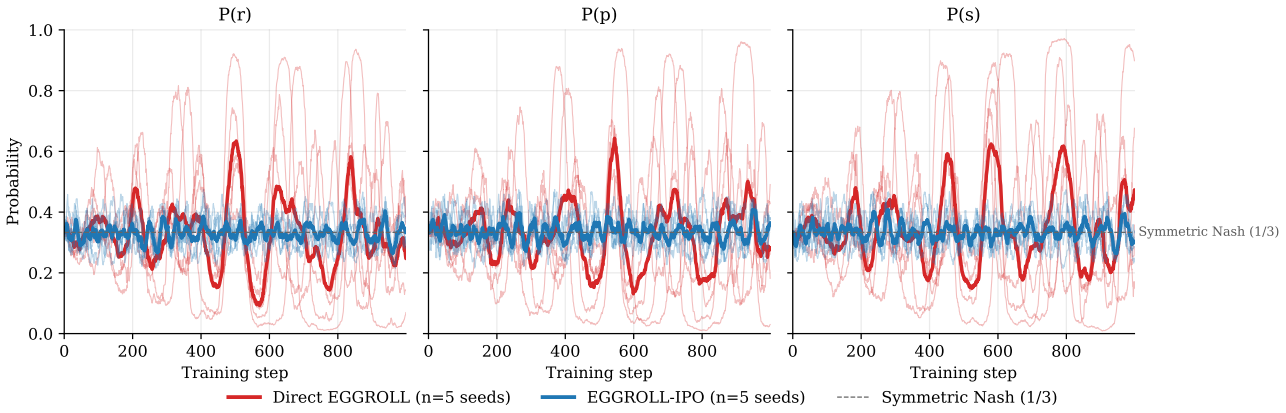

Figure 3. Rock-paper-scissors with EGGROLL. Per-candidate probability mass over training; mean across 5 seeds with ± 1 standard-deviation band. The raw ± 1 fitness collapses to a pure strategy on 5/5 seeds; Online IPO maintains a mixed strategy on 5/5 seeds.

match of each generation against the four candidate strings, returning $p = 0.95$ if y_1 matches the preferred candidate and y_2 does not, $p = 0.05$ for the reverse, and $p = 0.5$ otherwise.

In the IPO arm, soft preferences enter the loss as the convex combination $p \cdot \mathcal{L}_{\text{IPO}}(y_1 \succ y_2) + (1-p) \cdot \mathcal{L}_{\text{IPO}}(y_2 \succ y_1)$. The naive arm uses $f_i = 2p_i - 1$ directly. For the appendix-only Community Alignment experiments (Section 5.5), we additionally use a trained RWKV-7 G1 0.1B (Peng et al., 2025) judge fit to the multi-annotator English subset, with an approximately 13% first-slot position bias corrected by averaging predictions across both orderings, $\hat{p} = \frac{1}{2}\big[p(y_1, y_2) + (1 - p(y_2, y_1))\big]$.

5. Experiments

We evaluate the Online IPO loss against a raw ± 1 preference fitness across three controlled experiments. Each isolates a different failure mode of raw-fitness optimisation: saddle-trapping on a cyclic preference game (Section 5.2), violation of the IIA axiom (Section 5.3), and instability on a transitive sequence-level convergence task (Section 5.4). In all three the EGGROLL noiser is held fixed; only the per-worker fitness function differs. Section 5.5 reports a negative result on real preference data and discusses what is needed to bridge the regime gap.

5.1. Setup

All three experiments use RWKV-7 G1 0.1B (Peng et al., 2025) as the base policy, with EGGROLL configured at $N_{\text{workers}} = 1024$ and $\sigma = 10^{-3}$ throughout. The two arms differ only in the per-worker fitness: the naive arm uses $f_i = 2p_i - 1$ where $p_i \in [0, 1]$ is the soft preference probability supplied by the judge; the IPO arm uses $f_i = -\mathcal{L}_{\text{IPO}}$ from Equation (2) with reference subtraction against a frozen π_{ref}. Per-experiment hyperparameters (learning-rate scale, τ^{-1}, training steps) are reported in each subsection. We evaluate primarily on probability mass assigned to the Condorcet winner under the experiment's preference structure, alongside auxiliary metrics noted in context.

5.2. Cyclic Preferences

This experiment tests behaviour under intransitive preferences, where the symmetric Nash is a mixed strategy and no Condorcet winner exists. We use rock-paper-scissors with the standard zero-sum cycle: R beats S, P beats R, S beats P. The unique symmetric Nash equilibrium is uniform $(1/3, 1/3, 1/3)$, which coincides with the maximal lottery. To disentangle algorithmic behaviour from base-model prior bias, we run two initialisation conditions. **Condition A** starts from the raw RWKV-7 G1 0.1B base, which assigns 87%/9%/4% of valid-action mass to $R/P/S$. **Condition B** starts from a 100-step uniform-target SFT pass that brings the prior to roughly $(1/3, 1/3, 1/3)$. Both arms train for 1000 steps with population size 1024, $\sigma = 10^{-3}$, learning-rate scale 0.125, and (for the IPO arm) $\tau^{-1} = 10$, across 5 seeds per condition. The smaller learning rate and target log-ratio (relative to the other two experiments) reflect that no transitive ranking exists for the loss to drive the policy toward.

Figure 3 reports per-strategy probability mass under Condition B. EGGROLL-IPO converges close to the symmetric Nash on 5/5 seeds with mean final probabilities $(0.33, 0.37, 0.30)$ and per-strategy cross-seed standard deviation at most 0.04. Direct EGGROLL fails to converge under the same conditions: 2/5 seeds collapse to pure S, 3/5 stay mixed but drift, and the cross-seed standard deviation on $P(S)$ reaches 0.27. Under the prior-biased Condition A, all 5 direct-EGGROLL seeds collapse to a pure strategy

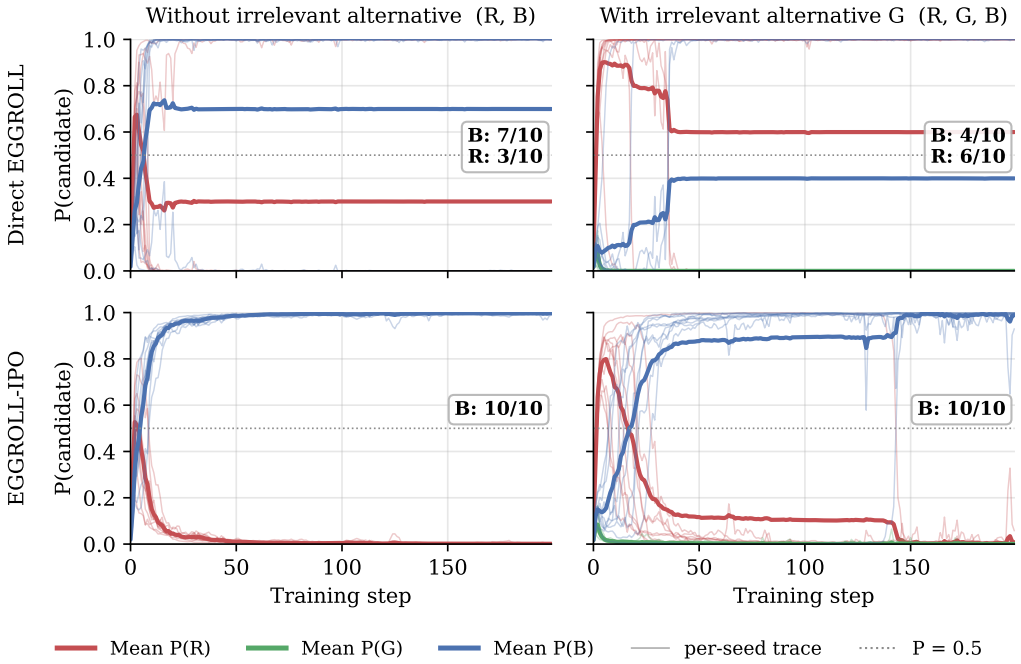

Figure 4. IIA experiment with EGGROLL. Top row: raw ± 1 preference fitness; bottom: Online IPO loss. Left: $k = 2$ alternatives; right: $k = 3$ with irrelevant alternative G. $N = 10$ seeds; per-seed traces shown to expose bistability, with attractor split annotated per panel. The raw fitness shows 3/10 R-attractor seeds at $k = 2$ from prior pull alone, doubling to 6/10 at $k = 3$ once G shifts the Borda landscape; Online IPO holds B on 10/10 in both conditions.

(3 to R, 1 to P, 1 to S), while EGGROLL-IPO settles at $(0.36, 0.46, 0.17)$, the best response to the R-heavy prior. The two conditions together rule out two natural alternative readings of these results: that EGGROLL-IPO's mixed strategy merely inherits the base-model prior (refuted by Condition B converging to Nash even though Condition A was asymmetric), and that direct EGGROLL's collapse is driven by the R-prior alone (refuted by Condition B still collapsing or drifting on most seeds). Across both conditions, EGGROLL-IPO's per-strategy cross-seed standard deviation is up to $12\times$ tighter than direct EGGROLL's at the same hyperparameters.

5.3. Independence of Irrelevant Alternatives

Following Maura-Rivero et al. (2025), we test IIA in the smallest setting where the axiom can fail. The model emits a single token after a fixed prefix asking it to choose a colour. With $k = 2$ alternatives (R, B) and 60% of voters preferring $B \succ R$, B is both the Borda and Condorcet winner. Adding an irrelevant alternative G to give $k = 3$ alternatives (R, G, B), with 40% of voters ranking $R \succ G \succ B$ and 60% ranking $B \succ R \succ G$, B remains the Condorcet winner, but R becomes the Borda winner: G drains R's pairwise losses without itself contesting any majority. A method satisfies IIA iff its winner distribution is invariant to this set change.

We train an RWKV-7 G1 0.1B base for 200 epochs across 10 seeds per cell, with learning-rate scale 0.5 and (for the IPO arm) $\tau^{-1} = 100$. Figure 4 reports per-candidate probability mass over training. EGGROLL-IPO converges to B on 10/10 seeds in both conditions: invariance to the irrelevant alternative. The raw-fitness baseline shows two distinct failures. At $k = 2$, prior pull from the base RWKV-7 distribution drives 3 of 10 seeds to the R-attractor despite B being the unambiguous winner. At $k = 3$, the Borda landscape now favours R and combines with prior pull to put 6 of 10 seeds in the R-attractor: adding the irrelevant alternative doubles the R-rate. With the optimiser held at EGGROLL, the IIA shift is therefore attributable to the loss.

5.4. Sequence-Level Convergence

We apply EGGROLL-IPO to multi-token policies on a transitive preference structure with no IIA confound, isolating convergence stability. We use a single fixed prompt with four length-matched candidate strings (A, B, C, D) and a synthetic judge with the strict total order $D \succ A \succ B \succ C$. The Condorcet winner D is chosen to have the lowest base-model prior so that the policy must actively learn to prefer it; a one-shot SFT warmup gives the initialisation roughly uniform mass across the four candidates. Both arms train for 1000 EGGROLL steps with population size 1024, $\sigma = 10^{-3}$, learning-rate scale 0.5, and (for the IPO

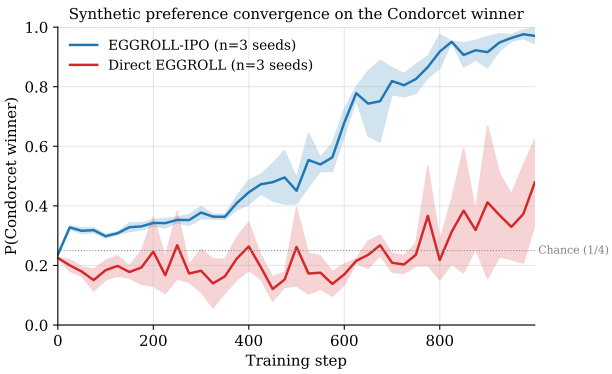

Figure 5. Sequence-level convergence on the four-candidate transitive task. Probability mass on the Condorcet winner D over training; mean across 3 seeds with ± 1 standard-deviation band. EGGROLL-IPO converges monotonically to $P(D) \approx 0.97$; the raw ± 1 fitness oscillates and most seeds finish below the SFT initialisation.

arm) $\tau^{-1} = 100$. We report Condorcet-winner probability mass $P(D)$ and pairwise-preference accuracy across the 6 ordered pairs implied by the total order.

Figure 5 reports $P(D)$ over training. EGGROLL-IPO climbs monotonically and converges to $P(D) = 0.97 \pm 0.03$ across 3 seeds, with KL drift from the reference of 1.35 and pairwise-preference accuracy of 0.78 ± 0.08 across the 6 ordered pairs. The raw-fitness baseline reaches $P(D) = 0.48 \pm 0.14$, more than $5\times$ wider cross-seed than EGGROLL-IPO and roughly $2\times$ lower in mean. The pairwise-accuracy gap is much smaller (0.72 ± 0.08 for the raw fitness): the raw fitness usually orders the four candidates correctly but does not concentrate mass on the Condorcet winner, while EGGROLL-IPO does both. The raw fitness also exhibits roughly $5\times$ less KL drift from the reference (mean 0.28 vs 1.35), consistent with the policy remaining near the SFT initialisation rather than concentrating on D. Even on a clean transitive task with no cyclic structure, the raw ± 1 fitness is markedly less reliable than the Online IPO loss at the same hyperparameters.

5.5. Real-Data Limitation

We attempted the same comparison on the Community Alignment dataset (Zhang et al., 2026), restricting to the multi-annotator English subset (12,866 training rows over 388 prompts; 575 aggregated validation pairs over 80 evaluation prompts). Optimisation stalled across all configurations we swept (population sizes $\in \{64, 256, 1024\}$, learning-rate scales $\in \{0.5, 2, 4\}$, $\sigma \in \{5\times10^{-4}, 10^{-3}\}$): the IPO loss decreased by less than 2% over 500 steps in the best run, and Condorcet-winner mass moved positively in only one configuration. Diagnostics rule out judge mis-specification (the trained judge prefers the labelled Condorcet winner on 80% of held-out pairs) and distribution shift between train and evaluation prompts. The bottleneck appears to be judge informativeness on novel policy-generated pairs: roughly 40% of even labelled pairs land in the uncertain band $p \in [0.4, 0.6]$, and EGGROLL's z-scored fitness aggregation cancels weakly correlated signals across the population, starving the update. We document this fully in Appendix B as an honest negative result. Bridging the regime gap likely requires far longer training, judge-variance-weighted worker reweighting, or a hybrid that uses backprop-based preference optimisation as a warm start before the EGGROLL phase.

6. Discussion

EGGROLL-IPO communicates a single scalar per worker per evaluation, which makes preference-level decentralisation practical in a way that gradient-based pipelines do not: compute donors can contribute compute, queries, and preferences without exchanging gradient tensors.

Several limitations qualify these results. We characterise the decentralised protocol but do not run it empirically; all experiments are centralised on a single GPU. We work at 0.1B parameters, and while EGGROLL has been validated at billion-parameter scale (Sarkar et al., 2025), EGGROLL-IPO has not. And although our results are consistent with convergence to maximal-lottery-like outcomes on synthetic preference structures, we do not prove that the EGGROLL+IPO update approximates the maximal-lottery fixed point in expectation.

Several concrete next steps follow. The most immediate is bridging the regime gap on real preference data, where length-normalised log-probabilities applied throughout the loss formulation, rather than the pair-level truncation patch we used, plausibly resolves the largest obstacle (Section B). A human-subject pilot using prompt-level overlap from multiple raters per prompt would replace the judge with the eventual target signal and test convergence on real population preferences. Demonstrating the algorithm at billion-parameter scale, and proving that the update approximates the maximal lottery in expectation, are natural milestones once the data-regime gap is closed.

7. Conclusion

EGGROLL-IPO is a variant of preference-based RLHF that uses an evolution-strategies optimiser and a voting-theoretic loss. The algorithm empirically tracks Online IPO's social-choice-theoretic guarantees and EGGROLL's scalar-fitness communication, the latter making pluralism a property of the training protocol rather than post-hoc aggregation. In three controlled experiments holding the optimiser fixed at EGGROLL, the Online IPO loss converges to the maximal

lottery on each task while a raw ± 1 preference fitness fails distinctly on each. Three directions remain open: pilots with real users via prompt-level overlap from human-subject studies, differential-privacy guarantees from noise on the scalar fitness channel, and scaling to larger base models.

Impact Statement

This paper presents work whose goal is to advance methods for aligning large language models with diverse human preferences. By reducing the compute and memory cost of pluralistic post-training, our approach lowers the barrier to safety-relevant fine-tuning by actors outside major AI labs, and provides a path toward decentralised alignment protocols where compute donors contribute their preferences and queries directly. We do not anticipate immediate negative societal consequences beyond those well-established for alignment research generally, but flag that any preference-aggregation scheme inherits the biases of the population it samples from, and that judge models used to provide online preference signal can themselves be misaligned with the population they purport to represent.

References

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., and Kaplan, J. Constitutional AI: Harmlessness from AI Feedback, December 2022. URL http://arxiv.org/abs/2212.08073. arXiv:2212.08073 [cs]. 1

Brandl, F. and Brandt, F. Arrovian Aggregation of Convex Preferences. *Econometrica*, 88(2):799–844, 2020. ISSN 1468-0262. doi: 10.3982/ECTA15749. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA15749. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA15749. 2, 3

Brandl, F., Brandt, F., and Seedig, H. G. Consistent Probabilistic Social Choice, July 2016. URL http://arxiv.org/abs/1503.00694. arXiv:1503.00694 [cs]. 3

Calandriello, D., Guo, Z. D., Munos, R., Rowland, M., Tang, Y., Pires, B. A., Richemond, P. H., Lan, C. L., Valko, M., Liu, T., Joshi, R., Zheng, Z., and Piot, B. Human Alignment of Large Language Models through Online Preference Optimisation. June 2024. URL https://openreview.net/forum?id=2RQqg2Y7Y6. 2, 3

Douillard, A., Feng, Q., Rusu, A. A., Chhaparia, R., Donchev, Y., Kuncoro, A., Ranzato, M., Szlam, A., and Shen, J. Diloco: Distributed low-communication training of language models, 2024. URL https://arxiv.org/abs/2311.08105. 3

Fishburn, P. C. Probabilistic Social Choice Based on Simple Voting Comparisons. *The Review of Economic Studies*, 51(4):683–692, October 1984. ISSN 0034-6527. doi: 10.2307/2297786. URL https://doi.org/10.2307/2297786. 3

Glaese, A., McAleese, N., Trębacz, M., Aslanides, J., Firoiu, V., Ewalds, T., Rauh, M., Weidinger, L., Chadwick, M., Thacker, P., Campbell-Gillingham, L., Uesato, J., Huang, P.-S., Comanescu, R., Yang, F., See, A., Dathathri, S., Greig, R., Chen, C., Fritz, D., Elias, J. S., Green, R., Mokrá, S., Fernando, N., Wu, B., Foley, R., Young, S., Gabriel, I., Isaac, W., Mellor, J., Hassabis, D., Kavukcuoglu, K., Hendricks, L. A., and Irving, G. Improving alignment of dialogue agents via targeted human judgements, September 2022. URL http://arxiv.org/abs/2209.14375. arXiv:2209.14375 [cs]. 1

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. October 2021. URL https://openreview.net/forum?id=nZeVKeeFYf9. 4

Kreweras, G. Aggregation of preference orderings. In *Mathematics and Social Sciences I: Proceedings of the seminars of Menthon-Saint-Bernard, France (1–27 July 1960) and of Gösing, Austria (3–27 July 1962)*, pp. 73–79, 1965. 3

Malladi, S., Gao, T., Nichani, E., Damian, A., Lee, J. D., Chen, D., and Arora, S. Fine-tuning language models with just forward passes. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc. 3

Maura-Rivero, R.-R., Lanctot, M., Visin, F., and Larson, K. Jackpot! Alignment as a Maximal Lottery, January 2025. URL http://arxiv.org/abs/2501.19266. arXiv:2501.19266 [cs]. 2, 3, 6

Munos, R., Valko, M., Calandriello, D., Azar, M. G., Rowland, M., Guo, Z. D., Tang, Y., Geist, M., Mesnard,

T., Fiegel, C., Michi, A., Selvi, M., Girgin, S., Momchev, N., Bachem, O., Mankowitz, D. J., Precup, D., and Piot, B. Nash Learning from Human Feedback. June 2024. URL https://openreview.net/forum?id=Y5AmNYiyCQ. 2, 3

OpenAI. OpenAI Model Spec, December 2025. URL https://model-spec.openai.com/2025-12-18.html. 1

Peng, B., Zhang, R., Goldstein, D., Alcaide, E., Du, X., Hou, H., Lin, J., Liu, J., Lu, J., Merrill, W., Song, G., Tan, K., Utpala, S., Wilce, N., Wind, J. S., Wu, T., Wuttke, D., and Zhou-Zheng, C. RWKV-7 "Goose" with Expressive Dynamic State Evolution, March 2025. URL http://arxiv.org/abs/2503.14456. arXiv:2503.14456 [cs]. 5

Sarkar, B., Fellows, M., Duque, J. A., Letcher, A., Villares, A. L., Sims, A., Wibault, C., Samsonov, D., Cope, D., Liesen, J., Li, K., Seier, L., Wolf, T., Berdica, U., Mohl, V., Goldie, A. D., Courville, A., Sevegnani, K., Whiteson, S., and Foerster, J. N. Evolution Strategies at the Hyperscale, November 2025. URL https://arxiv.org/abs/2511.16652v2. 1, 4, 7

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]. 4

Siththaranjan, A., Laidlaw, C., and Hadfield-Menell, D. Distributional Preference Learning: Understanding and Accounting for Hidden Context in RLHF. October 2023. URL https://openreview.net/forum?id=0tWTxYYPnW. 2

Tang, X., Panda, A., Nasr, M., Mahloujifar, S., and Mittal, P. Private fine-tuning of large language models with zeroth-order optimization. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=3Y3o0yFZfu. 3

Zhang, L. H., Milli, S., Jusko, K., Smith, J., Amos, B., Bouaziz, W., Revel, M., Kussman, J., Sheynin, Y., Titus, L., Radharapu, B., Yu, J., Sarma, V., Rose, K., and Nickel, M. Cultivating Pluralism In Algorithmic Monoculture: The Community Alignment Dataset, February 2026. URL http://arxiv.org/abs/2507.09650. arXiv:2507.09650 [cs]. 7, 12

Appendix

A. Experimental Details

A.1. Independence of Irrelevant Alternatives

Preference structure. With voter populations $R \succ G \succ B$ (40%) and $B \succ R \succ G$ (60%), the pairwise majorities are: B beats R on 60% of voters, B beats G on 60%, and R beats G unanimously. Hence B is the Condorcet winner: it beats every alternative pairwise. The Borda points (expected pairwise wins per voter) are $R = 0.4 \cdot 2 + 0.6 \cdot 1 = 1.4$, $B = 0.4 \cdot 0 + 0.6 \cdot 2 = 1.2$, $G = 0.4 \cdot 1 + 0.6 \cdot 0 = 0.4$, making R the Borda winner. Removing the irrelevant alternative G collapses both voter rankings to $B \succ R$, so B becomes both Borda and Condorcet winner. IIA invariance requires the choice rule's winner distribution to be unchanged by this set change; raw-fitness EGGROLL fails this test (Section 5.3).

A.2. Sequence-Level Convergence

Prompt and candidates. The fixed prompt is `user: Choose your favourite: 'the apple is red', 'the banana is yellow', 'the cherry is red', 'the date is brown'\nassistant:`. The four candidate completions (A = "the apple is red", B = "the banana is yellow", C = "the cherry is red", D = "the date is brown") are 4 tokens each. The synthetic judge implements the strict total order $D \succ A \succ B \succ C$, evaluated via case-insensitive substring match (see Section 4.2).

SFT warmup. A 100-step pass with random per-step candidate selection (uniform over the 4 candidates) at learning rate 10^{-4}, taking approximately 5 minutes on RWKV-7 G1 0.1B. Post-SFT generation match rates over 50 samples at temperature 1.0: A 28%, B 36%, C 12%, D 20%, no match 4%. The Condorcet winner D has the lowest match rate of the four candidates by design, forcing the policy to actively learn to prefer it during the EGGROLL pass.

Per-seed final metrics. Table 1 reports each seed's final $P(D)$ (Condorcet-winner mass), pairwise-preference accuracy across the 6 ordered pairs implied by the total order, and KL drift to the reference policy after 1000 EGGROLL steps.

Loss	Seed	$P(D)$	Pref. acc.	KL drift
EGGROLL-IPO	0	0.994	4/6	1.448
	1	0.933	5/6	1.184
	2	0.986	5/6	1.404
Direct EGGROLL	0	0.425	4/6	0.117
	1	0.675	5/6	0.580
	2	0.336	4/6	0.156

Table 1. Sequence-level per-seed final metrics after 1000 EGGROLL steps.

Naive instability example. Direct EGGROLL seed 1 (the highest-finishing of the three) traversed $P(D) \approx 0.225$ at the start, 0.078 at step 600, 0.598 at step 775, 0.140 at step 800, 0.598 at step 825, and 0.675 at step 999. Seeds 0 and 2 traced similarly non-monotonic paths, finishing at 0.425 and 0.336. EGGROLL-IPO seeds were monotonically ascending and finished within ± 0.03 of $P(D) = 0.97$.

A.3. Cyclic Preferences

Payoff matrix. The judge returns the row vs column payoff with no per-call randomness:

	r	p	s	invalid
r	0	-1	$+1$	$+1$
p	$+1$	0	-1	$+1$
s	-1	$+1$	0	$+1$
invalid	-1	-1	-1	0

Symmetric on the three valid actions, with cyclic off-diagonals and zero on the diagonal; any valid action beats any invalid

sample. The output surfaces to the trainer as $p \in \{0, 0.5, 1\}$.

Condition B SFT recipe. A 100-step pass with cross-entropy loss against a uniform target on $\{r, p, s\}$ at learning rate 10^{-4}, taking approximately 5 minutes on RWKV-7 G1 0.1B. The post-SFT valid-action distribution is $P(R, P, S \mid \text{valid}) = (0.326, 0.347, 0.326)$.

Per-seed final probabilities. Table 2 reports each seed's final $P(R), P(P), P(S)$ at step 999 across both conditions and arms.

Condition	Loss	Seed	$P(R)$	$P(P)$	$P(S)$
Cond A (raw prior)	Direct EGGROLL	0	0.000	0.999	0.000
		1	1.000	0.000	0.000
		2	0.911	0.048	0.035
		3	0.000	0.000	1.000
		4	1.000	0.000	0.000
	EGGROLL-IPO	0	0.339	0.471	0.177
		1	0.344	0.441	0.193
		2	0.383	0.469	0.136
		3	0.418	0.388	0.176
		4	0.300	0.510	0.168
Cond B (uniform init)	Direct EGGROLL	0	0.069	0.030	0.898
		1	0.142	0.182	0.676
		2	0.381	0.262	0.358
		3	0.286	0.429	0.286
		4	0.359	0.491	0.150
	EGGROLL-IPO	0	0.392	0.368	0.238
		1	0.312	0.354	0.332
		2	0.279	0.382	0.337
		3	0.311	0.376	0.311
		4	0.367	0.345	0.286

Table 2. Rock-paper-scissors per-seed final probabilities at step 999.

Cross-seed standard deviations. Table 3 summarises per-strategy cross-seed standard deviations across both conditions and arms. EGGROLL-IPO's variation is up to $12\times$ tighter than direct EGGROLL's at matched hyperparameters.

Condition	Loss	σ_R	σ_P	σ_S
Cond A (raw prior)	Direct EGGROLL	0.476	0.395	0.397
	EGGROLL-IPO	0.040	0.040	0.019
Cond B (uniform init)	Direct EGGROLL	0.122	0.167	0.274
	EGGROLL-IPO	0.041	0.014	0.036

Table 3. Cross-seed standard deviations of final per-strategy probabilities, $n = 5$ seeds per cell.

A.4. Compute Summary

The three controlled experiments consumed approximately 35 GPU-hours total: IIA 20.0 GPU-h (40 runs at roughly 0.5 GPU-h each), sequence-level IPO 3.5 GPU-h (3 seeds at 1.17 GPU-h), sequence-level naive 8.5 GPU-h (3 seeds at 2.82 GPU-h average; two of three seeds ran on slower pods), RPS Condition A 1.6 GPU-h (10 seeds at 0.16 GPU-h), and RPS Condition B 1.6 GPU-h (10 seeds plus a 5-minute SFT warmup). The Community Alignment experiments in Section B consumed an additional approximately 30 GPU-h across the Community Alignment exploration and the 1.5B-parameter scaling check.

B. Community Alignment Negative Result

We attempted to extend the EGGROLL-IPO comparison to real preference data using Community Alignment v2 (Zhang et al., 2026), restricting to the multi-annotator English subset (at least 5 annotators per prompt with agreement at least 0.7): 12,866 training rows over 388 prompts, with 575 aggregated validation pairs over 80 evaluation prompts (4 candidates each, in the Condorcet-winner-mass evaluation set). We trained an RWKV-7 G1 0.1B judge on annotator-agreement fractions and explored configurations informally rather than as a full factorial sweep, varying population size in $\in \{64, 256, 1024\}$, learning-rate scale $\in \{0.5, 2, 4\}$, and $\sigma \in \{5 \times 10^{-4}, 10^{-3}\}$, with $\tau^{-1} = 100$. Optimisation stalled across all configurations we tried, varying population size in $\{64, 256, 1024\}$, learning-rate scale in $\{0.5, 2, 4\}$, and $\sigma \in \{5{\times}10^{-4}, 10^{-3}\}$ informally rather than as a full grid: the IPO loss decreased by less than 2% over 500 steps in the best run, and Condorcet-winner mass moved positively in only one configuration.

Length-normalisation confound. The primary diagnosed obstacle is a length confound in the IPO log-ratio. Real CA completions vary substantially in token length, and the IPO loss as written in Equation (2) compares unnormalised sequence log-probabilities, which scale roughly linearly with sequence length. When the two completions in a pair differ in length by Δ tokens, the log-ratio $\log[\pi(y_1)/\pi(y_2)] - \log[\pi_{\text{ref}}(y_1)/\pi_{\text{ref}}(y_2)]$ is dominated by an order-Δ term in expected per-token log-probability rather than by the per-token preference content. A diagnostic accuracy curve over held-out pairs binned by Δ shows roughly 70% accuracy at $\Delta \approx 0$ and decreases monotonically toward chance as Δ grows: the trainer is largely tracking which completion is longer rather than which is preferred. We applied a naive truncation fix (cutting the longer completion to the length of the shorter before computing log-probabilities), which removes the bin-wise degradation, but this alone does not lift optimisation into a regime where the IPO loss decreases meaningfully. A more principled length normalisation, applied throughout the loss formulation rather than as a pair-level patch, is left to future work.

Optimiser stall. Across all 9 sweep configurations, the per-step IPO loss barely moves: less than 2% reduction over 500 steps in the best run. At $\tau^{-1} = 100$, the IPO target log-ratio is $\tau^{-1}/2 = 50$, corresponding to loss $= 0$; in practice the trainer sat at loss ≈ 2500 (log-ratio ≈ 0). The mean log-ratio stays near 0 while its standard deviation grows as workers diverge, indicating that workers explore but their preferences cancel on average. Only one configuration (population 64, $\sigma = 5{\times}10^{-4}$, learning-rate scale 2) showed positive mid-run movement on both Condorcet-winner mass and pairwise-preference accuracy; no configuration drove the IPO loss meaningfully toward zero.

Secondary factor: judge informativeness on novel generations. On labelled CA pairs where a Condorcet winner exists, roughly 40% of pairs land in the uncertain band $p \in [0.4, 0.6]$. On novel worker-generated (y_1, y_2) pairs from the policy distribution, the informative-pair rate is even lower (the policy's outputs are out-of-distribution relative to the labelled CA pairs the judge was trained on). EGGROLL z-scores per-worker fitness across the population, which cancels weakly correlated signals and starves the gradient when most pairs are uncertain. This compounds the length confound: even when length is controlled by truncation, the per-pair signal from the judge is too weak across most pairs to drive optimisation.

The training signal itself is sound. Independent of these aggregation issues, the trained CA judge prefers the labelled Condorcet winner on 80% of held-out (CW, non-CW) pairs (mean $p(\text{CW} \succ \text{non-CW}) = 0.656$ across 507 pairs; argmax correct on 406/507). Train and evaluation prompt distributions match: zero overlap, matching word-length statistics, so the failure is not a distribution-shift issue. A roughly 13% A-slot position bias is corrected by averaging predictions across both pair orderings, $\hat{p} = \frac{1}{2}\big[p(y_1, y_2) + (1 - p(y_2, y_1))\big]$.

Scale partially helps but does not fix it. A 1.5B-parameter SFT-warmed initialisation begins training at $P(\text{CW}) = 0.378$ and pairwise accuracy 0.802, compared with 0.275 and 0.62 at 0.1B. But the gain is in the SFT phase, not in the EGGROLL pass: from either starting point, the EGGROLL phase fails to make meaningful further progress.

What would be needed. Three plausible directions to bridge the regime gap, none of which we attempted in this work: (i) more principled length normalisation than the naive truncation fix, for example length-normalised log-probabilities applied throughout the loss formulation, or pair construction that draws length-matched candidates; (ii) judge-variance-weighted worker reweighting that downweights pairs with p near 0.5; or (iii) a hybrid that uses backprop-based preference optimisation as a warm start before the EGGROLL phase. The EGGROLL framework retains its scalar-fitness communication advantages in all three cases.