CAN MORE LLM USERS REDUCE THE WELFARE COST OF SELFISH ADVICE?

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) increasingly guide decisions whose consequences extend beyond their users. We study a shared LLM *advisor* that learns from user feedback while other players follow no regret learning in the *volunteer's dilemma*, a multiplayer public good game. We compare a *selfish* LLM benchmark that optimizes its users' expected payoffs with a *selfless* benchmark that optimizes total expected welfare of the system. Our central quantity is the social welfare gap, the selfless benchmark's total welfare minus the selfish benchmark's at the same adoption level. To analyze this gap, we first establish convergence of the other players' response and the advisor's simplified mean learning dynamics. When the other players continue volunteering, both objectives select the same policy, so the gap is zero. When LLM users alone supply the public good, this gap has a unique maximum as adoption varies. The gap first widens and then narrows when the maximum is interior. Broader adoption brings more of those who benefit from the public good into the selfish objective. Higher contribution costs relative to the benefit each player receives increase the gap and move an interior maximum toward lower adoption level. Our experiments on Qwen3 8B reveal dynamics close to theoretical predictions and generally smaller average welfare gaps at higher adoption levels. These findings suggest that wider LLM adoption can reduce the welfare cost of serving users alone, so losses at limited adoption need not persist as more users join.

1 INTRODUCTION

Large language models (LLMs) are increasingly used to support decisions in daily life and work (Bick et al., 2026; Chatterji et al., 2025). Applications range from practical advice and information gathering to helping customer support workers resolve problems and helping professionals develop product proposals (Chatterji et al., 2025; Brynjolfsson et al., 2025; Dell'Acqua et al., 2026). These applications bring LLM advice into decisions that affect colleagues, customers, and collaborators. In particular, team applications connect individual contributions to a shared outcome. LLMs are also being studied as players in repeated cooperation and coordination games alongside humans and programmed strategies (Akata et al., 2025). Together, these developments motivate understanding how decisions guided by LLMs affect the people with whom their users interact.

However, wider use does not imply uniform adoption. Recent surveys have identified concerns about privacy and suitability for particular tasks among the reasons for retaining classic decision-making methods (OECD, 2025). A natural setting is therefore a mixed population in which players guided by LLMs coexist with people relying on their own judgment or conventional algorithms. As the number of LLM players grows, two challenges arise. First, LLM advice changes users' actions and hence the incentives facing other players, but benefits to current users need not translate into benefits for the whole population. Second, the groups may adapt at different timescales. No regret algorithms can revise strategies after each interaction (Cesa-Bianchi & Lugosi, 2006), whereas a *shared LLM advisor* can be updated periodically from feedback pooled across its users (Cao & Hu, 2026). One service update can consequently change many users' decisions while other players continue adjusting to them. The social effect of adoption therefore depends on both population composition and the responses that emerge over time. Evaluating the advisor in isolation, or before these responses settle, can miss that interaction.

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

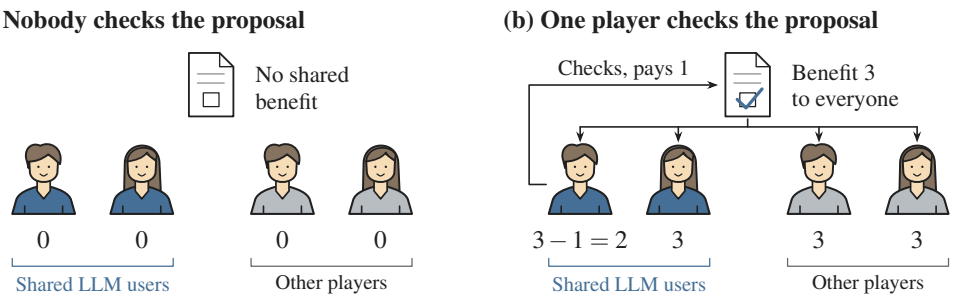

Figure 1: **Example of the Volunteer's Dilemma Diekmann (1985).** Checking a shared proposal costs 1 and gives every player benefit 3. Blue players use the shared advisor. In (b), the arrows show one player's check and the benefit reaching all four players. Numbers below players give net payoffs for these two illustrative action outcomes.

As the two groups adapt to each other, an advisor optimizing only its users' payoffs can affect benefits that its objective does not include. We therefore compare two policy benchmarks. A *selfish* LLM optimizes the expected payoffs of its users, whereas a *selfless* LLM optimizes the total expected payoff of all players. These terms describe whose payoffs enter the objective. At the same adoption level, the total welfare achieved by selfless advice minus that achieved by selfish advice is the *social welfare gap* studied in this paper. Section 2.5 defines this comparison precisely.

This gap measures the additional total welfare available from including other players' payoffs in the objective, holding the game, adoption level, and feasible shared policies fixed. Using the best policy for each objective separates this cost from losses caused by incomplete optimization. This distinction matters because total welfare can change with adoption even under the selfless benchmark. However, wider adoption also brings more players into the group whose payoffs the selfish objective includes, while changing contribution incentives and the optimal policies themselves. It is therefore unclear whether the welfare gap grows or shrinks. Thus, we ask

Can wider LLM adoption narrow the welfare gap created by optimizing advice only for its users?

To answer this, we study the *volunteer's dilemma* (Diekmann, 1985; Palfrey & Rosenthal, 1984), a classical multiplayer public-good game in which contributors bear individual costs while the benefits extend to the whole group. For example, a team may need someone to check a shared proposal; everyone benefits from a successful check, but the reviewer bears the effort. Each member may therefore prefer that someone else take responsibility, even though the team needs the task completed. This simple yet insightful setting isolates how advice that saves effort for its users can affect a benefit shared with other players. Figure 1 illustrates the payoffs in a single interaction with four players, two of whom use the same advisor.

We consider a finite population of no-regret learners and LLM-guided players. No-regret learners update their strategies using full-information reward feedback (Cesa-Bianchi & Lugosi, 2006). LLM-guided players report their actions and realized rewards to a shared advisor (a publicly available LLM serving multiple players), which periodically updates its recommendations from these reports. We first analyze how these learning processes interact and how the advisor's initial tendency affects their long-run behavior. We then evaluate social welfare under the volunteering behavior of both groups after learning settles, across different game payoffs and LLM penetration rates.

Volunteering is costly and affects provision[1] only when nobody else volunteers, creating the classical incentive to rely on others (Diekmann, 1985). Under fixed advice, Lemma 1 derives the other players' limiting response. As LLM users' volunteering probability rises, the others' limiting probability decreases continuously to zero at a threshold set by game payoffs and adoption. Below it, both groups contribute (*type 2*); at or above it, only LLM users contribute (*type 1*). These *contribution regimes*, formalized in Definition 3, emerge from the learning dynamics and determine who bears contribution costs and how reliably the public good is provided.

Our central finding is that wider adoption can eventually narrow the welfare gap created by optimizing advice for current users, in particular for the type 1 regime. To characterize the social welfare gap

[1]Note that *provision* means that at least one player volunteers, so all players receive the shared benefit.

Table 1: Main results on learning (Section 3) and the social welfare gap (Section 4).

Setting	Reference
No-regret learning (fixed advice)	Lemma 1
Advisor learning (strong regularization)	Theorem 3.1
Type 1 regime	Lemma 2(i)
Type 2 regime	Lemma 2(ii)
LLM penetration rate (type 1)	Theorem 4.1(ii), Corollary 4.2
Contribution cost ratio (type 1)	Theorem 4.1(i), Corollary 4.1

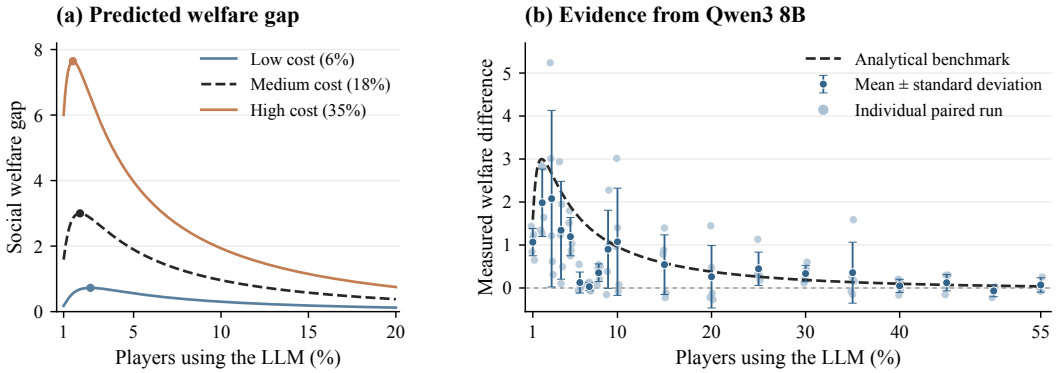

Figure 2: **Wider LLM penetration can narrow the social welfare gap.** We compare total welfare under advice optimized for users and for all players ($N = 200$). **(a) Theory in type 1.** The gap decreases beyond its peak. Higher contribution costs raise the gap and shift an interior peak toward lower penetration (Theorem 4.1 and Corollary 4.1). **(b) Qwen3 8B at 18% cost.** All 91 paired observations compare learned policies approximating the two benchmarks. We show **means ± one sample standard deviation**. Costs and welfare gaps are normalized by each player's benefit. Appendix E.3 gives the experimental protocol and data provenance.

(see Definition 2), we compare the total welfare under advice optimized for LLM users with the best total welfare attainable while those users remain the only contributors. Interestingly, we show that the combined effect can first widen and then narrow the social welfare gap, as the number of LLM players increases (see Figure 2). Thus, a large loss at limited adoption need not persist at higher adoption, even though the advisor continues to optimize only for its users. Figure 2 demonstrates this phenomenon by varying the *LLM penetration rate*, i.e., the fraction of players using the advisor. The *social welfare gap* on the y-axis measures the welfare cost of the selfish objective. Each curve fixes a *cost* in the game, and lets learning settle by increasing the numbers of LLM players. For type 2, the best advice for LLM users also maximizes total welfare within that regime, so this gap is zero.

We first establish the learning foundation for this welfare comparison. Then, we characterize the adoption effect and extend the analysis to welfare across contribution regimes. Table 1 maps these results to the learning analysis in Section 3 and the welfare analysis in Section 4.

(i) **Learning dynamics and convergence.** Lemma 1 first determines the other players' limiting response to fixed advice. Substituting this response into the advisor's conditional expected update gives its reduced mean dynamics (Proposition 1). Under strong regularization, these dynamics converge to a unique attracting policy (Theorem 3.1). Lemma 2 then shows that, in type 1, the policy maximizing users' welfare and the policy maximizing total welfare can each be reached by the reduced mean dynamics when learning is regularized toward a suitably chosen pretrained policy. This provides a foundation for the welfare analysis.

(ii) **Wider adoption can narrow the welfare gap.** In type 1, the gap has a unique maximum over adoption, possibly at a boundary due to integer population limitation. If this maximum is interior, the gap first widens and then narrows (Theorem 4.1). Higher contribution cost ratios increase the gap and move an interior maximum toward lower penetration (Corollary 4.1). Corollary 4.2 gives a sufficient penetration level for a prescribed gap. In

type 2, the benchmark policies coincide and the gap is zero. Thus, a large welfare gap at limited adoption need not persist as more users join, even though advice remains optimized for users alone (Figure 2).

(iii) **Experimental results.** Experiments show generally smaller average welfare differences at higher LLM penetration. Qwen3 8B policies also yield volunteering long run dynamics near the theoretical predictions (Section 5). The welfare experiment compares policies learned from different references approximating the two benchmarks, both trained with individual payoff feedback. These finite training comparisons provide exploratory evidence and retain substantial variation across runs (Figure 2(b)).

(iv) **Extension to welfare across contribution regimes.** We also compare the best payoff per LLM user in type 1 and type 2. This ranking depends on contribution costs and penetration, and can reverse twice as adoption grows (Theorem B.1, Appendix B.6, and Figure 4).

2 Game and Learning Model

2.1 Stage Game and Population Composition

In the repeated volunteer's dilemma (Diekmann, 1985), N players simultaneously choose volunteering or defection from $\mathcal{A} \coloneqq \{V, D\}$. One volunteer gives everyone benefit G, while each volunteer pays c, where $G > c > 0$. A volunteer receives $G - c$, whereas a defector receives G if any other volunteers and zero otherwise. Free riding saves costs but risks the shared benefit.

The population contains $N\rho$ users of a shared LLM advisor and $N(1-\rho)$ no-regret (NR) learners. The *LLM penetration rate* $\rho \in (0, 1)$ and group membership remain fixed during whole dynamic process. We write $\mathcal{V} \coloneqq (N, \rho, c, G)$ for the game, state which entries vary in each comparison, and suppress this argument within a fixed game.

Assumption 1 (Non-degenerate subpopulations)**.** *We assume $N\rho \geq 2$ and $N(1-\rho) \geq 2$.*

Note that this restriction keeps opponent counts within each player's group positive in the response and welfare formulas. Singleton groups and homogeneous populations require separate analysis.

2.2 No Regret Learners

We use exponential weights, which favors actions with higher accumulated rewards (Cesa-Bianchi & Lugosi, 2006; Arora et al., 2012). Let z_t be the volunteering score minus the defection score at the start of period t. Each NR player volunteers with probability $\sigma(z_t)$, where $\sigma(z) \coloneqq (1 + e^{-z})^{-1}$, and defects with probability $1 - \sigma(z_t)$. The population shares one score, updated from a representative player's *full information feedback*, namely the rewards $r_{V,t}$ and $r_{D,t}$ of both actions against the same realized opponents. With reward difference $\delta r_t \coloneqq r_{V,t} - r_{D,t}$, learning rate $\tilde{\eta}_t > 0$, and finite initial score z_1, exponential weights gives (Arora et al., 2012, Section 2.1)

$$z_{t+1} = z_t + \tilde{\eta}_t \delta r_t. \tag{1}$$

A positive reward difference therefore increases volunteering probability.

Assumption 2 (Aggregate no-regret sampling and step sizes)**.** *The process uses the conditional sampling rule in Section 2.4 and positive deterministic step sizes satisfying $\sum_{t=1}^{\infty} \tilde{\eta}_t = \infty$ and $\sum_{t=1}^{\infty} \tilde{\eta}_t^2 < \infty$.*

We use a common representative score to approximate the population. Standard stochastic approximation step conditions permit continued adjustment while controlling sampling noise (Robbins & Monro, 1951; Benaim, 1999). Appendix C.1.1 derives this representation and clarifies its scope.

2.3 LLM Users and the Shared Advisor

Every batch $b = 0, 1, \ldots$ contains B periods. Shared logits $\mu_b \coloneqq (\mu_{V,b}, \mu_{D,b})$ determine the recommendation policy $\mathbf{q}_b$ and users' effective action policy $\mathbf{p}_b$ through response sensitivity $\lambda > 0$,

$$\mathbf{q}_b \coloneqq \mathrm{softmax}(\mu_b), \qquad \mathbf{p}_b \coloneqq \mathrm{softmax}(\lambda \mu_b). \tag{2}$$

Bold symbols denote action distributions with components $q_{a,b}$ and $p_{a,b}$. The scalars $q_b \coloneqq q_{V,b}$ and $p_b \coloneqq p_{V,b}$ give recommended and actual volunteering probabilities, equal when $\lambda = 1$. Appendix C.1.5 details the mapping for other λ.

Remark 1 (Action abstraction). *Mapping language recommendations to game actions follows LLM game experiments (Akata et al., 2025; Lorè & Heydari, 2024; Robinson & Burden, 2025). A common λ summarizes all users' responses.*

Assumption 3 (Batched shared policy). *The logits μ_b and policies $\mathbf{q}_b, \mathbf{p}_b$ remain fixed during batch b. The advisor updates after its B periods.*

Each guided user $n = 1, \ldots, N\rho$ reports its realized action $a_{n,t}$ and payoff $r_{n,t}$ for $t = 1, \ldots, B$, yielding $BN\rho$ *bandit feedback* reports. From finite μ_0, freeze $\mathbf{q}_0, \mathbf{p}_0$ as references. Subtract a sampled KL penalty toward $\mathbf{p}_0$ with strength β/λ ($\beta \geq 0$), then center rewards by their batch average to form advantages $A_{n,t}$ (Appendix C.1.4). With e_a the unit vector for action a and advisor learning rate $\eta_b > 0$, the update is

$$g_b := \sum_{t=1}^{B} \sum_{n=1}^{N\rho} A_{n,t}(e_{a_{n,t}} - \mathbf{p}_b), \qquad \mu_{b+1} = \mu_b + \eta_b g_b. \tag{3}$$

Positive advantages favor the reported action. Absorbing the score derivative factor λ into η_b, the summed reports give effective batch step $\gamma_b := \eta_b BN\rho$ (Section 3.2). The separate NR period step $\tilde{\eta}_t$ appears in (1).

Assumption 4 (Advisor step sizes). *The advisor uses positive deterministic step sizes η_b, possibly depending on B, satisfying $\gamma_b \to 0$ as $b \to \infty$ and $\sum_{b=0}^{\infty} \gamma_b = \infty$ for each fixed B. For fixed $\mathcal{V}$, the large-batch reduction takes $B \to \infty$ with $\eta_b BN\rho \to 0$, equivalently $\eta_b = o(B^{-1})$.*

The shrinking effective step controls the finite-update error, while its divergent sum gives an unbounded learning-time horizon.

Remark 2 (Feedback and regularization). *The sampled KL penalty retains the frozen reference's influence, following Ziegler et al. (2020, Eq. (2)), while the update weights log probability gradients by advantages as in Schulman et al. (2017, Eq. (1)). It uses only each reporting user's own payoff. Appendices C.1.3–C.1.5 give the reward construction and score scaling. Algorithm 1 specifies the complete update.*

2.4 Joint Sampling and Information

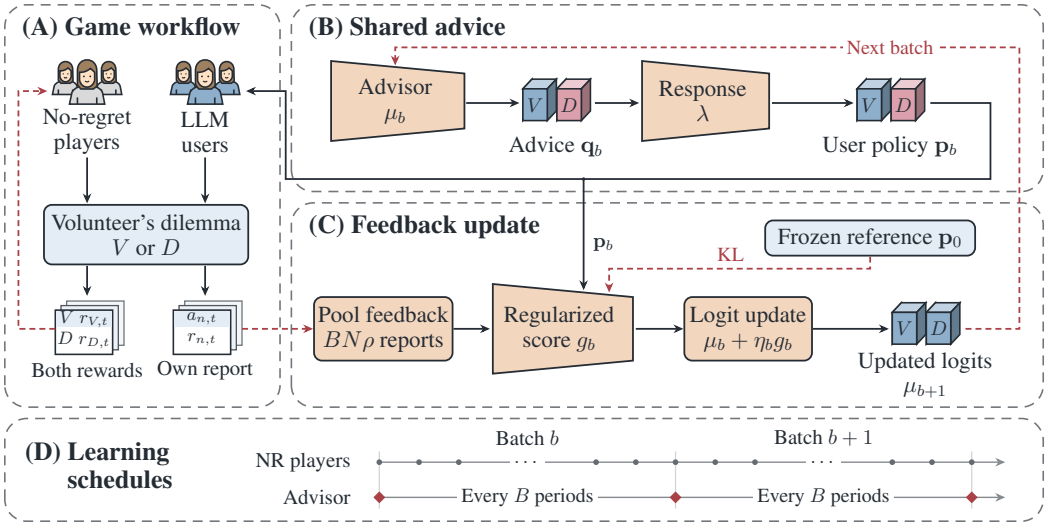

Figure 3: **Shared advisor and two learning timescales. (A)** Both populations play the volunteer's dilemma. **(B)** Sensitivity λ maps recommendations to the effective policy, fixed within a batch. **(C)** The advisor centers users' realized rewards after KL regularization toward the frozen reference. **(D)** NR players update from both action rewards each period. The advisor updates from $BN\rho$ action and payoff reports every B periods. Red dashed arrows carry feedback and updated logits.

Figure 3 shows the learning schedules. NR learning continues without restarting between batches. In local period t of batch b, batch notation $z_t, \tilde{\eta}_t$ means $z_{bB+t}, \tilde{\eta}_{bB+t}$. Appendix C.1.2 explains batching, and Appendix C.3 bounds NR tracking error.

Let $\mathcal{F}_b$ be the history, including current states, at every batch start, and let $\mathcal{H}_{b,t}$ add the first $t-1$ periods' actions and feedback, so $\mathcal{H}_{b,1} = \mathcal{F}_b$. Conditional on $\mathcal{H}_{b,t}$, players sample actions independently with fresh randomness each period. Thus the probability of no volunteer factors into individual defection probabilities. Realized payoffs remain dependent through the shared outcome.

2.5 Welfare Benchmarks and the Social Welfare Gap

Advice that aims to save users' costs may reduce others' shared benefit. We measure this externality by comparing optimal advice for users with optimal advice for everyone, after NR learning settles.

Fix $\mathcal{V} = (N, \rho, c, G)$ from Section 2.1. For fixed guided volunteering probability $p \in (0,1)$, let $\sigma^\star(p;\mathcal{V}) \coloneqq \lim_{t\to\infty} \sigma(z_t)$ be the limiting NR volunteering probability established in Lemma 1, extended continuously to $p = 0, 1$ for benchmark evaluation. By the sampling rule in Section 2.4, the probability of at least one volunteer is $P(p;\mathcal{V}) \coloneqq 1-(1-p)^{N\rho}[1-\sigma^\star(p;\mathcal{V})]^{N(1-\rho)}$. The expected payoff per LLM user and total welfare subtract contribution costs from shared benefits and exclude the training KL penalty,

$$\begin{aligned} w(p;\mathcal{V}) &\coloneqq GP(p;\mathcal{V}) - cp, \\ W^{\mathrm{tot}}(p;\mathcal{V}) &\coloneqq NGP(p;\mathcal{V}) - Nc\big[\rho p + (1-\rho)\sigma^\star(p;\mathcal{V})\big]. \end{aligned} \tag{4}$$

Definition 1 (Selfish and selfless LLM benchmarks)**.** *For game $\mathcal{V}$, fix a common policy set $\mathcal{I} \subseteq [0,1]$ on which both maximizers below exist and are unique. The **selfish** and **selfless** benchmarks optimize payoffs for users (*usr*) and society (*soc*), respectively,*

$$p^{\mathrm{usr}}_{\mathcal{I}}(\mathcal{V}) \coloneqq \arg\max_{p\in\mathcal{I}} w(p;\mathcal{V}), \qquad p^{\mathrm{soc}}_{\mathcal{I}}(\mathcal{V}) \coloneqq \arg\max_{p\in\mathcal{I}} W^{\mathrm{tot}}(p;\mathcal{V}).$$

Definition 2 (Social welfare gap)**.** *For the benchmarks in Definition 1, the **social welfare gap**, measured in units of G, is*

$$\Delta_{\mathcal{I}}(\mathcal{V}) \coloneqq \frac{W^{\mathrm{tot}}(p^{\mathrm{soc}}_{\mathcal{I}}(\mathcal{V});\mathcal{V}) - W^{\mathrm{tot}}(p^{\mathrm{usr}}_{\mathcal{I}}(\mathcal{V});\mathcal{V})}{G} \geq 0.$$

Both optima use the same game and policy set, so the gap measures the welfare cost of excluding other players' payoffs from the objective. Note that a smaller gap need *not imply* higher total welfare. Definition 3 supplies the regime domains, Section 4.1 establishes their benchmarks, and Section 4.2 analyzes the gap.

3 Game Dynamics: Convergence and Stability

3.1 Aggregate no-regret response

LLM users' rewards depend on NR player's actions. Fix $\mathcal{V} = (N, \rho, c, G)$ from Section 2.1 and batch index b. Hold $p \coloneqq p_b$ fixed in batch since LLM's fixed policy, continuing the NR score and learning rates while t counts periods from batch start. Lemma 1 identifies the limit $\sigma^\star(p;\mathcal{V})$ of $\sigma(z_t)$ (Section 2.2) needed for the advisor's mean update and welfare benchmarks. Let $p_{\mathrm{bd}}(\mathcal{V}) \coloneqq 1-(c/G)^{1/(N\rho)}$ denote the response threshold. Appendix C.2 derives the conditional rewards and separates their mean incentive from sampling noise.

Lemma 1 (Aggregate no-regret response)**.** *Under Assumptions 1 and 2, consider the NR score update* (1) *with $p \in (0,1)$ held fixed throughout the recursion and a finite initial score z_1 known at batch start. The NR volunteering probability satisfies, as $t \to \infty$,*

$$\sigma(z_t) \longrightarrow \sigma^\star(p) = \begin{cases} 1 - \left[\dfrac{c}{G(1-p)^{N\rho}}\right]^{1/[N(1-\rho)-1]}, & p < p_{\mathrm{bd}}, \\ 0, & p \geq p_{\mathrm{bd}}, \end{cases} \qquad \textit{almost surely},$$

Appendix C.2.1 proves the result. Since $p > 0$, only guided users have positive limiting volunteering probability when $\sigma^\star = 0$. Both groups do when $\sigma^\star > 0$. We extend regime labels continuously to $p = 0, 1$ for benchmark and limiting policies. Section 4.1 uses these intervals to maximize user and total welfare.

Definition 3 (Contribution regimes)**.** *For a game $\mathcal{V} = (N, \rho, c, G)$, the policy sets for the **type 1** and **type 2** regimes are $\mathcal{I}_1(\mathcal{V}) \coloneqq [p_{\mathrm{bd}}(\mathcal{V}), 1]$ and $\mathcal{I}_2(\mathcal{V}) \coloneqq [0, p_{\mathrm{bd}}(\mathcal{V}))$, respectively.*

3.2 BRANCHWISE LLM MEAN FLOW

Lemma 1 establishes the NR population's long-run response, characterizing the feedback sampled by the LLM advisor, which remains stochastic since players' sampling. Proposition 1 reduces the stochastic batch update to a mean-drift description under assumption 4, allowing the study of the LLM advisor's convergence through mean dynamics.

Proposition 1 (Large-batch mean-flow reduction)**.** *Under Assumptions 1, 3, and 4, fix the game $\mathcal{V}$, and recall the effective batch step $\gamma_b = \eta_b BN\rho$ from Section 2.3. For the branch-reduced batch update obtained by substituting the response in Lemma 1,*

$$\frac{\mathbb{E}[p_{b+1} - p_b \mid \mathcal{F}_b]}{\gamma_b} = f_R(p_b) + O\left(\frac{1}{BN\rho}\right) + O(\gamma_b),$$

where $\mathcal{F}_b$ is the history at batch start and f_R is the mean drift associated with regime $R \in \{1, 2\}$ in Definition 3. The expansion holds uniformly on compact subsets of either open branch.

The left side is the expected volunteering probability change per unit effective step. The remainders describe empirical baseline bias and error from the finite update step. Both vanish in the large-batch scaling of Assumption 4. The leading drift defines the *reduced mean dynamics* $\dot{p} = f_R(p)$, with time measured in accumulated effective batch steps. Its sign determines whether volunteering rises or falls. Appendix C.4 gives the drift, proof, and time scaling (Kushner & Yin, 2003; Borkar, 2008). Appendix C.3 separately bounds the error from replacing the adapting NR population by its limiting response.

3.3 CONVERGENCE UNDER STRONG REGULARIZATION

Welfare after learning depends on which advisor policy persists. In the reduced mean dynamics, sufficiently strong KL regularization makes the incentive to volunteer decrease with its probability, yielding a unique attracting policy. For a fixed game $\mathcal{V}$ and response sensitivity λ, define $\beta_c := \lambda c(1 + \frac{N\rho}{N(1-\rho)-1})[(G/c)^{1/(N\rho)} - 1]$. For $\beta > 0$, let q_{cut} be the unique pretrained recommendation probability for which the reduced mean drift vanishes at p_{bd}. Appendix C.5 derives this cutoff and the critical level β_c.

Theorem 3.1 (Convergence of the reduced mean dynamics)**.** *Suppose Assumption 1 holds, $\beta \geq \beta_c$, and $0 < q_0 < 1$. The reduced mean dynamics in Section 3.2 are continuous across p_{bd} and converge from any initial probability in $(0, 1)$ to a unique attracting fixed point $p^\star \in (0, 1)$. The limit lies in type 2 if $q_0 < q_{\mathrm{cut}}$, equals p_{bd} if $q_0 = q_{\mathrm{cut}}$, and lies in the interior of type 1 if $q_0 > q_{\mathrm{cut}}$.*

Appendix C.5 proves this result through the geometry of the expected reward gaps. Theorem D.1 in Appendix D.1 analyzes weak regularization. The welfare analysis below uses these reduced mean limits to select references that sustain its benchmark policies.

4 WELFARE RESULTS

4.1 BRANCH WELFARE AND BENCHMARK POLICIES

For fixed $\mathcal{V}$, subscript $R \in \{1, 2\}$ restricts user and total welfare w_R, W_R^{tot} in (4) to regime $\mathcal{I}_R(\mathcal{V})$ (Definition 3). Their optima $p_R^{\mathrm{usr}}, p_R^{\mathrm{soc}}$ follow Definition 1, suppressing the fixed game argument. Substituting Lemma 1's response yields the benchmarks below.

Lemma 2 (Welfare benchmarks and attainability)**.** *Suppose Assumption 1 holds.*

- *(i)* ***Type 1.*** *Both w_1 and W_1^{tot} are strictly concave on $\mathcal{I}_1$, with unique maximizers $p_1^{\mathrm{usr}} = 1 - (c/(GN\rho))^{1/(N\rho-1)}$ and $p_1^{\mathrm{soc}} = 1 - (c/(NG))^{1/(N\rho-1)} > p_1^{\mathrm{usr}}$.*
- *(ii)* ***Type 2.*** *Both w_2 and W_2^{tot} are strictly decreasing on $\mathcal{I}_2$, with the common unique maximizer $p_2^{\mathrm{usr}} = p_2^{\mathrm{soc}} = 0$.*

For $\beta \geq \beta_c$, separately chosen pretrained references in $(q_{\mathrm{cut}}, 1)$ sustain each type 1 optimum as the unique attracting reduced mean limit. In type 2, this limit approaches zero as $q_0 \downarrow 0$ through positive references, and both welfare values converge to their benchmarks.

Appendices B.1 and B.2 prove these welfare claims. Note that the type 1 user optimum satisfies $N\rho G(1-p)^{N\rho-1} = c$, whereas the social optimum satisfies $NG(1-p)^{N\rho-1} = c$. The social

objective also counts benefits to the $N(1-\rho)$ other players and therefore calls for more volunteering. Increasing adoption brings more of these beneficiaries into the user objective, but also changes both optimal policies.

The type 2 gap vanishes even though other players remain outside the advisor's objective. Their limiting mixed response equalizes the two action payoffs at $G - c$, independently of the guided volunteering probability within this regime. Excluding these constant payoffs therefore leaves the optimal advice unchanged. In type 1, the other players stop volunteering, but their payoffs increase with guided users' contributions. The selfish objective omits this additional benefit, producing distinct optima and a positive gap (Appendix B.2).

4.2 Wider adoption and the social welfare gap

For type 1, write $\Delta(c/G, \rho)$ for $\Delta_{\mathcal{I}_1(\mathcal{V})}(\mathcal{V})$ from Definition 2. We fix N throughout this comparison. Normalization by G leaves dependence only on c/G and ρ. Both policies are reoptimized whenever $\mathcal{V} = (N, \rho, c, G)$ changes and evaluated in that same game. Adoption comparisons vary ρ with c, G fixed. Actual group sizes are integers, while adoption derivatives use the continuous extension of the welfare formulas.

The type 1 gap can persist even at the user optimum. The following theorem quantifies how adoption and contribution costs change this gap (Figure 2(a)).

Theorem 4.1 (Social welfare gap)**.** *Under Assumption 1,*

(i) For fixed N and ρ, Δ is strictly increasing and strictly convex in c/G.
(ii) For fixed N and c/G, Δ is single-peaked in ρ, with a unique maximizing penetration rate $\rho^\star \in [2/N, 1 - 2/N]$.

An interior maximum separates a widening gap from a narrowing one as adoption grows. The maximum may also occur at the left boundary due to the integer population. The next corollary relates its height and location to contribution costs.

Corollary 4.1 (Peak height and peak location)**.** *Under the hypotheses of Theorem 4.1, fix N and G. Over any range of c/G with an interior maximizer $\rho^\star(c/G)$, increasing c/G strictly raises $\Delta(c/G, \rho^\star(c/G))$ and strictly lowers $\rho^\star(c/G)$.*

Thus more severe welfare losses can peak while the advisor still serves a smaller share of the population. Appendices B.3 and B.4 prove the theorem and corollary. We next bound the remaining gap at every admissible penetration rate.

Corollary 4.2 (Adoption bound for the social welfare gap)**.** *Under the hypotheses of Theorem 4.1, every admissible penetration rate satisfies $\Delta(c/G, \rho) \le (c/G)[c/(GN\rho)]^{1/(N\rho-1)}(1-\rho)^2/\rho \le (c/G)(1-\rho)^2/\rho$. For any $\epsilon > 0$, an admissible penetration rate satisfying $\rho > 2/\big(2 + \epsilon G/c + \sqrt{(\epsilon G/c)^2 + 4\epsilon G/c}\big)$ has $\Delta(c/G, \rho) < \epsilon$.*

The bound gives sufficient adoption for a target gap ϵ. Appendix B.5 proves it, and Table 2 summarizes the welfare results.

Remark 3 (Total welfare)**.** *Lemma 2 also characterizes total welfare in both regimes. Appendices B.2 and B.3 give its values at the user and social benchmarks (Figure 4(b)). For $N = 200$, $G = 1$, and $c = 0.18$, both levels fall as ρ increases from 0.90 to 0.99, while their gap shrinks.*

Extension across contribution regimes. For fixed N, define $\Gamma(c/G, \rho) := [w_2(p_2^{\text{usr}}) - w_1(p_1^{\text{usr}})]/G$, evaluating both user optima from Lemma 2 in the same game $\mathcal{V} = (N, \rho, c, G)$. Unlike the total welfare gap Δ, Γ compares users' best payoffs across regimes which captures the policy preference. Their preferred regime can reverse twice as adoption grows (Figure 4(a)). Theorem B.1 in Appendix B.6 characterizes this dependence on adoption and contribution costs.

5 Experiments

5.1 Final Qwen policies and theoretical predictions

We train Qwen3-8B (Yang et al., 2025) at $N = 100$, $G = 1$, $c = 0.90$, $\beta = 0.35$, and $\lambda = 1$. Five seeds at each $\rho \in \{0.02, 0.05, 0.10, 0.30, 0.90\}$ give 25 runs, each evaluated using 2,048 fresh

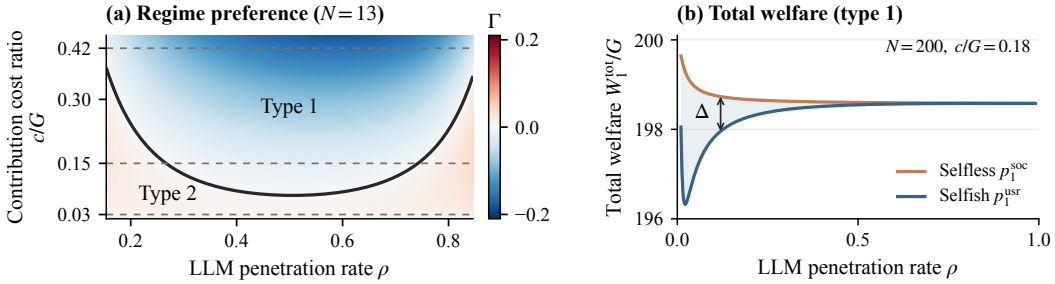

Figure 4: **Two welfare comparisons.** (a) User regime preference for $N = 13$. Blue favors type 1 ($\Gamma < 0$), red type 2 ($\Gamma > 0$), and the boundary marks equal user payoffs. (b) Type 1 total welfare at the selfish and selfless benchmarks for $N = 200$ and $c/G = 0.18$. Their vertical separation is the social welfare gap Δ.

frozen policy generations. Supervised fine tuning (SFT) on example responses provides the shared initialization and frozen reference ($q_0 = 0.6860$). Proximal policy optimization (PPO) then learns from users' game payoffs while penalizing deviation from this reference. Appendix E.1 details training and the action mapping, including uniform random fallback for invalid outputs.

Figure 5 shows all estimates and means with 95% Student-t intervals. The mean absolute error against theory is 2.23 percentage points, with 24 of 25 runs within five points. Mean offsets remain 2.48 and 2.79 points above theory at $\rho = 0.30$ and 0.90. Appendix E.1 provides the full training trajectories, numerical comparisons, and late training diagnostics.

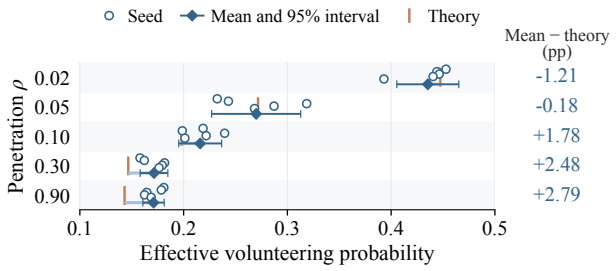

Figure 5: **Final Qwen policies.** All 25 runs appear as circles. Diamonds show means with 95% **Student** t **intervals** across five seeds. Ticks mark attracting probabilities of the reduced mean dynamics (Section 3.2). The pretrained reference stays fixed across penetration rates. Data from Figure 7(f) in Appendix E.1.

5.2 Social welfare gap in Qwen experiments

Figure 2(b) compares 91 pairs approximating the type 1 benchmarks at $N = 200$, $G = 1$, $c = 0.18$, and $\beta = 0.35$, across 19 penetration rates from 0.01 to 0.55. Both arms use individual payoff PPO feedback with different calibrated SFT references targeting the two optima. The social reference arm receives no total welfare reward. Each point subtracts user arm total welfare from social arm total welfare, normalized by G, and bars show means $\pm$ one sample standard deviation.

Mean gaps generally decrease with penetration but retain substantial variation. Training variation, imperfect references, and finite NR learning make this exploratory evidence, not a test of asymptotic convergence or strict unimodality. Appendix E.3 gives training, evaluation, and data provenance. Appendix E.2 (Figure 8) separately isolates welfare sampling error. Code, experiment configurations, and numerical figure inputs are available in the anonymous repository.

6 Discussion and Conclusion

In the volunteer's dilemma, selfish advice optimizes its users' payoffs without accounting for benefits to other players. As more players become users, the advisor also takes their benefits into account. A large welfare gap relative to selfless advice therefore need not persist as adoption grows. Two extensions may strengthen our current results. First, heterogeneous players and competing advisors could test this insight beyond a symmetric game with one advisor. Second, it is worth exploring how users join, leave, or switch advisors during learning, especially whether temporary losses discourage participation before adoption reaches levels where the welfare gap narrows.

AI USE STATEMENT

Qwen was the model trained and evaluated in our experiments. Separately, generative AI tools assisted the authors with mathematical consistency checks, paper polishing, and preparation of figures and associated code. The authors take responsibility for the final text, mathematical claims, experimental results, and accompanying artifacts.

ETHICS STATEMENT

This work studies a mathematical model and computational experiments using simulated game interactions and generated actions. The experiments do not involve collecting human participant data. The welfare conclusions are interpreted within the stated model assumptions.

REPRODUCIBILITY STATEMENT

The anonymous repository provides code, experiment configurations, and numerical figure inputs. Section 2 specifies the game, learning rules, and assumptions. Proofs and supporting analyses appear in Appendices B, C, and D. Appendix E describes experimental configurations, evaluation procedures, and uncertainty estimates. Appendix E.3 documents the provenance and recovery precision of the archived welfare observations, including the unavailability of their original response and training logs.

REFERENCES

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. *Nature Human Behaviour*, 9:1380–1390, 2025. doi: 10.1038/s41562-025-02172-y. URL https://www.nature.com/articles/s41562-025-02172-y.

Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: A meta-algorithm and applications. *Theory of Computing*, 8(6):121–164, 2012. doi: 10.4086/toc.2012.v008a006. URL https://theoryofcomputing.org/articles/v008a006/.

Michel Benaim. Dynamics of stochastic approximation algorithms. In *Seminaire de Probabilites XXXIII*, volume 1709 of *Lecture Notes in Mathematics*, pp. 1–68. Springer, 1999. doi: 10.1007/BFb0096509.

Alexander Bick, Adam Blandin, and David J. Deming. The rapid adoption of generative AI. *Management Science*, 2026. doi: 10.1287/mnsc.2025.02523. URL https://pubsonline.informs.org/doi/10.1287/mnsc.2025.02523. Published online January 20, 2026.

Jessica Y. Bo, Sophia Wan, and Ashton Anderson. To rely or not to rely? evaluating interventions for appropriate reliance on large language models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2025. doi: 10.1145/3706598.3714097. URL https://doi.org/10.1145/3706598.3714097.

Rishi Bommasani, Kathleen A. Creel, Ananya Kumar, Dan Jurafsky, and Percy Liang. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? In *Advances in Neural Information Processing Systems*, volume 35, pp. 3663–3678, 2022. doi: 10.52202/068431-0265. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/17a234c91f746d9625a75cf8a8731ee2-Abstract-Conference.html.

Vivek S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. Generative AI at work. *The Quarterly Journal of Economics*, 140(2):889–942, 2025. doi: 10.1093/qje/qjae044. URL https://academic.oup.com/qje/article/140/2/889/7990658.

Shengyu Cao and Ming Hu. Supracompetitive pricing under AI monoculture, 2026. URL https://arxiv.org/abs/2601.01279v3. Version 3.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006. doi: 10.1017/CBO9780511546921. URL https://doi.org/10.1017/CBO9780511546921.

Aaron Chatterji, Thomas Cunningham, David J. Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. How people use ChatGPT. Working Paper 34255, National Bureau of Economic Research, September 2025. URL https://www.nber.org/papers/w34255.

Fabrizio Dell'Acqua, Charles Ayoubi, Hila Lifshitz, Raffaella Sadun, Ethan Mollick, Lilach Mollick, Yi Han, Jeff Goldman, Hari Nair, Stew Taub, and Karim R. Lakhani. The cybernetic teammate: A field experiment on generative AI and teamwork. *Organization Science*, 37(4):1217–1242, 2026. doi: 10.1287/orsc.2025.20702. URL https://pubsonline.informs.org/doi/10.1287/orsc.2025.20702.

Andreas Diekmann. Volunteer's dilemma. *Journal of Conflict Resolution*, 29(4):605–610, 1985. doi: 10.1177/0022002785029004003. URL https://doi.org/10.1177/0022002785029004003.

Seyed A. Esmaeili, Kevin Lim, Kshipra Bhawalkar, Zhe Feng, Di Wang, and Haifeng Xu. How to strategize human content creation in the era of GenAI? In *Proceedings of the ACM Web Conference 2026*, pp. 285–296. ACM, 2026. doi: 10.1145/3774904.3792475. URL https://doi.org/10.1145/3774904.3792475.

Kanishk Gandhi, Dorsa Sadigh, and Noah D. Goodman. Strategic reasoning with language models, 2023. URL https://arxiv.org/abs/2305.19165.

Matthew J. Grawitch, Steven L. Winton, and Srikanth P. Mudigonda. Who listens to ChatGPT and when should they? a two-study examination of AI-assisted decision making. *Computers in Human Behavior: Artificial Humans*, 7:100283, 2026. doi: 10.1016/j.chbah.2026.100283. URL https://doi.org/10.1016/j.chbah.2026.100283.

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. Way off-policy batch deep reinforcement learning of human preferences in dialog, 2020. URL https://openreview.net/forum?id=rJl5rRVFvH.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond). In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. doi: 10.52202/085713-2732. URL https://openreview.net/forum?id=saDOrrnNTz. Datasets and Benchmarks Track.

Elliot Myunghoon Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 30038–30066. PMLR, 2025. URL https://proceedings.mlr.press/v267/kim25e.html.

Jon Kleinberg and Manish Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, 2021. doi: 10.1073/pnas.2018340118. URL https://doi.org/10.1073/pnas.2018340118.

Harold J. Kushner and G. George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2 edition, 2003. doi: 10.1007/b97441.

Jiacai Liu, Chaojie Wang, Chris Yuhao Liu, Liang Zeng, Rui Yan, Yiwen Sun, Yang Liu, and Yahui Zhou. Improving multi-step reasoning abilities of large language models with direct advantage policy optimization, 2024. URL https://arxiv.org/abs/2412.18279.

Nunzio Lorè and Babak Heydari. Strategic behavior of large language models and the role of game structure versus contextual framing. *Scientific Reports*, 14:18490, 2024. doi: 10.1038/s41598-024-69032-z. URL https://www.nature.com/articles/s41598-024-69032-z.

OECD. *Generative AI and the SME Workforce: New Survey Evidence*. OECD Publishing, Paris, 2025. doi: 10.1787/2d08b99d-en. URL https://www.oecd.org/en/publications/generative-ai-and-the-sme-workforce_2d08b99d-en.html.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Thomas R. Palfrey and Howard Rosenthal. Participation and the provision of discrete public goods: A strategic analysis. *Journal of Public Economics*, 24(2):171–193, 1984. doi: 10.1016/0047-2727(84)90023-9. URL https://doi.org/10.1016/0047-2727(84)90023-9.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior, 2023. URL https://arxiv.org/abs/2304.03442.

Manish Raghavan. Competition and diversity in generative AI, 2024. URL https://arxiv.org/abs/2412.08610v3. Version 3, revised June 11, 2026.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. doi: 10.1214/aoms/1177729586.

Isaac Robinson and John Burden. Framing the game: How context shapes LLM decision-making, 2025. URL https://arxiv.org/abs/2503.04840.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback, 2020. URL https://arxiv.org/abs/2009.01325.

Haoran Sun, Yusen Wu, Yukun Cheng, and Xu Chu. Game theory meets large language models: A systematic survey. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 10669–10677, 2025. doi: 10.24963/ijcai.2025/1184. URL https://www.ijcai.org/proceedings/2025/1184.

Boaz Taitler and Omer Ben-Porat. Braess's paradox of generative AI. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(13):14139–14147, 2025a. doi: 10.1609/aaai.v39i13.33548. URL https://ojs.aaai.org/index.php/AAAI/article/view/33548.

Boaz Taitler and Omer Ben-Porat. Selective response strategies for GenAI. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 58189–58226. PMLR, 2025b. URL https://proceedings.mlr.press/v267/taitler25a.html.

Lex Weaver and Nigel Tao. The optimal reward baseline for gradient-based reinforcement learning, 2013. URL https://arxiv.org/abs/1301.2315.

Kaixuan Xu, Jiajun Chai, Sicheng Li, Yuqian Fu, Yuanheng Zhu, and Dongbin Zhao. DipLLM: Fine-tuning LLM for strategic decision-making in diplomacy. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 69078–69095. PMLR, 2025. URL https://proceedings.mlr.press/v267/xu25b.html.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 Technical Report, 2025. URL https://arxiv.org/abs/2505.09388.

Fan Yao, Chuanhao Li, Denis Nekipelov, Hongning Wang, and Haifeng Xu. Human vs. generative AI in content creation competition: Symbiosis or conflict? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 56885–56913. PMLR, 2024. URL https://proceedings.mlr.press/v235/yao24b.html.

Yaolong Yu, Fan Yao, and Sinno Jialin Pan. Beyond self-interest: How group strategies reshape content creation in recommendation platforms? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 73326–73349. PMLR, 2025. URL https://proceedings.mlr.press/v267/yu25t.html.

Qinlin Zhao, Jindong Wang, Yixuan Zhang, Yiqiao Jin, Kaijie Zhu, Hao Chen, and Xing Xie. CompeteAI: Understanding the competition dynamics of large language model-based agents. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 61092–61107. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhao24q.html.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020. URL https://arxiv.org/abs/1909.08593.

APPENDIX GUIDE

PAPER OUTLINE

Main paper

1 Introduction
2 Game and learning model
3 Convergence and stability
4 Welfare results
5 Experiments
6 Discussion and conclusion

Appendix

A Related work
B Welfare proofs and extension
C Dynamic reduction and finite batch analysis
D Additional dynamic cases
E Experimental details

OVERVIEW OF THE ANALYSIS

Learning foundation. Lemma 1 gives the other players' limiting response under fixed advice. Proposition 1 uses this response to obtain the advisor's conditional mean update, and Theorem 3.1 establishes convergence of the reduced mean dynamics under strong regularization. Appendix C also bounds the error from collecting feedback before the other players reach their limit. Appendix D treats weaker or absent regularization and the symmetric Nash benchmark.

Policy benchmarks. Lemma 2 identifies the user and total welfare optima over the same policy sets. Under strong regularization, separately chosen references sustain the two type 1 optima. In type 2, both objectives select the boundary policy with zero user volunteering, approached as the reference tends to zero. Appendix B proves these results.

Adoption and welfare. Theorem 4.1 and Corollaries 4.1 and 4.2 characterize the social welfare gap and its dependence on adoption and costs. This compares the two objectives at the same adoption level. The extension in Theorem B.1 instead compares users' best payoffs across regimes. The corresponding proofs and detailed summary tables appear in Appendix B.

Experimental evidence. Appendix E records the Qwen training and evaluation protocols, individual outcomes, and fixed policy simulations. These experiments distinguish agreement in action probabilities from agreement in welfare differences. They provide finite training comparisons with the theory. Appendix A places the work in the literature.

A Related Work

A.1 Empirical studies of LLMs in games

Experiments with LLMs in games examine cooperation, coordination, and context-sensitive decisions, as reviewed by Sun et al. (2025). In repeated games, Akata et al. (2025) find systematic differences across models and games, including coordination failures and changes from cooperation to unforgiving behavior. Prompting for social reasoning can improve repeated interaction, indicating that cooperation also depends on how the task is presented. Lorè & Heydari (2024) compare sensitivity to game structure and contextual frames. Holding the strategic game fixed, Robinson & Burden (2025) document substantial changes in choices under different descriptions. Together, these studies connect strategic behavior to both payoff incentives and the context in which a model encounters them.

Interactive simulations connect individual decisions to population-level patterns. Park et al. (2023) study emergent social coordination supported by memory, reflection, and planning. In CompeteAI, Zhao et al. (2024) examine restaurant agents competing for customer agents in a virtual environment. Strategic-agent work also develops reasoning and training methods: Gandhi et al. (2023) use language-based demonstrations to support reasoning about states, values, and beliefs in tasks with hidden information and negotiation, while Xu et al. (2025) fine-tune DipLLM for multiplayer decisions in Diplomacy using an equilibrium-policy learning objective. Section 5.1 compares final Qwen3-8B action probabilities with the structured-model predictions. Appendix E.1 provides the full training trajectories.

A.2 Theoretical studies of GenAI competition and welfare

GenAI competition models distinguish platform revenue, creator utility, and welfare. Taitler & Ben-Porat (2025a) show how a revenue-maximizing GenAI platform can reduce user welfare by drawing participation away from a human forum and weakening its network benefits. Taitler & Ben-Porat (2025b) study selective answering: demand retained by the forum generates data that improve future AI quality, linking response choices to long-run revenue and welfare constraints. Yao et al. (2024) model multiple creators competing for traffic across topics, including an inclusive setting where creators strategically choose GenAI use. Esmaeili et al. (2026) optimize a human creator's production over time and topics against a GenAI competitor whose future capabilities improve through human output. These models identify how present participation and production affect later competitive opportunities.

Competition and group objectives also shape population welfare. Raghavan (2024) studies equilibrium diversity, social welfare, and the price of anarchy in a multiplayer model of AI-assisted content production. Competition rewards differentiation, and a tool's relative performance depends on the competitive environment. Yu et al. (2025) study creators who coordinate to maximize group utility on recommendation platforms. Group size changes equilibrium content allocation, while effects on consumer welfare depend on the reward mechanism. Both analyses connect population structure and strategic objectives to outcomes beyond individual content choices.

When recommendations follow a common policy, shared updates can change decisions across a strategic group. The theory of Kleinberg & Raghavan (2021) shows that individually attractive algorithm use can worsen collective selection outcomes. Bommasani et al. (2022) study homogenization from shared components and its dependence on model adaptation. Empirically, Jiang et al. (2025) document within-model repetition and cross-model homogeneity on open-ended tasks, while Kim et al. (2025) examine correlated errors and downstream consequences. Reference-regularized feedback fine-tuning provides the update background: Ziegler et al. (2020) and Ouyang et al. (2022) adapt existing checkpoints with a KL penalty relative to a reference policy. Algorithm 1 uses these ingredients at the decision layer with user-payoff feedback.

Cao & Hu (2026) provide the closest shared-policy comparison: performance feedback updates a common advisor in a pricing duopoly, with output fidelity capturing adherence and generation variation. They analyze long-run dynamics and finite-batch selection. Their recommendations condition on a shared latent mode; our actions are independent conditional on the pre-period history. Our finite mixed population combines guided users with an aggregate no-regret response, and the LLM penetration rate ρ varies the population served. Non-users respond to advice-mediated play and receive public-good benefits omitted from the guided-user objective. Changing the penetration rate

therefore changes both the scope of the user objective and the division of provision costs through the responding non-user population. Our main result quantifies the resulting social welfare gap Δ between the best policies for users and for society within a fixed contribution regime (Theorem 4.1). An extension studies the difference Γ in the best payoff per guided user across regimes (Theorem B.1).

B Welfare Proofs and Extension

Table 2: The social welfare gap and its dependence on adoption and cost.

Setting	**Conclusion**	**Result**
Type 1	$p_1^{\mathrm{usr}} < p_1^{\mathrm{soc}}, \quad \Delta > 0$	Lemma 2
Type 2	$p_2^{\mathrm{usr}} = p_2^{\mathrm{soc}} = 0, \quad \Delta_{\mathcal{I}_2} = 0$	Lemma 2
Adoption ρ Fixed $N, c/G$	Unique peak $\rho^\star$, decreasing afterward	Theorem 4.1
Cost ratio c/G Fixed N, ρ	Δ strictly increasing and strictly convex	Theorem 4.1
Interior peak $\rho^\star$ Fixed N, G	Height increases, location decreases with c/G	Corollary 4.1
Adoption bound	$\Delta \le \frac{c}{G}\frac{(1-\rho)^2}{\rho}$	Corollary 4.2

The last four rows concern type 1. The peak $\rho^\star$ may be at an endpoint. Attainability uses $\beta \ge \beta_{\mathrm{c}}$ and separately chosen references, with the type 2 optimum approached as $q_0 \downarrow 0$.

B.1 User welfare and attainability in Lemma 2

For the regime comparison in Appendix B.6, define the best payoff per guided user by $w_R^{\max}(\mathcal{V}) := \max_{p \in \mathcal{I}_R(\mathcal{V})} w_R(p; \mathcal{V})$, for $R \in \{1, 2\}$. We suppress the fixed game argument below. For $\beta \ge \beta_{\mathrm{c}}$, write $p^\star(q_0)$ for the attracting reduced mean limit in Theorem 3.1, with $\mathcal{V}, \lambda, \beta$ fixed.

Proof. For guided and NR volunteering probabilities p and σ, the payoff per guided user can be written as

$$w_{\mathrm{LLM}}(p, \sigma) := p(G - c) + (1 - p)\left(1 - (1 - p)^{N\rho - 1}(1 - \sigma)^{N(1-\rho)}\right) G. \tag{5}$$

Evaluating at $\sigma = \sigma^\star(p)$ recovers $w(p)$ in (4). By Theorem 3.1, when $\beta \ge \beta_{\mathrm{c}}$ the mean flow converges to the type 1 regime, with $\sigma^\star = 0$, or to the type 2 regime, where

$$\sigma^\star = 1 - \left(\frac{c}{G(1 - p^\star)^{N\rho}}\right)^{\frac{1}{N(1-\rho)-1}}.$$

We analyze the two branches separately.

For the type 1 regime, substituting $\sigma^\star = 0$ gives

$$w_1(p) = p(G - c) + (1 - p)\left(1 - (1 - p)^{N\rho - 1}\right) G = G - cp - G(1 - p)^{N\rho},$$

for $p \in [p_{\mathrm{bd}}, 1]$. Differentiating for $p < 1$ yields

$$w_1'(p) = -c + GN\rho(1 - p)^{N\rho - 1}, \qquad w_1''(p) = -GN\rho(N\rho - 1)(1 - p)^{N\rho - 2} < 0.$$

Hence w_1 is strictly concave, so any critical point is its unique global maximizer. Solving $w_1'(p) = 0$ yields $p_1^{\mathrm{usr}} = 1 - (c/(GN\rho))^{1/(N\rho - 1)}$. Moreover,

$$w_1'(p_{\mathrm{bd}}) = c\left[N\rho\left(\frac{G}{c}\right)^{\frac{1}{N\rho}} - 1\right] > 0, \qquad w_1'(1) = -c < 0,$$

so indeed $p_1^{\mathrm{usr}} \in (p_{\mathrm{bd}}, 1)$.

Let q_1^{usr} denote the unique model-policy volunteering probability satisfying

$$\log \frac{q_1^{\mathrm{usr}}}{1-q_1^{\mathrm{usr}}} = \frac{1}{\lambda} \log \frac{p_1^{\mathrm{usr}}}{1-p_1^{\mathrm{usr}}}.$$

Since $w_1'(p_1^{\mathrm{usr}}) = 0$, we have $G(1-p_1^{\mathrm{usr}})^{N\rho-1} = c/(N\rho)$. Substituting this identity into the type 1 drift gives

$$\Psi_1(p_1^{\mathrm{usr}}) = -c\left(1 - \frac{1}{N\rho}\right) - \beta\left(\log \frac{q_1^{\mathrm{usr}}}{1-q_1^{\mathrm{usr}}} - \log \frac{q_0}{1-q_0}\right).$$

Therefore, if $q_0 = q_{0,1}^{\mathrm{usr}}$ is chosen so that

$$\log \frac{q_{0,1}^{\mathrm{usr}}}{1-q_{0,1}^{\mathrm{usr}}} = \log \frac{q_1^{\mathrm{usr}}}{1-q_1^{\mathrm{usr}}} + \frac{c}{\beta}\left(1 - \frac{1}{N\rho}\right),$$

then $\Psi_1(p_1^{\mathrm{usr}}) = 0$. Since Ψ_1 is strictly decreasing by Proposition 2, this makes p_1^{usr} the unique fixed point in the type 1 regime.

To verify feasibility, note that the threshold in Theorem 3.1 satisfies

$$\log \frac{q_{\mathrm{cut}}}{1-q_{\mathrm{cut}}} = \frac{1}{\lambda} \log \frac{p_{\mathrm{bd}}}{1-p_{\mathrm{bd}}} - \frac{c}{\beta}\left(\left(\frac{G}{c}\right)^{\frac{1}{N\rho}} - 1\right).$$

Hence

$$\log \frac{q_{0,1}^{\mathrm{usr}}}{1-q_{0,1}^{\mathrm{usr}}} - \log \frac{q_{\mathrm{cut}}}{1-q_{\mathrm{cut}}} = \frac{1}{\lambda}\left(\log \frac{p_1^{\mathrm{usr}}}{1-p_1^{\mathrm{usr}}} - \log \frac{p_{\mathrm{bd}}}{1-p_{\mathrm{bd}}}\right) + \frac{c}{\beta}\left(\left(\frac{c}{G}\right)^{-\frac{1}{N\rho}} - \frac{1}{N\rho}\right).$$

The first term is positive because $p_1^{\mathrm{usr}} > p_{\mathrm{bd}}$, and the second term is also positive because $(c/G)^{-1/(N\rho)} > 1 > 1/(N\rho)$. Thus $q_{0,1}^{\mathrm{usr}} > q_{\mathrm{cut}}$. Theorem 3.1 then guarantees convergence to the type 1 regime, and the uniqueness of the fixed point implies convergence to $\left(p_1^{\mathrm{usr}}, 0\right)$.

Evaluating w_1 at the maximizer yields

$$w_1(p_1^{\mathrm{usr}}) = G - c + c\left(1 - \frac{1}{N\rho}\right)\left(\frac{c}{GN\rho}\right)^{\frac{1}{N\rho-1}}.$$

This gives $w_1^{\max}$.

For the type 2 regime, Lemma 1 gives $1-\sigma^\star = [c/(G(1-p)^{N\rho})]^{1/[N(1-\rho)-1]}$. Substituting this identity into (5) yields

$$\begin{aligned} w_2(p) &= p(G-c) + (1-p)\left(1 - (1-p)^{N\rho-1}(1-\sigma^\star)^{N(1-\rho)}\right) G \\ &= G\left[1 - \frac{c}{G}p - \left(\frac{c}{G}\right)^{\frac{N(1-\rho)}{N(1-\rho)-1}} (1-p)^{-\frac{N\rho}{N(1-\rho)-1}}\right], \end{aligned}$$

for $p \in [0, p_{\mathrm{bd}})$. Differentiating gives

$$w_2'(p) = -c - G\frac{N\rho}{N(1-\rho)-1}\left(\frac{c}{G}\right)^{\frac{N(1-\rho)}{N(1-\rho)-1}} (1-p)^{-\frac{N\rho}{N(1-\rho)-1}-1} < 0.$$

Hence w_2 is strictly decreasing on $[0, p_{\mathrm{bd}})$, and therefore

$$\max_{0 \le p < p_{\mathrm{bd}}} w_2(p) = w_2(0) = G\left[1 - \left(\frac{c}{G}\right)^{\frac{N(1-\rho)}{N(1-\rho)-1}}\right].$$

This gives $w_2^{\max}$.

To relate the boundary benchmark to the mean dynamics, hold N, ρ, G, c, λ, and $\beta \ge \beta_{\mathrm{c}}$ fixed and let $q_0 \downarrow 0$ through positive references. For any fixed $\varepsilon \in (0, p_{\mathrm{bd}})$, the only reference-dependent term in $\Psi_2(\varepsilon)$ is $\beta \log[q_0/(1-q_0)]$, which tends to $-\infty$. For each positive reference, $\Psi_2(0+) = +\infty$ and Ψ_2 is strictly decreasing on the type 2 interval. Hence, for all sufficiently small q_0, its unique zero satisfies $0 < p^\star < \varepsilon$. By Theorem 3.1, this zero is the attracting limit from the corresponding pretrained initialization. Since ε is arbitrary, $p^\star \to 0$ as $q_0 \downarrow 0$. Continuity of w_2 at zero gives $w_2(p^\star) \to w_2(0)$. □

B.2 Total welfare and attainability in Lemma 2

Proof. In the type 1 regime,

$$W_1^{\mathrm{tot}}(p) = NG\left(1-(1-p)^{N\rho}\right) - N\rho cp.$$

Differentiating gives

$$\frac{d}{dp}W_1^{\mathrm{tot}}(p) = N\rho\left(NG(1-p)^{N\rho-1} - c\right)$$

and

$$\frac{d^2}{dp^2}W_1^{\mathrm{tot}}(p) = -N^2\rho G(N\rho-1)(1-p)^{N\rho-2} < 0, \qquad p<1,$$

where the strict inequality uses Assumption 1. Hence W_1^{tot} is strictly concave. Its unique critical point solves $NG(1-p)^{N\rho-1} = c$, so

$$p_1^{\mathrm{soc}} = 1 - \left(\frac{c}{NG}\right)^{\frac{1}{N\rho-1}}.$$

It remains to check that this point lies in the type 1 regime. Write $x \coloneqq c/G \in (0,1)$ and $m \coloneqq N\rho > 1$. Then

$$\left(\frac{c}{NG}\right)^{\frac{1}{N\rho-1}} < \left(\frac{c}{G}\right)^{\frac{1}{N\rho}}$$

is equivalent, after raising both sides to the positive power $m(m-1)$, to $\left(\frac{x}{N}\right)^m < x^{m-1}$, or $x < N^m$, which is immediate. Thus $p_1^{\mathrm{soc}} > p_{\mathrm{bd}}$. Also, since $\rho < 1$, $\frac{c}{NG} < \frac{c}{GN\rho}$, and therefore

$$p_1^{\mathrm{soc}} = 1 - \left(\frac{c}{NG}\right)^{\frac{1}{N\rho-1}} > 1 - \left(\frac{c}{GN\rho}\right)^{\frac{1}{N\rho-1}} = p_1^{\mathrm{usr}}.$$

At the social optimum, $G(1-p_1^{\mathrm{soc}})^{N\rho-1} = c/N$. Choose the pretrained reference $q_{0,1}^{\mathrm{soc}}$ by

$$\log\frac{q_{0,1}^{\mathrm{soc}}}{1-q_{0,1}^{\mathrm{soc}}} = \frac{1}{\lambda}\log\frac{p_1^{\mathrm{soc}}}{1-p_1^{\mathrm{soc}}} + \frac{c}{\beta}\left(1-\frac{1}{N}\right).$$

The right-hand side is finite, so this specifies a reference in $(0,1)$. Substitution into Ψ_1 gives $\Psi_1(p_1^{\mathrm{soc}}) = 0$. Since Ψ_1 is strictly decreasing and $p_1^{\mathrm{soc}} > p_{\mathrm{bd}}$, we have $\Psi_1(p_{\mathrm{bd}}) > 0$. The reference-cutoff relation therefore implies $q_{0,1}^{\mathrm{soc}} > q_{\mathrm{cut}}$. Theorem 3.1 makes p_1^{soc} the unique attracting mean-flow limit under this reference and its corresponding pretrained initialization.

Substituting the first-order condition into the welfare profile gives

$$\begin{aligned} W_1^{\mathrm{tot}}(p_1^{\mathrm{soc}}) &= NG\left[1-\left(\frac{c}{NG}\right)^{\frac{N\rho}{N\rho-1}}\right] - N\rho c\left[1-\left(\frac{c}{NG}\right)^{\frac{1}{N\rho-1}}\right] \\ &= NG - N\rho c + c(N\rho-1)\left(\frac{c}{NG}\right)^{\frac{1}{N\rho-1}}. \end{aligned}$$

In the type 2 regime,

$$W_2^{\mathrm{tot}}(p) = N\rho\, w_2(p) + N(1-\rho)(G-c).$$

Appendix B.1 shows that w_2 is strictly decreasing on $[0, p_{\mathrm{bd}})$. Hence W_2^{tot} is also strictly decreasing on this interval, and its maximum over the type 2 regime is attained at $p=0$. Evaluating at $p=0$ yields

$$W_2^{\mathrm{tot}}(0) = N\rho\, G\left[1-\left(\frac{c}{G}\right)^{\frac{N(1-\rho)}{N(1-\rho)-1}}\right] + N(1-\rho)(G-c).$$

The attracting policies constructed in Appendix B.1 satisfy $p^\star \to 0$ as $q_0 \downarrow 0$. Using $W_2^{\mathrm{tot}}(p) = N\rho\, w_2(p) + N(1-\rho)(G-c)$ and continuity at zero gives $W_2^{\mathrm{tot}}(p^\star) \to W_2^{\mathrm{tot}}(0)$.

This proves both claims. □

B.3 Proof of Theorem 4.1

Proof. By Lemma 2, W_1^{tot} is strictly concave and has the unique maximizer p_1^{soc}. Since $p_1^{\mathrm{soc}} > p_1^{\mathrm{usr}}$, the LLM-local type 1 optimum is not the total-welfare maximizer. Hence $\Delta > 0$.

Using the value derived in Appendix B.2,

$$W_1^{\mathrm{tot}}(p_1^{\mathrm{soc}}) = NG - N\rho c + c(N\rho - 1)\left(\frac{c}{NG}\right)^{\frac{1}{N\rho-1}}.$$

At the LLM-local type 1 optimum, Appendix B.9 gives

$$W_1^{\mathrm{tot}}(p_1^{\mathrm{usr}}) = NG - N\rho c + c\left(N\rho - \frac{1}{\rho}\right)\left(\frac{c}{GN\rho}\right)^{\frac{1}{N\rho-1}}.$$

Subtracting and normalizing by G yields

$$\Delta = N^{-\frac{1}{N\rho-1}}\left(\frac{c}{G}\right)^{1+\frac{1}{N\rho-1}}\left[N\rho - 1 - \left(N\rho - \frac{1}{\rho}\right)\left(\frac{1}{\rho}\right)^{\frac{1}{N\rho-1}}\right].$$

For fixed N and ρ, the bracketed term times $N^{-1/(N\rho-1)}$ is positive because $\Delta > 0$. Thus $\Delta = K_0(c/G)^\kappa$ with $K_0 > 0$ and $\kappa := 1 + 1/(N\rho - 1) > 1$, proving strict monotonicity and strict convexity in c/G.

For the adoption comparison, fix N, c, G and set $a := N\rho$. Since $da/d\rho = N > 0$, the sign of $\partial\Delta/\partial\rho$ is the same as the sign of $\partial\Delta/\partial a$. In this coordinate,

$$\Delta(c/G, a/N) = \frac{c}{G}\left(\frac{c}{NG}\right)^{\frac{1}{a-1}} C(a),$$

where

$$C(a) := a - 1 - \left(a - \frac{N}{a}\right)\left(\frac{N}{a}\right)^{\frac{1}{a-1}}.$$

The positivity of $C(a)$ follows from $\Delta > 0$. Differentiating gives

$$\frac{\partial\Delta}{\partial a} = \frac{c}{G}\left(\frac{c}{NG}\right)^{\frac{1}{a-1}}\frac{C(a)}{(a-1)^2}\left[h(a) - \log\frac{c}{NG}\right],$$

where $h(a) := (a-1)^2\frac{C'(a)}{C(a)}$. Thus any interior critical point must solve $h(a) = \log\frac{c}{NG}$. It remains to show that h is strictly decreasing.

Set $t := 1/(a-1)$ and $s := N/a = Nt/(1+t)$. Then $dt/da = -t^2 < 0$, and the same function in the t coordinate is

$$\widehat{C}(t) := C(1 + 1/t) = \frac{1}{t} - \left(1 + \frac{1}{t} - s\right)s^t = (s-1)s^t - \frac{s^t - 1}{t}.$$

Hence

$$h(a) = -\frac{d}{dt}\log\widehat{C}(t), \qquad \frac{dh(a)}{da} = t^2\frac{d^2}{dt^2}\log\widehat{C}(t).$$

Therefore h is decreasing if $\widehat{C}(t)$ is log-concave.

Let $y := \log s > 0$. A direct calculation gives

$$\widehat{C}(t)^2\frac{d^2}{dt^2}\log\widehat{C}(t) = \frac{\mathcal{K}(t,y)}{(1+t)t^4},$$

where

$$\begin{aligned}\mathcal{K}(t,y) := {} & t^2(1+t)y^2\left(te^{(t+1)y} - (1+t)e^{ty}\right) \\ & + 4t^2(1+t)y\left(e^{(t+1)y} - e^{ty}\right) \\ & - t^2e^{2(t+1)y} - (2t^2+4t)e^{(2t+1)y} + (3t^2+4t+1)e^{2ty} \\ & + (5t^2+4t)e^{(t+1)y} - (5t^2+5t+2)e^{ty} + t + 1.\end{aligned}$$

Expanding $\mathcal{K}(t,y)=\sum_{n\ge 0}\mathcal{A}_n(t)y^n$ gives $\mathcal{A}_0(t)=\mathcal{A}_1(t)=0$. For $\mathcal{B}_n(t)\coloneqq n!\mathcal{A}_n(t)$,

$$\begin{aligned}\mathcal{B}_n(t) &= n(n-1)t^2(1+t)\left[t(1+t)^{n-2}-(1+t)t^{n-2}\right]\\ &\quad+4nt^2(1+t)\left[(1+t)^{n-1}-t^{n-1}\right]\\ &\quad-t^2(2t+2)^n-(2t^2+4t)(2t+1)^n+(3t^2+4t+1)(2t)^n\\ &\quad+(5t^2+4t)(1+t)^n-(5t^2+5t+2)t^n.\end{aligned}$$

For $n=2$, $\mathcal{B}_2(t)=-t^2(1+t)<0$. For $n\ge 3$, set $\omega\coloneqq t/(1+t)\in(0,1)$. Then

$$\mathcal{B}_n(t)=-\frac{\omega^2(1-\omega)P_n(\omega)}{(1-\omega)^{n+2}},$$

where $P_n(\omega)\coloneqq\sum_{j=0}^{n-2}a_{n,j}\omega^j$ has positive coefficients:

$$a_{n,0}=2^n-3>0,\qquad a_{n,n-2}=2^{n+1}-2n-5>0,$$

and, for $1\le j\le n-3$,

$$a_{n,j}=2\sum_{k=0}^{j}\binom{n}{k}+4\binom{n}{j+1}+2^n-n^2-3n-5\ge 0.$$

The only boundary case is $n=3$, for which $P_3(\omega)=5+5\omega>0$. Hence $P_n(\omega)>0$ for all $n\ge 3$, so $\mathcal{B}_n(t)<0$ for all $n\ge 2$. It follows that $\mathcal{K}(t,y)<0$ for $t>0$ and $y>0$, and therefore $\widehat{C}(t)$ is strictly log-concave. Consequently $h'(a)<0$.

Because h is strictly decreasing, the critical-point equation $h(a)=\log\frac{c}{NG}$ has at most one solution. If such a solution exists, the sign of $\partial\Delta/\partial a$ changes from positive to negative at that point, so it is the unique interior maximizer. If no solution exists, the gap is monotone on the admissible interval and its maximum is attained at a boundary. □

B.4 Proof of Corollary 4.1

Proof. Fix N and G, and take $0<c_1<c_2<G$ in a range where the maximizers are interior. Since the penetration interval is common to both costs, optimality at c_2 and pointwise strict increase from Theorem 4.1 imply

$$\begin{aligned}\Delta(c_2/G,\rho^\star(c_2/G)) &\ge \Delta(c_2/G,\rho^\star(c_1/G))\\ &> \Delta(c_1/G,\rho^\star(c_1/G)).\end{aligned}$$

This proves strict increase of the peak gap.

For the location, reuse the local function h from Appendix B.3. Its critical-point equation is $h(N\rho^\star)=\log(c/(NG))$, and $h'<0$. Implicit differentiation gives

$$\frac{d\rho^\star}{d(c/G)}=\frac{1}{N(c/G)\,h'(N\rho^\star)}<0.$$

Thus the interior peak moves to a strictly lower penetration rate as c/G increases. □

B.5 Proof of Corollary 4.2

Proof. Direct substitution of the two policies into $W_1^{\mathrm{tot}}(p)=NG[1-(1-p)^{N\rho}]-N\rho cp$ gives

$$\begin{aligned}W_1^{\mathrm{tot}}(p_1^{\mathrm{soc}}) &= NG\left[1-\left(\frac{c}{NG}\right)^{\frac{N\rho}{N\rho-1}}\right]-N\rho c\left[1-\left(\frac{c}{NG}\right)^{\frac{1}{N\rho-1}}\right]\\ &= NG-N\rho c+c(N\rho-1)\left(\frac{c}{NG}\right)^{\frac{1}{N\rho-1}},\\ W_1^{\mathrm{tot}}(p_1^{\mathrm{usr}}) &= NG\left[1-\left(\frac{c}{GN\rho}\right)^{\frac{N\rho}{N\rho-1}}\right]-N\rho c\left[1-\left(\frac{c}{GN\rho}\right)^{\frac{1}{N\rho-1}}\right]\\ &= NG-N\rho c+c\left(N\rho-\frac{1}{\rho}\right)\left(\frac{c}{GN\rho}\right)^{\frac{1}{N\rho-1}}.\end{aligned}$$

Subtracting these values and dividing by G yields the exact identity

$$\Delta(c/G,\rho) = \frac{c}{G}\left(\frac{c}{GN\rho}\right)^{\frac{1}{N\rho-1}} \cdot \left[\frac{1}{\rho} - 1 - (N\rho-1)\left(1-\rho^{\frac{1}{N\rho-1}}\right)\right]. \tag{6}$$

For fixed N and ρ, the exponent $1/(N\rho-1)$ lies in $(0,1]$. Concavity of the power function and its tangent at one give

$$\rho^{\frac{1}{N\rho-1}} \le 1 - \frac{1-\rho}{N\rho-1}, \qquad (N\rho-1)\left(1-\rho^{\frac{1}{N\rho-1}}\right) \ge 1-\rho.$$

Thus the bracket in (6) satisfies

$$\frac{1}{\rho} - 1 - (N\rho-1)\left(1-\rho^{\frac{1}{N\rho-1}}\right) \le \frac{1}{\rho} - 2 + \rho = \frac{(1-\rho)^2}{\rho},$$

which proves the first upper bound. Since $c/(GN\rho) < 1$, its positive power is at most one, giving the second. Strict positivity follows from the welfare-optimality comparison in Lemma 2.

For $\epsilon > 0$, the simpler upper bound is less than ϵ precisely when

$$\rho^2 - (2 + \epsilon G/c)\rho + 1 < 0.$$

The smaller root is

$$\frac{2+\epsilon G/c - \sqrt{(\epsilon G/c)^2 + 4\epsilon G/c}}{2} = \frac{2}{2+\epsilon G/c + \sqrt{(\epsilon G/c)^2 + 4\epsilon G/c}}.$$

The larger root exceeds one, so every admissible penetration rate above the smaller root satisfies $\Delta(c/G,\rho) < \epsilon$. □

B.6 Extension to welfare across contribution regimes

The main result quantifies the social welfare gap within a fixed contribution regime. We now extend the analysis to users' welfare across regimes. Their payoff also depends on the contribution pattern supported by the pretrained reference. Contributing requires them to bear costs, while relying on the other players leaves their payoff dependent on others' contributions. Using the best expected user payoffs $w_R^{\max}$ defined in Appendix B.1, we compare regimes through the normalized difference introduced in Section 4.2,

$$\Gamma(c/G,\rho) = \frac{w_2^{\max}(\mathcal{V}) - w_1^{\max}(\mathcal{V})}{G} = \frac{c}{G} - (N\rho-1)\left(\frac{c}{GN\rho}\right)^{\frac{N\rho}{N\rho-1}} - \left(\frac{c}{G}\right)^{\frac{N(1-\rho)}{N(1-\rho)-1}}.$$

To distinguish the comparisons, Γ measures the difference in best payoff per guided user across regimes, in units of G, whereas Δ compares total welfare under two policies within type 1. Thus $\Gamma(c/G,\rho) > 0$ favors the type 2 benchmark for users, while $\Gamma(c/G,\rho) < 0$ favors type 1. From Assumption 1, the admissible LLM penetration rate is $\rho \in [2/N, 1-2/N]$.

Theorem B.1 (User welfare across contribution regimes)**.** *For every $\rho \in [2/N, 1-2/N]$, $\Gamma(\cdot,\rho)$ is strictly concave in c/G, has a unique maximizer, and admits a unique cutoff $x_c(\rho) \in (0,1)$ such that the type 2 regime is preferred for $c/G < x_c(\rho)$ and the type 1 regime is preferred for $c/G > x_c(\rho)$. The cutoff map $\rho \mapsto x_c(\rho)$ is either monotone on $[2/N, 1-2/N]$ or has a unique interior minimum. Consequently, for each fixed $c/G \in (0,1)$, the set $E_{c/G} := \{\rho \in [2/N, 1-2/N] : \Gamma(c/G,\rho) < 0\}$ is either empty or an interval, and this interval expands as c/G increases.*

Theorem B.1 identifies when guided users gain more at the type 1 user optimum than at the type 2 user optimum. In this range, the additional expected benefit outweighs the contribution costs they avoid in type 2. A pretrained tendency toward contributing can therefore support higher user welfare, with the ranking determined by the cost-benefit ratio and the LLM penetration rate.

Table 3 in Appendix B.8 summarizes users' preferred regime.

Figure 4 illustrates how users' preferred regime changes with adoption and contribution costs. The proof is in Appendices B.7 and B.8.

As a complementary coordination result, Appendix B.9 shows that at the branchwise user optima the normalized residual cost is $O(\log N)$ in type 1, compared with $\Theta(N)$ in type 2 and N at the symmetric mixed Nash benchmark.

B.7 Sufficient Empty Set Upper Bound

For this subsection, write $x \coloneqq c/G$ and use the preference set E_x from Theorem B.1.

Lemma 3 (A sufficient empty-set region and its N-dependence)**.** *If $x = c/G$ satisfies $x < 2^{-(N-3)}$, then $E_x = \varnothing$. Moreover, the sufficient bound $2^{-(N-3)}$ is strictly decreasing in N.*

Lemma 3 gives a sufficient region in which the type 2 regime yields larger welfare to the LLM population at every admissible ρ. This guaranteed region shrinks with N.

Proof. Fix $x = c/G \in (0, 1)$ and, within this proof, let $a \coloneqq N\rho \in [2, N-2]$. The welfare difference factors as

$$\Gamma(x, \rho) = x \left[1 - \frac{a-1}{a} \left(\frac{x}{a}\right)^{\frac{1}{a-1}} - x^{\frac{1}{N-a-1}}\right].$$

Both $a - 1$ and $N - a - 1$ lie in $[1, N - 3]$. Since $0 < x < 1$ and $\frac{a-1}{a} a^{-1/(a-1)} < 1$, we have

$$\frac{a-1}{a} \left(\frac{x}{a}\right)^{\frac{1}{a-1}} \le x^{\frac{1}{N-3}}, \qquad x^{\frac{1}{N-a-1}} \le x^{\frac{1}{N-3}}.$$

Consequently,

$$\Gamma(x, \rho) \ge x \left[1 - 2x^{\frac{1}{N-3}}\right] > 0 \qquad \text{if } x < 2^{-(N-3)}.$$

Thus $E_x = \varnothing$ under this sufficient condition, whose upper bound is strictly decreasing in N. □

B.8 Proof of Theorem B.1

Table 3: Optimal user payoff across regimes at fixed ρ (Theorem B.1).

Cost ratio	$c/G < x_c(\rho)$	$c/G = x_c(\rho)$	$c/G > x_c(\rho)$
Welfare difference Γ	> 0	$= 0$	< 0
Preferred regime	Type 2	Indifferent	Type 1

$\Gamma = (w_2^{\max} - w_1^{\max})/G$ compares the best payoff per LLM user across regimes.

Proof. Within this proof, write $x \coloneqq c/G$ and $a \coloneqq N\rho \in [2, N-2]$, retaining the preference set E_x from the theorem. Substituting the branch optima from Appendix B.1 gives

$$\begin{aligned}\Gamma(x, \rho) &= x - x\frac{a-1}{a} \left(\frac{x}{a}\right)^{\frac{1}{a-1}} - x^{\frac{N-a}{N-a-1}} \\ &= x - (a-1) \left(\frac{x}{a}\right)^{\frac{a}{a-1}} - x^{\frac{N-a}{N-a-1}}.\end{aligned}$$

For fixed a, differentiation yields

$$\partial_x \Gamma = 1 - \left(\frac{x}{a}\right)^{\frac{1}{a-1}} - \frac{N-a}{N-a-1} x^{\frac{1}{N-a-1}},$$

and

$$\partial_x^2 \Gamma = -\frac{1}{a(a-1)} \left(\frac{x}{a}\right)^{\frac{2-a}{a-1}} - \frac{N-a}{(N-a-1)^2} x^{\frac{2-(N-a)}{N-a-1}} < 0.$$

All denominators are positive on the admissible domain. Thus $\Gamma(\cdot,\rho)$ is strictly concave on $(0,1)$. Moreover, $\Gamma(0^+,\rho)=0$, $\partial_x\Gamma(0^+,\rho)=1$, and

$$\Gamma(1,\rho)=-(a-1)\left(\frac{1}{a}\right)^{\frac{a}{a-1}}<0.$$

It follows that $\Gamma(\cdot,\rho)$ has a unique maximizer and a unique cutoff $x_{\mathrm{c}}(\rho)\in(0,1)$, with positive gap below the cutoff and negative gap above it.

To study the penetration rate, set

$$F(a,x)\coloneqq\frac{a-1}{a}\left(\frac{x}{a}\right)^{\frac{1}{a-1}}+x^{\frac{1}{N-a-1}},$$

so that $\Gamma(x,\rho)=x(1-F(a,x))$. In the local a coordinate, write $x_{\mathrm{c}}(a)$ for $x_{\mathrm{c}}(a/N)$. The cutoff satisfies $F(a,x_{\mathrm{c}}(a))=1$, and $F_x>0$ makes it smooth. When $N=4$, the penetration interval is a singleton and the fixed-cost conclusions already prove the result. Suppose henceforth that $N>4$.

For fixed $x\in(0,1)$, write

$$f_1(a)\coloneqq\frac{a-1}{a}\left(\frac{x}{a}\right)^{\frac{1}{a-1}},\qquad f_2(a)\coloneqq x^{\frac{1}{N-a-1}}.$$

Their logarithmic derivatives are

$$\varrho(a)\coloneqq\frac{f_1'(a)}{f_1(a)}=\frac{\log(a/x)}{(a-1)^2}>0,\qquad \varsigma(a)\coloneqq-\frac{f_2'(a)}{f_2(a)}=\frac{-\log x}{(N-a-1)^2}>0.$$

Thus f_1 is strictly increasing and f_2 is strictly decreasing. Direct differentiation gives

$$\varrho'(a)=\frac{1}{a(a-1)^2}-\frac{2\varrho(a)}{a-1},\qquad \varsigma'(a)=\frac{2\varsigma(a)}{N-a-1}.$$

Consider an arbitrary interior stationary point of x_{c} and evaluate the following quantities at $(a,x_{\mathrm{c}}(a))$, with the derivatives of $f_1,f_2,\varrho,\varsigma$ taken at fixed x. The cutoff identity and its first derivative give

$$f_1+f_2=1,\qquad f_1\varrho=f_2\varsigma.$$

Since $\log(a/x)>\log a>(a-1)/a$, we obtain

$$\varrho'<-\frac{\varrho}{a-1}.$$

The definitions also give

$$(a-1)\varrho=\log(1-1/a)-\log f_1,\qquad (N-a-1)\varsigma=-\log f_2.$$

For positive f_1,f_2 summing to one, $\log(1+t)<t$ yields $-\log f_1<f_2/f_1$ and $-\log f_2<f_1/f_2$. Combining these inequalities with $f_1\varrho=f_2\varsigma$ gives

$$(a-1)\varsigma=\frac{f_1}{f_2}\big[\log(1-1/a)-\log f_1\big]<1,$$
$$(N-a-1)\varrho=\frac{f_2}{f_1}(-\log f_2)<1.$$

Hence $\varsigma<1/(a-1)$ and $\varrho<1/(N-a-1)$.

Differentiating $F=f_1+f_2$ twice and using the stationary-point identity,

$$\begin{aligned}F_{aa}&=f_1(\varrho'+\varrho^2)+f_2(\varsigma^2-\varsigma')\\&=f_1\left[\varrho'+\varrho\left(\varrho+\varsigma-\frac{2}{N-a-1}\right)\right]\\&<f_1\varrho\left[-\frac{1}{a-1}+\frac{1}{N-a-1}+\frac{1}{a-1}-\frac{2}{N-a-1}\right]\\&=-\frac{f_1\varrho}{N-a-1}<0.\end{aligned}$$

Implicit differentiation at this stationary point therefore gives

$$x_{\mathrm{c}}''(a) = -\frac{F_{aa}}{F_x} > 0.$$

Every interior stationary point is a strict local minimum. Two such minima would require an intervening interior maximum, which would be a stationary point that is not a strict minimum. Hence there is at most one interior stationary point: the cutoff is monotone or decreases to one interior minimum and then increases. Since $a = N\rho$ is affine and increasing, the same conclusion holds in ρ.

Finally, $E_x = \{\rho \in [2/N, 1 - 2/N] : x > x_{\mathrm{c}}(\rho)\}$ is a strict sublevel set of this cutoff. It is therefore empty or an interval, and these intervals are nested and expand as x increases. □

B.9 Population-size scaling at the user optima

The branchwise user optima also allow a comparison of how coordination losses scale with population size. Let $W_R^{\mathrm{LLM,max}} := N\rho\, w_R^{\max}$ be the aggregate user welfare at the regime R optimum, for $R \in \{1, 2\}$. We measure the gap from the benchmark NG through the normalized residual cost.

$$W_1^{\mathrm{tot}}(p_1^{\mathrm{usr}}) = W_1^{\mathrm{LLM,max}} + N(1-\rho)G\left[1 - \left(\frac{c}{GN\rho}\right)^{\frac{N\rho}{N\rho-1}}\right], \tag{7}$$

$$W_2^{\mathrm{tot}}(0) = W_2^{\mathrm{LLM,max}} + N(1-\rho)(G-c). \tag{8}$$

The residual cost $NG - W$ includes contribution costs and the provision benefits lost when nobody volunteers. Exactly one volunteer gives first-best welfare $W^* := NG - c$, with normalized residual cost one. At the branchwise user optima, define

$$R_1 := \frac{NG - W_1^{\mathrm{tot}}(p_1^{\mathrm{usr}})}{c}, \qquad R_2 := \frac{NG - W_2^{\mathrm{tot}}(0)}{c}.$$

Lemma 4 (Normalized residual costs at the branchwise optima)**.** *Under the assumptions of Lemma 2, for fixed $c/G \in (0,1)$ and $\rho \in (0,1)$, the normalized residual costs satisfy $R_1 = O(\log N)$ whereas $R_2 = \Theta(N)$ as $N \to \infty$ along a sequence satisfying Assumption 1.*

Proof. At the type 1 optimizer $(p_1^{\mathrm{usr}}, 0)$, each no-regret player's limiting policy assigns probability one to defection, with payoff

$$w_{\mathrm{NR},1}(p_1^{\mathrm{usr}}) = G\left[1 - (1 - p_1^{\mathrm{usr}})^{N\rho}\right].$$

Using

$$1 - p_1^{\mathrm{usr}} = \left(\frac{c}{GN\rho}\right)^{\frac{1}{N\rho-1}},$$

we obtain

$$(1 - p_1^{\mathrm{usr}})^{N\rho} = \left(\frac{c}{GN\rho}\right)^{\frac{N\rho}{N\rho-1}}.$$

Hence

$$\begin{aligned} W_1^{\mathrm{tot}}(p_1^{\mathrm{usr}}) &= W_1^{\mathrm{LLM,max}} + N(1-\rho)G\left[1 - \left(\frac{c}{GN\rho}\right)^{\frac{N\rho}{N\rho-1}}\right] \\ &= NG - N\rho c + c\left(N\rho - \frac{1}{\rho}\right)\left(\frac{c}{GN\rho}\right)^{\frac{1}{N\rho-1}}. \end{aligned}$$

Removing the common term NG and normalizing by c yields

$$R_1 = \frac{NG - W_1^{\mathrm{tot}}(p_1^{\mathrm{usr}})}{c} = N\rho - \left(N\rho - \frac{1}{\rho}\right)\left(\frac{c}{GN\rho}\right)^{\frac{1}{N\rho-1}}.$$

At the type 2 optimizer $p = 0$, the aggregate no-regret response is the quasi-equilibrium that equalizes volunteering and defecting, so a representative no-regret player receives $G - c$. Therefore

$$\begin{aligned} W_2^{\mathrm{tot}}(0) &= W_2^{\mathrm{LLM,max}} + N(1-\rho)(G-c) \\ &= NG - Nc\left[(1-\rho) + \rho\left(\frac{c}{G}\right)^{\frac{1}{N(1-\rho)-1}}\right]. \end{aligned}$$

Removing the common term NG and normalizing by c gives

$$R_2 = \frac{NG - W_2^{\mathrm{tot}}(0)}{c} = N\left[(1-\rho) + \rho\left(\frac{c}{G}\right)^{\frac{1}{N(1-\rho)-1}}\right].$$

We now study the cost trend as $N \to \infty$, holding $c/G \in (0,1)$ and $\rho \in (0,1)$ fixed. The following expansions are therefore used for sufficiently large N. For the asymptotics of R_1, set $L_N \coloneqq \log(GN\rho/c)$. Since $L_N/(N\rho - 1) \to 0$, Taylor expansion of the exponential gives

$$\left(\frac{c}{GN\rho}\right)^{\frac{1}{N\rho-1}} = \exp\left(-\frac{L_N}{N\rho-1}\right) = 1 - \frac{L_N}{N\rho-1} + O\left(\frac{L_N^2}{N^2}\right),$$

where the remainder is valid because $L_N/(N\rho - 1)$ is small for sufficiently large N. Substituting into the formula for R_1 yields

$$\begin{aligned} R_1 &= N\rho - \left(N\rho - \frac{1}{\rho}\right)\left[1 - \frac{L_N}{N\rho-1} + O\left(\frac{L_N^2}{N^2}\right)\right] \\ &= \frac{1}{\rho} + \left(N\rho - \frac{1}{\rho}\right)\frac{L_N}{N\rho-1} + O\left(\frac{L_N^2}{N}\right) \\ &= \log\left(\frac{GN\rho}{c}\right) + \frac{1}{\rho} + o(1). \end{aligned}$$

Similarly, because $\log(c/G)/(N(1-\rho)-1) \to 0$,

$$\left(\frac{c}{G}\right)^{\frac{1}{N(1-\rho)-1}} = \exp\left(\frac{\log(c/G)}{N(1-\rho)-1}\right) = 1 + \frac{\log(c/G)}{N(1-\rho)-1} + O\left(\frac{1}{N^2}\right),$$

so

$$\begin{aligned} R_2 &= N\left[1 + \rho\frac{\log(c/G)}{N(1-\rho)-1} + O\left(\frac{1}{N^2}\right)\right] \\ &= N + \frac{\rho\log(c/G)}{1-\rho} + o(1). \end{aligned}$$

This proves the stated asymptotic comparison. At the full-population symmetric mixed Nash benchmark, the probability that all other players defect is c/G. Volunteering gives $G - c$, and defection gives $G(1 - c/G) = G - c$. Each player's expected payoff is therefore $G - c$, so the normalized residual cost is $(NG - N(G-c))/c = N$. □

At the type 1 user optimum, the common action distribution spreads contribution probabilities across guided users, who sample independently conditional on that policy. Its normalized residual cost has $O(\log N)$ scaling, while the type 2 user optimum retains a linear contribution from the no-regret population's indifference payoff $G - c$.

C Dynamic Reduction and Finite-Batch Analysis

We use the game, conditional sampling, and update conventions of Section 2. The calculations below retain the finite-population dependence within each period and the empirical baseline used by the advisor.

Table 4: Limiting responses and advisor convergence.

Setting	Condition	Outcome	Result
Fixed advice p	$p < p_{\mathrm{bd}}$	Type 2, $\sigma^\star > 0$	Lemma 1
	$p \geq p_{\mathrm{bd}}$	Type 1, $\sigma^\star = 0$	
KL strength β	$0 < \beta < \beta_{\mathrm{c}}$	Multiple fixed points possible	Theorem D.1
	$\beta \geq \beta_{\mathrm{c}}$	Unique attracting limit	Theorem 3.1
Reference q_0 ($\beta \geq \beta_{\mathrm{c}}$)	$q_0 < q_{\mathrm{cut}}$	Type 2	Theorem 3.1
	$q_0 = q_{\mathrm{cut}}$	Type 1 at the boundary	
	$q_0 > q_{\mathrm{cut}}$	Type 1 in the interior	

The KL and reference rows describe convergent advisor mean dynamics. All entries use the assumptions of the cited results.

C.1 Learning rules and coordinate identities

We give the exponential weights interpretation of the NR score, the complete advisor algorithm, and the coordinate identities used in the mean dynamics. Policy vectors and their scalar volunteering components follow the convention in Section 2.3.

C.1.1 Exponential weights and the aggregate score

The logistic probability in Section 2.2 is the binary form of exponential weights (Cesa-Bianchi & Lugosi, 2006; Arora et al., 2012). To see this, let $h_{a,t}$ be the reward score of action $a \in \{V, D\}$, with update

$$h_{a,t+1} = h_{a,t} + \tilde{\eta}_t r_{a,t}.$$

Sampling actions with probabilities proportional to $e^{h_{a,t}}$ and setting $z_t = h_{V,t} - h_{D,t}$ gives

$$\frac{e^{h_{V,t}}}{e^{h_{V,t}} + e^{h_{D,t}}} = \frac{1}{1 + e^{-z_t}} = \sigma(z_t), \qquad z_{t+1} = z_t + \tilde{\eta}_t(r_{V,t} - r_{D,t}).$$

Equivalently, the action weights are multiplied by $e^{\tilde{\eta}_t r_{a,t}}$ at each update, and

$$z_t = z_1 + \sum_{s=1}^{t-1} \tilde{\eta}_s \delta r_s.$$

This identity explains both the logistic map and the role of the learning rates. The aggregate model uses the representative player's reward difference to update one common score. All NR players sample independently conditional on the current score and history, as specified in Section 2.4. This is an explicit population approximation. The convergence proof below applies to this process under Assumption 2; the exponential weights representation alone does not establish a regret guarantee for an arbitrary learning rate sequence or a population of private scores. The score z_t is the log odds of volunteering, so zero gives equal action probabilities and positive values favor volunteering. A positive reward difference raises these odds, with $\tilde{\eta}_t$ controlling the size of the adjustment. In Assumption 2, the divergent step sum permits continued adjustment, while the finite squared sum controls accumulated sampling noise, as in classical stochastic approximation (Robbins & Monro, 1951; Benaim, 1999).

C.1.2 Complete advisor update

Algorithm 1 implements the policy maps and score update in Section 2.3, using the rewards and advantages defined in Appendices C.1.3 and C.1.4. Each batch uses $BN\rho$ reports, with the batch mean computed from the same observations used by the score estimate.

Remark 4 (Batched updates)**.** *Batching separates data collection from policy revision, as in policy gradient methods such as PPO (Schulman et al., 2017). NR players continue learning each period, so freezing the advisor does not assume that they have reached their limiting response. Appendix C.3 bounds the resulting tracking error.*

Algorithm 1 KL-LLM-PG update for the shared LLM advisor

Input. batch length B, $N\rho$ guided users, scores μ_b, frozen model-policy reference $\mathbf{q}_0$, response sensitivity λ, regularization parameter β, learning rate η_b

Output. updated scores μ_{b+1}, model policy $\mathbf{q}_{b+1}$, and effective policy $\mathbf{p}_{b+1}$

$\mathbf{q}_b \leftarrow \mathrm{softmax}(\mu_b)$ and $\mathbf{p}_b \leftarrow \mathrm{softmax}(\lambda\mu_b)$ ▷ freeze the advisor during batch b

$p_{a,0} \leftarrow q_{a,0}^{\lambda} / \sum_{a' \in \mathcal{A}} q_{a',0}^{\lambda}$ for each $a \in \mathcal{A}$

$\mathcal{D}_b \leftarrow \textsc{BatchFeedback}(B, \mathbf{p}_b, \mathbf{p}_0, \beta, \lambda)$

$g_b \leftarrow \textsc{CenteredGradient}(\mathcal{D}_b, \mathbf{p}_b)$

$\mu_{b+1} \leftarrow \mu_b + \eta_b g_b$

$\mathbf{q}_{b+1} \leftarrow \mathrm{softmax}(\mu_{b+1})$

$\mathbf{p}_{b+1} \leftarrow \mathrm{softmax}(\lambda\mu_{b+1})$

return $(\mu_{b+1}, \mathbf{q}_{b+1}, \mathbf{p}_{b+1})$

BatchFeedback$(B, \mathbf{p}_b, \mathbf{p}_0, \beta, \lambda)$

Run B periods of the joint game with the current no-regret process. Collect the realized action and payoff of each of the $N\rho$ guided users each period, apply (9), and return the records $\mathcal{D}_b = \{(a_{n,t}, \tilde{r}_{n,t})\}_{n,t}$.

CenteredGradient$(\mathcal{D}_b, \mathbf{p}_b)$

From the same records, compute the empirical baseline $\bar{r}_b$ and empirical advantages $A_{n,t}$ in (10). Sum their score contributions over all $BN\rho$ records and return g_b.

C.1.3 Reward Design

Recall the realized action $a_{n,t}$ and game payoff $r_{n,t}$ of guided user n in period t of batch b from Section 2.3. The frozen references are $\mathbf{q}_0 = \mathrm{softmax}(\mu_0)$ and $\mathbf{p}_0 = \mathrm{softmax}(\lambda\mu_0)$ for finite initial logits μ_0. With model-scale regularization strength $\beta \geq 0$, the adjusted reward is

$$\tilde{r}_{n,t} \coloneqq r_{n,t} - \frac{\beta}{\lambda} \log \frac{p_{a_{n,t},b}}{p_{a_{n,t},0}}. \tag{9}$$

The log ratio compares the current and reference probabilities of the observed action. Its expectation is $D_{\mathrm{KL}}(\mathbf{p}_b \| \mathbf{p}_0)$, so the penalty discourages departure from the initial action distribution. The coefficient β/λ expresses regularization in effective policy coordinates. Appendix C.1.5 derives the connection to the model policy and coefficient β.

This sampled KL penalty follows Ziegler et al. (2020, Eq. (2)) and Stiennon et al. (2020, Section 3.4), consistent with reward regularization in LLM training (Jaques et al., 2020; Ouyang et al., 2022). For $\beta > 0$, payoff feedback competes with the frozen reference's action preferences at every update. Because only bandit feedback is available, each adjusted reward supplies one Monte Carlo contribution to the regularized score estimate.

C.1.4 Advantage Function

We compare each adjusted reward with the average in its batch. Using the rewards in (9), define the empirical baseline and advantages by

$$\bar{r}_b \coloneqq \frac{1}{BN\rho} \sum_{t=1}^{B} \sum_{n=1}^{N\rho} \tilde{r}_{n,t}, \qquad A_{n,t} \coloneqq \tilde{r}_{n,t} - \bar{r}_b. \tag{10}$$

An advantage is positive when the adjusted reward exceeds its batch average and negative when it falls below it. The baseline averages the reports, whereas g_b in (3) sums their score contributions, giving the effective step $\gamma_b = \eta_b BN\rho$.

This construction parallels the Group Relative Policy Optimization (GRPO) approach of Shao et al. (2024). Because all LLM players share a common policy and the repeated game has no state transitions, $\bar{r}_b$ serves as a natural Monte Carlo estimate of the expected reward under the current effective action distribution, without requiring an explicit value-function estimator. The advantage function is also a standard variance-reduction technique in reinforcement learning, widely applied in LLM fine-tuning and reasoning Liu et al. (2024).

When $A_{n,t} > 0$, that report contributes toward reinforcing the sampled action; when $A_{n,t} < 0$, it contributes in the opposite direction. The net update depends on the sum over all reports.

Under stationary sampling, $\bar{r}_b$ estimates the regularized reward. Reusing the same observations to construct both the baseline and the score creates finite-batch dependence in the gradient estimate. The large-batch conditional mean accounts for the resulting correction explicitly. A baseline estimated from an independent batch—or, under independent stationary periods, a leave-one-period-out baseline—would decouple the focal period's strategically interdependent reports from the score term Weaver & Tao (2013).

Combining the KL-adjusted reward with the centered score-function estimate gives a batched KL-regularized policy-gradient-style update, denoted KL-LLM-PG.

C.1.5 Normalized Score Update

The effective action distribution has score

$$\nabla_{\mu_b} \log p_{a_{n,t},b} = \lambda\big(e_{a_{n,t}} - \mathbf{p}_b\big).$$

Applying learning rate η_b/λ to the advantage weighted sum of these derivatives gives exactly the update (3), following the policy gradient estimator in Schulman et al. (2017, Eq. (1)). Thus Section 2.3 absorbs the fixed factor λ into η_b. Algorithm 1 then sums the scaled score contributions defined in Section 2.3.

For each $a \in \mathcal{A}$, the policy components satisfy

$$p_{a,b} = \frac{q_{a,b}^{\lambda}}{\sum_{a' \in \mathcal{A}} q_{a',b}^{\lambda}}.$$

A value $0 < \lambda < 1$ moves the effective distribution toward equal probabilities, $\lambda = 1$ preserves the model distribution, and $\lambda > 1$ increases the probability of the action with the larger model score. A single common λ is a parsimonious response model. It does not represent all differences in prompt wording, trust, or deployment context that may affect responses to LLM advice (Robinson & Burden, 2025; Cao & Hu, 2026; Bo et al., 2025; Grawitch et al., 2026). The same coordinate relation holds at $b = 0$. In particular,

$$p_{V,0} = \frac{q_{V,0}^{\lambda}}{q_{V,0}^{\lambda} + (1 - q_{V,0})^{\lambda}}, \qquad \log \frac{p_{V,0}}{1 - p_{V,0}} = \lambda \log \frac{q_{V,0}}{1 - q_{V,0}}.$$

The centered estimator also gives an exact finite-batch bridge to the model-policy coordinates used later. Softmax normalization implies that, for a scalar C_b independent of the sampled action,

$$\frac{1}{\lambda} \log \frac{p_{a,b}}{p_{a,0}} = \log \frac{q_{a,b}}{q_{a,0}} + C_b, \qquad C_b \coloneqq \frac{1}{\lambda} \log \frac{\sum_{a' \in \mathcal{A}} q_{a',0}^{\lambda}}{\sum_{a' \in \mathcal{A}} q_{a',b}^{\lambda}}.$$

The additive term $-\beta C_b$ disappears when $\bar{r}_b$ is subtracted. Moreover, $\sum_{n,t} A_{n,t} = 0$, and hence

$$\sum_{n,t} A_{n,t}(e_{a_{n,t}} - \mathbf{p}_b) = \sum_{n,t} A_{n,t}(e_{a_{n,t}} - \mathbf{q}_b).$$

Consequently, the effective-policy update in Algorithm 1 is samplewise identical to the alternative representation using the model-policy log ratio and centering vector. This identity is why the subsequent drift formulas can be written compactly using the volunteering components q_b and q_0 and the coefficient β while the deployed-policy interpretation remains internally consistent.

Each summand uses the reporting user's own regularized payoff while holding the other users' realized actions fixed. Pooling therefore shares statistical information and model parameters. Although the empirical baseline induces finite-batch cross-user covariance terms, the update does not include the full cross-user derivatives required for an unbiased gradient of aggregate LLM-user welfare or total social welfare. We interpret KL-LLM-PG as a representative-user, independent-agent score update. The welfare objectives in Section 2.5 are analytical benchmarks and need not coincide with the update's implicit learning objective.

C.2 AGGREGATE NO-REGRET RESPONSE

Fix $\mathcal{V} = (N, \rho, c, G)$ and continue batch b indefinitely with $p = p_b$ fixed, as in Section 3.1. Recall that $\sigma(z_t)$ is each NR player's volunteering probability in Section 2.2, and $\mathcal{H}_{b,t}$ is the history before period t from Section 2.4, extended to this continuation. Conditional independence gives the probability that all $N\rho$ guided and $N(1-\rho)-1$ NR opponents defect. Hence the action rewards satisfy

$$\mathbb{E}[r_{V,t} \mid \mathcal{H}_{b,t}] = G - c, \quad \mathbb{E}[r_{D,t} \mid \mathcal{H}_{b,t}] = G\big[1 - (1-p)^{N\rho}(1-\sigma(z_t))^{N(1-\rho)-1}\big]. \tag{11}$$

For the reward difference δr_t defined in Section 2.2, the expected incentive to volunteer is

$$h_b(z_t) \coloneqq \mathbb{E}[\delta r_t \mid \mathcal{H}_{b,t}] = G(1-p)^{N\rho}(1-\sigma(z_t))^{N(1-\rho)-1} - c. \tag{12}$$

Setting $\xi_{t+1} \coloneqq \delta r_t - h_b(z_t)$ separates this incentive from sampling noise, whose conditional mean is zero. The score update (1) becomes

$$z_{t+1} = z_t + \tilde{\eta}_t\big(h_b(z_t) + \xi_{t+1}\big), \tag{13}$$

where $\tilde{\eta}_t$ is the NR learning rate. The step sizes satisfy Assumption 2. This representation lets the proof of Lemma 1 below distinguish the systematic incentive from accumulated sampling noise.

Remark 5 (Classical no-LLM benchmark)**.** *Lemma 1 also recovers the standard no-regret benchmark when the LLM effect is removed. Although the boundary case $\rho = 0$ lies outside Assumption 1, formally setting $\rho = 0$ and $p = 0$ in the no-regret response gives $\sigma^\star = 1 - (c/G)^{1/(N-1)}$, the symmetric mixed Nash volunteering probability of the classical N-player volunteer's dilemma. Thus the mixed-population model nests the usual all-no-regret case as the no-LLM limit, while the subsequent analysis studies how a positive LLM-guided population changes this benchmark.*

C.2.1 PROOF OF LEMMA 1

Proof. The noise ξ_{t+1} in (13) has zero conditional mean given the history $\mathcal{H}_{b,t}$ and is $\mathcal{H}_{b,t+1}$-measurable. Since δr_t takes the two values $-c$ and $G-c$, its centered value satisfies $|\xi_{t+1}| \le G$. The martingale series $\sum_t \tilde{\eta}_t \xi_{t+1}$ converges almost surely by Assumption 2.

The derivative calculation in Appendix C.3.1 gives $|h_b'(z)| \le G[N(1-\rho)-1]/4$, so h_b is Lipschitz continuous. It is strictly decreasing, with

$$\lim_{z\to+\infty} h_b(z) = -c < 0, \qquad \lim_{z\to-\infty} h_b(z) = G(1-p)^{N\rho} - c.$$

Solving $G(1-p)^{N\rho} = c$ gives the response boundary $p_{\mathrm{bd}} = 1 - (c/G)^{1/(N\rho)}$. If $G(1-p)^{N\rho} > c$, h_b has a unique finite zero $z_b^\star$. The drift is positive below this zero and negative above it. Outside a sufficiently large interval it is bounded away from zero and directed inward; convergence of the martingale series and $\tilde{\eta}_t \to 0$ then give almost-sure boundedness of the recursion. The Robbins–Monro conditions and this attracting scalar drift yield $z_t \to z_b^\star$ almost surely (Robbins & Monro, 1951; Benaim, 1999; Borkar, 2008). Solving $h_b(z_b^\star) = 0$ gives

$$\sigma(z_b^\star) = 1 - \left[\frac{c}{G(1-p)^{N\rho}}\right]^{1/[N(1-\rho)-1]}.$$

If $G(1-p)^{N\rho} \le c$, then $h_b(z) < 0$ for every finite z. In the decomposition

$$z_{t+1} = z_1 + \sum_{j=1}^{t} \tilde{\eta}_j h_b(z_j) + \sum_{j=1}^{t} \tilde{\eta}_j \xi_{j+1},$$

the drift sums are nonincreasing and the noise sums converge. Thus z_t has a limit in $\mathbb{R} \cup \{-\infty\}$. A finite limit would leave $h_b(z_t)$ bounded away from zero, and $\sum_t \tilde{\eta}_t = \infty$ would force the drift sum to diverge to $-\infty$. Therefore $z_t \to -\infty$ and $\sigma(z_t) \to 0$ almost surely, including at the response boundary. □

A finite initial score keeps $\sigma(z_t)$ in $(0,1)$ at every finite period. The continuous extension of the probability update fixes the endpoints zero and one.

C.3 FINITE-BATCH TRACKING

For appendix calculations, write

$$\alpha \coloneqq \frac{1}{N(1-\rho)-1}, \qquad \chi(p) \coloneqq \frac{c}{G(1-p)^{N\rho}}, \qquad \chi_b \coloneqq \chi(p_b).$$

Lemma 1 is asymptotic. To quantify the tracking accuracy after B updates, specialize to $\tilde{\eta}_t = a/(t+t_0)$ with $a, t_0 > 0$. The offset t_0 includes any elapsed learning time; this notation does not require restarting the step-size schedule at a batch boundary. All expectations and probabilities in the following bounds are conditional on the batch-start history $\mathcal{F}_b$, so z_1 and p_b are fixed.

In the type 2 regime, the equilibrium log-odds and restoring slope are

$$z_b^\star = \log\frac{1-\chi_b^\alpha}{\chi_b^\alpha}, \qquad -h_b'(z_b^\star) = \frac{c}{\alpha}(1-\chi_b^\alpha). \tag{14}$$

Choose an interval around $z_b^\star$ on which

$$-\frac{3c}{2\alpha}(1-\chi_b^\alpha) < h_b'(z) < -\frac{c}{2\alpha}(1-\chi_b^\alpha). \tag{15}$$

Such an interval exists by continuity. The local tracking condition requires $z_1, \dots, z_B$ to remain in it almost surely. It is an explicit hypothesis, not conditioning on future confinement or a finite-time consequence of Lemma 1. For $1 < (ac/\alpha)(1-\chi_b^\alpha) < t_0 + 1$, the type 2 estimate is

$$\begin{aligned}&\mathbb{E}\big[|\sigma(z_{B+1}) - (1-\chi_b^\alpha)|^2 \mid \mathcal{F}_b\big] \\ &\quad\le \frac{1}{16(B+t_0+1)} \max\left\{(t_0+1)(z_1-z_b^\star)^2, \frac{G^2a^2}{(ac/\alpha)(1-\chi_b^\alpha)-1}\right\}.\end{aligned} \tag{16}$$

For $p_b > p_{\mathrm{bd}}$, $\chi_b > 1$ and $h_b(z) \le -c(1-\chi_b^{-1}) < 0$ for all z. For a probability tolerance $\varepsilon \in (0,1)$ and failure level $\bar{\delta} \in (0,1)$, write $[x]_+ \coloneqq \max\{x, 0\}$. The corresponding tail estimate and sufficient batch-length condition are

$$\mathbb{P}\big(\sigma(z_{B+1}) > \varepsilon \mid \mathcal{F}_b\big) \le \exp\left\{-\frac{\left[\log\frac{\varepsilon}{1-\varepsilon} - z_1 + c(1-\chi_b^{-1})\sum_{t=1}^B \tilde{\eta}_t\right]_+^2}{2G^2\sum_{t=1}^B \tilde{\eta}_t^2}\right\}, \tag{17}$$

$$B \ge (t_0+1)\exp\left\{\left[\frac{z_1 - \log\frac{\varepsilon}{1-\varepsilon} + Ga\sqrt{2\log(1/\bar{\delta})/t_0}}{ac(1-\chi_b^{-1})}\right]_+\right\} - t_0 - 1. \tag{18}$$

Lemma 5 (Finite-batch aggregate no-regret tracking)**.** *Suppose Assumptions 1–3 hold, and consider the symmetric finite-population aggregate response and harmonic step sizes specified above. For every integer $B \ge 1$:*

- *(i) If $p_b < p_{\mathrm{bd}}$, the local tracking condition holds, and the stated restrictions on a, t_0 are satisfied, then the mean-square bound* (16) *holds.*
- *(ii) If $p_b > p_{\mathrm{bd}}$, the tail bound* (17) *holds. Under* (18)*, the volunteering probability at the end of the batch is at most ε with probability at least $1-\bar{\delta}$.*

The proof is deferred to Appendix C.3.1. The population dependence is explicit: the type 2 restoring slope is $c((1-\rho)N-1)(1-\chi_b^\alpha)$, whereas the type 1 drift margin is $c(1-\chi_b^{-1}) = c - G(1-p_b)^{N\rho}$. Both approach zero as $\chi_b \to 1$ from their respective branches, so no uniform finite-B tracking guarantee is claimed at the response boundary. Away from it, the type 2 estimate is of order B^{-1} in mean square under local tracking, and (18) quantifies the type 1 batch requirement. These dependencies also show why a batch length chosen for one population size need not provide the same accuracy for another.

For the outer update, batch-average error must be controlled as well as the terminal error. Appendix C.3.1 derives the corresponding bounds, including the initial transient. Under the local type 2 condition, average policy error is of order $B^{-1/2}$, and the induced average LLM-user payoff-gap error is at most $G(1-\rho)N(1-p_b)^{N\rho-1}$ times that error. Thus the finite-batch accuracy relevant to the outer drift depends on both N and B, in addition to the empirical-centering and linearization errors specified in Proposition 1.

C.3.1 Proof of Lemma 5

Proof. Condition throughout on $\mathcal{F}_b$. In the symmetric finite-population aggregate response, the conditional distribution of the next reward gap depends on the past only through z_t and the frozen p_b. Recall its centered noise $\xi_{t+1} = \delta r_t - h_b(z_t)$. Since δr_t takes the two values $-c$ and $G - c$, separated by G,

$$\mathbb{E}[\xi_{t+1} \mid z_t, \mathcal{F}_b] = 0, \qquad |\xi_{t+1}| \le G, \qquad \mathbb{E}[\xi_{t+1}^2 \mid z_t, \mathcal{F}_b] \le G^2/4.$$

The same identities hold conditional on the full pre-period history $\mathcal{H}_{b,t}$, so the noise is a martingale difference. Also, $|\delta r_t| \le G$ because $0 < c < G$.

Type 2 regime. Differentiation gives

$$h_b'(z) = -\frac{G}{\alpha}(1-p_b)^{N\rho}\sigma(z)(1-\sigma(z))^{1/\alpha}.$$

At $z_b^\star$, the equilibrium identity $G(1-p_b)^{N\rho}(\chi_b^\alpha)^{1/\alpha} = c$ gives $h_b'(z_b^\star) = -(c/\alpha)(1-\chi_b^\alpha)$. For the proof, write $e_t \coloneqq z_t - z_b^\star$. Under the local tracking condition, the segment joining z_t and $z_b^\star$ lies in the interval where (15) holds. The mean-value theorem implies

$$2e_t h_b(z_t) \le -\frac{c}{\alpha}(1-\chi_b^\alpha)e_t^2.$$

Expanding the squared update, taking conditional expectation, and using $|\delta r_t| \le G$ yields

$$\mathbb{E}[e_{t+1}^2 \mid z_t, \mathcal{F}_b] \le \left(1 - \frac{c}{\alpha}(1-\chi_b^\alpha)\tilde{\eta}_t\right) e_t^2 + G^2\tilde{\eta}_t^2.$$

For the induction only, denote the maximum on the right-hand side of (16) by K. The base case is $\mathbb{E}[e_1^2 \mid \mathcal{F}_b] \le K/(t_0+1)$. The coefficient of e_t^2 is nonnegative by the restriction on t_0. If $\mathbb{E}[e_t^2 \mid \mathcal{F}_b] \le K/(t+t_0)$, then

$$\begin{aligned}\mathbb{E}[e_{t+1}^2 \mid \mathcal{F}_b] &\le \frac{K}{t+t_0} - \frac{(ac/\alpha)(1-\chi_b^\alpha)K - G^2a^2}{(t+t_0)^2} \\ &\le \frac{K}{t+t_0} - \frac{K}{(t+t_0)^2} \le \frac{K}{t+t_0+1}.\end{aligned}$$

Induction and the $1/4$-Lipschitz property of σ prove (16). This calculation uses almost-sure local tracking as a hypothesis. Conditioning instead on future confinement would not preserve the martingale argument; without the hypothesis, an unconditional localized estimate must retain a neighborhood-exit term.

Type 1 regime. For $p_b > p_{\mathrm{bd}}$, $h_b(z) \le -c(1-\chi_b^{-1}) < 0$ for all z. Summing the updates gives

$$z_{B+1} \le z_1 - c(1-\chi_b^{-1})\sum_{t=1}^{B}\tilde{\eta}_t + \sum_{t=1}^{B}\tilde{\eta}_t\xi_{t+1}.$$

The final sum is a martingale whose increments have absolute value at most $G\tilde{\eta}_t$. Since $\sigma(z_{B+1}) > \varepsilon$ is equivalent to $z_{B+1} > \log(\varepsilon/(1-\varepsilon))$, Azuma–Hoeffding yields (17) when the threshold is positive. Otherwise, the trivial bound one gives the positive-part formulation. For harmonic steps, integral comparison gives

$$\sum_{t=1}^{B}\tilde{\eta}_t \ge a\log\frac{B+t_0+1}{t_0+1}, \qquad \sum_{t=1}^{B}\tilde{\eta}_t^2 \le \frac{a^2}{t_0}.$$

Thus a sufficient condition for the tail probability to be at most $\bar{\delta}$ is

$$ac(1-\chi_b^{-1})\log\frac{B+t_0+1}{t_0+1} \ge z_1 - \log\frac{\varepsilon}{1-\varepsilon} + Ga\sqrt{2\log(1/\bar{\delta})/t_0}.$$

Solving for B, including the case of a nonpositive right-hand side, proves (18).

Batch-average tracking and population dependence. Under the type 2 local tracking condition, Cauchy–Schwarz gives $\mathbb{E}[|\sigma(z_t) - (1-\chi_b^\alpha)| \mid \mathcal{F}_b] \le \sqrt{K}/(4\sqrt{t+t_0})$. Therefore,

$$\begin{aligned}&\frac{1}{B}\sum_{t=1}^{B}\mathbb{E}[|\sigma(z_t) - (1-\chi_b^\alpha)| \mid \mathcal{F}_b] \\ &\quad\le \frac{\sqrt{B+t_0}-\sqrt{t_0}}{2B}\sqrt{\max\left\{(t_0+1)(z_1-z_b^\star)^2, \frac{G^2a^2}{(ac/\alpha)(1-\chi_b^\alpha)-1}\right\}},\end{aligned} \tag{19}$$

where we used $\sum_{t=1}^{B}(t+t_0)^{-1/2} \le 2(\sqrt{B+t_0}-\sqrt{t_0})$. This is an $O(B^{-1/2})$ bound with the displayed population-dependent constant, subject to local tracking.

For $p_b > p_{\mathrm{bd}}$, $\mathbb{E}[\sigma(z_{t+1}) \mid \mathcal{F}_b] \le \varepsilon + \mathbb{P}(\sigma(z_{t+1}) > \varepsilon \mid \mathcal{F}_b)$. Applying (17) to every prefix and bounding the initial policy by one gives

$$\begin{aligned}&\frac{1}{B}\sum_{t=1}^{B}\mathbb{E}[\sigma(z_t) \mid \mathcal{F}_b]\\ &\quad\le \varepsilon + \frac{1}{B} + \frac{1}{B}\sum_{t=1}^{B-1}\exp\left\{-\frac{\left[\log\frac{\varepsilon}{1-\varepsilon} - z_1 + c(1-\chi_b^{-1})\sum_{j=1}^{t}\tilde{\eta}_j\right]_+^2}{2G^2\sum_{j=1}^{t}\tilde{\eta}_j^2}\right\}.\end{aligned} \tag{20}$$

Unlike the terminal bound alone, this estimate retains the initial within-batch transient. Finally, at a common no-regret volunteering probability σ, the LLM-user payoff gap is $G(1-p_b)^{N\rho-1}(1-\sigma)^{(1-\rho)N} - c$. Its derivative in σ has absolute value at most $G(1-\rho)N(1-p_b)^{N\rho-1}$. Multiplying the appropriate policy-error bound by this factor controls the average payoff-gap error. The frozen KL term does not depend on the no-regret policy, so the same error bound applies to the KL-adjusted gap. This controls the fast-response approximation, not the separate empirical centering and Taylor errors in the outer update. □

C.4 Proof of Proposition 1

Section 3.2 defines the branchwise mean drifts f_R obtained by substituting the limiting response from Lemma 1. The corresponding KL-adjusted reward gaps are denoted by Ψ_R in Section 3.3. We now prove Proposition 1.

Proof. Fix a compact subset of one open branch, condition on $\mathcal{F}_b$, and substitute the aggregate response from Lemma 1. Let $s_{n,t} \coloneqq e_{a_{n,t}} - (p_b, 1-p_b)^\top$ be the vector score factor. Then $\mathbb{E}[s_{n,t} \mid \mathcal{F}_b] = 0$ and

$$\mathbb{E}[\tilde{r}_{n,t}s_{n,t} \mid \mathcal{F}_b] = p_b(1-p_b)\Psi_R(p_b)(e_V - e_D).$$

The batch score is

$$S_b \coloneqq \sum_{t=1}^{B}\sum_{n=1}^{N\rho}(\tilde{r}_{n,t} - \bar{r}_b)s_{n,t}.$$

Under the branch-reduced sampling, periods are independent conditional on $\mathcal{F}_b$, while users' payoffs within a period remain strategically linked. Thus the empirical-centering correction is

$$\mathbb{E}\left[\bar{r}_b\sum_{t,n}s_{n,t}\,\middle|\,\mathcal{F}_b\right] = \frac{1}{BN\rho}\sum_{t=1}^{B}\sum_{n,m=1}^{N\rho}\mathbb{E}[\tilde{r}_{n,t}s_{m,t} \mid \mathcal{F}_b] = O(1)$$

for fixed N and ρ. Consequently,

$$\mathbb{E}[S_b \mid \mathcal{F}_b] = BN\rho\, p_b(1-p_b)\Psi_R(p_b)(e_V - e_D) + O(1).$$

The bounds are uniform on the chosen compact subset, where the KL-adjusted rewards are bounded.

Write $\delta_b \coloneqq \mu_{V,b} - \mu_{D,b}$, so $p_b = \sigma(\lambda\delta_b)$. Algorithm 1 gives a score increment bounded by a constant times $\eta_b BN\rho$. The bounded second derivative of the sigmoid therefore yields

$$\begin{aligned}\mathbb{E}[p_{b+1} - p_b \mid \mathcal{F}_b] &= 2\eta_b BN\rho\lambda[p_b(1-p_b)]^2\Psi_R(p_b)\\ &\quad + O(\eta_b) + O\big((\eta_b BN\rho)^2\big).\end{aligned}$$

Dividing by $\gamma_b = \eta_b BN\rho$ gives

$$\frac{\mathbb{E}[p_{b+1} - p_b \mid \mathcal{F}_b]}{\gamma_b} = 2\lambda[p_b(1-p_b)]^2\Psi_R(p_b) + O\left(\frac{1}{BN\rho}\right) + O(\gamma_b).$$

The $O((BN\rho)^{-1})$ term is the conditional-mean bias from using the same batch to estimate the empirical baseline and the policy update. The $O(\gamma_b)$ term is the logit-to-policy Taylor remainder after division by γ_b. For fixed N and ρ, both corrections vanish under the large-batch scaling in Assumption 4: $B \to \infty$ and $\gamma_b = \eta_b BN\rho \to 0$. □

These corrections concern the branch-reduced update; finite-B tracking of the actual no-regret process is a separate approximation treated in Appendix C.3.

We use γ_b as the learning-time increment between successive batch updates, giving the cumulative clock $\tau_b := \sum_{j<b} \gamma_j$. Assumption 4 gives $\gamma_b > 0$, $\gamma_b \to 0$, and $\sum_{b=0}^{\infty} \gamma_b = \infty$, so the clock increases without bound. On this time scale, the leading drift f_R defines the limiting conditional-mean dynamics

$$\frac{dp}{d\tau} = 2\lambda[p(1-p)]^2 \Psi_R(p).$$

C.5 KL GEOMETRY AND THE STRONG-REGULARIZATION LIMIT

The sign of the reduced mean drift is determined by the KL-adjusted expected reward gap Ψ_R derived in Appendix C.4. Its game component reflects contribution incentives, while its KL component pulls the policy toward the pretrained reference.

In type 1, the NR population's limiting volunteering probability is zero. Increasing p makes another guided user's contribution more likely, reducing the focal user's reward difference between V and D. Both the game and KL components decrease with p, so Ψ_1 is strictly decreasing. In type 2, NR volunteering decreases with p to keep V and D equally rewarding. This response makes the guided user's game reward gap increase with p, opposing the KL component. The regularization needed to offset this increase is greatest at the response boundary, giving β_c in Section 3.3.

The two gaps agree at the response boundary. For $\beta \geq \beta_c$, they therefore join into a decreasing function across both regimes. Increasing q_0 raises the gap at the boundary and moves its zero toward higher volunteering. The equation $\Psi_1(p_{bd}) = 0$ defines the cutoff q_{cut}. The calculation below establishes its unique value and the geometry used in Theorem 3.1.

Proposition 2 (Branchwise KL reward-gap geometry)**.** *Suppose Assumption 1 holds, $\beta > 0$, and $0 < q_0 < 1$. With the cutoff q_{cut} and critical level β_c above, the branch gaps have the following geometry.*

- *(i) On $p_{bd} < p < 1$, Ψ_1 is strictly decreasing, tends to $-\infty$ as $p \uparrow 1$, and has at most one zero. Its boundary value is positive, zero, or negative according as q_0 is above, equal to, or below q_{cut}.*
- *(ii) On $0 < p < p_{bd}$, Ψ_2 is continuous and tends to $+\infty$ as $p \downarrow 0$. It is strictly decreasing when $\beta \geq \beta_c$. When $0 < \beta < \beta_c$, it strictly decreases to a unique minimum and then strictly increases.*

The boundary values satisfy $\Psi_2(p_{bd}) = \Psi_1(p_{bd})$.

Proof. Type 1 regime. Lemma 1 sets the aggregate no-regret response to zero on this branch. The game reward gap is $G(1-p)^{N\rho-1} - c$. For a model-policy probability q with effective probability p, the log-odds identity $\log \frac{p}{1-p} = \lambda \log \frac{q}{1-q}$ shows that the effective-policy KL gap is exactly the model-coordinate penalty used in Appendix C.1.5. Thus the branch gap is $\Psi_1(p)$ from Section 3.3. Its derivative is

$$\Psi_1'(p) = -G(N\rho - 1)(1-p)^{N\rho-2} - \frac{\beta}{\lambda p(1-p)} < 0.$$

Since $\Psi_1(p) \to -\infty$ as $p \uparrow 1$, its boundary value determines whether a zero occurs on the open branch.

At the response boundary, the gap equals

$$\Psi_1(p_{bd}) = c\left[\left(\frac{G}{c}\right)^{1/(N\rho)} - 1\right] - \beta\left\{\frac{1}{\lambda}\log\left[\left(\frac{G}{c}\right)^{1/(N\rho)} - 1\right] - \log\frac{q_0}{1-q_0}\right\}.$$

This expression is strictly increasing in q_0 and tends to $-\infty$ and $+\infty$ at the two anchor endpoints. Its unique zero is

$$q_{\mathrm{cut}} = \frac{\left(\left(\frac{c}{G}\right)^{-1/(N\rho)} - 1\right)^{1/\lambda} \exp\left[-\frac{c}{\beta}\left(\left(\frac{c}{G}\right)^{-1/(N\rho)} - 1\right)\right]}{1 + \left(\left(\frac{c}{G}\right)^{-1/(N\rho)} - 1\right)^{1/\lambda} \exp\left[-\frac{c}{\beta}\left(\left(\frac{c}{G}\right)^{-1/(N\rho)} - 1\right)\right]}.$$

The boundary gap is positive, zero, or negative according as q_0 is above, equal to, or below q_{cut}. Strict decrease of Ψ_1 gives part (i). A positive boundary gap gives a unique open-branch zero, with positive gap below it and negative gap above it, so this zero is attracting. A zero boundary gap places the zero at the response boundary; a negative boundary gap makes the gap negative throughout the type 1 regime. At the boundary, $\chi(p_{\mathrm{bd}}) = 1$, so the type 2 and type 1 gap formulas agree.

Type 2 regime. Substitute $\sigma^\star(p) = 1 - \chi(p)^\alpha$ from Lemma 1 into the LLM user's defect payoff. The resulting game gap is $-c + G(1-p)^{N\rho-1}\chi(p)^{N(1-\rho)\alpha}$, which yields $\Psi_2(p)$ after the same KL adjustment. For this proof, set

$$K \coloneqq G\left(\frac{c}{G}\right)^{\frac{N(1-\rho)}{N(1-\rho)-1}}, \qquad \nu \coloneqq 1 + \frac{N\rho}{N(1-\rho)-1} > 1.$$

Then

$$\Psi_2(p) = -c + K(1-p)^{-\nu} - \frac{\beta}{\lambda}\log\frac{p}{1-p} + \frac{\beta}{\lambda}\log\frac{p_0}{1-p_0},$$

and

$$\Psi_2'(p) = K\nu(1-p)^{-\nu-1} - \frac{\beta}{\lambda p(1-p)}.$$

The derivative vanishes exactly when $K\nu p(1-p)^{-\nu} = \beta/\lambda$. The left-hand side is strictly increasing, since its derivative is $K\nu(1-p)^{-\nu-1}[1 + (\nu-1)p] > 0$. At the boundary it equals

$$K\nu p_{\mathrm{bd}}(1-p_{\mathrm{bd}})^{-\nu} = c\left(1 + \frac{N\rho}{N(1-\rho)-1}\right)\left[\left(\frac{G}{c}\right)^{1/(N\rho)} - 1\right] = \frac{\beta_{\mathrm{c}}}{\lambda}.$$

For $\beta \geq \beta_{\mathrm{c}}$, $\Psi_2'(p) < 0$ in the open type 2 regime, including when equality holds at the boundary. Since $\Psi_2(0^+) = +\infty$, the common boundary sign gives the three strong-KL cases.

For $0 < \beta < \beta_{\mathrm{c}}$, the same derivative calculation gives a unique interior minimum, with strict decrease to its left and strict increase to its right. This completes part (ii).

The mean-flow prefactor is positive on $(0, 1)$, so its direction agrees with the reward-gap sign. The gaps agree at the response boundary, and uniqueness of scalar ODE solutions prevents crossing a fixed point. In the strong-KL case, these signs give convergence to the unique attracting zero, including the boundary zero when $q_0 = q_{\mathrm{cut}}$. This proves Theorem 3.1. □

The displayed critical ratio $\beta_{\mathrm{c}}/\lambda$ is independent of λ. Responsiveness converts between model-scale and effective-policy regularization and also changes the effective initial policy and the speed of the mean flow.

D Additional Dynamic Cases

D.1 Full KL phase portrait

Proposition 2 supplies the branchwise shapes, and Theorem 3.1 handles strong regularization. We now state the weak regularization result and describe its basins.

In the weak-KL case, a zero minimum gives the tangential zero $p^\star_{\mathrm{tan}}$. When the minimum is negative, write $p^\star_{\mathrm{lo}}$ for the lower zero and $p^\star_{\mathrm{hi}}$ for the upper zero when it exists. The upper zero lies in the open type 2 regime exactly when the boundary gap is positive. These shapes give the following long-run classification.

Theorem D.1 (Convergence under weak regularization)**.** *Suppose Assumption 1 holds,* $0 < \beta < \beta_c$*, and* $0 < q_0 < 1$*. The reduced mean dynamics in Section 3.2 are continuous across* p_{bd} *and converge to a fixed point* $p^\star \in (0, 1)$*. The type 2 reward gap decreases to a unique minimum and then increases, so multiple fixed points are possible. Starting from the pretrained effective policy* p_0*, the trajectory moves in the direction of its reward gap and converges to the first fixed point encountered. If* p_0 *is itself a fixed point, the trajectory remains there.*

Proof. Appendix C.5 establishes the branchwise geometry and strong regularization result of Theorem 3.1. For weaker regularization, we describe the basins from any starting point in $(0, 1)$ and then apply them to the initialization $p(0) = p_0$. The positive mean drift prefactor preserves the reward gap signs, and uniqueness prevents trajectories from crossing a zero.

For $0 < \beta < \beta_c$, there is a unique minimizer $p^{\min} \in (0, p_{bd})$, with negative derivative to its left and positive derivative to its right. If $\Psi_2(p^{\min}) > 0$, the gap is positive throughout the branch and the boundary gap is positive, so trajectories enter the type 1 regime.

If $\Psi_2(p^{\min}) = 0$, the minimizer is the tangential zero $p^\star_{tan}$. The gap is positive on either side of it, making the tangency attracting from the left and repelling from the right. Trajectories starting to its left increase toward it, and uniqueness of the scalar ODE prevents crossing; trajectories starting to its right move toward the type 1 attractor. The trajectory at the tangency is stationary.

If $\Psi_2(p^{\min}) < 0$, continuity and strict decrease to the minimum give exactly one lower zero $p^\star_{lo}$, with sign change from positive to negative. On the increasing part of the branch, a second zero exists exactly when the boundary gap is positive. This upper zero $p^\star_{hi}$ has a negative-to-positive sign change, is repelling, and separates the lower attractor from the type 1 attractor.

If the boundary gap is zero, there is no second zero in the open type 2 regime. The gap is negative between $p^\star_{lo}$ and p_{bd}, and the type 1 gap is negative to the right of the boundary. Hence the boundary repels trajectories on its left and attracts those on its right. If the boundary gap is negative, there is only the lower zero, and trajectories from both branches converge to it. This proves Theorem D.1, including the boundary and stationary cases. □

The type 2 regime places more of the provision burden on no-regret users, consistent with a self-serving tendency of the advisor. Within that branch, increasing guided-user volunteering reduces the no-regret contribution response and raises the guided user's game reward gap. Once no-regret volunteering reaches zero, additional guided-user volunteering makes provision by another guided user more likely and lowers the focal user's return to contributing. The type 1 regime thus supports a more cooperative tendency among guided users, who bear the provision costs there.

D.2 Unregularized learning and the symmetric Nash benchmark

Setting $\beta = 0$ retains the same game, shared-policy update, conditional sampling, and population response, while removing the KL reference penalty. The initial policy still determines the starting state $p(0) = p_0$; the reference no longer enters the drift. Which contribution patterns do game payoffs alone generate, and does the full-population symmetric mixed Nash benchmark attract the shared update? We use the branch-reduced mean ODE associated with Proposition 1 to answer these questions.

Type 1 regime. With $\beta = 0$, the gap in Section 3.3 reduces to $\Psi_1(p) = G(1-p)^{N\rho-1} - c$.

Lemma 6 (LLM drift in the type 1 regime)**.** *Suppose Assumption 1 holds and* $\beta = 0$*, and consider the branchwise mean ODE associated with Proposition 1. The type 1 regime is forward-invariant and has the unique attracting fixed point*

$$p^\star = 1 - \left(\frac{c}{G}\right)^{1/(N\rho-1)}.$$

Proof of Lemma 6. The gap Ψ_1 is strictly decreasing and vanishes at $p^\star = 1-(c/G)^{1/(N\rho-1)}$. Since $0 < c/G < 1$ and $N\rho - 1 < N\rho$, this zero lies strictly above p_{bd}. The gap is positive between the response boundary and $p^\star$, and negative between $p^\star$ and 1. The mean ODE associated with Proposition 1 has a positive prefactor on $(0, 1)$, so trajectories in the interior of the type 1 regime

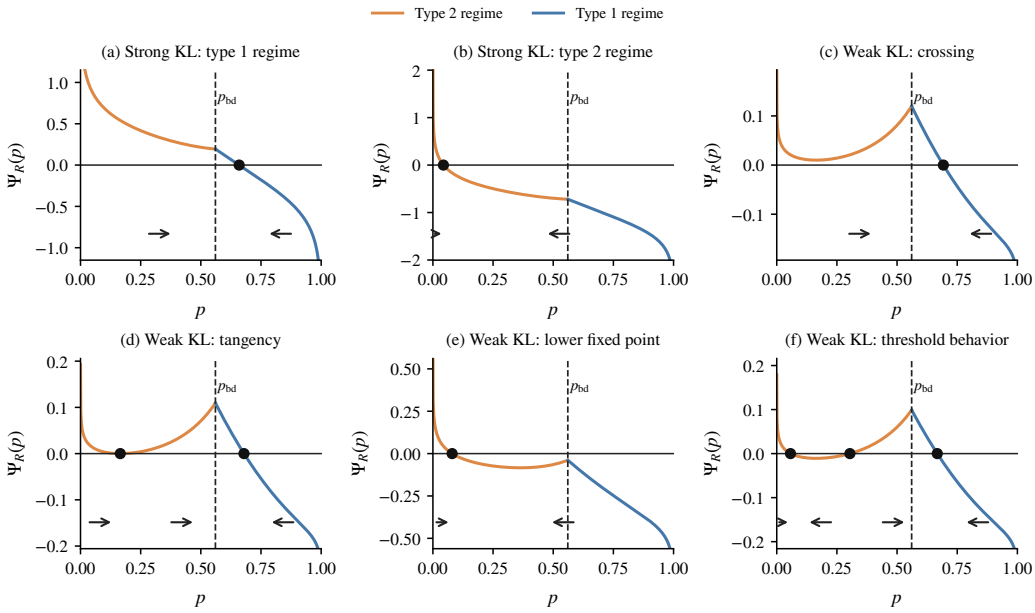

Figure 6: KL-adjusted reward-gap geometries. The vertical axis is $\Psi_R(p)$; the positive probability prefactor in the complete drift is omitted. Logarithmic endpoint tails extend beyond the displayed vertical ranges. The horizontal coordinate is the effective volunteering probability p. The dashed line marks the response boundary p_{bd}, black dots mark zeros, and arrows show the mean-flow direction. Panels (a)–(b) illustrate the monotone outcomes for anchors above and below q_{cut}. Panels (c)–(f) show weak-KL crossing, a tangency attracting from the left, a single attracting interior zero, and an upper interior basin boundary. Proposition 2 states the reward-gap shapes; Theorem 3.1 gives the strong regularization limits, while Theorem D.1 and its proof in Appendix D.1 give the weak regularization limits, including the response-boundary equality cases.

move toward this zero. The positive boundary drift and uniqueness of the scalar flow prevent exit through the response boundary; the vanishing endpoint prefactor prevents exit through 1. Thus the branch is forward-invariant and its zero is attracting. □

Type 2 regime. The full-population symmetric mixed Nash volunteering probability is $p_{\mathrm{NE}} := 1 - (c/G)^{1/(N-1)}$. Assumption 1 gives $N - 1 > N\rho$, hence $0 < p_{\mathrm{NE}} < p_{\mathrm{bd}}$.

Lemma 7 (LLM drift with an interior aggregate no-regret response)**.** *Suppose Assumption 1 holds and $\beta = 0$, and consider the branchwise mean ODE associated with Proposition 1. In the type 2 regime, the drift is negative for $p < p_{\mathrm{NE}}$ and positive for $p > p_{\mathrm{NE}}$. Hence p_{NE} is an unstable fixed point separating the two directions of motion.*

Proof of Lemma 7. Setting $\beta = 0$ in Ψ_2 and collecting powers gives

$$\Psi_2(p) = -c + G\left(\frac{c}{G}\right)^{\frac{N(1-\rho)}{N(1-\rho)-1}} (1-p)^{\frac{1-N}{N(1-\rho)-1}}.$$

The exponent of $1 - p$ is negative, so this gap is strictly increasing. Its zero satisfies

$$(1-p)^{\frac{N-1}{N(1-\rho)-1}} = \left(\frac{c}{G}\right)^{\frac{1}{N(1-\rho)-1}},$$

which is equivalent to $p = p_{\mathrm{NE}}$. The gap is therefore negative below the Nash point and positive above it. The positive mean-flow prefactor on the open branch preserves these signs, proving instability. □

Global outcomes. The two attracting unregularized limit profiles are

$$(p_1^\star, \sigma_1^\star) \coloneqq \left(1 - \left(\frac{c}{G}\right)^{\frac{1}{N\rho - 1}}, 0\right), \tag{21}$$

$$(p_2^\star, \sigma_2^\star) \coloneqq \left(0, 1 - \left(\frac{c}{G}\right)^{\frac{1}{(1-\rho)N - 1}}\right). \tag{22}$$

Theorem D.2 (Asymptotic regimes of the unregularized mean flow)**.** *Suppose Assumption 1 holds, consider the branchwise mean ODE associated with Proposition 1, and let $\beta = 0$. Its asymptotic outcomes are as follows.*

(i) If $p_0 \geq p_{\mathrm{bd}}$, then the mean flow converges to the type 1 limit in (21)*, with the boundary case $p_0 = 1$ yielding $(p^\star, \sigma^\star) = (1, 0)$.*

(ii) If $p_0 < p_{\mathrm{NE}}$, then the mean flow converges to the type 2 limit in (22)*.*

(iii) If $p_0 = p_{\mathrm{NE}}$, then the mean flow remains at this stationary but unstable threshold.

(iv) If $p_{\mathrm{NE}} < p_0 < p_{\mathrm{bd}}$, then the trajectory first crosses into the type 1 regime and then converges to the limit in (21)*.*

Proof. Below p_{NE}, a nonstationary trajectory decreases to zero: any positive limit would have strictly negative drift. Above p_{NE} and below p_{bd}, the complete drift is continuous and strictly positive on $[p_0, p_{\mathrm{bd}}]$, so its positive minimum gives finite-time crossing of the response boundary. Lemma 6 then gives convergence in the invariant type 1 regime, including for trajectories starting at the response boundary. The Nash point is stationary, and the vanishing probability prefactor keeps the pure initial states $p_0 = 0$ and $p_0 = 1$ fixed. Substitution into the aggregate response gives the stated limiting profiles. □

The two attracting profiles place contribution costs on the no-regret population or on guided users, respectively. Appendix D.3 examines how the KL reference and penalty strength affect the stationary Nash point and its local stability.

D.3 Nash stationarity and local stability under KL regularization

The payoff-only symmetric mixed Nash point is unstable in the branch-reduced mean flow. With $\beta > 0$, which reference policies make it stationary, and when is it locally attracting? The reference condition makes the Nash point a zero of the drift; the derivative at that zero determines local stability.

The symmetric Nash volunteering propensity is $p_{\mathrm{NE}} = 1 - (c/G)^{1/(N-1)}$. The corresponding model-policy anchor is

$$q_{\mathrm{NE}} \coloneqq \frac{\left(\left(\frac{G}{c}\right)^{\frac{1}{N-1}} - 1\right)^{1/\lambda}}{1 + \left(\left(\frac{G}{c}\right)^{\frac{1}{N-1}} - 1\right)^{1/\lambda}}.$$

This mapping preserves the distinction between model and effective probabilities when $\lambda \neq 1$.

Lemma 8 (Necessary pretrained anchor for Nash-equilibrium convergence)**.** *Suppose Assumption 1 holds, consider the branchwise mean ODE associated with Proposition 1, and let $\beta > 0$. If this mean flow converges to the symmetric mixed Nash equilibrium p_{NE}, then necessarily $q_0 = q_{\mathrm{NE}}$.*

Appendix D.3.1 proves the claim by evaluating the remaining KL term at the zero of the game reward gap.

For local stability, hold the reference $q_0 = q_{\mathrm{NE}}$ and the resulting vector field fixed while perturbing the initial effective state. The derivative changes sign at

$$\beta_{\mathrm{NE}} \coloneqq \lambda c \left(1 + \frac{N\rho}{N(1-\rho) - 1}\right)\left(1 - \left(\frac{c}{G}\right)^{\frac{1}{N-1}}\right). \tag{23}$$

At this value of β, the Nash point coincides with the local minimum of the type 2 reward gap under the exact Nash anchor.

Theorem D.3 (Exact Nash anchor and Nash stability)**.** *Suppose Assumption 1 holds and consider the branchwise mean ODE associated with Proposition 1. If $0 < \beta < \beta_c$ and $q_0 = q_{\mathrm{NE}}$, then p_{NE} is a zero of the type 2 mean drift. Moreover:*

(i) If $0 < \beta < \beta_{\mathrm{NE}}$, then p_{NE} is the upper zero $p^\star_{\mathrm{hi}}$ of the two-zero geometry. It is an unstable threshold, so nearby trajectories are pushed away from Nash equilibrium.

(ii) If $\beta = \beta_{\mathrm{NE}}$, then p_{NE} is the tangential zero of the type 2 drift. This point is not a robust attracting limit, so exact anchoring alone still does not guarantee Nash convergence.

(iii) If $\beta_{\mathrm{NE}} < \beta < \beta_c$, then p_{NE} is the attracting lower zero $p^\star_{\mathrm{lo}}$. Mean-flow trajectories initialized in its basin converge to Nash equilibrium; if an upper zero also lies in the type 2 regime, that upper zero is the threshold separating this Nash basin from trajectories that move toward the competing regime.

Appendix D.3.2 gives the derivative and basin arguments. Under the specified initialization $p(0) = p_0$, the exact reference $q_0 = q_{\mathrm{NE}}$ also initializes the effective state at p_{NE}. Its deterministic trajectory remains stationary, including in the unstable and tangential cases. Local attraction concerns the nearby initial states described above. The monotone strong-KL case is covered by Theorem 3.1.

D.3.1 Proof of Lemma 8

Proof. The Nash point lies in the type 2 regime, and at this point the game-theoretic part of the type 2 drift vanishes. Substituting p_{NE} into Ψ_2 therefore leaves only the KL-anchor term:

$$\Psi_2(p_{\mathrm{NE}}) = -\frac{\beta}{\lambda} \log \left(\left(\frac{G}{c} \right)^{\frac{1}{N-1}} - 1 \right) + \beta \log \frac{q_0}{1-q_0}.$$

If the limiting mean flow converges to p_{NE}, this point must be a zero of the limiting drift. Hence $\Psi_2(p_{\mathrm{NE}}) = 0$. Because $\beta > 0$, this condition is equivalent to

$$\frac{q_0}{1-q_0} = \left(\left(\frac{G}{c} \right)^{\frac{1}{N-1}} - 1 \right)^{1/\lambda}.$$

Solving for q_0 gives

$$q_0 = \frac{\left(\left(\frac{G}{c} \right)^{\frac{1}{N-1}} - 1 \right)^{1/\lambda}}{1 + \left(\left(\frac{G}{c} \right)^{\frac{1}{N-1}} - 1 \right)^{1/\lambda}} = q_{\mathrm{NE}},$$

where the last equality uses the definition of q_{NE}. This proves the claim. □

D.3.2 Proof of Theorem D.3

Proof. Because $0 < \beta < \beta_c$, Proposition 2 is in the non-monotone case where the type 2 drift may have a tangency or two-zero geometry. When $q_0 = q_{\mathrm{NE}}$, the calculation in the proof of Lemma 8 gives $\Psi_2(p_{\mathrm{NE}}) = 0$. Thus the Nash point is a fixed point of the limiting type 2 drift. The remaining question is whether it is attracting.

The proof of Proposition 2 characterizes the unique critical point of Ψ_2, when it exists, by $K\nu p(1-p)^{-\nu} = \beta/\lambda$, where

$$K = G \left(\frac{c}{G} \right)^{\frac{N(1-\rho)}{N(1-\rho)-1}}, \qquad \nu = 1 + \frac{N\rho}{N(1-\rho)-1}.$$

The left-hand side is strictly increasing in p. Evaluating it at $p = p_{\mathrm{NE}} = 1 - (c/G)^{1/(N-1)}$ gives $\beta_{\mathrm{NE}}/\lambda$, with β_{NE} defined in (23). Therefore, $\beta = \beta_{\mathrm{NE}}$ is precisely the condition under which the local minimum of Ψ_2 is located at the Nash point.

First, if $0 < \beta < \beta_{\mathrm{NE}}$, the critical point lies to the left of p_{NE}. Since $\Psi_2(0^+) = +\infty$ and $\Psi_2(p_{\mathrm{NE}}) = 0$, the value at the local minimum must be negative. Hence the Nash point is the upper

zero $p^\star_{\mathrm{hi}}$ of the two-zero geometry. The sign pattern is negative immediately to the left of $p^\star_{\mathrm{hi}}$ and positive immediately to the right, so p_{NE} is an unstable threshold rather than an attracting limit.

Second, if $\beta = \beta_{\mathrm{NE}}$, then p_{NE} coincides with the local minimum of Ψ_2. Because the minimum value is zero, the Nash point is the tangential zero described in Appendix D.1. The drift does not change from positive to negative across this point, so the tangency is not a robust attracting equilibrium.

Third, if $\beta_{\mathrm{NE}} < \beta < \beta_{\mathrm{c}}$, the critical point lies to the right of p_{NE}. Hence the drift changes from positive to negative at the Nash point, so p_{NE} is the attracting lower zero $p^\star_{\mathrm{lo}}$. If an upper interior zero exists, it is the basin boundary. If the reward gap is zero at the response boundary, that boundary separates the Nash basin from trajectories approaching the boundary from the right. If the boundary gap is negative, both branches converge to the Nash point.

The comparison between β_{NE} and β_{c} follows from

$$1 - \left(\frac{c}{G}\right)^{\frac{1}{N-1}} < \left(\left(\frac{G}{c}\right)^{\frac{1}{N\rho}} - 1\right),$$

which holds under Assumption 1. Therefore $\beta_{\mathrm{NE}} < \beta_{\mathrm{c}}$, so the cases above are ordered consistently. □

E EXPERIMENTAL DETAILS

E.1 QWEN3-8B CONVERGENCE PROTOCOL

Figure 5 in the main text compares final policies. Figure 7 adds their full training histories.

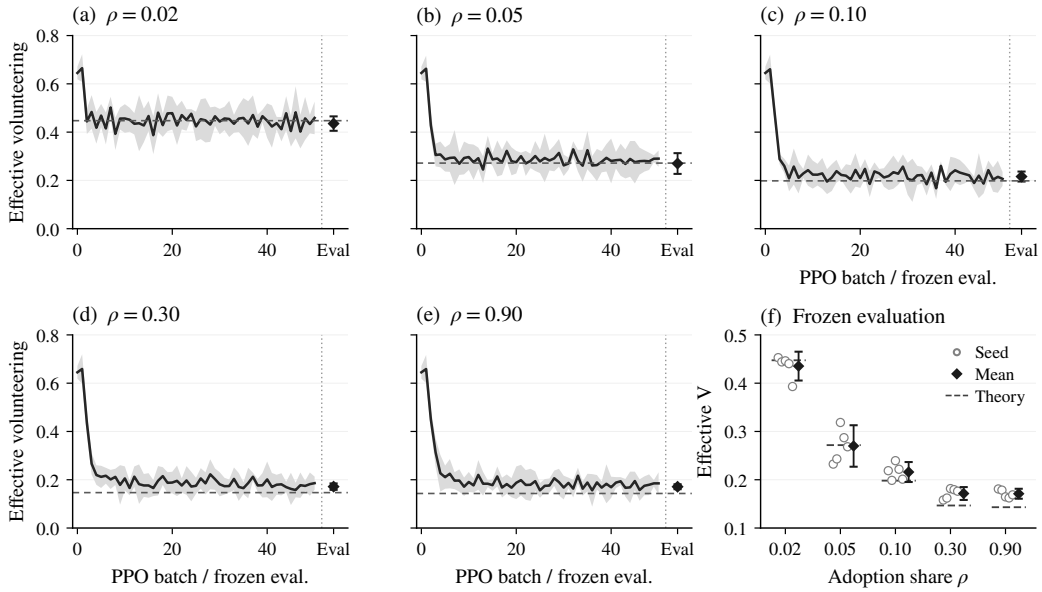

Figure 7: Qwen3-8B training trajectories and frozen-policy outcomes. Panels (a)–(e) show effective volunteering probabilities, with seed means (solid lines), pointwise 95% Student-t intervals (shading), and fixed-anchor theoretical roots (dashed lines). Batch 0 is a pre-game probe; batches 1–50 use pre-update response pools weighted by their use in the game. Diamonds after the dotted separators show independent frozen evaluations. Panel (f) retains all five final seed estimates at each penetration rate (open circles), their means and intervals (diamonds), and the theoretical roots (dashed segments).

Initialization and feedback training. The experiment uses `Qwen/Qwen3-8B-MLX-4bit`, distributed under Apache-2.0, with LoRA feedback training in MLX.[2] Rank-8 adapters with scale 20 and zero dropout are applied to the final four Transformer layers, giving 2,424,832 trainable parameters out of 8,190,735,360 total parameters (0.0296%). The output head and remaining base-model weights stay frozen. A common seed-0 adapter is obtained by ten epochs of natural-language SFT with six volunteer and six defect demonstrations, learning rate 10^{-5}, and minibatch size 4. The same saved adapter initializes all 25 runs and supplies their frozen reference. The five seeds are $101, 202, 303, 404, 505$, reused across the five penetration rates. Their variation reflects language generation, game sampling, and PPO conditional on one SFT checkpoint.

The prompt describes the contribution decision in natural language. Responses contain at most 24 generated tokens at temperature 1 and are classified with the third version of the semantic parser. Each PPO batch uses 128 new responses, resampled across guided-player decisions during the $B = 300$ game periods. Each run therefore uses 6,400 training generations over 15,000 game periods. The no-regret learning rate is $0.5/t^{0.6}$. Feedback training uses two PPO epochs per batch, clip parameter 0.2, minibatch size 4, learning rate 5×10^{-7}, and gradient-norm cap 1. Invalid outputs receive a $0.25G$ format penalty. The reference penalty subtracts 0.35 times the sum of generated-token log-probability ratios from each response's reward. The structured comparison sets $\lambda = 1$, so its effective-policy regularization coefficient is $\beta/\lambda = 0.35$.

[2]Official checkpoint and model card: `https://huggingface.co/Qwen/Qwen3-8B-MLX-4bit`.

Policy measurement and theoretical comparison. The initial SFT evaluation contains 659 volunteer, 278 defect, and 87 invalid recommendations among 1,024 generations. Counting half of the invalid outputs as effective volunteering gives the fixed anchor $q_0 = 0.68603515625$. Per-run initialization probes measure the common policy again and provide stage 0 in Figure 7. The structured roots use this initial estimate throughout. For each penetration rate, the analysis computes the attracting zero of the KL-adjusted structured reward gap from the recorded game parameters. All five settings have a unique attracting root in the type 1 regime. This calculation is performed after training.

The training trajectories measure semantic volunteering plus half the invalid frequency, weighting each response by its number of uses in that batch. The frozen endpoint applies the same action mapping to 2,048 fresh generations with equal weights. Figure 5 and panel (f) of Figure 7 compare these endpoints with the structured roots and retain every seed. A further 6,000 game periods provide environmental diagnostics, with the final 3,000 periods used for game-tail summaries. These game periods resample the frozen response pool.

For five seeds, the pointwise 95% interval uses the sample standard deviation multiplied by $2.776/\sqrt{5}$. The reported theory comparison holds the reference estimate q_0 fixed. Absolute discrepancies are calculated within each run before averaging. The root-mean-square distance across all 25 runs is 2.70 percentage points. Figure 7 displays the recorded batch sequence without smoothing and separates the frozen endpoint from the training observations. Table 5 summarizes discrepancies by penetration rate, while Table 6 retains the individual frozen estimates.

For the compact display in Figure 5, we recover the 25 frozen estimates from the archived vector figure and check them against Table 6. The replot preserves the archived means and Student-t intervals. The extraction records the source checksum and numerical verification errors.

Table 5: Qwen3-8B frozen-evaluation discrepancies across five seeds per penetration rate. Mean gap is the signed difference from the theoretical root; MAE averages absolute differences within runs. Both are in percentage points (pp). Invalid rates are measured before random fallback. Figure 5 shows the probability levels and uncertainty.

ρ	Mean gap (pp)	MAE (pp)	Invalid (%)
0.02	-1.21	1.43	7.71
0.05	-0.18	2.67	6.27
0.10	+1.78	1.78	7.06
0.30	+2.48	2.48	6.15
0.90	+2.79	2.79	6.76

Recorded game outcomes. Table 7 summarizes the final 3,000 periods of the frozen-policy game evaluation. Provision and payoffs are reconstructed from the recorded joint actions. Total welfare is divided by NG, and the guided-user payoff by G. These observed outcomes describe play under the learned policies; the analytical social welfare gap compares total welfare under two separately optimized policies at each penetration rate.

Late-training diagnostics. Table 8 summarizes batches 41–50. Within each run, an ordinary least-squares slope measures the late-batch trend, and the frozen endpoint is compared with the mean of these ten batches. Both diagnostics are then summarized across five seeds. The slope intervals include zero at every penetration rate, and the frozen evaluations are compatible with the late-training means. Together with the trajectories, these measurements support stabilization over the observed training horizon; the positive final offsets at higher penetration rates remain visible in Table 5.

Table 6: Qwen3-8B frozen outcomes for all 25 runs. Effective V uses 2,048 fresh generations per run and the common random-fallback mapping. Absolute gaps are measured against the theoretical root at fixed q_0 in percentage points (pp).

ρ	Seed	Theory	Effective V	Abs. gap (pp)	Invalid (%)
0.02	101	0.447	0.453	0.54	6.59
0.02	202	0.447	0.444	0.33	7.76
0.02	303	0.447	0.446	0.11	8.01
0.02	404	0.447	0.440	0.70	8.40
0.02	505	0.447	0.393	5.44	7.81
0.05	101	0.272	0.232	3.93	5.27
0.05	202	0.272	0.243	2.85	5.96
0.05	303	0.272	0.319	4.69	7.08
0.05	404	0.272	0.287	1.54	6.84
0.05	505	0.272	0.268	0.34	6.20
0.10	101	0.198	0.219	2.06	6.74
0.10	202	0.198	0.199	0.06	6.35
0.10	303	0.198	0.240	4.14	7.96
0.10	404	0.198	0.222	2.36	8.11
0.10	505	0.198	0.201	0.31	6.15
0.30	101	0.147	0.158	1.13	5.32
0.30	202	0.147	0.162	1.54	6.05
0.30	303	0.147	0.182	3.50	6.64
0.30	404	0.147	0.179	3.28	7.28
0.30	505	0.147	0.176	2.96	5.47
0.90	101	0.143	0.181	3.78	6.40
0.90	202	0.143	0.178	3.54	6.79
0.90	303	0.143	0.165	2.15	7.32
0.90	404	0.143	0.162	1.93	8.06
0.90	505	0.143	0.169	2.56	5.22

Table 7: Recorded Qwen3-8B game outcomes in the frozen-evaluation tail. Entries give means $\pm$ 95% Student-t interval half-widths across five seeds. Each run contributes one average over periods 3,001–6,000; these periods resample its frozen response pool.

ρ	Provision rate	Guided payoff $/G$	Total welfare $/(NG)$
0.02	0.682 ± 0.026	0.288 ± 0.005	0.674 ± 0.025
0.05	0.786 ± 0.061	0.544 ± 0.019	0.774 ± 0.059
0.10	0.911 ± 0.023	0.717 ± 0.007	0.891 ± 0.021
0.30	0.996 ± 0.002	0.843 ± 0.009	0.950 ± 0.001
0.90	1.000 ± 0.000	0.846 ± 0.010	0.862 ± 0.009

Table 8: Late-training diagnostics for Qwen3-8B. Entries give seed means $\pm$ 95% Student-t interval half-widths. Slopes use batches 41–50; the final column subtracts their mean effective volunteering rate from the independent frozen endpoint within each run.

ρ	Tail slope (pp/batch)	Frozen $-$ tail mean (pp)
0.02	$+0.223 \pm 0.680$	-0.34 ± 4.34
0.05	$+0.075 \pm 0.187$	-1.22 ± 4.19
0.10	-0.168 ± 0.378	$+0.17 \pm 2.63$
0.30	$+0.089 \pm 0.661$	-0.24 ± 1.37
0.90	-0.026 ± 0.225	-0.83 ± 2.25

E.2 Finite-batch sampling protocol

We validate the type 1 welfare formula using paired action rollouts at the user and social optima, with $N = 12$, $G = 1$, and $c = 0.18$. The no-regret population defects throughout. Across nine

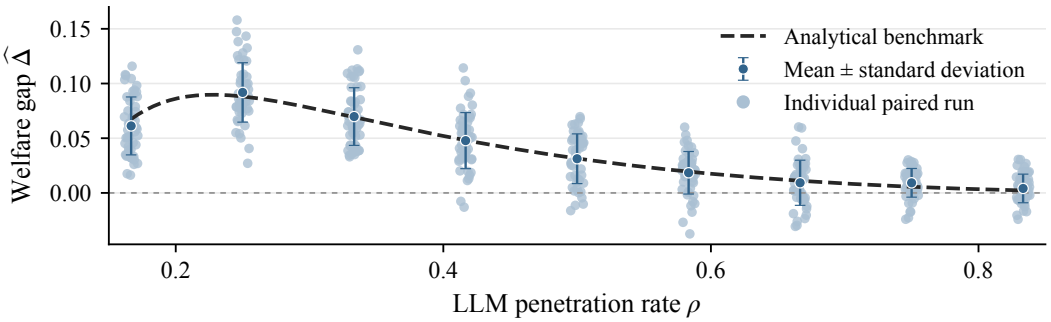

Figure 8: **Welfare gap at the fixed type 1 user and social optima** for $N = 12$, $G = 1$, and $c = 0.18$. Each of nine penetration rates uses 50 seeds and 2,000 periods per seed. Pale points show all 450 estimates, blue markers and bars show means $\pm$ one sample standard deviation, and the dashed curve is the analytical gap.

integer guided population sizes and 50 seeds per setting, the largest absolute discrepancy between the empirical mean and analytical gap is 0.0063 in units of G (Figure 8). These fixed policy rollouts isolate sampling error. The errors of a learning update are analyzed in Appendix C, while Section 5.2 compares learned Qwen policies.

The generator uses NumPy seeds $10000, \ldots, 10049$. For each seed and each guided-population size $N\rho = 2, \ldots, 10$, it draws a $2000 \times N\rho$ array of independent uniform random variables. Thresholding the same array at p_1^{usr} and p_1^{soc} gives paired volunteer counts. Total welfare in each period is N times the provision indicator minus c times the volunteer count, with $G = 1$ and $c = 0.18$. The period-average social-minus-user welfare difference estimates Δ. All 50 seeds enter each mean. Figure 8 shows one sample standard deviation across seeds, computed with denominator 49, on either side of the mean. The archived observations were recovered from the vector figure using the procedure described below. Only $B = 2000$ is studied.

The normalized illustration uses $\Delta / \max_{\rho'} \Delta$, with the maximum evaluated over $[2/N, 1 - 2/N]$. The same analytical normalizer is applied to every sampled gap and its sample standard deviation. This ratio measures the social welfare gap relative to its analytical maximum at the given cost; raw total welfare is measured by the period payoffs above.

E.3 Open-language paired welfare protocol

The experiment in Figure 2(b) uses the 4-bit Qwen3-8B model with LoRA/PPO feedback training, natural-language game descriptions and unrestricted natural-language recommendations up to 24 tokens at temperature 1. The semantic action mapping and independent uniform fallback for invalid recommendations are the same as in Appendix E.1. It fixes $N = 200$, $G = 1$, $c = 0.18$, $\beta = 0.35$, and $\lambda = 1$. The penetration-rate grid is $\{0.01, 0.02, \ldots, 0.10\} \cup \{0.15, 0.20, \ldots, 0.55\}$. Seeds $101, 202, 303, 404, 505$ are paired across the two reference arms at each rate through 0.45; rates 0.50 and 0.55 use seeds $101, 202, 303$. This yields 182 trained-policy evaluations and 91 paired welfare differences. The follow-up grid and reference refinements were chosen after inspecting earlier runs, so this is an exploratory comparison.

The user and social labels identify the two theory-targeted initial references, not different PPO rewards. Both arms learn from guided users' individual payoffs with the same token-level reference penalty. Each run uses 50 batches of $B = 300$ game periods, initially 128 fresh language responses per batch, with additional sampling up to 4,096 when required for action support. PPO uses two epochs, minibatch size 4, learning rate 5×10^{-7}, reference coefficient 0.35, and invalid-output penalty $0.25G$. Each selected SFT adapter initializes its policy and remains its frozen reference. The calibrated anchors approximate the ideal optima; nominal calibration misses for the user-reference arm at $\rho = 0.02$ and 0.50 are retained rather than excluded.

Frozen evaluation draws 1,024 new responses per trained policy and resamples this fixed response pool during 6,000 game periods. Ordinary players restart with volunteering probability $1/2$ and

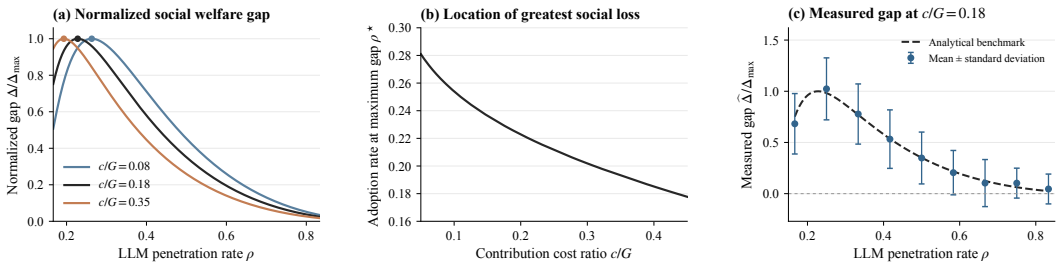

Figure 9: Normalized social welfare gap and the location of its maximum. Panel (a) fixes $N = 12$ and plots $\Delta/\max_{\rho'} \Delta$ against the LLM penetration rate ρ for $c/G \in \{0.08, 0.18, 0.35\}$. Each curve is divided by its own maximum, so the panel compares shapes; Figure 2(a) shows the unnormalized gaps. Dots mark maximizing adoption levels $\rho^\star$. Panel (b) shows how the maximizing adoption level varies with c/G. An interior maximum moves toward lower adoption as costs rise (Corollary 4.1). Panel (c) uses the paired estimates from Figure 8, with $c/G = 0.18$, 50 seeds per setting, and 2000 periods per seed. Sampled means and sample standard deviations are divided by the same analytical maximum. Bars show means $\pm$ one sample standard deviation, and the dashed curve is the normalized analytical gap. Larger values in panels (a) and (c) indicate larger social losses relative to this normalizer.

update with step size $0.5/t^{0.6}$. Total welfare is reconstructed from joint actions over periods 3,001–6,000. This restart and finite horizon need not realize the vanishing no-regret contribution assumed by the type 1 limiting formula. Thus a welfare discrepancy may reflect the evaluation transient as well as learned-policy and anchor differences. The ideal analytical curve is a benchmark, not the data-generating formula for the observed welfare estimates.

For paired seed j, let $\widehat{\Delta}_j$ be the total welfare under the social reference arm minus that under the user reference arm, divided by G. At each rate we plot the mean $\overline{\Delta}$ and sample standard deviation

$$s_\Delta := \left[\frac{1}{s-1} \sum_{j=1}^{s} (\widehat{\Delta}_j - \overline{\Delta})^2 \right]^{1/2}, \qquad s \in \{3, 5\}.$$

The bars extend from $\overline{\Delta} - s_\Delta$ to $\overline{\Delta} + s_\Delta$. They describe dispersion across paired training runs, conditional on the chosen SFT references, rather than confidence intervals for the mean. They omit uncertainty in SFT calibration. Every estimate, including negative differences, is retained; horizontal offsets separate overlapping points. The observations are not smoothed or interpolated.

Data provenance. The figures reuse the recorded observations embedded in the archived vector welfare figure. The original numerical records underlying these observations, including their seed identifiers, are available under `figures/data/` in the anonymous repository.[3] For the figure reconstruction, we recovered all 91 Qwen differences and 450 fixed-policy estimates from their vector coordinates, calibrated against the labeled axes, and checked them against the original numerical records. Recovered means and the original interval endpoints agree with the archived markers within 10^{-7} welfare units. A conservative tolerance for coordinate recovery is 10^{-6} welfare units. The accompanying extraction script records the source checksum, validates counts and summaries, and exports the recovered observations. Observation indices in the recovered files identify points within each rate rather than numeric training seeds; seed identifiers are retained in the original numerical records. No new language responses or simulated Qwen observations were generated for this figure reconstruction.

E.4 Additional analytical illustrations

The following curves evaluate the welfare formulas or integrate the deterministic mean update. They complement the sampled estimates in Figure 8.

[3] `https://anonymous.4open.science/r/llm-welfare-code-23F3/`

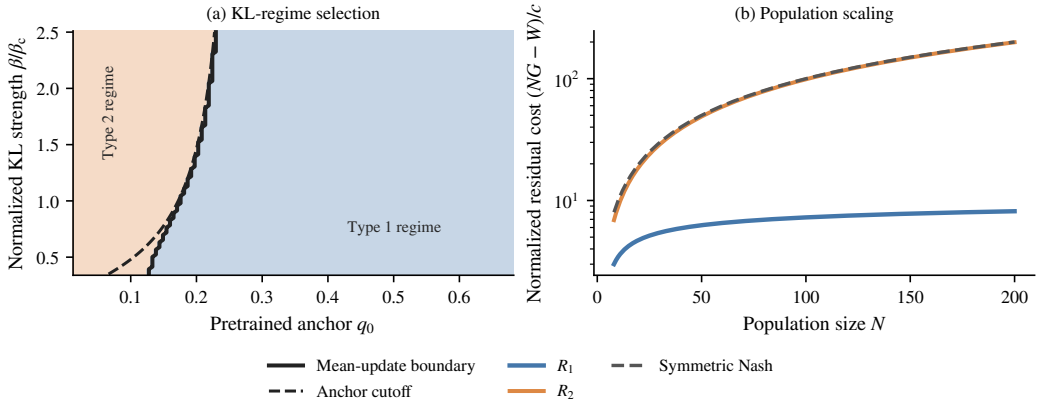

Figure 10: Additional theory-matched numerical diagnostics. Panel (a) simulates the deterministic mean LLM update for $N = 20$, $\rho = 0.4$, $c/G = 0.08$, and $\lambda = 1$ over the model-policy anchor q_0 and normalized model-scale parameter β/β_c. At $\lambda = 1$, the model-policy and effective-policy anchor coordinates coincide, as do the model-scale and effective KL coefficients. Colors mark the selected limiting branch, the solid curve is the empirical basin boundary for initialization at the pretrained policy, and the dashed curve is the pretrained-anchor cutoff q_{cut} from Proposition 2. The two boundaries agree in the monotone KL regime; Theorem D.1 and Appendix D.1 give the basin-dependent weak-KL outcomes. Panel (b) plots the normalized residual costs R_1, R_2, and the symmetric mixed Nash benchmark N on a log scale for $c/G = 0.2$ and $\rho = 0.4$, illustrating Lemma 4.

Regime selection. We simulated the mean LLM update at $\lambda = 1$ over model-policy anchors q_0 and normalized model-scale parameters β/β_c, initializing the LLM at its pretrained model policy. At this normalization, $q_0 = p_0$ and β is also the effective KL coefficient. Panel (a) of Figure 10 shows that low anchors select the type 2 regime, while higher anchors select the type 1 regime. The empirical branch boundary moves right as the KL penalty grows, consistent with the sign patterns in Proposition 2 and the anchor-dependent basins in Theorem D.1 and Appendix D.1; the dashed curve is q_{cut}, where the KL-adjusted reward gap at the response boundary vanishes. For $\beta \geq \beta_c$, the anchor transition is exactly q_{cut}; for weaker KL, the initial policy's basin also matters.

Social welfare gap. For fixed N and each $x = c/G$, we plot $\Delta(x, \rho)/\max_{\rho'} \Delta(x, \rho')$ and locate $\rho^\star(x) \in \arg\max_\rho \Delta(x, \rho)$. Panel (a) of Figure 9 shows the unique maximum predicted by Theorem 4.1. Normalization preserves the location of the maximum and allows the curve shapes to be compared across costs, while the absolute severity of the gap is shown in Figure 2(a). Panel (b) shows that an interior maximizing adoption level moves lower as c/G increases, until it reaches the admissible boundary. Panel (c) applies the same normalization to the recorded fixed policy estimates, retaining their sampling variation around the analytical gap.

Population scaling. We varied N while holding c/G and ρ fixed, and compared normalized residual costs at the two branchwise user optima. Panel (b) of Figure 10 shows that R_1 grows slowly and stays below 10 across the plotted range, whereas R_2 grows nearly linearly and the symmetric mixed Nash residual equals N. This illustrates the population-size scaling in Lemma 4; Appendix B.9 gives the exact formulas and expansions.

E.5 Reproducibility and Assets

The Qwen3-8B experiment records its configuration, common SFT adapter, initial policy evaluation, seeds, generated responses, within-batch usage, game trajectories, and frozen-evaluation summaries. The analysis reconstructs effective action probabilities, checks them against saved trajectories and summaries, and recomputes structured roots and seedwise discrepancies. Figures and tables use these verified records, with one completed run per penetration-rate and seed pair.

Analytical plots and paired rollouts use NumPy/Matplotlib on CPU, and language experiments use MLX/MLX-LM with the reported LoRA/PPO settings. Original licenses apply. Balanced SFT demonstrations and game rewards are generated without participant data.