MARGINAL-CONTRIBUTION POLICY GRADIENTS UNDER FILTERED FEEDBACK FOR MULTI-AGENT LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

We develop a unified treatment of credit assignment for RL training in multi-agent LLM systems. We show that observed reward alone cannot distinguish an agent that determines it from one that never affects it, and that standard shared-reward training performs exact gradient ascent on each agent's private utility rather than system performance. Moreover, we prove no single scalar per agent can consistently account for joint performance once agents interact. We thus develop the unique background-dependent notion of marginal contribution satisfying natural consistency requirements. From it we derive gradient-correct marginal contribution training signals, identify them from filtered feedback, and optimally allocate a budget of exact counterfactual evaluations against learned-signal error. Instantiated in GRPO, our signal improves routed GSM8K accuracy over winner-take-all training at no extra generation cost.

1 INTRODUCTION

Modern LLMs increasingly use multi-agent systems, in which a mechanism sits between the model and the user: several agents generate outputs that a system-level rule selects, combines, or filters into one deployed answer. In *routed* systems, agents compete: each proposes a candidate and a router selects one (e.g., a pool of solvers that attacks a mathematics problem). In *collaborative* systems, agents hold complementary roles (e.g., planner, solver, critic) and jointly produce the output. When such systems are RL fine-tuned, a natural default is to use methods built for single policies: observe reward on the deployed output and allocate it back to the participating agents.

This fails as a method of individual credit assignment. The reward is a result of the system's output, not any agent's action, and we show this leads to structural failures. The reward's distribution can be identical whether an agent determines it or never affects it, so the reward distribution alone cannot identify credit (Proposition 1). Worse, the standard winner-take-all-credit ascent in routed systems is not a noisy estimate of the system gradient but an exact gradient of each agent's private utility, with an irreducible bias (Theorem 1). Repairing this requires separating three questions. *Optimization*: which objective does a training rule actually ascend? *Attribution*: what does it mean for an agent to have contributed? *Identification*: which contribution quantities does the data determine?

We answer these in order, beginning with a unified construction for any mechanism (Section 3). We characterize contribution axiomatically and find that once agents interact, no single scalar per agent can consistently account for agent contribution, while requiring contributions to sum to every group's value uniquely determines a field of marginal contributions conditioned on which other agents are present (Section 4). We determine which entries of this field are gradient-correct training signals (Section 5). Exact contributions are expensive, hence we learn them from the training data, including under routing, where only the selected candidate's reward is observed (Sections 6–6.1). We then optimally allocate a fixed budget of exact counterfactual evaluations across the learned contributions (Section 7). Instantiated in GRPO, the removal signal identified from routing data improves GSM8K accuracy over winner-take-all training at no extra generation cost (Section 8).

2 RELATED WORK

Credit under shared rewards. That a system-level reward need not reveal individual contribution, and that allocating it back to agents can misalign private incentives with the system objective, are

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.

classical observations on team production and collectives (Alchian & Demsetz, 1972; Holmström, 1982; Wolpert & Tumer, 2001). Multi-agent RL addresses the first problem with counterfactual signals: COMA, difference rewards, and action-dependent baselines remove teammate variation while preserving the team policy gradient (Foerster et al., 2018; Wu et al., 2018; Castellini et al., 2021); Lemma 1 gives the corresponding condition in our setting. Cooperative game theory treats attribution through marginal contributions and scalar summaries such as the Shapley value (Shapley, 1953; Wang et al., 2020; Li et al., 2026b). We sharpen both threads for routed LLM training: winner-take-all training exactly optimizes each agent's private utility, with bias equal to the displaced competitor return (Theorem 1), and once agents interact no single scalar per agent can consistently account for contribution, so we retain the full family of group-dependent contributions, of which Shapley is a lossy average.

Multi-agent LLM training and evaluation budgets. Recent multi-agent LLM methods obtain counterfactual credit through fixed-context replay, agent-removal rollouts, shared-prefix comparisons, or separate router–expert signals (Chen et al., 2026; Li et al., 2026c; Liu et al., 2026; Li et al., 2026a). We address the questions that precede any estimation method: which counterfactual quantity is valid for training, and which quantities the observed feedback identifies. Ordinary training data identifies the conditional expected reward given any chosen subset of the observed outputs. Under routing, propensity correction identifies the effect of deleting one candidate without additional rollouts (Dudík et al., 2011) (Corollary 3). Separately, work on allocating rollout compute in RL with verifiable rewards targets gradient variance (Nguyen et al., 2026; Zou et al., 2026); we instead allocate a budget of exact counterfactual evaluations against learned-signal error, obtaining the classical square-root allocation (Neyman, 1934) with a matching lower bound (Section 7).

3 Mechanisms, Filtered Feedback, and Classical Failures

In this section we develop a general theory of multi-agent LLM systems. The development is deliberately abstract, and nothing in it is specific to language models: the results apply to any RL system of differentiable stochastic policies, composed by a mechanism and trained by policy gradients. The LLM setting is where the theory is necessary and novel, as action spaces are too large to marginalize, and counterfactuals are too expensive to run exhaustively. We thus instantiate in LLMs, and focus on prevalent cases such as softmax routing where a result requires them.

Notation. Random variables and population-level scalars are uppercase; realizations and deterministic functions are lowercase; estimators carry a hat. Mechanism randomness ξ is the only exception. Layer subscripts are dropped when clear from context.

A *system* consists of agents arranged in *layers* $\ell \in \{1, \dots, L\}$, a *mechanism* M that maps their actions to one deployed output, and a scalar reward on that output. Thus, the mechanism produces *filtered feedback*: the observed reward attaches to the deployed output, and what it reveals about any agent's action is determined by M. At layer ℓ, the active agents $N_\ell = \{1, \dots, K_\ell\}$ observe the history H_ℓ, which contains the prompt and all outputs of earlier layers, and act conditionally independently: agent i samples $A_\ell^{(i)} \sim \pi_{\theta_i}(\cdot \mid H_\ell)$, so that

$$\mathbb{P}_\theta\left(A_\ell^{(N_\ell)} = a^{(N_\ell)} \mid H_\ell = h\right) = \prod_{i \in N_\ell} \pi_{\theta_i}\left(a^{(i)} \mid h\right). \tag{1}$$

The factorization applies within a layer only. Across layers, agents interact through the history: sequential pipelines such as planner–solver–critic are chains of singleton layers, so later agents observe all earlier outputs through H_ℓ, while a routing or voting pool is a single layer of parallel agents acting without sight of one another. Any mixture of the two, such as parallel solvers followed by an editor, is a sequence of layers of the corresponding widths.

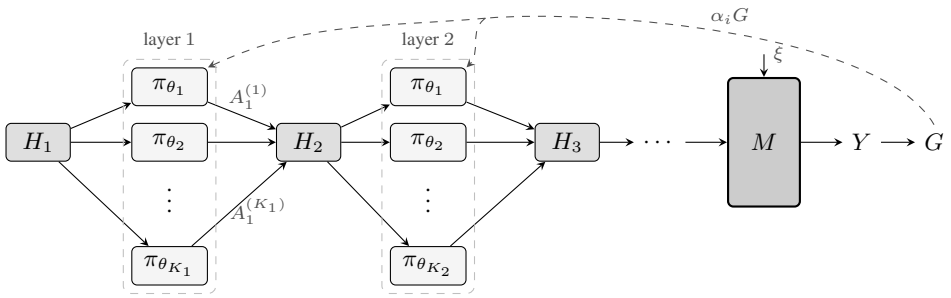

Figure 1: Agents act in layers, conditionally independently given the history H_ℓ; the mechanism M deploys one output Y, which alone is rewarded, and $\alpha_i G$ is returned to each agent (dashed).

Let $A = \left(A_\ell^{(N_\ell)}\right)_{\ell=1}^{L}$ collect agent actions across all layers. The mechanism determines the deployed output $Y = M(H_1, A, \xi)$, where H_1 is the initial prompt and ξ is any mechanism randomness. The system receives bounded reward $G = r(H_1, Y)$ with objective $J_{\text{sys}}(\theta) = \mathbb{E}_\theta[G]$. The mechanism also specifies a reward allocation rule $\alpha_i(H_1, A, \xi) \in [0, 1]$ with $\sum_i \alpha_i = 1$, defining agent i's private utility $U_i(\theta) = \mathbb{E}_\theta[\alpha_i G]$. Note that the allocation rule partitions the system objective into $\sum_\ell K_\ell$ private ones, namely $J_{\text{sys}} = \sum_i U_i$. For $i \in N_\ell$, define the score $\Psi_i = \nabla_{\theta_i} \log \pi_{\theta_i}\left(A_\ell^{(i)} \mid H_\ell\right)$. Throughout, rewards are bounded, each policy π_{θ_i} is differentiable in θ_i, $\mathbb{E}\left\|\Psi_i\right\|^2 < \infty$ for every agent, and relevant gradients may be interchanged with expectation in J_{sys} and U_i. All gradient identities are on-policy. All proofs are deferred to the appendix.

We begin with the failure that motivates the paper. Training each agent on a shared reward presumes the reward says something about that agent in particular. The next result shows it need not: the reward can have exactly the same distribution whether an agent is the sole cause of it or entirely irrelevant to it, so no amount of reward data alone can tell the cases apart. The quantity we use, displayed in (2) below, is how much the agent's realized action moves the expected reward. This is one natural *marginal contribution* notion among several developed in Section 4. The result needs only that it separates an agent who determines the reward from one who never affects it, which any reasonable notion must.

Proposition 1 (The reward's law does not identify contribution)**.** *There exist two systems with identical agent policies and identical deployed-reward distributions such that*

$$\mathbb{E}[G \mid H, A^{(1)}] - \mathbb{E}[G \mid H] \tag{2}$$

is nonconstant in one and vanishes almost surely in the other. Thus the reward distribution alone does not identify this contribution quantity. The construction embeds in any system with at least two agents, in the same or different layers.

3.1 Routed Systems

We study *routed systems* in particular, which have a single layer of parallel agents $N = \{1, \ldots, K\}$, and whose mechanism M selects one candidate for deployment. Given the agents' actions $A^{(1:K)}$, the mechanism draws $I \sim p(\cdot \mid H, A^{(1:K)})$ and deploys $Y = A^{(I)}$. We consider the ubiquitous choice of softmax routing with uniform exploration,

$$p_j = (1 - \epsilon)\, \frac{e^{s(H, A^{(j)})/\tau}}{\sum_{k=1}^{K} e^{s(H, A^{(k)})/\tau}} + \frac{\epsilon}{K}, \tag{3}$$

where s scores each candidate, $\tau > 0$ is a temperature, and $\epsilon \in [0, 1]$ mixes in uniform exploration. Writing $R^{(j)} := r(H, A^{(j)})$, the deployed reward is $G = R^{(I)} = \sum_{j=1}^{K} \mathbf{1}_{\{I=j\}} R^{(j)}$. Commonly, *winner-take-all* is the allocation rule $\alpha_i = \mathbf{1}_{\{I=i\}}$, under which $U_i = \mathbb{E}[\mathbf{1}_{\{I=i\}} G]$. Winner-take-all trains the selected agent on the observed reward. We now show this update is not a noisy estimate of the system gradient, but the exact gradient of private utility, with bias equaling the return agent i's

improvement displaces from its competitors, whose total selection probability its own gain necessarily lowers. We define the displaced competitor return and its conditional means

$$D_i := \sum_{j \neq i} p_j R^{(j)}, \qquad d_i(H, A^{(i)}) := \mathbb{E}[D_i \mid H, A^{(i)}], \qquad \bar{d}_i(H) := \mathbb{E}\left[D_i \mid H\right],$$

so that D_i is the competitors' expected return, d_i and $\bar{d}_i$ average it with and without agent i's output held fixed, and their difference $d_i - \bar{d}_i$ is how much that output moves the competitors' return, as (2) does for G.

Theorem 1 (Winner-take-all bias is displaced return)**.** *Fix i and suppose the router depends on θ_i only through the candidates. Then*

$$\mathbb{E}\left[\mathbf{1}_{\{I=i\}} \Psi_i G\right] = \nabla_{\theta_i} U_i, \tag{4}$$

so the winner-take-all update performs exact gradient ascent on private utility, and its bias for the system gradient, $b_i := \mathbb{E}[\mathbf{1}_{\{I=i\}} \Psi_i G] - \nabla_{\theta_i} J_{\mathrm{sys}}$, satisfies

$$b_i = -\nabla_{\theta_i} \sum_{j \neq i} U_j = -\mathbb{E}\left[\Psi_i \left(d_i(H, A^{(i)}) - \bar{d}_i(H)\right)\right]. \tag{5}$$

In particular, $b_i = 0$ if and only if $d_i(H, A^{(i)}) - \bar{d}_i(H)$ is L^2-orthogonal to every coordinate of Ψ_i. Two score-free conditions also suffice. The bias vanishes if agent i's output carries no information about the displaced return, $\mathbb{E}[D_i \mid H, A^{(i)}] = \mathbb{E}[D_i \mid H]$. Under (1)*, it also vanishes if the router is candidate-blind for agent i: no p_j, $j \neq i$, changes as $A^{(i)}$ varies with $(H, A^{(-i)})$ fixed.*

Theorem 1 locates the bias in the correlation between the score and the displaced return. A router that selects from the prompt alone never reads $A^{(i)}$, is candidate-blind, and has $b_i = 0$; a router that rates candidates displaces return, and $b_i \neq 0$ additionally requires score correlation. Three refinements are deferred to Appendix C: selection makes private utility risk-sensitive, so candidate-reward distributions with equal conditional means can induce different private utilities (Proposition 6); conditional mean independence is also necessary under score completeness (Remark 2); and when the router ranks candidates by reward, the bias inflates every direction along which the policy raises its own reward (Proposition 7).

4 Marginal Contribution

We now develop machinery for attributing marginal contribution to agents. We proceed axiomatically, in cooperative game-theoretic terms; rather than proposing a formula, we state the properties required of a contribution notion: it must capture each agent's added value and decompose every group's value among that group's members. We then let these requirements determine the object we adopt. Thus, consider a system with agents $N = \{1, \ldots, K\}$ and a set function $v : 2^N \to \mathbb{R}$ which assigns a *value* to each group of agents. We deliberately leave v abstract for now: later, an evaluation rule ρ will specify the counterfactual under which a group S is evaluated, thereby inducing a concrete valuation v^ρ. Because $v(S)$ is the value of all agents $i \in S$, it captures higher-order value that exists only through interactions among agents and would be invisible agent by agent. We do not yet know how this value divides among the members of S, which is what any notion of marginal contribution must determine.

Suppose first that it is possible to assign each agent $i \in N$ a single scalar Δ_i that completely characterizes its marginal contribution to the task, regardless of which other agents are present. We enforce the following properties for any such notion of marginal contribution:

1. *Nullity.* If $v(S \cup \{i\}) = v(S)$ for every $S \subseteq N \setminus \{i\}$, then $\Delta_i = 0$.
2. *Totality.* Every group's value decomposes into individual marginal contributions:

$$\sum_{i \in S} \Delta_i = v(S) \qquad \text{for every } S \subseteq N.$$

Totality gives the desirable property of $v(\varnothing) = 0$ and, for each singleton, $\Delta_i = v(\{i\})$. Thus, the formula for contributions is determined by these axioms. Note that nullity follows from totality since if agent i is null, $v(\{i\}) = v(\varnothing) = 0$, so $\Delta_i = 0$. The remaining question is whether such scalar contributions can exist in a general system, which the following proposition answers.

Proposition 2 (Scalar credit exists only without interaction)**.** *For any $v : 2^N \to \mathbb{R}$ with $v(\varnothing) = 0$, the following are equivalent:*

(i) There exist scalars $\{\Delta_i : i \in N\}$ satisfying totality (and nullity) for every $S \subseteq N$.

(ii) Agents never interact, in the sense that adding one agent changes the value by the same amount regardless of which others are present:

$$v(S \cup \{i, j\}) - v(S \cup \{i\}) = v(S \cup \{j\}) - v(S) \tag{6}$$

for every distinct $i, j \in N$ and every $S \subseteq N \setminus \{i, j\}$.

(iii) v is additive: $v(S) = \sum_{i \in S} v(\{i\})$ for every $S \subseteq N$.

When these conditions hold, the scalar contributions are unique and satisfy $\Delta_i = v(\{i\})$ for every $i \in N$.

We thus see that scalar contributions cannot in general account for interactions (Remark 1 gives a two-agent witness). Therefore, we retain totality but generalize it in the natural way: an agent's contribution need not be the same regardless of which other agents are present. For $i \in N$ and $S \subseteq N \setminus \{i\}$, let $\Delta_{i|S}$ denote the contribution of agent i given the background S. A group of agents can be built by adding its members in any order. We require that along every group-building path, the successive contributions sum to the value of the resulting group, independently of the order in which its members are added. We call this *ordered totality*. That is, for every $S = \{i_1, \ldots, i_m\}$ and every ordering $(i_1, \ldots, i_m)$ of its members,

$$\sum_{t=1}^{m} \Delta_{i_t|\{i_1,\ldots,i_{t-1}\}} = v(S).$$

Equivalently, ordered totality is a path-independence requirement on the lattice of agent subsets: every path from $\varnothing$ to S accumulates the same total contribution, namely $v(S)$.

Proposition 3 (Set-conditioned contribution is unique)**.** *For any $v : 2^N \to \mathbb{R}$ with $v(\varnothing) = 0$, ordered totality determines the contributions uniquely:*

$$\Delta_{i|S} = v(S \cup \{i\}) - v(S) \qquad \text{for every } i \in N,\ S \subseteq N \setminus \{i\},$$

and conversely these value differences satisfy ordered totality for every such v.

Ordered totality therefore replaces one scalar per agent with the indexed family $\left\{\Delta_{i|S} : i \in N,\ S \subseteq N \setminus \{i\}\right\}$, which we call the *contribution field*. Variation of $\Delta_{i|S}$ across backgrounds S is how interactions appear. Proposition 2 shows that this field collapses to one background-independent scalar per agent exactly when v is additive.

The familiar Shapley value is a scalar compression of this field: it averages $\Delta_{i|S}$ over the backgrounds encountered when agents are added in uniformly random order (Shapley, 1953). Averaging does not remove the background dependence created by interaction, and the resulting allocation is efficient but satisfies totality only when v is additive. We retain the full field: its variation across backgrounds is where interaction appears, and the sections that follow determine which entries are valid training signals and which are estimable from data.

Everything so far holds for an arbitrary set of agents N and valuation v. We now return to the layered multi-agent system and fix a layer ℓ, so that $N = N_\ell$. We instantiate the preceding theory by specifying what it means to evaluate a given background $S \subseteq N_\ell$. An *evaluation rule* ρ specifies the counterfactual experiment associated with each such S: given the history and the realized outputs of the agents in S, it determines how the system is evaluated, including how agents outside S are treated, $\rho_{\ell,S} : \left(H_\ell, A_\ell^{(S)}, \text{fresh randomness}\right) \longmapsto G^\rho_{\ell,S}$. Because the evaluation may involve fresh randomness, $\rho_{\ell,S}$ induces a conditional distribution over terminal rewards; we need only its mean, $q^\rho_\ell(S) := \mathbb{E}\left[G^\rho_{\ell,S} \mid H_\ell, A_\ell^{(S)}\right]$, the expected reward under the counterfactual experiment specified by ρ. The valuation used above is its centered version,

$$\begin{aligned} v^\rho_\ell(S) &:= q^\rho_\ell(S) - q^\rho_\ell(\varnothing), \\ \Delta^\rho_{i|S} &:= v^\rho_\ell(S \cup \{i\}) - v^\rho_\ell(S) = q^\rho_\ell(S \cup \{i\}) - q^\rho_\ell(S). \end{aligned} \tag{7}$$

The centering gives $v_\ell^\rho(\varnothing) = 0$ as required for the previous theorems, while it cancels from every marginal contribution.

Under the *retention* rule, omitted agents are averaged under their current policies, so

$$q_\ell^{\mathrm{ret}}(S) := \mathbb{E}\left[G \mid H_\ell, A_\ell^{(S)}\right]. \tag{8}$$

This recovers the quantity used earlier: agent i's contribution at the empty background is exactly (2) from Proposition 1.

Under the *removal* rule, omitted agents are instead deleted and the reduced system is evaluated with the realized outputs in S held fixed. Writing $G_{\ell,S}^{\mathrm{rem}}$ for the resulting terminal reward, $q_\ell^{\mathrm{rem}}(S) := \mathbb{E}\left[G_{\ell,S}^{\mathrm{rem}} \mid H_\ell, A_\ell^{(S)}\right]$, whenever the reduced system is well defined. Unlike literal removal, the retention valuation remains well defined for pipelines that may cease to function when an intermediate agent is deleted, such as a planner–solver–critic system.

5 Valid Training Signals

In the following section we develop our theory for the purpose of reinforcement learning. In particular, we wish to understand when entries of the contribution field $\{\Delta_{i|S}^\rho\}$ can be used as gradient-correct training signals for our system. We begin with a lemma that identifies how total reward can be changed without changing an agent's policy gradient.

Lemma 1 (Baseline invariance)**.** *Fix $i \in N_\ell$ and let a baseline B_i be any square-integrable random variable satisfying*

$$\mathbb{E}[B_i \mid H_\ell, A_\ell^{(i)}] = \mathbb{E}[B_i \mid H_\ell]. \tag{9}$$

Then

$$\mathbb{E}[\Psi_i (G - B_i)] = \nabla_{\theta_i} J_{\mathrm{sys}}. \tag{10}$$

The baseline may also depend on a randomly sampled variable S. This remains valid provided $S \perp A_\ell^{(i)} \mid H_\ell$ and, conditional on S, the baseline remains action-independent: $\mathbb{E}[B_i(S) \mid H_\ell, A_\ell^{(i)}, S] = \mathbb{E}[B_i(S) \mid H_\ell, S]$. Thus S may, for example, be a randomly sampled background without biasing the expected policy gradient.

Not every entry of a contribution field is gradient-correct.[1] Under removal, both endpoints of an entry evaluate systems with agents deleted, and their difference need not recover the factual policy gradient. Retention avoids this, and we use it for the remainder of the paper: every entry of its field is gradient-correct.

Corollary 1 (Every entry of the retention field is a valid signal)**.** *Under the retention rule, for every $S \subseteq N_\ell \setminus \{i\}$, fixed or drawn conditionally independently of $A_\ell^{(i)}$ given H_ℓ,*

$$\mathbb{E}\left[\Psi_i \, \Delta_{i|S}^{\mathrm{ret}}\right] = \mathbb{E}\left[\Psi_i \left(G - q_\ell^{\mathrm{ret}}(S)\right)\right] = \nabla_{\theta_i} J_{\mathrm{sys}}. \tag{11}$$

Moreover,

$$\mathbb{E}\left[G - q_\ell^{\mathrm{ret}}(S) \mid H_\ell, A_\ell^{(S\cup\{i\})}\right] = \Delta_{i|S}^{\mathrm{ret}},$$

so the above residual is an unbiased estimate of the marginal contribution.[2]

We now turn to the reason for using contribution-based signals in place of the raw reward, namely, variance reduction in G, unrelated to agent i's contribution.

Proposition 4 (Reward variance reduction)**.** *For nested backgrounds $S \subseteq T \subseteq N_\ell$,*

$$\mathbb{E}\left[\mathrm{Var}\left(G \mid H_\ell, A_\ell^{(T)}\right)\right] \le \mathbb{E}\left[\mathrm{Var}\left(G \mid H_\ell, A_\ell^{(S)}\right)\right]. \tag{12}$$

Therefore, conditioning on more agents removes more variation from the reward.

[1] In general, $\Delta_{i|S}^\rho$ recovers the factual gradient whenever $q_\ell^\rho(S)$ is an admissible baseline under Lemma 1 and $\mathbb{E}[\Psi_i \, q_\ell^\rho(S \cup \{i\})] = \mathbb{E}[\Psi_i G]$.

[2] Using $G - v_\ell^{\mathrm{ret}}(S)$ instead preserves the expected gradient but shifts this conditional mean by the history-measurable offset $q_\ell^{\mathrm{ret}}(\varnothing)$.

6 LEARNING MARGINAL CONTRIBUTIONS

Computing marginal contributions requires the conditional values $q_\ell(S)$ across many backgrounds, and evaluating all 2^{K_ℓ} of them directly is exponentially expensive. Instead we learn them from training data: reveal a randomly chosen subset of agent outputs to a critic and train it to predict the observed reward from that subset. At the population optimum, one mask-conditioned critic recovers $q_\ell(S)$ for every background in the mask distribution. Let $m(S) \in \{0,1\}^{K_\ell}$ denote the mask of S, and let $\hat{q}_{\ell,\phi}(H_\ell, m(S), A_\ell^{(S)})$ be a mask-conditioned critic with parameters ϕ. Given data $(H_\ell, A_\ell^{(N_\ell)}, G)$, sample $S \subseteq N_\ell$ independently of $(A_\ell^{(N_\ell)}, G)$ given H_ℓ, and minimize the mean squared error

$$\mathcal{L}(\phi) = \mathbb{E}\left[\left(G - \hat{q}_{\ell,\phi}\left(H_\ell, m(S), A_\ell^{(S)}\right)\right)^2\right]. \tag{13}$$

Every background whose value the critic must identify has to carry positive mask probability: if $P(S = s) = 0$, critics agreeing on sampled backgrounds have equal population loss regardless of their values at s.

Lemma 2. *Suppose a mask is sampled as detailed above and the critic class contains the conditional mean of G given its inputs. Then for every $S \subseteq N_\ell$ in its support, any population minimizer of* (13) *satisfies*

$$\hat{q}_{\ell,\phi}\left(H_\ell, m(S), A_\ell^{(S)}\right) = \mathbb{E}\left[G \mid H_\ell, A_\ell^{(S)}\right] \tag{14}$$

almost surely.

The critic thus learns the retention values (8) directly and we omit the superscript hereafter. Define the learned marginal contributions $\widehat{\Delta}_{i|S} := G - \hat{q}_{\ell,\phi}(S)$. As agent i's baseline, the critic excludes $A_\ell^{(i)}$ and is held fixed during the actor update. Cross-fitting additionally prevents the trajectory being updated from entering the critic through its training sample.

Proposition 5 (Gradient correctness and attribution error)**.** *Let $S \subseteq N_\ell \setminus \{i\}$ be fixed or drawn conditionally independently of $A_\ell^{(i)}$ given H_ℓ, and let $\hat{q}_{\ell,\phi}$ be any fixed or cross-fitted square-integrable critic that excludes $A_\ell^{(i)}$. Then*

$$\mathbb{E}\left[\Psi_i \, \widehat{\Delta}_{i|S}\right] = \nabla_{\theta_i} J_{\mathrm{sys}}, \tag{15}$$

and for every such S,

$$\mathbb{E}\left[\widehat{\Delta}_{i|S} \,\middle|\, H_\ell, A_\ell^{(S\cup\{i\})}\right] - \Delta_{i|S} = \mathbb{E}\left[G \mid H_\ell, A_\ell^{(S)}\right] - \hat{q}_{\ell,\phi}(S). \tag{16}$$

Critic accuracy is therefore unnecessary for gradient correctness. For attribution, the critic's prediction error at S passes through unchanged. For diagnostics, $\hat{q}_{\ell,\phi}(S \cup \{i\}) - \hat{q}_{\ell,\phi}(S)$ estimates any entry of the contribution field with two critic evaluations. We do not use this difference as the actor signal: its first endpoint depends on $A_\ell^{(i)}$, so critic error there can enter the gradient. Training therefore uses the factual residual $G - \hat{q}_{\ell,\phi}(S)$, while reporting uses critic differences.

6.1 LEARNING REMOVAL CONTRIBUTIONS UNDER ROUTING

Routing adds an identification problem. The training data contains every candidate output, so the retention values above remain learnable by masked regression, but only the selected candidate is rewarded; asking how much candidate i contributes relative to a router in which it was unavailable requires reasoning about unobserved rewards. Recall $R^{(j)} := r(H, A^{(j)})$ and assume selection does not change a candidate's conditional expected reward,

$$\mathbb{E}\left[R^{(j)} \mid H, A^{(1:K)}, I = j\right] = \mathbb{E}\left[R^{(j)} \mid H, A^{(j)}\right] =: g_j(H, A^{(j)}), \tag{17}$$

so that, conditional on the realized actions, the router has expected reward $\sum_{j=1}^K p_j g_j$. Let p_j^{-i} denote the router probabilities renormalized over $j \neq i$. The removal contribution

$$\Delta_i^{\mathrm{rem}} := \sum_{j=1}^K p_j g_j - \sum_{j \neq i} p_j^{-i} g_j$$

is the value of making candidate i available to the router on the realized actions. This is also valid for training: the reward after removing candidate i does not depend on $A^{(i)}$ and is therefore a valid baseline.

Corollary 2 (Removal is a valid routing signal)**.** *Let G^{-i}_{route} be a draw from the router after candidate i is removed, conditional on the realized remaining candidates. Then*

$$\mathbb{E}\left[\Psi_i \left(G - G^{-i}_{\text{route}}\right)\right] = \nabla_{\theta_i} J_{\text{sys}}. \tag{18}$$

The identity remains valid if G^{-i}_{route} is replaced by any fixed or cross-fitted square-integrable predictor based only on H and $A^{(-i)}$.

The remaining problem is estimation: g_j is observed only when candidate j is selected. Under overlap, we can recover it from the same training data using a doubly robust estimator

$$\widehat{g}_j = \widehat{\mu}_j + \frac{\mathbf{1}_{\{I=j\}}}{p_j}\left(G - \widehat{\mu}_j\right), \qquad \widehat{\Delta}^{\text{rem}}_i = \sum_{j=1}^{K} p_j \widehat{g}_j - \sum_{j \neq i} p_j^{-i} \widehat{g}_j. \tag{19}$$

The propensity term corrects for the fact that candidate j's reward is seen only on the fraction of examples in which j is selected (Dudík et al., 2011). Appendix D gives the removed propensities and the error decomposition.

Corollary 3 (The removal contribution is identified from training data)**.** *Under* (17)*, suppose $p_j \geq p_{\min} > 0$ for every candidate and treat p_j and p_j^{-i} as known. Conditional on H and the realized actions, $\widehat{\Delta}^{\text{rem}}_i$ is unbiased for Δ^{rem}_i for any fixed or cross-fitted outcome model $\widehat{\mu}_j$.*

7 Optimal Allocation of Counterfactual Evaluations

The preceding section gives cheap, gradient-correct contribution estimates; their limitation is attribution error, since learned critics need not recover the exact counterfactual. Exact evaluation removes the error by rerunning the counterfactual, at the price of additional model calls. We therefore choose evaluation probabilities to minimize gradient error under a fixed budget. For each marginal contribution $\Delta_{i|S}$ we may evaluate, let $\widehat{\Delta}_{i|S}$ denote its learned estimate. An exact evaluation reruns the counterfactual experiment defining $\Delta_{i|S}$ and returns $Y_{i|S} = G^{\rho}_{S\cup\{i\}} - G^{\rho}_{S}$, with mean $\Delta_{i|S}$ and variance $\text{Var}(Y_{i|S}) = \sigma^2_{i|S}$. Each exact evaluation costs $c_{i|S} > 0$. We run this evaluation with probability $p_{i|S}$, with $Z_{i|S} \sim \text{Bernoulli}(p_{i|S})$ indicating whether it is run. Write $\eta_{i|S} := \widehat{\Delta}_{i|S} - \Delta_{i|S}$ for the learned-estimate error and $a_i := \|\Psi_i\|$. Conditioning on the data and fitted estimate makes $\eta_{i|S}$ fixed when the $p_{i|S}$ are chosen. The theorem below gives the optimal allocation.

Theorem 2 (Optimal allocation of counterfactual evaluations)**.** *Condition on the factual batch and the learned estimates, and assume the pairs $(Z_{i|S}, Y_{i|S})$ are independent across marginal contributions, $Z_{i|S}$ is independent of $Y_{i|S}$, and $p_{i|S} > 0$. The correction*

$$\widetilde{\Delta}_{i|S} = \widehat{\Delta}_{i|S} + \frac{Z_{i|S}}{p_{i|S}}\left(Y_{i|S} - \widehat{\Delta}_{i|S}\right) \tag{20}$$

is unbiased for $\Delta_{i|S}$ for every choice of p, with mean-squared error

$$\mathbb{E}\left[\left(\widetilde{\Delta}_{i|S} - \Delta_{i|S}\right)^2\right] = \frac{\sigma^2_{i|S} + \eta^2_{i|S}}{p_{i|S}} - \eta^2_{i|S}. \tag{21}$$

Restrict to the marginal contributions with $a_i^2(\sigma^2_{i|S} + \eta^2_{i|S}) > 0$; the rest are never worth evaluating. Minimizing the gradient-weighted risk $\mathcal{R}(p) = \sum a_i^2\, \mathbb{E}[(\widetilde{\Delta}_{i|S} - \Delta_{i|S})^2]$ over $0 < p_{i|S} \leq 1$ under the budget $\sum c_{i|S}\, p_{i|S} \leq C_{\max}$ gives $p^{\star}_{i|S} = 1$ everywhere when $C_{\max} \geq \sum c_{i|S}$, and otherwise the unique optimum

$$p^{\star}_{i|S} = \min\left\{1,\ \lambda\, \frac{a_i\sqrt{\sigma^2_{i|S} + \eta^2_{i|S}}}{\sqrt{c_{i|S}}}\right\}, \tag{22}$$

with λ set by the budget.

Equation (22) gives the optimal allocation of exact evaluations. We evaluate a marginal contribution more often when it has greater gradient leverage, larger learned-estimate error or evaluation variance, and lower evaluation cost. The budget affects the variance of the corrected estimate, not its unbiasedness. The risk attained at $p^\star$, and a matching lower bound over all allocations of the same expected cost, are given in Appendix B.12. The quantities $\sigma^2_{i|S}$ and $\eta^2_{i|S}$ are unknown in practice. They may be estimated from an initial set of exact evaluations and substituted into the allocation rule. The theorem assumes evaluations are independent across marginal contributions. If not, covariance terms enter the risk and the optimal allocation need not have the same form.

8 Experiment: Marginal-Contribution Signals in GRPO

Marginal contribution is independent of the RL fine-tuning algorithm: any gradient-correct contribution Δ_i may replace the raw reward or advantage in a standard RLFT update. We instantiate this principle in GRPO, natural here because GSM8K (Cobbe et al., 2021) has verifiable rewards and small models permit several completions per prompt. For routing, the signal is the leave-one-out removal contribution identified from selection-gated feedback by Corollary 3; we refer to this method as *removal-signal GRPO*.

Task and signal. Each GSM8K prompt is one history H. Three agents independently generate $A^{(1:3)}$, the router draws $I \sim p(\cdot \mid H, A^{(1:3)})$, and the deployed reward is $G = R^{(I)}$ with $R^{(j)} \in \{0, 1\}$ for exact numerical answer correctness, so mean reward equals accuracy. Because $R^{(j)}$ is deterministic given prompt and candidate, condition (17) holds automatically. Winner-take-all scores each agent by $\mathbf{1}_{\{I=i\}}G$, which by Theorem 1 ascends private utility; removal-signal GRPO scores by $\widehat{\Delta}^{\mathrm{rem}}_i$ from (19), with exactly logged factual propensities and overlap from uniform exploration. Each agent samples $m = 4$ completions per prompt, the gth completions forming one slate; GRPO standardizes each agent's scores within its group and applies the same clipped loss, so the routed methods differ only in credit signal and the experiment compares the practical signals rather than the unnormalized score-gradient identities.

Setup. All agents share a `meta-llama/Llama-3.2-3B-Instruct` backbone with separate LoRA adapters (Hu et al., 2022), differing in prompting role, sampling temperature, and a keyword-assigned specialty that induces weak heterogeneity and defines the specialization diagnostic. The router and the outcome model share one network: its output s_j enters the softmax rule (3) as the score, and $\widehat{\mu}_j = \mathrm{sigmoid}(s_j)$ is the outcome estimate; it learns only from selected feedback by inverse-propensity-weighted cross-entropy. We train for 150 updates on 512 problems and evaluate on 128 held-out problems; each actor update uses its fresh rollout and is on-policy. Full architecture, hyperparameters, and schedules are in Appendix E. We compare a frozen single agent, winner-take-all GRPO, and removal-signal GRPO.

Evaluation and results. At evaluation each agent generates one deterministic candidate and the router deploys the highest-scoring one. We report exact-match accuracy, routing entropy, deployed-candidate Brier score, and specialization, the fraction routed to the heuristic-matched specialty. We also score all three candidates only for an oracle diagnostic: oracle accuracy is the fraction with at least one correct candidate and $\mathrm{Regret} = A_{\mathrm{oracle}} - A_{\mathrm{router}}$.

Method	Acc. ↑	Entropy	Brier ↓	Spec. ↑	Oracle ↑	Regret ↓
Frozen single-agent	0.352	–	–	–	–	–
Winner-take-all GRPO	0.37	0.68	0.50	0.46	0.45	0.08
Removal-signal GRPO (ours)	**0.41**	0.79	**0.43**	**0.58**	**0.47**	**0.06**

Table 1: GSM8K results and full-feedback oracle diagnostic. Accuracy equals mean deployed reward. Removal-signal GRPO improves deployed accuracy by 4 points over winner-take-all and 5.8 over the frozen reference, with higher routing entropy, lower Brier score, higher specialization, higher oracle accuracy, and lower regret (Table 1): better candidates and better use of the available slate. Only the agents' credit signal differs between the two methods; the router and outcome model are trained identically, so their improved diagnostics are downstream effects of the signal. This is a small mechanism-level study rather than a benchmark result. We still find that replacing winner-take-all private reward by the removal contribution improves deployed performance without additional counterfactual generation.

AI USE STATEMENT

All research ideas, mathematical arguments, and manuscript writing are the authors' own. Generative AI tools were used only in a limited capacity to assist with code implementation and debugging for the experiments. All resulting code was reviewed and validated by the authors; AI tools were not used to generate the paper's claims, proofs, analyses, or conclusions.

REPRODUCIBILITY STATEMENT

Complete proofs of all theoretical results are provided in the appendix. Full experimental details are provided in Appendix E, and code for reproducing the experiments is included in the supplementary material.

REFERENCES

Armen A. Alchian and Harold Demsetz. Production, information costs, and economic organization. *American Economic Review*, 62(5):777–795, 1972.

Jacopo Castellini, Sam Devlin, Frans A. Oliehoek, and Rahul Savani. Difference rewards policy gradients. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 1475–1477, 2021. Extended version in Neural Computing and Applications, 2022.

Yanjun Chen, Yirong Sun, Hanlin Wang, Jinghan Wang, Xinming Zhang, Xiaoyu Shen, Wenjie Li, and Wei Zhang. Exact is easier: Credit assignment for cooperative LLM agents. *arXiv preprint arXiv:2603.06859*, 2026.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pp. 1097–1104, 2011. arXiv:1103.4601.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Bengt Holmström. Moral hazard in teams. *Bell Journal of Economics*, 13(2):324–340, 1982.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Peiwen Li, Shiyang Zhang, Yangtian Zhang, Sizhuang He, David van Dijk, and Rex Ying. MoRSE: Task-oriented multi-agent system with mixture of role-subtask experts. *arXiv preprint arXiv:2608.09251*, 2026a.

Yanming Li, Xuelin Zhang, WenJie Lu, Ziye Tang, Maodong Wu, Haotian Luo, Tongtong Wu, Zijie Peng, Hongze Mi, Yibo Feng, Naiqiang Tan, Chao Huang, Lian Peng, and Li Shen. Who deserves the reward? SHARP: Shapley credit-based optimization for multi-agent system. *arXiv preprint arXiv:2602.08335*, 2026b.

Zhongyi Li, Wan Tian, Jinju Chen, Huiming Zhang, Yang Liu, Yikun Ban, and Fuzhen Zhuang. Counterfactual credit policy optimization for multi-agent collaboration. *arXiv preprint arXiv:2603.21563*, 2026c.

Yiqing Liu, Zihao Wang, Hantao Yao, Wu Liu, and Yongdong Zhang. Shared prefixes, better credit: Adaptive routing for multi-agent reasoning. *arXiv preprint arXiv:2608.02291*, 2026.

Jerzy Neyman. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4): 558–625, 1934.

Hieu Trung Nguyen, Bao Nguyen, Wenao Ma, Yuzhi Zhao, Ruifeng She, and Viet Anh Nguyen. Adaptive rollout allocation for online reinforcement learning with verifiable rewards. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2602.01601.

Thomas J. Rothenberg. Identification in parametric models. *Econometrica*, 39(3):577–591, 1971.

Lloyd S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games, Volume II*, volume 28 of *Annals of Mathematics Studies*, pp. 307–317. Princeton University Press, 1953.

Jianhong Wang, Yuan Zhang, Tae-Kyun Kim, and Yunjie Gu. Shapley Q-value: A local reward approach to solve global reward games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7285–7292, 2020.

David H. Wolpert and Kagan Tumer. Optimal payoff functions for members of collectives. *Advances in Complex Systems*, 4(2–3):265–279, 2001.

Cathy Wu, Aravind Rajeswaran, Yan Duan, Vikash Kumar, Alexandre M. Bayen, Sham Kakade, Igor Mordatch, and Pieter Abbeel. Variance reduction for policy gradient with action-dependent factorized baselines. In *International Conference on Learning Representations (ICLR)*, 2018.

H. Zou, Q. Wang, Y. Qu, Y. Jiang, L. Cai, Y. Mao, R. Peng, X. Xu, W. Liu, K. Yang, S. Yang, and X. Ji. TRACE: A unified rollout budget allocation framework for efficient agentic reinforcement learning. *arXiv preprint arXiv:2606.11119*, 2026.

A Safety and Alignment

As AI systems gain increasing capability and power, the authors of this paper are deeply concerned with the alignment of modern ML systems to humanity and its goals, as well as the safety concerns which it raises. We believe one crucial part of alignment is mechanistic understanding of these systems, which explains their behavior and their decisions. This belief is a main motivation for this paper: as ML systems become more agentic, autonomous, and decentralized, causally understanding the roles and contributions of individual agents is necessary for keeping them safe.

Our results sharpen this concern. In multi-agent training, the credit signal defines what each agent is trained to want. Under the default winner-take-all signal, agents are trained to want selection, not system success. This pressure can also favor risky outputs over good safe ones, and none of this need surface in reward metrics. Our signals remove this specific incentive mismatch, and we consider this a minimal alignment property of any decomposition of a shared reward.

In the future, we plan to investigate the mean-field limit of multi-agent systems: whether the displacement bias survives the many-agent limit, where each agent's bias shrinks because no single candidate can move the router much while the population as a whole may still be trained toward the wrong collective behavior, a misalignment invisible agent by agent; and whether the exponentially many entries of the contribution field collapse, for large populations of agents, to a function of the background's size and its mix of policies.

B Proofs of Main Results

B.1 Proof of Proposition 1

Proof. Consider two systems with mechanisms M and M'. Each has at least two agents $\{1, 2\}$ either in the same layer or in separate layers. The agents are identical in both systems. In addition, the systems have an identical constant initial history $H_1 = h$, and independent actions

$$A^{(1)}, A^{(2)} \sim \text{Bernoulli}(1/2).$$

Define the identity reward function in both systems,

$$r(h, y) := y \in \{0, 1\}.$$

In the first system, M deploys agent 1's output:

$$Y = M(H_1, A) = A^{(1)}, \qquad G = r(H_1, Y) = A^{(1)}.$$

In the second system, the mechanism deploys agent 2's output:

$$Y' = M'(H_1, A) = A^{(2)}, \qquad G' = r(H_1, Y') = A^{(2)}.$$

Both deployed rewards are distributed as $\text{Bernoulli}(1/2)$, so

$$G \overset{d}{=} G'.$$

In the first system,

$$\mathbb{E}[G \mid H_1, A^{(1)}] - \mathbb{E}[G \mid H_1] = A^{(1)} - \frac{1}{2} \in \left\{ \pm \frac{1}{2} \right\}.$$

In the second system, conditional independence gives

$$\mathbb{E}[G' \mid H_1, A^{(1)}] - \mathbb{E}[G' \mid H_1] = \mathbb{E}[A^{(2)} \mid H_1, A^{(1)}] - \frac{1}{2} = 0$$

almost surely.

Together, the two systems have identical policies and identical reward laws, yet agent 1's action determines the reward in the first and never affects it in the second, as witnessed by (2). A marginal contribution measure must therefore assign them different values, while any mapping defined on the deployed-reward law alone receives the same input in both cases, and thus will return the same value. This follows the classical treatment of identification as in (Rothenberg, 1971), where identifiability of a quantity from given data means the data determines it. □

B.2 Proof of Theorem 1

Proof. Consider a single routed layer with agents $N = \{1, \dots, K\}$. Since θ_i enters the system only through agent i's policy, the score-function identity gives

$$\nabla_{\theta_i} U_i = \nabla_{\theta_i} \mathbb{E}[\mathbf{1}_{\{I=i\}} G] = \mathbb{E}\left[\mathbf{1}_{\{I=i\}} \Psi_i G\right].$$

Equivalently for the system

$$\nabla_{\theta_i} J_{\text{sys}} = \mathbb{E}\left[\Psi_i G\right].$$

Using $J_{\text{sys}} = \sum_j U_j$,

$$b_i = \nabla_{\theta_i} U_i - \nabla_{\theta_i} J_{\text{sys}} = -\nabla_{\theta_i} \sum_{j \neq i} U_j.$$

From the above we get

$$b_i = -\mathbb{E}\left[\Psi_i \mathbf{1}_{\{I \neq i\}} G\right].$$

We now average out the router randomness. Conditional on the realized history and candidates $(H, A^{(N)})$, the router selects candidate j with probability p_j, and on the event $\{I = j\}$ the deployed reward is $G = R^{(j)}$. Thus

$$\mathbf{1}_{\{I \neq i\}} G = \sum_{j \neq i} \mathbf{1}_{\{I=j\}} R^{(j)},$$

and therefore

$$\begin{aligned}
\mathbb{E}\left[\mathbf{1}_{\{I \neq i\}} G \mid H, A^{(N)}\right] &= \sum_{j \neq i} R^{(j)} \mathbb{E}\left[\mathbf{1}_{\{I=j\}} \mid H, A^{(N)}\right] \\
&= \sum_{j \neq i} R^{(j)} \mathbb{P}\left(I = j \mid H, A^{(N)}\right) \\
&= \sum_{j \neq i} p_j R^{(j)} = D_i.
\end{aligned}$$

Since Ψ_i is determined by $(H, A^{(i)})$, and hence is measurable with respect to $(H, A^{(N)})$, the tower property gives

$$\begin{aligned}
\mathbb{E}\left[\Psi_i \mathbf{1}_{\{I \neq i\}} G\right] &= \mathbb{E}\left[\mathbb{E}\left[\Psi_i \mathbf{1}_{\{I \neq i\}} G \mid H, A^{(N)}\right]\right] \\
&= \mathbb{E}\left[\Psi_i \mathbb{E}\left[\mathbf{1}_{\{I \neq i\}} G \mid H, A^{(N)}\right]\right] \\
&= \mathbb{E}\left[\Psi_i D_i\right].
\end{aligned}$$

Hence

$$b_i = -\mathbb{E}\left[\Psi_i D_i\right].$$

We next average out the competitors' actions while retaining $(H, A^{(i)})$. Define

$$d_i(H, A^{(i)}) = \mathbb{E}\left[D_i \mid H, A^{(i)}\right].$$

Since Ψ_i is measurable with respect to $(H, A^{(i)})$, another application of the tower property gives

$$\begin{aligned}
\mathbb{E}\left[\Psi_i D_i\right] &= \mathbb{E}\left[\mathbb{E}\left[\Psi_i D_i \mid H, A^{(i)}\right]\right] \\
&= \mathbb{E}\left[\Psi_i \mathbb{E}\left[D_i \mid H, A^{(i)}\right]\right] \\
&= \mathbb{E}\left[\Psi_i d_i(H, A^{(i)})\right].
\end{aligned}$$

Moreover, the score identity gives

$$\mathbb{E}\left[\Psi_i \mid H\right] = 0.$$

Thus, considering

$$\bar{d}_i(H) = \mathbb{E}\left[D_i \mid H\right]$$

which depends only on H,

$$\mathbb{E}\left[\Psi_i \bar{d}_i(H)\right] = 0.$$

Subtracting this zero term yields

$$b_i = -\mathbb{E}\left[\Psi_i \left(d_i(H, A^{(i)}) - \bar{d}_i(H)\right)\right],$$

which proves the result. It follows that $b_i = 0$ if and only if

$$d_i(H, A^{(i)}) - \bar{d}_i(H)$$

is L^2-orthogonal to every coordinate of Ψ_i. In particular, if

$$\mathbb{E}\left[D_i \mid H, A^{(i)}\right] = \mathbb{E}\left[D_i \mid H\right],$$

then

$$d_i(H, A^{(i)}) = \bar{d}_i(H)$$

almost surely, and therefore $b_i = 0$.

Finally, suppose the router is candidate-blind for agent i. Under the layer-wise conditional independence assumption,

$$A^{(-i)} \perp A^{(i)} \mid H.$$

Candidate-blindness ensures that, with $(H, A^{(-i)})$ fixed, the competitors' routing probabilities p_j, $j \neq i$, do not vary with $A^{(i)}$. Hence

$$\mathbb{E}\left[D_i \mid H, A^{(i)}\right] = \mathbb{E}\left[D_i \mid H\right],$$

and therefore $b_i = 0$. □

B.3 Proof of Proposition 2

Proof. We show that (i) and (iii) are equivalent, and that (ii) and (iii) are equivalent.

First suppose (i): scalars $\{\Delta_i\}_{i \in N}$ satisfy totality. The singleton cases already determine every credit: taking $S = \{i\}$ gives

$$\Delta_i = v(\{i\}).$$

Thus the credits are unique. Substituting these forced values back into totality gives

$$v(S) = \sum_{i \in S} v(\{i\}) \qquad \text{for every } S \subseteq N,$$

which is (iii).

Conversely, suppose (iii). Define $\Delta_i := v(\{i\})$. Then

$$\sum_{i \in S} \Delta_i = \sum_{i \in S} v(\{i\}) = v(S),$$

so totality holds. Recall that nullity holds as well: if i is null, then $v(\{i\}) = v(\varnothing) = 0$, and hence $\Delta_i = 0$. Thus (i) holds.

It remains to show that (ii) and (iii) are equivalent. If v is additive, then adding agent j to any coalition not containing it changes the value by exactly $v(\{j\})$. Hence both sides of (6) equal $v(\{j\})$, so (ii) holds.

Now suppose (ii). We prove additivity by induction on $|S|$. The claim is immediate for $|S| \leq 1$. Suppose it holds for all coalitions of size at most m, and let $|S| = m + 1$. Choose distinct $i, j \in S$ and write $T = S \setminus \{i, j\}$. By (6),

$$v(S) = v(T \cup \{i\}) + v(T \cup \{j\}) - v(T).$$

All three coalitions on the right have size at most m, so the inductive hypothesis gives

$$v(S) = \sum_{k \in T} v(\{k\}) + v(\{i\}) + v(\{j\}) = \sum_{k \in S} v(\{k\}).$$

Thus (iii) holds, completing the proof. □

Remark 1 (A complementary pair). Take $N = \{1, 2\}$ with $v(\{1\}) = v(\{2\}) = 0$ and $v(\{1, 2\}) = 1$: the agents are valuable only together. No scalars satisfy totality, since $\Delta_1 + \Delta_2$ would need to equal both 0 and 1, and condition (6) fails. The contribution field records the interaction instead: $\Delta_{1|\varnothing} = 0$ while $\Delta_{1|\{2\}} = 1$.

B.4 Proof of Proposition 3

Proof. We prove the following: for every group $T \subseteq N$ and every ordering $i_1, \dots, i_m$ of its elements,

$$\sum_{k=1}^{m} \Delta_{i_k|\{i_1,\dots,i_{k-1}\}} = v(T), \tag{23}$$

so that crediting the agents one at a time, each against those already accounted for, recovers the group's value regardless of the order.

($\Rightarrow$) Suppose the candidates satisfy (23), and fix any $i \in N$ and $S \subseteq N \setminus \{i\}$. The pair (i, S) appears as the final step of a chain: order the elements of S arbitrarily as $i_1, \dots, i_m$ and append i as element i_{m+1}, which is an ordering of the group $S \cup \{i\}$. Apply (23) twice, once to $S \cup \{i\}$ with this ordering and once to S with the same ordering less its final element:

$$\sum_{k=1}^{m} \Delta_{i_k|\{i_1,\dots,i_{k-1}\}} + \Delta_{i|S} = v(S \cup \{i\}), \qquad \sum_{k=1}^{m} \Delta_{i_k|\{i_1,\dots,i_{k-1}\}} = v(S).$$

The two chains share every term except the last, so subtracting leaves

$$\Delta_{i|S} = v(S \cup \{i\}) - v(S).$$

Since i and S were arbitrary, this holds for every contribution.

($\Leftarrow$) Conversely, define $\Delta_{i|S} := v(S \cup \{i\}) - v(S)$ for all admissible pairs and let T and an ordering $i_1, \dots, i_m$ be given. Writing $T_k := \{i_1, \dots, i_k\}$ for the group after k steps, with $T_0 = \varnothing$, each summand is $v(T_k) - v(T_{k-1})$, and the sum telescopes

$$\sum_{k=1}^{m} \Delta_{i_k|T_{k-1}} = \sum_{k=1}^{m} \big(v(T_k) - v(T_{k-1})\big) = v(T_m) - v(T_0) = v(T),$$

using $v(\varnothing) = 0$. The differences therefore satisfy the demand for every v, and for every ordering, completing the proof. □

B.5 Proof of Lemma 1

Proof. Throughout, fix the layer ℓ of agent i and abbreviate $H := H_\ell$, $A^{(i)} := A_\ell^{(i)}$. The proof has three steps: all expectations exist; the score has conditional mean zero, which yields the gradient identity for the raw reward; and the baseline condition (9) makes the subtracted term vanish.

Step 0: integrability. The reward is bounded and $\mathbb{E}\,\|\Psi_i\|^2 < \infty$ by the standing assumption, so $\mathbb{E}\,\|\Psi_i G\|$ is finite. For the baseline term, Cauchy–Schwarz gives

$$\mathbb{E}\,\|\Psi_i B_i\| \le \left(\mathbb{E}\,\|\Psi_i\|^2\right)^{1/2} \left(\mathbb{E}\left[B_i^2\right]\right)^{1/2} < \infty,$$

since B_i is square-integrable. Every expectation below is therefore well-defined. The standing assumption permits differentiation under the expectations defining the system objective.

Step 1: the score has conditional mean zero. Conditionally on H, the action $A^{(i)}$ has density $\pi_{\theta_i}(\cdot \mid H)$, so

$$\mathbb{E}\left[\Psi_i \mid H\right] = \int \nabla_{\theta_i} \log \pi_{\theta_i}(a \mid H)\, \pi_{\theta_i}(a \mid H)\, da = \int \nabla_{\theta_i} \pi_{\theta_i}(a \mid H)\, da = \nabla_{\theta_i} \int \pi_{\theta_i}(a \mid H)\, da = 0,$$

where the second equality is the identity $\pi \nabla \log \pi = \nabla \pi$ and the last holds because the density integrates to one for every θ_i. The same computation with a θ_i-independent integrand gives the gradient identity: since θ_i enters the joint law only through $\pi_{\theta_i}(A^{(i)} \mid H)$, by the factorization (1),

$$\nabla_{\theta_i} J_{\mathrm{sys}} = \nabla_{\theta_i} \mathbb{E}\left[G\right] = \mathbb{E}\left[\Psi_i G\right].$$

Step 2: the baseline vanishes. It remains to show $\mathbb{E}[\Psi_i B_i] = 0$. Condition on $(H, A^{(i)})$, with respect to which Ψ_i is measurable, and apply the tower property together with the hypothesis (9):

$$\mathbb{E}\left[\Psi_i B_i\right] = \mathbb{E}\left[\Psi_i\, \mathbb{E}\left[B_i \mid H, A^{(i)}\right]\right] = \mathbb{E}\left[\Psi_i\, \mathbb{E}\left[B_i \mid H\right]\right] = \mathbb{E}\left[\mathbb{E}\left[B_i \mid H\right]\, \mathbb{E}\left[\Psi_i \mid H\right]\right] = 0,$$

where the third equality conditions on H and uses that $\mathbb{E}[B_i \mid H]$ is a function of H alone, and the last is Step 1. Subtracting, $\mathbb{E}[\Psi_i(G - B_i)] = \mathbb{E}[\Psi_i G] = \nabla_{\theta_i} J_{\mathrm{sys}}$, which is (10).

Step 3: indexed baselines. Let $B_i = B_i(S)$ satisfy the indexed conditions in the lemma. Then

$$\begin{aligned}\mathbb{E}\left[B_i(S) \mid H, A^{(i)}\right] &= \mathbb{E}\left[\mathbb{E}\left[B_i(S) \mid H, A^{(i)}, S\right]\middle| H, A^{(i)}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[B_i(S) \mid H, S\right]\middle| H, A^{(i)}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[B_i(S) \mid H, S\right]|H\right] \\ &= \mathbb{E}\left[B_i(S) \mid H\right].\end{aligned}$$

The second equality is the indexed baseline condition, and the third uses $S \perp A^{(i)} \mid H$. Thus the mixture satisfies (9), and Step 2 applies. The same draw of S may serve every agent for which these two indexed conditions hold. □

B.6 Proof of Corollary 1

Proof. Fix $S \subseteq N_\ell \setminus \{i\}$ and abbreviate $H := H_\ell$. The baseline

$$q_\ell^{\mathrm{ret}}(S) = \mathbb{E}\left[G \mid H, A_\ell^{(S)}\right]$$

is bounded and depends only on H and $A_\ell^{(S)}$. By (1), $A_\ell^{(S)} \perp A_\ell^{(i)} \mid H$, so $q_\ell^{\mathrm{ret}}(S)$ satisfies (9). Lemma 1 therefore gives

$$\mathbb{E}\left[\Psi_i\left(G - q_\ell^{\mathrm{ret}}(S)\right)\right] = \nabla_{\theta_i} J_{\mathrm{sys}}.$$

By (7) and (8),

$$\Delta_{i|S}^{\mathrm{ret}} = q_\ell^{\mathrm{ret}}(S \cup \{i\}) - q_\ell^{\mathrm{ret}}(S).$$

Since Ψ_i is measurable with respect to $\left(H, A_\ell^{(S\cup\{i\})}\right)$, the tower property gives

$$\mathbb{E}\left[\Psi_i q_\ell^{\mathrm{ret}}(S \cup \{i\})\right] = \mathbb{E}[\Psi_i G].$$

The admissible-baseline argument gives $\mathbb{E}[\Psi_i q_\ell^{\mathrm{ret}}(S)] = 0$, hence

$$\mathbb{E}[\Psi_i \Delta_{i|S}^{\mathrm{ret}}] = \mathbb{E}[\Psi_i G] = \nabla_{\theta_i} J_{\mathrm{sys}}.$$

This proves (11). The same definitions also give

$$\mathbb{E}\left[G - q_\ell^{\mathrm{ret}}(S) \mid H, A_\ell^{(S\cup\{i\})}\right] = q_\ell^{\mathrm{ret}}(S \cup \{i\}) - q_\ell^{\mathrm{ret}}(S) = \Delta_{i|S}^{\mathrm{ret}}.$$

For a random S, the argument holds conditionally on S and then averages, using the indexed clause of Lemma 1. □

B.7 Proof of Proposition 4

Proof. The reward is bounded, so G is square-integrable and all variances below are finite. Fix nested backgrounds $S \subseteq T \subseteq N_\ell$ and write $\mathcal{F}_S := \sigma\left(H_\ell, A_\ell^{(S)}\right)$ and $\mathcal{F}_T := \sigma\left(H_\ell, A_\ell^{(T)}\right)$, so that $\mathcal{F}_S \subseteq \mathcal{F}_T$. The law of total variance, applied to G with the two nested conditionings, gives

$$\mathbb{E}\left[\mathrm{Var}\left(G \mid \mathcal{F}_S\right)\right] = \mathbb{E}\left[\mathrm{Var}\left(G \mid \mathcal{F}_T\right)\right] + \mathbb{E}\left[\mathrm{Var}\left(\mathbb{E}\left[G \mid \mathcal{F}_T\right] \mid \mathcal{F}_S\right)\right].$$

The second term on the right is nonnegative, being an expected variance, and the result follows. □

B.8 Proof of Lemma 2

Proof. Abbreviate $H := H_\ell$, $A := A_\ell^{(N_\ell)}$. The proof has two steps: the squared loss is minimized by a conditional mean, and the mask identifies that conditional mean as the one we want.

Step 1: the minimizer is a conditional mean. The loss (13) is an expected squared error of a predictor of G from the inputs $\left(H, m(S), A^{(S)}\right)$. For any such prediction problem, the population minimizer

over all square-integrable functions of the inputs is the conditional mean $\mathbb{E}\left[G \mid H, S, A^{(S)}\right]$, and the risk decomposes as

$$\mathcal{L}(\phi) = \mathbb{E}\left[\left(G - \mathbb{E}\left[G \mid H, S, A^{(S)}\right]\right)^2\right] + \mathbb{E}\left[\left(\mathbb{E}\left[G \mid H, S, A^{(S)}\right] - \hat{q}_\phi\right)^2\right],$$

so if the critic class contains this conditional mean, any population minimizer agrees with it almost surely on the support of the inputs; this is where positive mask probability on S is used.

Step 2: independence removes S from the conditional mean. It remains to show $\mathbb{E}[G \mid H, S, A^{(S)}] = \mathbb{E}[G \mid H, A^{(S)}]$. By hypothesis S is drawn with fresh randomness, conditionally independently of (A, G) given H. Fix a background s in the support. Conditionally on H, the event $\{S = s\}$ is independent of (A, G), so conditioning on it changes no conditional law of $(A^{(s)}, G)$. Hence

$$\mathbb{E}\left[G \mid H, S = s, A^{(s)}\right] = \mathbb{E}\left[G \mid H, A^{(s)}\right] \quad \text{almost surely.}$$

In other words, the mask tells the critic which outputs it may see, and nothing else. Combining the steps, any population minimizer equals $\mathbb{E}[G \mid H, A^{(S)}]$ at every supported background, which is (14). Had the mask been allowed to depend on the actions or the return, Step 2 would fail and the target would be $\mathbb{E}[G \mid H, S, A^{(S)}]$, a mask-contaminated quantity; the independent mask is not a technicality but the identifying assumption. □

B.9 Proof of Proposition 5

Proof. The gradient identity (15) is an instance of Lemma 1 with $B_i = \hat{q}_\phi(S)$; we verify its two hypotheses.

Square-integrability holds by assumption on the critic.

Conditional mean independence. Fix first a deterministic $S \subseteq N_\ell \setminus \{i\}$. A fixed or cross-fitted critic is a deterministic function of its inputs on the trajectory where it is applied (cross-fitting ensures the function was built without that trajectory, so no dependence on the current $A_\ell^{(i)}$ enters through ϕ). The inputs $(H_\ell, m(S), A_\ell^{(S)})$ exclude $A_\ell^{(i)}$, and by the within-layer factorization (1) the retained outputs $A_\ell^{(S)}$ are conditionally independent of $A_\ell^{(i)}$ given H_ℓ. Hence

$$\mathbb{E}\left[\hat{q}_\phi(S) \mid H_\ell, A_\ell^{(i)}\right] = \mathbb{E}\left[\hat{q}_\phi(S) \mid H_\ell\right],$$

which is (9). Lemma 1 then gives $\mathbb{E}\left[\Psi_i \left(G - \hat{q}_\phi(S)\right)\right] = \nabla_{\theta_i} J_{\text{sys}}$, which is (15). For S drawn conditionally independently of $A_\ell^{(i)}$ given H_ℓ, the indexed clause of the lemma applies, and one draw serves every excluded agent. Nowhere was accuracy of the critic used: the argument is identical for a perfect critic and an untrained one.

For (16), condition on $(H_\ell, A_\ell^{(S\cup\{i\})})$. The critic term $\hat{q}_\phi(S)$ is a function of $(H_\ell, A_\ell^{(S)})$, which the conditioning contains, so it passes through:

$$\mathbb{E}\left[G - \hat{q}_\phi(S) \,\middle|\, H_\ell, A_\ell^{(S\cup\{i\})}\right] = \mathbb{E}\left[G \mid H_\ell, A_\ell^{(S\cup\{i\})}\right] - \hat{q}_\phi(S).$$

Subtract the marginal contribution

$$\Delta_{i|S} = \mathbb{E}\left[G \mid H_\ell, A_\ell^{(S\cup\{i\})}\right] - \mathbb{E}\left[G \mid H_\ell, A_\ell^{(S)}\right].$$

The first conditional expectation cancels, leaving

$$\mathbb{E}\left[G \mid H_\ell, A_\ell^{(S)}\right] - \hat{q}_\phi(S),$$

which is (16): the conditional attribution error is the critic's error at the baseline background. □

B.10 Proof of Corollary 2

Proof. Both claims are instances of Lemma 1; we verify its hypothesis for each baseline.

The removed-router draw. By construction, G^{-i}_{route} is generated from the history, the remaining realized candidates, and fresh router randomness: it is a function of $\left(H, A^{(-i)}, \xi'\right)$ with ξ' drawn independently of everything else. Under independent candidate generation (the within-layer factorization (1) applied to the routing layer), $A^{(-i)}$ is conditionally independent of $A^{(i)}$ given H, and ξ' carries no information about $A^{(i)}$ at all. Hence

$$\mathbb{E}\left[G^{-i}_{\text{route}} \mid H, A^{(i)}\right] = \mathbb{E}\left[G^{-i}_{\text{route}} \mid H\right] :$$

what the reduced slate would deliver does not depend on the output of the candidate no longer in it. The draw is bounded, hence square-integrable, and Lemma 1 gives (18).

A predictor. A fixed or cross-fitted predictor built from $\left(H, A^{(-i)}\right)$ is a deterministic function of those inputs on the trajectory where it is applied, cross-fitting ensuring that its parameters were trained without that trajectory. The same conditional independence gives the same conditional-mean identity, and the lemma applies again. Accuracy of the predictor is used nowhere. □

B.11 Proof of Corollary 3

Proof. Condition on $\left(H, A^{(1:K)}\right)$. The router weights p_j and p_j^{-i} are then fixed. By (17),

$$\mathbb{E}\left[\mathbf{1}_{\{I=j\}}\left(G - \widehat{\mu}_j\right) \mid \left(H, A^{(1:K)}\right)\right] = p_j\left(g_j - \widehat{\mu}_j\right).$$

Therefore,

$$\mathbb{E}\left[\widehat{g}_j \mid \left(H, A^{(1:K)}\right)\right] = \widehat{\mu}_j + \frac{1}{p_j} p_j\left(g_j - \widehat{\mu}_j\right) = g_j.$$

Overlap makes the ratio well-defined. Taking the difference,

$$\mathbb{E}\left[\widehat{\Delta}_i^{\text{rem}} \mid \left(H, A^{(1:K)}\right)\right] = \sum_j p_j g_j - \sum_{j \neq i} p_j^{-i} g_j = \Delta_i^{\text{rem}}.$$

The outcome model may therefore be misspecified without changing the conditional mean of the corrected contrast. □

B.12 Proof of Theorems 2 and 3

We begin by stating Theorem 3.

Theorem 3 (Matching lower bound)**.** *In the setting of Theorem 2, write $\mathcal{A}$ for the marginal contributions with $p^\star_{i|S} = 1$, $\mathcal{U}$ for the rest, and $C_{\mathcal{U}} = C_{\max} - \sum_{(i,S)\in\mathcal{A}} c_{i|S}$. Every allocation of the same expected cost satisfies*

$$\mathcal{R}(p) \geq \sum_{(i,S)\in\mathcal{A}} a_i^2 \sigma^2_{i|S} + \frac{\left(\sum_{(i,S)\in\mathcal{U}} a_i \sqrt{\sigma^2_{i|S} + \eta^2_{i|S}}\sqrt{c_{i|S}}\right)^2}{C_{\mathcal{U}}} - \sum_{(i,S)\in\mathcal{U}} a_i^2 \eta^2_{i|S}, \tag{24}$$

with equality at $p^\star$.

The following proof addresses both theorems.

Proof. Condition throughout on the data and learned estimates, so $\Delta_{i|S}$, $\eta_{i|S}$, a_i, and $c_{i|S}$ are fixed.

First, since $Z_{i|S}$ is independent of $Y_{i|S}$ and $\mathbb{E}[Z_{i|S}] = p_{i|S}$,

$$\mathbb{E}\left[\widetilde{\Delta}_{i|S}\right] = \widehat{\Delta}_{i|S} + \frac{\mathbb{E}[Z_{i|S}]}{p_{i|S}} \mathbb{E}\left[Y_{i|S} - \widehat{\Delta}_{i|S}\right] = \Delta_{i|S}.$$

Thus the correction is unbiased.

Next,

$$\widetilde{\Delta}_{i|S} - \Delta_{i|S} = \eta_{i|S} + \frac{Z_{i|S}}{p_{i|S}}\left(Y_{i|S} - \widehat{\Delta}_{i|S}\right).$$

Since

$$\mathbb{E}\left[Y_{i|S} - \widehat{\Delta}_{i|S}\right] = -\eta_{i|S}, \qquad \mathbb{E}\left[\left(Y_{i|S} - \widehat{\Delta}_{i|S}\right)^2\right] = \sigma_{i|S}^2 + \eta_{i|S}^2,$$

and $Z_{i|S}^2 = Z_{i|S}$, we obtain

$$\begin{aligned}\mathbb{E}\left[\left(\widetilde{\Delta}_{i|S} - \Delta_{i|S}\right)^2\right] &= \eta_{i|S}^2 + 2\eta_{i|S}\mathbb{E}\left[\frac{Z_{i|S}}{p_{i|S}}\left(Y_{i|S} - \widehat{\Delta}_{i|S}\right)\right] \\ &\quad + \mathbb{E}\left[\frac{Z_{i|S}}{p_{i|S}^2}\left(Y_{i|S} - \widehat{\Delta}_{i|S}\right)^2\right] \\ &= \eta_{i|S}^2 - 2\eta_{i|S}^2 + \frac{\sigma_{i|S}^2 + \eta_{i|S}^2}{p_{i|S}} \\ &= \frac{\sigma_{i|S}^2 + \eta_{i|S}^2}{p_{i|S}} - \eta_{i|S}^2.\end{aligned}$$

It remains to optimize. Write $\mathcal{M}_+$ for the set of marginal contributions with $a_i^2(\sigma_{i|S}^2 + \eta_{i|S}^2) > 0$; inactive pairs contribute zero risk for every p and may be skipped. If the budget permits evaluating every active marginal contribution, then $p_{i|S}^\star = 1$ for all $(i, S) \in \mathcal{M}_+$. Otherwise the budget binds, and terms independent of p may be dropped, leaving

$$\min_p \sum_{(i,S)\in\mathcal{M}_+} \frac{a_i^2\left(\sigma_{i|S}^2 + \eta_{i|S}^2\right)}{p_{i|S}}$$

subject to

$$\sum_{(i,S)\in\mathcal{M}_+} c_{i|S}p_{i|S} = C_{\max}, \qquad 0 < p_{i|S} \le 1.$$

The objective is strictly convex on $\mathcal{M}_+$. Let $\nu > 0$ be the multiplier for the budget constraint and $\mu_{i|S} \ge 0$ the multiplier for $p_{i|S} \le 1$. Stationarity gives

$$-\frac{a_i^2\left(\sigma_{i|S}^2 + \eta_{i|S}^2\right)}{p_{i|S}^2} + \nu c_{i|S} + \mu_{i|S} = 0.$$

When $p_{i|S} < 1$, complementary slackness gives $\mu_{i|S} = 0$, so

$$p_{i|S} = \frac{1}{\sqrt{\nu}} \frac{a_i\sqrt{\sigma_{i|S}^2 + \eta_{i|S}^2}}{\sqrt{c_{i|S}}}.$$

When this value exceeds one, the constraint binds. Setting $\lambda := 1/\sqrt{\nu}$ gives (22), with λ chosen so that the budget binds. Strict convexity gives uniqueness.

It remains to evaluate the risk at the optimum. Let $\mathcal{A}$ contain the marginal contributions with $p_{i|S}^\star = 1$, and let $\mathcal{U} = \mathcal{M}_+ \setminus \mathcal{A}$. On $\mathcal{A}$,

$$a_i^2\left(\frac{\sigma_{i|S}^2 + \eta_{i|S}^2}{p_{i|S}^\star} - \eta_{i|S}^2\right) = a_i^2\sigma_{i|S}^2.$$

On $\mathcal{U}$,

$$p_{i|S}^\star = \lambda \frac{a_i\sqrt{\sigma_{i|S}^2 + \eta_{i|S}^2}}{\sqrt{c_{i|S}}}.$$

Since the remaining budget is

$$C_{\mathcal{U}} = C_{\max} - \sum_{(i,S)\in\mathcal{A}} c_{i|S},$$

the binding budget constraint gives

$$\lambda = \frac{C_{\mathcal{U}}}{\sum_{(i,S)\in\mathcal{U}} a_i \sqrt{\sigma^2_{i|S} + \eta^2_{i|S}}\sqrt{c_{i|S}}}.$$

Substituting into the risk yields

$$\mathcal{R}(p^\star) = \sum_{(i,S)\in\mathcal{A}} a_i^2 \sigma^2_{i|S} + \frac{\left(\sum_{(i,S)\in\mathcal{U}} a_i \sqrt{\sigma^2_{i|S} + \eta^2_{i|S}}\sqrt{c_{i|S}}\right)^2}{C_{\mathcal{U}}} - \sum_{(i,S)\in\mathcal{U}} a_i^2 \eta^2_{i|S}.$$

Because $p^\star$ is the unique global minimizer, $\mathcal{R}(p) \geq \mathcal{R}(p^\star)$ for every feasible allocation. This is (24), with equality at $p^\star$. □

C Refinements of Theorem 1

The results below refine Theorem 1; each is referenced at the end of Section 3.

Proposition 6 (Selection makes private utility risk-sensitive)**.** *Under softmax routing* (3)*, private utility is not determined by the expected candidate reward. Thus, there exist candidate-reward distributions with equal conditional means that induce different private utilities.*

Proof. Condition on a fixed history. Take $K = 2$, $\epsilon = 0$, $\tau = 1$, and $s(H, A^{(j)}) = R^{(j)}$, and let $R^{(2)} = c$ be deterministic. Compare $R^{(1)}_{\text{safe}} = \frac{1}{2}$ almost surely with $R^{(1)}_{\text{risky}} \sim \text{Bernoulli}(\frac{1}{2})$. Agent 1 is selected with probability

$$p_1(r) = \frac{e^r}{e^r + e^c} = \text{sigmoid}(r - c),$$

where $\text{sigmoid}(u) = (1 + e^{-u})^{-1}$. Under the constant policy,

$$U_1^{\text{safe}} = p_1\left(\frac{1}{2}\right)\frac{1}{2} = \frac{1}{2}\,\text{sigmoid}\left(\frac{1}{2} - c\right).$$

Under the Bernoulli policy, the zero-reward branch contributes nothing, so

$$U_1^{\text{risky}} = \frac{1}{2}\,p_1(1) = \frac{1}{2}\,\text{sigmoid}(1 - c).$$

The policies have the same mean reward, but $U_1^{\text{risky}} > U_1^{\text{safe}}$ because the sigmoid function is strictly increasing. Private utility is therefore not determined by expected candidate reward. □

Remark 2 (Converse under score completeness). The reverse implication in Theorem 1, from $b_i = 0$ to conditional mean independence, holds when the score is complete: when the closed linear span in L^2 of the coordinates of Ψ_i contains every square-integrable function of $(H, A^{(i)})$ with zero conditional mean given H. Completeness holds when each history–action probability has its own free parameter, as in a finite tabular softmax, but can fail when parameters are shared across histories, as in an autoregressive model.

Proposition 7 (Directional inflation under monotone routing)**.** *Assume the router* (3) *with score $s(H, A^{(j)}) = \kappa(R^{(j)})$ for an increasing κ, and conditionally independent nonnegative candidate rewards given H. Fix a direction w_i and let*

$$\zeta_{w_i}(h, r) := \mathbb{E}\left[w_i^\top \Psi_i \mid H = h,\ R^{(i)} = r\right].$$

If $r \mapsto \zeta_{w_i}(h, r)$ is nondecreasing for almost every h, then

$$w_i^\top b_i \geq 0, \qquad \mathbb{E}\left[\mathbf{1}_{\{I=i\}}\, w_i^\top \Psi_i G\right] \geq w_i^\top \nabla_{\theta_i} J_{\text{sys}}. \tag{25}$$

The inequality is strict when no factor degenerates: the router depends on scores ($\epsilon < 1$), $R^{(i)}$ varies on a set of histories of positive probability, κ and $\zeta_{w_i}(h, \cdot)$ are strictly increasing on its conditional support, and on those histories some competitor has positive reward with positive conditional probability.

Proof. Write $Z_i := w_i^\top \Psi_i$. By (5) and $\mathbb{E}[\Psi_i \, \bar{d}_i(H)] = 0$, as in the proof of Theorem 1,

$$w_i^\top b_i = -\mathbb{E}[Z_i D_i].$$

For fixed h, define

$$m_h(r) := \mathbb{E}\left[D_i \mid H = h,\ R^{(i)} = r \right].$$

Conditional independence of candidate rewards gives

$$m_h(r) = \mathbb{E}\left[(1-\epsilon) \frac{\sum_{j \neq i} e^{\kappa(R^{(j)})/\tau} R^{(j)}}{e^{\kappa(r)/\tau} + \sum_{k \neq i} e^{\kappa(R^{(k)})/\tau}} + \frac{\epsilon}{K} \sum_{j \neq i} R^{(j)} \;\middle|\; H = h \right].$$

Since rewards are nonnegative and κ is increasing, $m_h(\cdot)$ is nonincreasing.

By the layerwise factorization, conditioning further on $R^{(i)}$ preserves the independence of $A^{(-i)}$ from $A^{(i)}$. Under the softmax router (3) with score $\kappa(R^{(j)})$, D_i depends on agent i only through $R^{(i)}$, so

$$Z_i \perp D_i \mid \left(H, R^{(i)} \right).$$

Therefore,

$$\mathbb{E}\left[Z_i D_i \mid H, R^{(i)} \right] = \zeta_{w_i}\left(H, R^{(i)} \right) m_H\left(R^{(i)} \right).$$

The score identity also gives

$$\mathbb{E}\left[\zeta_{w_i}\left(H, R^{(i)} \right) \mid H \right] = \mathbb{E}[Z_i \mid H] = 0.$$

Let $\widetilde{R}^{(i)}$ be an independent copy of $R^{(i)}$ conditional on H. The covariance identity yields

$$\begin{aligned} 2\,\mathbb{E}\left[\zeta_{w_i}\left(H, R^{(i)} \right) m_H\left(R^{(i)} \right) \mid H \right] = \mathbb{E}\Big[&\big(\zeta_{w_i}(H, R^{(i)}) - \zeta_{w_i}(H, \widetilde{R}^{(i)}) \big) \\ &\cdot \big(m_H(R^{(i)}) - m_H(\widetilde{R}^{(i)}) \big) \mid H \Big] \leq 0, \end{aligned}$$

because the first factor is nondecreasing in the reward and the second is nonincreasing. Hence

$$\mathbb{E}[Z_i D_i] \leq 0, \qquad w_i^\top b_i \geq 0.$$

Under the stated strictness conditions, m_h is strictly decreasing on a nondegenerate conditional reward support for a set of histories of positive probability. The covariance above is then strictly negative, so $w_i^\top b_i > 0$. Finally,

$$\mathbb{E}\left[\mathbf{1}_{\{I=i\}} Z_i G \right] = w_i^\top \nabla_{\theta_i} J_{\mathrm{sys}} + w_i^\top b_i,$$

which proves the second inequality and its strict form. □

D Doubly Robust Routing Details

The routing observation is $O = \left(H, A^{(1:K)}, I, G, p_{1:K} \right)$, with $G = R^{(I)}$ the selected candidate's reward. Under the ϵ-mixed softmax with weights $w_j = e^{s(H, A^{(j)})/\tau}$, the factual and removed probabilities are

$$p_j = (1-\epsilon) \frac{w_j}{\sum_{k=1}^K w_k} + \frac{\epsilon}{K}, \qquad p_j^{-i} = (1-\epsilon) \frac{w_j}{\sum_{k \neq i} w_k} + \frac{\epsilon}{K-1} \quad (j \neq i),$$

so removal renormalizes both the softmax mass and the exploration mass over the surviving agents. Removal changes only the probabilities being evaluated, p_j versus p_j^{-i} in the two sums; the inverse-propensity correction always divides by the factual p_j, because that is the router that generated the data.

With $w_{ij} = p_j - \mathbf{1}_{\{j \neq i\}}\, p_j^{-i}$, the estimator admits the centered representation

$$\widehat{\Delta}_i^{\mathrm{rem}} - \Delta_i^{\mathrm{rem}} = \sum_{j=1}^{K} w_{ij} \frac{\mathbf{1}_{\{I=j\}}}{p_j} (G - g_j) + \sum_{j=1}^{K} w_{ij} \left(1 - \frac{\mathbf{1}_{\{I=j\}}}{p_j}\right) (\widehat{\mu}_j - g_j).$$

Conditional on the data, both sums have mean zero: the first is selected-outcome noise, and the second is the augmentation term correcting outcome-model error. This is the doubly robust form specialized to known factual propensities. Overlap governs both identification and variance, since the correction scales as $1/p_j$.

For actor training without queries, the direct baseline $G - \sum_{j \neq i} p_j^{-i} \widehat{\mu}_j$ is gradient-correct whenever the cross-fitted predictor excludes $A^{(i)}$ (Corollary 2), but it is not in general an unbiased estimate of the removal attribution; the augmented inverse-propensity contrast restores attribution under the known factual propensities for any fixed or cross-fitted outcome model and is the quantity selectively corrected in Section 7.

E Implementation Details

Agent roles are direct arithmetic, equation-first reasoning, and final-answer-only generation, with sampling temperatures $(0.35, 0.50, 0.72)$ and top-p $(0.80, 0.85, 0.95)$. We assign money/rate, geometry/measurement, and counting/logic specialties; a keyword heuristic labels prompts and informs each agent whether a prompt is near its specialty. Reward is exact match after extracting the required final line `#### <number>`.

LoRA adapters use rank 16, scaling 32, zero dropout, applied to attention and MLP projections; the shared backbone is loaded in 4-bit form when needed. The router outcome model mean-pools hashed token embeddings, concatenates an agent embedding, and applies a one-hidden-layer predictor; its sigmoid is $\widehat{\mu}_j$ and its logit s_j defines $p_j = (1-\epsilon)\exp(s_j/\tau)/\sum_k \exp(s_k/\tau) + \epsilon/3$. We anneal $\tau : 1.00 \to 0.70$ and $\epsilon : 0.05 \to 0.03$; removal renormalizes the same rule over two candidates. The outcome model learns only from selected feedback using a $50{,}000$-entry replay buffer, batches of 64, inverse-propensity-weighted binary cross-entropy, propensity floor 0.05, weight clip 3.0, and eight replay steps during warmup and four thereafter. The first 25 updates train only this model, and the removal signal is computed before the current rollout enters replay.

The run configuration (`signal_single`) is 512 train and 128 test examples, 150 updates, batch size 1, $m = 4$, maximum completion length 96, and seed 42. GRPO standardizes each agent's four scores as $\mathrm{Adv}_{i,g} = (X_{i,g} - \overline{X}_i)/\max\{\mathrm{sd}(X_{i,1:m}), 10^{-6}\}$ before the clipped loss. LoRA uses AdamW at 10^{-5}, gradient norm 1.0, GRPO clip 0.2, and KL coefficient 0.02; the outcome model uses AdamW at 10^{-3}. Each actor update uses its fresh rollout and is therefore on-policy; reusing a batch would require the usual importance ratio and clipping.

Experiment implementation repository available in supplementary material.