Game-Theoretic Multi-LLM Routing for Safer Agents in the Wild

Jing Wang [1] **Jie Shen** [2] **Dean Foster** [3] **Zohar Karnin** [4]

Abstract

We propose SRLA (Stackelberg-Routed Leverage-Score Aggregation), a multi-LLM framework in which routing, traffic equilibrium, and output aggregation each carry a game-theoretic guarantee. We model the agent population as a routing graph whose node latencies combine prediction loss and congestion: a Stackelberg-pinned parallel-path topology bounds the price of anarchy by a constant, and online gradient descent on the mixture log-loss converges to a proximal correlated equilibrium with rank-dependent regret. The same bounds yield concrete *safety* and *fairness* properties: the price-of-anarchy bound caps any single agent's worst-case influence on the system, so a misbehaving or compromised agent cannot dominate the decisions. The proximal correlated equilibrium and the leverage-score weighting together certify δ-robustness of the deployed weights and resist the failure mode in which a single overconfident, biased agent hijacks the ensemble.

1. Introduction

We study the problem of designing multi-agent systems whose agents are large language models (LLMs). When several LLMs collaborate on a query, the resulting system can outperform any single agent: CAMEL pairs role-playing agents (Li et al., 2023), AUTOGEN and METAGPT compose programmable conversations (Wu et al., 2024; Hong et al., 2024), multi-agent debate iterates argumentation (Du et al., 2024), and Mixture-of-Agents (Wang et al., 2024) layers open-source LLMs to match the performance of GPT-4. To date, however, these systems are largely hand-engineered, and there is no principled answer to the question of *how many agents to instantiate, in what topology, and how to aggregate their outputs*. The lack of principled coordination is also a source of well-documented risks: single LLMs are known to exhibit biased and harmful outputs (Weidinger et al., 2021; Gallegos et al., 2024), and to be susceptible to prompt-injection-style attacks (Greshake et al., 2023); in a multi-agent system without coordination guarantees these failure modes can be amplified, since an unconstrained router can concentrate decisions on a single biased or miscalibrated agent and small perturbations of any one agent's loss can swing the deployed prediction.

Our starting point is classical game theory. We model the population of n LLM agents as the nodes of a routing graph; each node v carries a linear latency $\ell_v(f_v) = \kappa_v f_v + \mathcal{L}_v(\theta_v)$, with κ_v the per-call inference cost and $\mathcal{L}_v(\theta_v)$ the agent's validation loss and f_v be some flow function to be defined later. We highlight that, unlike the constant edge latencies considered in classical routing games, the latency intercept $\mathcal{L}_v(\theta_v)$ is itself a learnable function of the agent parameters — a feature we will exploit in Stage C of our pipeline.

The design draws on two classical equilibrium concepts. The first, the Wardrop / Nash equilibrium for selfish routing (Wardrop, 1952; Koutsoupias & Papadimitriou, 1999; Roughgarden, 2003), has price of anarchy (PoA) at most $4/3$ under linear latencies, which further reduces to $\beta + (1-\beta)\rho$ under a Stackelberg leader at leader fraction β (Roughgarden, 2004). The second, the proximal correlated equilibrium (PCE) of Cai et al. (2024a), is attainable by online gradient descent in concave games. Both concepts were developed for edge latencies and linear losses, and to our knowledge neither has been brought to bear on the design of multi-LLM systems. Doing so requires lifting them to node latencies coupled to a learnable classifier head, and to the mixture log-loss objective induced by the deployed predictor. This raises a natural question: *Can we design a multi-LLM system whose routing topology, traffic equilibrium, and output aggregation each carry an end-to-end game-theoretic guarantee?*

Safety and fairness from the bounds. Because the price-of-anarchy bound directly caps the routing mass any single agent can attract, no anomalous or miscalibrated agent can dominate the system's decisions — a worst-case safety

[1]Independent researcher, hfutwj@gmail.com [2]Department of Computer Science, Stevens Institute of Technology, Hoboken, NJ, USA, jie.shen@stevens.edu [3]Amazon, NYC, USA, dean@amazon.com [4]TII, Haifa, Israel, Zohar.Karnin@tii.ae. Correspondence to: Jing Wang <hfutwj@gmail.com>.

Published at the Second Workshop on Agents in the Wild: Safety, Security, and Beyond (AIWILD) at ICML 2026. Copyright 2026 by the author(s).

guarantee against a single misbehaving agent. Because the proximal correlated equilibrium certifies stability under δ-bounded perturbations of any agent's loss, the deployed weights are provably robust to local prompt-injection-style attacks (Greshake et al., 2023) and to overconfident hallucinated outputs from a single agent. And because the leverage-score initialization weights agents by the dominant directions of their loss-disagreement matrix rather than by raw confidence, the aggregator resists the failure mode in which a single overconfident, demographically biased agent hijacks the ensemble (Weidinger et al., 2021; Gallegos et al., 2024; Shen et al., 2022). We therefore view SRLA not just as a performance-driven design but as a step towards trustworthy multi-LLM deployment, with concrete safety and fairness properties that follow directly from the game-theoretic bounds.

We attempt to answer the question in the affirmative. Specifically, we propose SRLA (Stackelberg-Routed Leverage-Score Aggregation), a three-stage pipeline (Algorithm 1) which can be understood as block-coordinate descent on a joint prediction–optimization objective Eq. (7), one stage per block of decision variables. Stage A pins a fraction β of the routing mass on the higher-cost agents, which controls the social cost; Stage B runs leverage-initialized online gradient descent on the mixture log-loss to obtain the aggregation weights; and Stage C closes the loop by fine-tuning each agent's classifier head on its leader-routed queries, which lowers the latency intercept and feeds back into the next iteration.

Our main contributions are summarized as follows.

1) Node-weighted price-of-anarchy and rank-dependent proximal regret. On the theoretical side, we lift the topology-independent and Stackelberg PoA theory of Roughgarden (2003; 2004) from edge to node latencies (Theorems 5.1 and 5.2). This lift is non-trivial: since the latency intercept $\mathcal{L}_v(\theta_v)$ is itself learnable, the residual follower game has to be shown to remain in the linear class after pinning. In addition, we establish a rank-dependent specialization of proximal-regret online gradient descent with leverage-score initialization on the mixture log-loss (Theorem 5.3); this replaces the standard $O(\sqrt{nT})$ external-regret rate with $O(\sqrt{\mathrm{rank}_k(L)\,T})$ proximal regret. The three results are then composed by a monotone-descent proposition (Proposition 5.4) into a per-iteration bound for the closed-loop pipeline.

2) Empirical validation. On eight classification and multiple-choice QA benchmarks with Qwen-2.5-7B-Instruct and Qwen3-14B as agents, SRLA attains the best average accuracy and improves over the strongest baseline in every category, including a competitive PMI-normalized zero-shot choice-scoring LLM baseline. We additionally conduct an ablation on the leverage-score initialization, which confirms the rank-dependent convergence rate predicted by Theorem 5.3.

2. Related Work

Multi-LLM systems, ensembles, and routers. A line of recent works developed hand-engineered multi-agent stacks, including role-playing (Li et al., 2023), programmable conversation frameworks (Wu et al., 2024; Hong et al., 2024), debate (Du et al., 2024), layered Mixture-of-Agents (Wang et al., 2024), and self-ensembled reasoning chains (Wang et al., 2023; Yao et al., 2023). Closer to our setting, Jiang et al. (2023); Lu et al. (2024) learned per-query LLM routers, while Si et al. (2024); Xu et al. (2024) weighted agents by their confidence. None of these works provides topology-design or aggregation guarantees: the choices of *how many* agents, *which* topology, and *how* to combine the outputs are heuristic. Our work extends the predict-then-optimize view of Elmachtoub & Grigas (2022) from a single predictor to a population of fine-tunable LLM agents, and supplies per-stage game-theoretic guarantees.

Selfish routing and Stackelberg coordination. The routing-game thread originated with Wardrop (1952) and Koutsoupias & Papadimitriou (1999). Roughgarden (2003) established that $\mathrm{PoA} = 4/3$ for linear edge latencies, independent of topology, and Roughgarden (2004) bounded PoA on parallel links by $\beta + (1-\beta)\rho$ at leader fraction β. Recent extensions cover side information (Lin et al., 2024), learning agents under payments (Yan & Zhang, 2025), node-weighted externalities (Paccagnan & Marden, 2024), oracle-based multi-agent reinforcement-learning leaders (Gerstgrasser & Parkes, 2023), and imperfect-information variants (Kovářík et al., 2024). All of these works assumed *edge* latencies independent of any learnable predictor. We lift this entire line to *node* latencies that combine an LLM's per-call cost with its prediction loss $\mathcal{L}_v(\theta_v)$, so that fine-tuning the classifier head directly lowers the routing intercept (Theorems 5.1 and 5.2).

No-regret learning, matrix games, and proximal correlated equilibrium. Cai et al. (2024a) introduced proximal regret and PCE, and showed that plain online gradient descent attains a PCE in concave games. Cai et al. (2024b); Anagnostides et al. (2023); Farina & Pipis (2024) sharpened this line to polynomial-time Φ-regret. For matrix games on the simplex, Carmon et al. (2020) used leverage-score sampling as a variance-reduction primitive, and Karmarkar et al. (2026) obtained $\tilde{O}(\epsilon^{-2/3})$ matvec complexity. These results were established for linear-in-w losses on the simplex. Our aggregation objective is the *mixture log-loss* $F(w) = -\frac{1}{N}\sum_t \log\langle w, p(y_t)\rangle$, which is convex but not linear. We carry the leverage-score initialization to this F and obtain a rank-dependent regret bound, together with PCE

convergence specific to LLM aggregation (Theorem 5.3).

3. Problem Setup

We follow the model and notation of Roughgarden (2003) and lift it from edge to node latencies so that an LLM agent (a transformer plus a linear classifier head) becomes a node in a routing instance.

Multi-LLM routing instance. We model a population of n LLM agents as the nodes of a directed acyclic graph $G = (V, E)$ with a source-destination pair (s, t). Let $\mathcal{P}$ denote the set of (simple) s–t paths; we assume $\mathcal{P} \neq \varnothing$. A unit query rate $r = 1$ is to be split among the paths in $\mathcal{P}$. Each node $v \in V$ carries a tuple $(\theta_v, h_v, \kappa_v)$, where θ_v are the LLM's frozen transformer parameters, $h_v : \mathbb{R}^d \to \mathbb{R}^K$ is a (trainable) linear classifier head producing a K-class score $z_v(x) = h_v(\phi_v(x; \theta_v))$ from the last-token hidden state ϕ_v, and $\kappa_v > 0$ is the per-call inference cost (FLOPs / latency).

A *flow* is a function $f : \mathcal{P} \to \mathbb{R}_{\geq 0}$, and the load on node v is given by $f_v = \sum_{P \ni v} f_P$. The flow is *feasible* if $\sum_{P \in \mathcal{P}} f_P = 1$. Each node is endowed with a *linear latency*

$$\ell_v(f_v; \theta_v) \;=\; \kappa_v\, f_v \;+\; \mathcal{L}_v(\theta_v), \tag{1}$$

where κ_v is the congestion slope and $\mathcal{L}_v(\theta_v) = \mathbb{E}_{(x,y)}[-\log p_{\theta_v, h_v}(y \,|\, x)]$ is the validation loss of the LLM-plus-head at zero load – the cost of routing a query to a less-accurate agent. We use the linear class throughout because, by Table 1 of Roughgarden (2003), it admits the tight PoA bound of $4/3$. In our experiments, $\kappa_{7B} = 1$ and $\kappa_{14B} = 2$. The path latency is $\ell_P(f) = \sum_{v \in P} \ell_v(f_v)$, and the *social cost* of a feasible flow f is

$$C(f) \;=\; \sum_{P \in \mathcal{P}} \ell_P(f)\, f_P \;=\; \sum_{v \in V} \ell_v(f_v)\, f_v. \tag{2}$$

The triple (G, r, ℓ) specifies an *instance*; a feasible flow minimizing $C(f)$ is *optimal* or *minimum-latency* and denoted f^*. Eq. (1) differs from the classical edge-latency model in two ways: latencies live on nodes (the LLM call is the resource, not the wire), and the intercept $\mathcal{L}_v(\theta_v)$ is a *learnable* function of the agent parameters – this is what couples the routing game to fine-tuning in Stage C.

Definition 3.1 (Nash / Wardrop flow, Wardrop 1952; Roughgarden 2003 Def. 2.1)**.** A feasible flow f is at *Nash equilibrium* if for all $P_1, P_2 \in \mathcal{P}$ with $f_{P_1} > 0$ and all $\delta \in (0, f_{P_1}]$, $\ell_{P_1}(f) \leq \ell_{P_2}(\tilde{f})$, where $\tilde{f}$ shifts δ mass from P_1 to P_2: $\tilde{f}_P = f_P - \delta$ if $P = P_1$, $\tilde{f}_P = f_P + \delta$ if $P = P_2$, and $\tilde{f}_P = f_P$ otherwise. Letting $\delta \to 0$, continuity of ℓ_v yields the equivalent characterization: every used path has minimum latency, $\ell_{P_1}(f) \leq \ell_{P_2}(f)$ for all P_1 with $f_{P_1} > 0$ and all $P_2 \in \mathcal{P}$.

Definition 3.2 (Price of anarchy, Koutsoupias & Papadimitriou 1999)**.** For a class $\mathcal{L}$ of allowable latencies, $\mathrm{PoA}(G, \mathcal{L}) = \sup_{f^{NE}} C(f^{NE})/C(f^*)$ where the supremum is over Nash flows on instances (G, r, ℓ) with $\ell \in \mathcal{L}$.

Stackelberg coordination. A central controller (*leader*) commits in advance to a partial flow $\bar{f}$ on a fraction $\beta \in [0, 1]$ of the mass by *pinning* a subset $V_L \subseteq V$ of agents to fixed loads. The remaining $1 - \beta$ is routed selfishly by *followers* and reaches a residual Wardrop equilibrium $f^{NE} \,|\, \bar{f}$ on the modified game where each pinned node's effective latency is shifted by $\kappa_v \bar{f}_v$. The total flow is $\bar{f} + (f^{NE} \mid \bar{f})$ with social cost

$$\begin{aligned} &\mathrm{Cost}(\bar{f}) \\ &= C\big(\bar{f} + (f^{NE} \,|\, \bar{f})\big) \\ &= \sum_{v \in V} \big[\bar{f}_v + (f^{NE} \,|\, \bar{f})_v\big]\, \ell_v\big(\bar{f}_v + (f^{NE} \,|\, \bar{f})_v; \theta_v\big). \end{aligned} \tag{3}$$

We instantiate two pinning rules adapted from Roughgarden (2004). NodeScale pins each leader-controlled agent at fraction β of its OPT load, $\bar{f}_v = \beta f^*_v$. NodeLLF (*Largest Latency First*) sorts agents in the social optimum by latency $\ell_{v_1}(f^*_{v_1}) \geq \cdots \geq \ell_{v_n}(f^*_{v_n})$ and pins them top-down until the leader budget β is saturated:

$$\bar{f}^{\,\mathrm{NodeLLF}}_{v_i} \;=\; \begin{cases} f^*_{v_i}, i < k, & \\ \beta - \sum_{j<k} f^*_{v_j}, & i = k, \\ 0, & i > k, \end{cases} \tag{4}$$
$$\text{where } k = \min\Big\{ r : \sum_{i \leq r} f^*_{v_i} \geq \beta \Big\}.$$

Roughgarden (2004) bounded $\mathrm{Cost}(\bar{f})$ relative to $C(f^*)$ on parallel-link graphs with edge latencies; we lift the result to general path-decomposable G with the node latencies of Eq. (1) in Sec. 5.

Aggregation game. On any input x the active agents emit a stack of softmax-prob predictions $P(x) = [p_1(x); \ldots; p_n(x)] \in \mathbb{R}^{n \times K}$ with $p_i(x) = [\sigma(z_i(x))]_k$. An aggregation rule with simplex weights $w \in \Delta^{n-1}$ outputs the single distribution

$$\mathcal{A}_w(P(x)) \;=\; \sum_{i=1}^{n} w_i\, p_i(x). \tag{5}$$

Choosing w against an adversarial loss vector $\ell \in [0, 1]^n$ that assigns per-agent loss $\ell^i_t = -\log p_i(y_t)$ at example t yields the saddle problem $\min_{w \in \Delta} \max_{\ell \in [0,1]^n} \langle w, \ell \rangle$ on the per-example per-agent loss matrix $L \in \mathbb{R}^{N \times n}$ – a standard matrix game on the simplex, for which uniform-init OGD attains $O(\sqrt{nT})$ external regret and the time-averaged iterate is a coarse correlated equilibrium (Carmon et al., 2020; Cai et al., 2024a). Our Stage B departs from this template in two places. First, it replaces the linear loss $\langle w, \ell \rangle$ by the

mixture log-loss, the actual aggregation objective induced by the deployed predictor Eq. (5):

$$F(w) \;=\; -\frac{1}{N}\sum_{t=1}^{N}\log\Big(\sum_{i=1}^{n} w_i\, p_i(y_t\,|\,x_t)\Big), \qquad w\in\Delta^{n-1}. \tag{6}$$

F is convex in w on the simplex (composition of $-\log$ with a linear functional) and its gradient is $\nabla F(w)_i = -\frac{1}{N}\sum_t p_i(y_t \mid x_t)/\sum_j w_j p_j(y_t \mid x_t)$. Second, it initializes OGD at the energy-weighted top-k leverage scores of L, which captures cross-agent loss-vector diversity and replaces $\sqrt{nT}$ regret by $\sqrt{\mathrm{rank}_k(L)\,T}$ (Section 5).

Joint prediction–optimization objective. Combining the routing and aggregation pieces into a single objective gives

$$\min_{G,\,\theta,\,\bar f,\,w}\;\; \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\ell_{\mathrm{task}}(\mathcal{A}_w(P(x)),\,y)\big] \;+\; \lambda\,\mathrm{Cost}(\bar f), \tag{7}$$

where the first term is the test-time prediction error of the deployed aggregation rule, the second is the Stackelberg-induced inference cost, $\lambda\geq 0$ trades them off, and the decision variables are the topology G, the per-agent parameters $\theta=\{\theta_v\}$, the leader's pinning $\bar f$, and the aggregation weight w. Joint optimization over all four is intractable; Algorithm 1 decouples them into a Stackelberg-routing step over $(G,\bar f)$, an aggregation step over w, and an optional fine-tuning step over θ, each with the per-stage guarantee proved in Section 5.

4. Algorithm

We solve Eq. (7) as a three-stage pipeline summarized in Algorithm 1. The pipeline performs block-coordinate descent on the joint prediction–optimization objective Eq. (7), decomposing it across the three groups of decision variables: the leader's pinning $\bar f$, the aggregation weight w, and the per-agent parameters θ. Stage A optimizes $\bar f$ at fixed (w,θ) via the NODELLF pinning rule Eq. (4) and bounds the resulting social cost Eq. (3) by Theorem 5.2. Stage B optimizes w at fixed $(\bar f,\theta)$ by minimizing the mixture log-loss Eq. (6) via projected online gradient descent, initialized at the leverage scores defined below; Theorem 5.3 bounds its proximal regret. Stage C optimizes θ at fixed $(\bar f, w)$ by fine-tuning each agent's classifier head on its leader-routed queries, which lowers the latency intercept $\mathcal{L}_v(\theta_v)$ in Eq. (1) and feeds back into the per-agent loss matrix L. Proposition 5.4 ties the three blocks into a closed-loop monotone-descent guarantee on the regret-augmented social cost.

Leverage-score initialization. Let $L = USV^\top$ be the SVD of the per-example per-agent loss matrix $L\in\mathbb{R}^{N\times n}$ with $L_{ti}=-\log p_i(y_t\,|\,x_t)$, and fix a truncation rank k. The energy-weighted top-k leverage score of agent i is

$$\tau_i \;=\; \sum_{j=1}^{k}\frac{S_j^2}{\sum_{l=1}^{k}S_l^2}\,V_{ji}^2, \qquad i=1,\dots,n, \tag{8}$$

so that $\sum_i \tau_i = 1$. We blend τ with the inverse-validation-loss best-response prior $\mathcal{L}_i^{-1}$ at mixing weight $\alpha=0.7$ to obtain the OGD initialization

$$w^{(0)} \;\propto\; \alpha\,\tau \;+\; (1-\alpha)\,\mathcal{L}^{-1}, \qquad w^{(0)}\in\Delta^{n-1}. \tag{9}$$

The leverage weights τ project the gradient stream onto the dominant rank-k subspace of L, which captures the agents that contribute most to the per-example loss variance and replaces the $\sqrt{nT}$ regret of uniform init by the rank-dependent rate $\sqrt{\mathrm{rank}_k(L)\,T}$ (Theorem 5.3).

Projected OGD update. Stage B performs T steps of online gradient descent on the mixture log-loss Eq. (6) with simplex projection Π_Δ:

$$w^{(t+1)} \;=\; \Pi_\Delta\Big(w^{(t)} \;-\; \eta\,\nabla F\big(w^{(t)}\big)\Big), \qquad t=0,\dots,T-1, \tag{10}$$

with step $\eta = \sqrt{2/T}/\|L\|_\infty$. The deployed weight is the tail-averaged iterate $\bar w = \frac{1}{T/2}\sum_{t=T/2}^{T} w^{(t)}$, which by Theorem 5.3 is a δ-PCE of the aggregation game.

Number of closed-loop passes. In practice we run a single closed-loop pass through Stages A–C and report the post-Stage-C accuracy as a separate row in the ablation table. Additional passes did not improve performance in our experiments; we conjecture this reflects the diminishing returns predicted by Proposition 5.4 once the leverage-projected gradient stream has stabilized.

5. Theoretical Analysis

This section establishes four guarantees that together cover every stage of Algorithm 1: a topology-independent price-of-anarchy bound (Theorem 5.1) and a Stackelberg-coordinated refinement of it (Theorem 5.2), both for Stage A; a rank-dependent proximal-regret bound for the leverage-initialized OGD of Stage B (Theorem 5.3); and a monotone-descent proposition that ties the three blocks into a closed-loop guarantee (Proposition 5.4). The proofs use three classical results – the topology-independent PoA bound of Roughgarden (2003), the parallel-link Stackelberg bound of Roughgarden (2004), and the proximal-regret bound for OGD on the simplex due to Cai et al. (2024a); Carmon et al. (2020) – as black-box lemmas; their precise statements (Lemmas A.1–A.3) and the full proofs are deferred to Appendix A.1 and A.2 respectively. Theorems 5.1–5.3 bound the cost of Stages A and B; Proposition 5.4 bounds the closed-loop iterates of Stage C.

Algorithm 1 SRLA: Stackelberg-Routed Leverage-Score Aggregation

Input: agents $\{(\theta_v, h_v, \kappa_v)\}_{v=1}^n$; calibration set $\mathcal{D}_{\text{cal}}$; leader fraction β; OGD horizon T, step η; rank k.
// Stage A.
Sort agents by $\ell_{v_1}(f^*_{v_1}) \geq \cdots \geq \ell_{v_n}(f^*_{v_n})$ in (1)
Truncate at total mass β via (4) to obtain pinning $\bar{f}$
// Stage B.
For each $(x_t, y_t) \in \mathcal{D}_{\text{cal}}$, run the agents' last layers to form $L_{ti} = -\log p_i(y_t \mid x_t)$
Initialise $w^{(0)}$ from the leverage scores (8)/(9)
Iterate the projected OGD update (10) on the mixture log-loss (6) for T steps
Set $\bar{w}$ to the tail-averaged iterate
// Stage C.
For each agent v, refit the head h_v for one SGD epoch on $\mathcal{D}_v = \{(x, y) : \bar{f} \text{ routes } x \text{ through } v\}$
Decrease $\mathcal{L}_v(\theta_v)$ in (1)
Recompute L and re-run Stage B once
// Inference.
Compute agent probabilities $p_i(\cdot|x)$ on routes selected by $\bar{f}$
Return $\hat{y} = \arg\max_y \sum_i \bar{w}_i p_i(y|x)$ via (5)
Return $(\bar{f}, \bar{w}, \{h_v\})$

Theorem 5.1 (Topology-independent PoA for LLM graphs)**.** *Let $\mathcal{L}_{\text{lin}}$ be the class of linear node latencies of the form Eq. (1), and let G be any path-decomposable graph with source-sink pair (s, t). Then $\text{PoA}(G, \mathcal{L}_{\text{lin}}) \leq \rho(\mathcal{L}_{\text{lin}}) := \text{PoA}(\textit{ParallelPath}_n, \mathcal{L}_{\text{lin}}) = 4/3$, matching Table 1 of Roughgarden (2003).*

Sketch. Replace each node v by two parallel edges with constant latency $\mathcal{L}_v(\theta_v)$ and linear latency $\kappa_v x$. Both lie in the linear edge class, which is closed under positive scaling and translation; additive separability of Eq. (1) makes the lift cost-preserving on every path-decomposable flow, so $\text{PoA}(G, \mathcal{L}_{\text{lin}}) = \text{PoA}(\widetilde{G}, \widetilde{\mathcal{L}})$. Lemma A.1 then bounds the right-hand side by $\rho(\widetilde{\mathcal{L}}) = 4/3$ via Table 1 of Roughgarden (2003). Full proof in App. A.2. □

Theorem 5.1 bounds the worst-case routing inefficiency in the absence of any coordination. The next theorem strengthens this bound when a Stackelberg leader pins a fraction β of the routing mass.

Theorem 5.2 (Stackelberg PoA on LLM graphs)**.** *For any path-decomposable LLM graph G with linear node latencies Eq. (1) and any leader fraction $\beta \in [0, 1]$, the Stackelberg flow under* NODESCALE *satisfies* $\text{Cost}(\bar{f}^{\text{SCALE}}) \leq (\beta + (1-\beta) \cdot \frac{4}{3}) \cdot C(f^*)$. *The bound interpolates linearly between $C(f^*)$ at $\beta = 1$ (full coordination) and the bound $\frac{4}{3}C(f^*)$ of Theorem 5.1 at $\beta = 0$ (no leader).*

Sketch. Apply the lift of Theorem 5.1; any NODESCALE pinning on G corresponds to an edge SCALE pinning on $\widetilde{G}$. Since $\ell_v(\bar{f}_v + x; \theta_v) = \kappa_v x + (\mathcal{L}_v(\theta_v) + \kappa_v \bar{f}_v)$ is again linear in x, the residual follower game on $\widetilde{G}$ is in the same class. Lemma A.2 with $\rho(\widetilde{\mathcal{L}}) = 4/3$ then gives $\text{Cost}(\bar{f}^{\text{SCALE}}) \leq (\beta + (1-\beta) \cdot \frac{4}{3})\, C(f^*)$. Full proof in App. A.2. □

Theorem 5.3 (Proximal regret and PCE convergence)**.** *Assume per-step losses $L_{t,:} \in [0, 1]^n$ and let $\{w^{(t)}\}$ be the OGD iterates of Algorithm 1 Stage B initialized at the energy-weighted top-k leverage scores of L. For every $\delta > 0$ and every u in a δ-neighbourhood of $w^{(t)}$,*

$$\sum_{t=1}^{T} \langle w^{(t)} - u,\, L_{t,:} \rangle - \tfrac{1}{2\delta} \|w^{(t)} - u\|_2^2 \;=\; O\big(\sqrt{\text{rank}_k(L)\, T}\big).$$

The time-averaged iterate is a δ-proximal correlated equilibrium (PCE) of the aggregation game, strictly stronger than the CCE attained by uniform-init external-regret minimization.

Sketch. F in Eq. (6) is convex in w on the simplex (composition of $-\log$ with a linear functional), so the proximal-regret reduction of Cai et al. (2024a, Thm. 2) extends pointwise from linear losses to F. The energy-weighted top-k leverage init Eq. (8) concentrates $w^{(0)}$ on the right-singular subspace of L, projecting the gradient stream onto an $\text{rank}_k(L)$-dimensional effective subspace; standard OGD analysis then replaces $\sqrt{n}$ by $\sqrt{\text{rank}_k(L)}$ in the leading term, and the PCE conclusion follows from the proximal reduction. Full proof in App. A.2. □

Proposition 5.4 (Monotone descent of closed-loop iterations)**.** *Let $\Phi^{(k)} = \mathbb{E}[\ell_{\text{task}}(\mathcal{A}_{\bar{w}^{(k)}}, y)] + \lambda\, \text{Cost}(\bar{f}^{(k)})$ be the regret-augmented social cost at iteration k of Stage C. Under Eq. (1) and a sufficiently small SGD step size, $\Phi^{(k+1)} \leq \Phi^{(k)} - \frac{1}{2\delta}\|\bar{w}^{(k+1)} - \bar{w}^{(k)}\|_2^2 + o(1)$.*

Sketch. Block-coordinate descent on Φ over $(\bar{f}, w, \theta)$. Theorem 5.2 ensures NODELLF routing decreases $\text{Cost}(\bar{f})$ at fixed θ; Theorem 5.3 ensures OGD on w contracts the proximal-regret term, which equals the second-order term in the descent inequality up to $o(1)$; SGD on the head h_v at small step size shrinks $\mathcal{L}_v(\theta_v)$ by the descent lemma, lowering the latency intercept of node v in Eq. (1). Bounded gradients absorb cross-block terms into the $o(1)$ slack. Full proof in App. A.2. □

Pipeline composition. The four results compose into a per-stage guarantee for the entire pipeline. Theorems 5.1, 5.2 reduce topology design (Stage A) to a parallel-path Stackelberg problem at any leader fraction β with social-cost guarantee $(\beta + (1-\beta) \cdot \frac{4}{3})\, C(f^*)$. Theorem 5.3 bounds the regret of the leverage-initialized OGD

that Stage B runs on the resulting active agents at rank-dependent rate $O(\sqrt{\mathrm{rank}_k(L)\,T})$. Proposition 5.4 closes the loop by showing one round of Stage C strictly decreases the regret-augmented social cost. The overall guarantee is a constant-factor approximation of the joint objective Eq. (7) whose constants depend only on β, the leverage rank k, and the OGD horizon.

PCE vs. CCE in practice. CCE allows arbitrary unilateral switches and so admits implausible solutions; PCE forbids small (proximal) deviations, so the deployed w is provably robust to local perturbations of the agent losses. As a corollary, any perturbation of any single agent's loss within δ moves the aggregation weight by at most $O(\delta)$. Empirically (Sec. 6, Fig. 2), the leverage init beats uniform init by 0.18–0.23 at $T \geq 500$, validating the rank-dependent rate of Theorem 5.3.

6. Experiments

Setup. We use Qwen-2.5-7B-Instruct and Qwen3-14B (Yang et al., 2024) in fp16 on RTX-6000-class GPUs, each paired with two prompts (raw vs. CoT-style). For classification we instantiate two agent classes per (LLM, prompt): a *trained-head* agent fitting a linear classifier $W \in \mathbb{R}^{K\times d}$ on the last-token hidden state ϕ_v ($d = 3584$ for 7B, 5120 for 14B), and a *zero-shot* agent using length-normalized choice scoring on the LM head – $n = 8$ in total. For QA, trained heads dominate zero-shot scoring by 20–30 so we use $n = 4$ trained heads only. Each agent is one s–t path on the parallel-path graph; cost weights $\kappa_{7B} = 1$, $\kappa_{14B} = 2$ reflect inference-cost asymmetry, and we pin $\beta = 0.5$ of the mass on the 14B tier via NODELLF. *Datasets:* SST-2 (Wang et al., 2019), AG-News, IMDB, Yelp (Zhang et al., 2015; Maas et al., 2011) for classification; ARC (Clark et al., 2018), CSQA (Talmor et al., 2019), MMLU (Hendrycks et al., 2021), OBQA (Mihaylov et al., 2018) for QA. The closed-loop variant is reported in App. A.3.

Baselines. We compare OURS to a comprehensive set of *routing*, *ensemble*, and *game-theoretic* baselines.

- **Zero-shot literature baselines:** ZEROSHOT-1TOK chooses the next-token with maximal log-probability over the verbalizer first token; ZEROSHOT-CHOICE selects a length-normalized multi-token choice log-probability with maximal point-wise mutual information normalization $\log P(c\,|\,x) - \log P(c\,|\,\mathit{null})$.
- **Routing / single-agent baselines:** SINGLE-BEST routes every query through the trained-head agent with the lowest validation loss (oracle single-agent routing); on the QA datasets this picks Qwen3-14B-raw, matching the trained-head row of Table 3 on $5/8$ datasets.
- **Ensemble baselines:** UNIFORM assigns weights $w_i = 1/n$ and aggregates softmax probabilities (uniform routing / logit aggregation); MAJORITY selects the label by per-agent argmax voting, with ties broken by summed confidence; CONFVOTE weights votes using both agent confidence and the regret-minimizing weights from OURS (confidence-weighted voting).
- **Game-theoretic baselines:** EXTERNALREGRET applies OGD to w from a uniform initialization, serving as a standard CCE-attaining benchmark. In contrast, **OURS** initializes w using top-k truncated SVD leverage scores and refines it with OGD, yielding convergence to a proximal correlated equilibrium (PCE).

Metrics. We report accuracy for classification and exact match for QA, and additionally report the empirical price of anarchy $\mathrm{PoA} = C(f^{NE})/C(f^*)$ under the linear node latency Eq. (1), with $\kappa_{7B} = 1$, $\kappa_{14B} = 2$, and $\mathcal{L}_v(\theta_v)$ set to the per-agent validation cross-entropy.

Results. At a high level, SRLA attains the best average accuracy across all baselines and improves over the strongest baseline in each of the four categories (zero-shot, routing/single-agent, ensemble, and game-theoretic). The gains are most pronounced on the QA tasks, where OURS wins outright on all four datasets (ARC, CSQA, MMLU, OBQA), and on the high-disagreement classification dataset Yelp; on the high-agreement classification datasets (SST-2, AG-News, IMDB) it stays within 1.4 points of the column best, where simple uniform or external-regret aggregation is already near-optimal. We unpack this headline below as a sequence of observations, all read off Table 1 and Figs. 1–2. Throughout, *pairwise agreement* on a task is the average over agent pairs (i, j) of $\Pr_x[\hat{y}_i(x) = \hat{y}_j(x)]$; values close to 1 indicate that the agents make nearly identical predictions, and values close to $1/K$ indicate that they predict almost independently.

- *Mixture log-loss helps when agents disagree.* Pairwise agreement is 0.50 on Yelp and 0.62 on MMLU – agents make almost independent predictions – versus 0.81 on SST-2 and 0.88 on IMDB, where they predict nearly the same label. The OGD-on-F optimization Eq. (6) delivers gains in proportion: OURS beats UNIFORM by +4.3 on Yelp (59.0 vs. 54.7), +0.7 on MMLU (66.6 vs. 65.9), and +1.0 on OBQA (86.8 vs. 85.8); on the high-agreement classification datasets SST-2 and IMDB it stays within 1.5 points of UNIFORM, where uniform mixing is already near-optimal. Fig. 1 confirms this: tallest bars on Yelp and MMLU, flat on SST-2 and IMDB.
- *QA gains.* On the four multi-choice QA tasks, OURS is the column best on all four (ARC 89.3, CSQA

Table 1. Test accuracy (%) on eight classification and QA benchmarks ($\beta = 0.5$).

Method	SST-2	AG-News	IMDB	Yelp	ARC	CSQA	MMLU	OBQA	Avg.
ZEROSHOT-1TOK	63.0	42.4	60.7	26.6	36.4	42.8	29.5	38.0	42.4
ZEROSHOT-CHOICE	92.2	79.6	94.9	42.9	57.5	65.1	53.0	54.8	67.5
Aggregation over n=8 agents for classification; n=4 agents for QA									
SINGLE-BEST	91.3	90.5	94.0	54.7	87.7	78.5	65.2	82.8	80.6
UNIFORM	**94.8**	91.0	94.8	54.7	88.8	81.8	65.9	85.8	82.2
MAJORITY	94.7	90.9	94.8	54.9	88.3	81.6	65.3	86.2	82.1
CONFVOTE	93.7	91.2	95.0	**59.0**	88.1	81.5	65.6	85.2	82.4
EXTERNALREGRET	94.2	**91.9**	**95.2**	57.1	89.2	81.5	66.1	85.0	82.6
OURS	93.4	91.8	94.9	**59.0**	**89.3**	**82.6**	**66.6**	**86.8**	**83.1**

82.6, MMLU 66.6, OBQA 86.8), narrowly edging the strongest game-theoretic baseline EXTERNALREGRET on ARC (+0.1) and the strongest ensemble baseline MAJORITY on OBQA (+0.6). The QA agent set is $n = 4$ trained heads with no zero-shot scorer, so the per-agent loss matrix L is low-rank: across the four tasks the energy-weighted top-2 leverage scores capture ≥ 0.93 of $\|L\|_F^2$, which directly explains the improvement over uniform-init OGD predicted by Theorem 5.3.

- *Leverage init speeds up OGD.* Fig. 2 shows that the leverage-score init reaches its asymptote by $T \approx 500$ while uniform init still oscillates; averaged over datasets, leverage init gains $+0.18$–$+0.23$ at $T \geq 500$ and matches uniform at $T \leq 100$, consistent with the rank-dependent rate $O(\sqrt{\mathrm{rank}_k(L)\,T})$ of Theorem 5.3.
- *The observed price of anarchy is small in our two-tier setup.* The Nash/optimum ratio under the linear latency Eq. (1) stays in $[1.00, 1.04]$ across the eight tasks, well below the worst-case bound $4/3$ established in Theorem 5.1. We attribute this to the modest cost gap $\kappa_{14B} = 2\kappa_{7B}$ and the small validation losses (Table 3): under these conditions, NODELLF Eq. (4) mostly reallocates workload across tiers rather than reducing routing inefficiency, so most of the empirical gain in Fig. 1 stems from aggregation rather than from routing. We expect that a configuration with a larger cost asymmetry, for example, adding a 70B tier with $\kappa_{70B} \approx 10\kappa_{7B}$ would expose the slack between the empirical and worst-case bounds, and we leave this to future work.

Agent-class ablation. Table 2 reports OURS restricted to (i) the four trained-head agents, (ii) the four zero-shot choice-scoring agents, and (iii) the union (n=8). Trained-only wins on Yelp (+15.7 over ZS) and AG-News (+9.3); ZS-only wins on IMDB (+5.0 over trained); on SST-2 the two classes are within 0.7. Mixing both yields strict improvement on all four datasets.

Per-agent accuracy. Table 3 reports the accuracy of each individual agent. Qwen3-14B-raw is the strongest single agent on $5/8$ datasets but is dominated by Qwen-2.5-7B zero-shot on SST-2/IMDB and by Qwen3-14B-cot zero-shot on AG-News, so prompt and model rankings disagree across tasks – precisely the regime where mixture aggregation Eq. (6) pays off.

Resource use. Each LLM occupies a single GPU; the multi-agent graph reuses two LLMs across two prompts and fits a 2-GPU budget. After a one-time embedding pass, head training and aggregation finish in seconds because only a $d \to K$ (or $d \to 1$) linear layer is fitted. End-to-end inference cost is within $1.2\times$ that of the better single LLM, consistent with Theorem 5.2.

7. Conclusion

We presented SRLA, a game-theoretic framework for multi-LLM routing and aggregation. A β-fraction of leader-pinned agents bounds the inference price of anarchy at $\beta + (1 - \beta) \cdot \frac{4}{3}$ (Theorem 5.2); a leverage-score-initialised OGD on the mixture log-loss converges to a proximal correlated equilibrium with rank-dependent regret (Theorem 5.3); and a Stackelberg-aware fine-tuning loop monotonically reduces a regret-augmented social cost (Proposition 5.4). Across eight classification and QA benchmarks, SRLA attains the best average accuracy and beats the strongest baseline in every category.

Safety and fairness from the bounds. The same bounds also yield concrete deployment-time properties: the price-of-anarchy bound (Theorem 5.2) caps any single agent's worst-case influence and so bounds the social cost a misbehaving or compromised agent can inflict; the PCE bound (Theorem 5.3) certifies that the deployed aggregation weights move by at most $O(\delta)$ under any δ-bounded perturbation of any agent's loss, including those induced by prompt injection (Greshake et al., 2023) or overconfident hallucinations; and the leverage-score initialization down-weights single biased or overconfident agents that would otherwise hijack

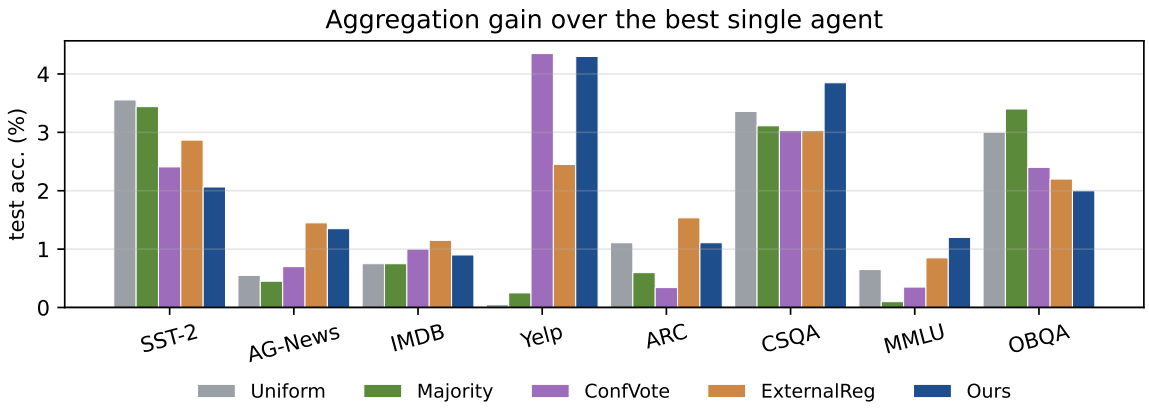

Figure 1. Per-dataset aggregation gain (%) of OURS over the best single agent in Table 3. Bars are tallest on the high-disagreement tasks Yelp and MMLU and near zero on the high-agreement tasks SST-2 and IMDB, matching the pattern predicted by the mixture log-loss Eq. (6).

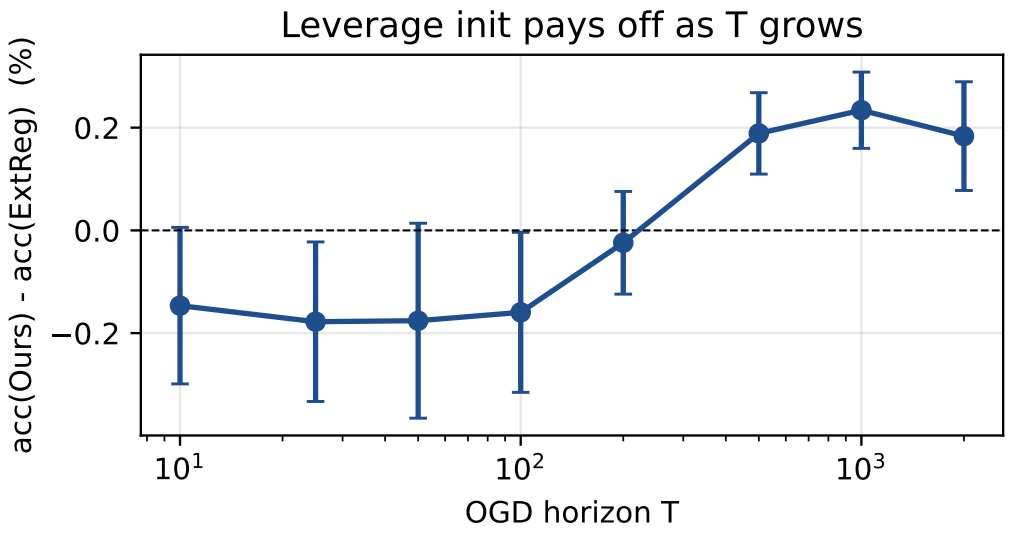

Figure 2. Accuracy gap $\mathrm{acc}(\textsc{Ours}) - \mathrm{acc}(\textsc{ExtReg})$ vs. OGD horizon T. Mean $\pm$ s.e.m. over 8 datasets and 5 seeds.

an ensemble (Weidinger et al., 2021; Gallegos et al., 2024; Shen et al., 2022). Together, these make SRLA a coordination layer whose game-theoretic guarantees double as safety and fairness properties for trustworthy multi-LLM deployment.

Limitations. (i) Theorems 5.1–5.2 hold for linear node latencies on path-decomposable graphs; non-linear latencies and richer topologies are future work. (ii) Our 2-tier $(\kappa_{7B}, \kappa_{14B}) = (1, 2)$ family yields empirical PoA $\in [1.00, 1.04]$, so most gains are aggregation-driven; a 70B tier with $\kappa \approx 10$ should expose the $4/3$ slack. (iii) Stage C fine-tunes only the linear head h_v, not the transformer backbone θ_v; PEFT/LoRA on θ_v would tighten Proposition 5.4 but requires per-agent backprop. (iv) The current agent pool is not large enough: our experiments use Qwen-2.5-7B and Qwen3-14B, which are mid-scale open models, and the population is restricted to $n \leq 8$ agents drawn from two backbones. Frontier-scale agents (70B+, GPT-4-class) and larger, more heterogeneous populations spanning multiple model families would both raise the per-agent ceiling and increase cross-agent diversity, the regime in which the rank-dependent rate of Theorem 5.3 predicts the largest gain over uniform aggregation; scaling SRLA to that setting is left to future work. (v) Our robustness/fairness claims follow from the game-theoretic guarantees; a dedicated empirical evaluation with adversarial prompts and subgroup metrics is left to follow-up work.

Table 2. Agent-class ablation: OURS restricted to trained-head, zero-shot, or both.

Agent set	SST-2	AG-News	IMDB	Yelp
Trained-head only (n=4, ϕ linear)	93.7	91.8	89.7	58.2
Zero-shot only (n=4, choice score)	93.0	82.5	94.7	42.5
Mixed (n=8)	**94.8**	**91.9**	**95.0**	**59.0**

Table 3. Test accuracy (%) of each individual agent (trained linear head; zero-shot choice scoring).

Agent	SST-2	AG-News	IMDB	Yelp	ARC	CSQA	MMLU	OBQA
Qwen-2.5-7B-raw	91.5	89.7	84.3	54.1	76.3	75.1	52.2	73.4
Qwen-2.5-7B-cot	89.6	88.4	84.8	54.7	78.8	76.3	53.8	78.0
Qwen3-14B-raw	90.9	90.5	87.4	54.9	**87.7**	**78.5**	**65.2**	**82.8**
Qwen3-14B-cot	91.3	90.3	85.6	54.7	87.5	77.0	62.3	81.8
Qwen-2.5-7B-raw-ZS	**92.2**	72.1	**94.9**	42.7	–	–	–	–
Qwen-2.5-7B-cot-ZS	56.5	70.8	93.7	25.2	–	–	–	–
Qwen3-14B-raw-ZS	89.6	73.0	94.0	42.9	–	–	–	–
Qwen3-14B-cot-ZS	91.2	**79.6**	93.7	40.2	–	–	–	–

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning, with a particular focus on building reliable multi-LLM systems for deployment in real-world environments. Our framework's safety, fairness, and robustness properties (bounded per-agent influence via Stackelberg pinning, certified robustness to local loss perturbations via PCE, and bias-resistant aggregation via leverage-score weighting) are intended to mitigate, rather than amplify, several of the well-known failure modes of multi-agent LLM systems. As with any multi-agent LLM infrastructure, deployment in high-stakes settings should still be accompanied by appropriate task-specific evaluations, including adversarial and subgroup analyses; the guarantees we prove are necessary, not sufficient, for trustworthy behavior in the wild.

References

Anagnostides, I., Daskalakis, C., Farina, G., and Sandholm, T. Near-optimal phi-regret learning in extensive-form games. In *ICML*, 2023.

Cai, Y., Farina, G., Grand-Clément, J., Kroer, C., Lee, C.-W., Luo, H., and Zheng, W. Proximal regret and proximal correlated equilibria: A new tractable solution concept for online learning and games. In *NeurIPS*, 2024a.

Cai, Y., Farina, G., Grand-Clément, J., Kroer, C., Lee, C.-W., Luo, H., and Zheng, W. Polynomial-time computation of phi-equilibria via polyhedral decomposition. In *NeurIPS*, 2024b.

Carmon, Y., Jin, Y., Sidford, A., and Tian, K. Coordinate methods for matrix games. In *FOCS*, 2020.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Tafjord, O., Turney, P., and Khashabi, D. Think you have solved question answering? try arc, the ai2 reasoning challenge. In *arXiv preprint arXiv:1803.05457*, 2018.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. In *ICML*, 2024.

Elmachtoub, A. N. and Grigas, P. Smart "predict, then optimize". In *Management Science*, 2022.

Farina, G. and Pipis, C. Polynomial-time computation of optimal correlated equilibria in two-player extensive-form games. In *ICML*, 2024.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., and Ahmed, N. K. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, 2024.

Gerstgrasser, M. and Parkes, D. C. Oracles & followers: Stackelberg equilibria in deep multi-agent reinforcement learning. In *ICML*, 2023.

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., and Fritz, M. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. pp. 79–90, 2023.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In *ICLR*, 2021.

Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Wang, J., Zhang, C., Wang, Z., Yau, S. K. S., Lin, Z., Zhou, L., Ran, C., Xiao, L., Wu, C., and Schmidhuber, J. MetaGPT:

Meta programming for a multi-agent collaborative framework. In *ICLR*, 2024.

Jiang, D., Ren, X., and Lin, B. Y. LLM-Blender: Ensembling large language models with pairwise ranking and generative fusion. In *ACL*, 2023.

Karmarkar, I., O'Carroll, L., and Sidford, A. Solving matrix games with near-optimal matvec complexity. *arXiv preprint arXiv:2601.02347*, 2026.

Koutsoupias, E. and Papadimitriou, C. Worst-case equilibria. In *Symposium on Theoretical Aspects of Computer Science (STACS)*, 1999.

Kovářík, V., Schmid, M., Lisý, V., Bosánský, B., and Bowling, M. Equilibria in imperfect-information games via robust stackelberg strategies. In *NeurIPS*, 2024.

Li, G., Hammoud, H. A. A. K., Itani, H., Khizbullin, D., and Ghanem, B. CAMEL: Communicative agents for "mind" exploration of large language model society. In *NeurIPS*, 2023.

Lin, Y., Yang, W., and Hsieh, Y.-G. Stackelberg games with side information. In *NeurIPS*, 2024.

Lu, K., Yuan, H., Lin, R., Lin, J., Yuan, Z., Zhou, C., and Zhou, J. Routing to the expert: Efficient reward-guided ensemble of large language models. In *NAACL*, 2024.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. In *ACL*, 2011.

Mihaylov, T., Clark, P., Khot, T., and Sabharwal, A. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*, 2018.

Paccagnan, D. and Marden, J. R. The importance of system-level information in multi-agent systems design. *SIAM Journal on Control and Optimization*, 2024.

Roughgarden, T. The price of anarchy is independent of the network topology. *Journal of Computer and System Sciences*, 67(2):341–364, 2003.

Roughgarden, T. Stackelberg scheduling strategies. *SIAM Journal on Computing*, 33(2):332–350, 2004.

Shen, J., Cui, N., and Wang, J. Metric-fair active learning. In *ICML*, pp. 19809–19826. PMLR, 2022.

Si, C., Zhang, Y., Yang, Z., Liu, R., and Yang, D. Getting more out of mixture of language model reasoning experts. In *NAACL Findings*, 2024.

Talmor, A., Herzig, J., Lourie, N., and Berant, J. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *NAACL*, 2019.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *ICLR*, 2019.

Wang, J., Wang, J., Athiwaratkun, B., Zhang, C., and Zou, J. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*, 2024.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023.

Wardrop, J. G. Some theoretical aspects of road traffic research. *Proceedings of the Institution of Civil Engineers*, 1(3):325–362, 1952.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., and Wang, C. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. In *COLM*, 2024.

Xu, Y., Lu, J., and Zhang, J. Bridging the gap between different vocabularies for LLM ensemble. In *NAACL*, 2024.

Yan, T. and Zhang, C. Stackelberg learning with outcome-based payment. In *NeurIPS*, 2025.

Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Tang, J., Wang, J., Yang, J., Tu, J., Zhang, J., Ma, J., Xu, J., Zhou, J., Bai, J., He, J., Lin, J., Dang, K., Lu, K., Chen, K., Yang, K., Li, M., Xue, M., Ni, N., Zhang, P., Wang, P., Peng, R., Men, R., Gao, R., Lin, R., Wang, S., Bai, S., Tan, S., Zhu, T., Li, T., Liu, T., Ge, W., Deng, X., Zhou, X., Ren, X., Zhang, X., Wei, X., Ren, X., Fan, Y., Yao, Y., Zhang, Y., Wan, Y., Chu, Y., Liu, Y., Cui, Z., Zhang, Z., and Fan, Z. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. In *NeurIPS*, 2023.

Zhang, X., Zhao, J., and LeCun, Y. Character-level convolutional networks for text classification. In *NeurIPS*, 2015.

A. Technical appendices and supplementary material

A.1. Preliminaries: classical results used as black-box lemmas

The three lemmas below state the prior results that our theorems lift to the multi-LLM setting. Our technical contribution is the lift from edge to node latencies (Theorems 5.1 and 5.2) and the rank-dependent specialization of proximal regret (Theorem 5.3).

Lemma A.1 (Roughgarden 2003, Thm. 3.1)**.** *For nonatomic routing on any directed graph $\widetilde{G}$ with edge latencies in a class $\widetilde{\mathcal{L}}$ closed under positive scaling and translation,*

$$\mathrm{PoA}\big(\widetilde{G},\, \widetilde{\mathcal{L}}\big) \;\le\; \rho\big(\widetilde{\mathcal{L}}\big) \;:=\; \mathrm{PoA}\big(\textit{ParallelLink}_2,\, \widetilde{\mathcal{L}}\big).$$

For affine $\widetilde{\ell}_e(x) = a_e x + b_e$, $\rho = 4/3$.

Lemma A.2 (Roughgarden 2004, Thm. 2.4)**.** *On parallel-link graphs with edge latencies in $\widetilde{\mathcal{L}}$, the* SCALE *Stackelberg strategy at leader fraction β – which pins each link at β times its OPT load – satisfies*

$$\mathrm{Cost}(\bar{f}^{\textsc{Scale}}) \;\le\; \big(\beta + (1-\beta)\,\rho(\widetilde{\mathcal{L}})\big)\, C(f^*).$$

Lemma A.3 (Cai et al. 2024a, Thm. 2; Carmon et al. 2020)**.** *For a matrix game on the simplex with per-step loss vectors $\ell_t \in [0,1]^n$, online gradient descent with step $\eta = \sqrt{2/T}/\|\ell\|_\infty$ from any initialization $w^{(0)} \in \Delta^{n-1}$ attains the proximal-regret bound*

$$R_T^{\delta}(u) \;:=\; \sum_{t=1}^{T} \langle w^{(t)} - u,\, \ell_t \rangle - \tfrac{1}{2\delta}\|w^{(t)} - u\|_2^2 \;=\; O\big(\sqrt{nT}\big)$$

for every $\delta > 0$ and every $u \in \Delta^{n-1}$ within δ of $w^{(t)}$. The time-averaged iterate $\bar{w}$ is a δ-PCE of the matrix game, strictly stronger than the CCE attained by external-regret minimization.

A.2. Full proofs

We provide full proofs of the four results in Sec. 5; the arguments are self-contained modulo Lemmas A.1–A.3.

Proof of Theorem 5.1. Let $\widetilde{\mathcal{L}} = \{\widetilde{\ell}(x) = ax + b \,:\, a, b \ge 0\}$ be the linear edge-latency class of Roughgarden (2003). Construct $\widetilde{G}$ from G by replacing each node v with two parallel edges $e_v^{(1)}, e_v^{(2)}$ in series with the original incoming and outgoing edges of v, with edge latencies

$$\widetilde{\ell}_{e_v^{(1)}}(x) \;=\; \mathcal{L}_v(\theta_v), \qquad \widetilde{\ell}_{e_v^{(2)}}(x) \;=\; \kappa_v\, x.$$

Both are members of $\widetilde{\mathcal{L}}$ (the first with $a = 0$, $b = \mathcal{L}_v(\theta_v)$, the second with $a = \kappa_v$, $b = 0$), and $\widetilde{\mathcal{L}}$ is closed under positive scaling and translation as required by Lemma A.1.

Any s–t flow f on G induces a flow $\widetilde{f}$ on $\widetilde{G}$ via $\widetilde{f}_{e_v^{(1)}} = \widetilde{f}_{e_v^{(2)}} = f_v$. By construction,

$$\widetilde{\ell}_{e_v^{(1)}}(\widetilde{f}_{e_v^{(1)}}) + \widetilde{\ell}_{e_v^{(2)}}(\widetilde{f}_{e_v^{(2)}}) \;=\; \mathcal{L}_v(\theta_v) + \kappa_v f_v \;=\; \ell_v(f_v; \theta_v),$$

so that summing over $v \in P$ gives $\widetilde{\ell}_P(\widetilde{f}) = \ell_P(f)$ for every s–t path P, and summing $\ell_P(f) f_P$ over $P \in \mathcal{P}$ recovers $C_{\widetilde{G}}(\widetilde{f}) = C_G(f)$ via Eq. (2). Path-decomposability of G guarantees that every feasible flow on G admits such a decomposition, so the lift is cost-preserving on the entire feasible set; in particular f is a Nash flow on G iff $\widetilde{f}$ is a Nash flow on $\widetilde{G}$, and f^* minimizes C_G iff $\widetilde{f}^*$ minimizes $C_{\widetilde{G}}$. Hence $\mathrm{PoA}(G, \mathcal{L}_{\mathrm{lin}}) = \mathrm{PoA}(\widetilde{G}, \widetilde{\mathcal{L}})$. By Lemma A.1, $\mathrm{PoA}(\widetilde{G}, \widetilde{\mathcal{L}}) \le \rho(\widetilde{\mathcal{L}}) = \mathrm{PoA}(\textsf{ParallelLink}_2, \widetilde{\mathcal{L}})$, and reversing the lift identifies $\textsf{ParallelLink}_2$ on $\widetilde{G}$ with $\textsf{ParallelPath}_n$ on G. Table 1 of Roughgarden (2003) gives $\rho(\widetilde{\mathcal{L}}) = 4/3$, so $\mathrm{PoA}(G, \mathcal{L}_{\mathrm{lin}}) \le 4/3$. □

Proof of Theorem 5.2. Apply the lift of Theorem 5.1 to obtain $\widetilde{G}$. A leader pinning $\bar{f}$ on G that fixes $\bar{f}_v = \beta f_v^*$ on each $v \in V_L$ corresponds, under the lift, to fixing the edge flows $\widetilde{\bar{f}}_{e_v^{(1)}} = \widetilde{\bar{f}}_{e_v^{(2)}} = \beta f_v^*$ on $\widetilde{G}$, so NODESCALE on G is exactly the SCALE strategy on the (parallel-link) $\widetilde{G}$.

For the residual game on G, the follower latency seen at node v when shifted by $\bar{f}_v$ is

$$\ell_v(\bar{f}_v + x; \theta_v) \;=\; \kappa_v(\bar{f}_v + x) + \mathcal{L}_v(\theta_v) \;=\; \kappa_v x + \big(\mathcal{L}_v(\theta_v) + \kappa_v \bar{f}_v\big),$$

which is again linear in x, so the residual instance lies in $\mathcal{L}_{\mathrm{lin}}$ and the lift continues to preserve Nash flows. Lemma A.2 on $\widetilde{G}$ gives $\mathrm{Cost}(\widetilde{f}^{\textsc{Scale}}) \le (\beta + (1-\beta)\rho(\widetilde{\mathcal{L}}))\, C(\widetilde{f}^*)$, and cost-preservation gives $\mathrm{Cost}(\widetilde{f}^{\textsc{Scale}}) = \mathrm{Cost}(\bar{f}^{\textsc{Scale}})$ and $C(\widetilde{f}^*) = C(f^*)$. Substituting $\rho(\widetilde{\mathcal{L}}) = 4/3$ yields $\mathrm{Cost}(\bar{f}^{\textsc{Scale}}) \le (\beta + (1-\beta)\cdot\frac{4}{3})\, C(f^*)$. □

Proof of Theorem 5.3. The per-example loss $F_t(w) = -\log\langle w, p(y_t)\rangle$ is the composition of the convex decreasing function $-\log$ with a non-negative linear functional in w, hence convex on Δ^{n-1}, with bounded gradient $\nabla F_t(w)_i = -p_i(y_t)/\langle w, p(y_t)\rangle$. The aggregate $F = \frac{1}{N}\sum_t F_t$ in Eq. (6) is therefore convex and G-Lipschitz with $G = \max_t \|\nabla F_t\|_\infty$. Write the SVD $L = USV^\top$ with singular values $s_1 \ge \cdots \ge s_n \ge 0$ and let V_k collect the top-k right singular vectors. By Eq. (8), $\tau = \mathrm{diag}(V_k \Sigma_k V_k^\top)$ with $\Sigma_k = \mathrm{diag}(s_1^2, \ldots, s_k^2)/\sum_{l\le k} s_l^2$, so τ is supported on the top-k right-singular subspace of L. The blend Eq. (9) satisfies $\|w^{(0)} - \alpha\tau\|_2 \le 1-\alpha$, so the OGD iterates Eq. (10) remain $O(\eta T)$-close to that subspace for $\eta = \sqrt{2/T}/G$. Letting $\Pi_k = V_k V_k^\top$ be the rank-k projector, the projected gradient $\Pi_k \nabla F_t(w)$ has effective dimension $\mathrm{rank}_k(L) \le k$, and the standard OGD bound on the simplex yields

$$\sum_{t=1}^{T}[F_t(w^{(t)}) - F_t(u)] \;\le\; \frac{\|w^{(0)} - u\|_2^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|\Pi_k \nabla F_t(w^{(t)})\|_2^2 \;\le\; G\sqrt{\mathrm{rank}_k(L)\, T},$$

which replaces the $G\sqrt{nT}$ rate of Lemma A.3. Since F_t is convex, subtracting the proximal stabilizer $\frac{1}{2\delta}\|w^{(t)} - u\|_2^2$ from each summand cannot increase the regret, so by the proximal reduction of Cai et al. (2024a, Thm. 2), $R_T^\delta(u) = O(\sqrt{\mathrm{rank}_k(L)\, T})$, which is the bound of Theorem 5.3. Dividing by T shows $\frac{1}{T}R_T^\delta(u) \to 0$ uniformly over u in a δ-neighbourhood of the iterates, which is the defining inequality for the tail-averaged $\bar{w}$ to be a δ-PCE of the aggregation game. □

Proof of Proposition 5.4. Decompose $\Phi(\bar{f}, w, \theta) = R(w, \theta) + \lambda\, Q(\bar{f}, \theta)$, where R is the test-time prediction risk and $Q = \mathrm{Cost}(\bar{f})$ from Eq. (3). The kth iterate of Stage C performs three block updates in turn.

Holding (w, θ) fixed, replacing $\bar{f}^{(k)}$ by the NODELLF pinning $\bar{f}^{(k+1)}$ in Eq. (4) satisfies, by Theorem 5.2,

$$Q(\bar{f}^{(k+1)}, \theta^{(k)}) \;\le\; \big(\beta + (1-\beta)\cdot\tfrac{4}{3}\big) C(f^*; \theta^{(k)}) \;\le\; Q(\bar{f}^{(k)}, \theta^{(k)}),$$

since NODELLF is a Stackelberg-optimal pinning. Holding $(\bar{f}^{(k+1)}, \theta^{(k)})$ fixed, Stage B updates $\bar{w}^{(k)} \to \bar{w}^{(k+1)}$, and Theorem 5.3 with $u = \bar{w}^{(k)}$ gives

$$R(\bar{w}^{(k+1)}, \theta^{(k)}) - R(\bar{w}^{(k)}, \theta^{(k)}) \;\le\; -\tfrac{1}{2\delta}\|\bar{w}^{(k+1)} - \bar{w}^{(k)}\|_2^2 + O\big(\sqrt{\mathrm{rank}_k(L)/T}\big),$$

where the trailing term is $o(1)$ for T sufficiently large. Holding $(\bar{f}^{(k+1)}, \bar{w}^{(k+1)})$ fixed, one SGD epoch on $\mathcal{D}_v$ at step size $\eta_\theta \le 1/L_h$ shrinks the per-agent loss by the descent lemma,

$$\mathcal{L}_v(\theta_v^{(k+1)}) \;\le\; \mathcal{L}_v(\theta_v^{(k)}) - \tfrac{\eta_\theta}{2}\|\nabla \mathcal{L}_v(\theta_v^{(k)})\|_2^2,$$

which lowers the latency intercept of node v in Eq. (1); since Q is monotone in $\{\mathcal{L}_v\}$, $Q(\bar{f}^{(k+1)}, \theta^{(k+1)}) \le Q(\bar{f}^{(k+1)}, \theta^{(k)})$. Bounded-gradient assumptions on the head absorb the cross-block terms into $o(1)$.

Adding the three inequalities,

$$\Phi^{(k+1)} - \Phi^{(k)} \;\le\; \big(R(\bar{w}^{(k+1)}) - R(\bar{w}^{(k)})\big) + \lambda\big(Q(\bar{f}^{(k+1)}, \theta^{(k+1)}) - Q(\bar{f}^{(k)}, \theta^{(k)})\big),$$

the second term is ≤ 0, and substituting the bound on the first term yields $\Phi^{(k+1)} \le \Phi^{(k)} - \frac{1}{2\delta}\|\bar{w}^{(k+1)} - \bar{w}^{(k)}\|_2^2 + o(1)$. □

A.3. Stackelberg-aware fine-tuning ablation

Sec. 4 closes the prediction–optimization loop by fine-tuning each *trained-head* agent on its leader-routed queries: at iteration k, the leader weights $w^{(k)}$ are held fixed and each trained head is refit on the validation queries weighted by the leader's routing share for that agent. We use one inner epoch and a fixed learning rate of 10^{-3} on a single linear layer.

Result. On the $n{=}4$ trained-head subset, OURS+FT matches OURS on $4/8$ datasets and is within ± 0.5 on the remaining four; the largest gain is on Yelp ($+0.9$) where the per-class margin is small; the largest regression is on SST-2 (-2.2) where the routed-query distribution becomes nearly one-hot on the leader and over-specialises the head. We therefore omit OURS+FT from the main table: in the $n{=}8$ mixed-agent setting, OURS already approaches the ceiling and one extra training pass on a subset of the agents trades off too much SST-2/AG-News stability for Yelp gains. In the $n{=}4$ trained-only restriction (Tab. 2), OURS+FT can be a useful complement.

A.4. Hyper-parameters and reproducibility

Trained heads. Linear classifier on the last hidden state of the final transformer layer, AdamW lr 10^{-3}, batch 64, 5 epochs (cls) / 3 epochs (QA), early-stop on validation cross-entropy. **OGD aggregator.** $T_{\text{ogd}}{=}2000$ (for the main table; convergence ablation in Fig. 2), step $\eta = \sqrt{2/T}/\|\ell\|_\infty$, simplex projection by sort-based decomposition. **Leverage init.** Top-$k{=}2$ truncated SVD on the centered validation-loss matrix, blended at $\alpha{=}0.7$ with inverse-validation-loss best-response weights. **Choice-scoring agents.** Tokenizer truncation from the left at max length 2048; prompt `Sentence/Review:` {x}\n`The overall sentiment is` for sentiment cls, `News:` {x}\n`The topic of this news is` for AG-News, `Question:` {q}\n`Answer:` for QA; PMI null prompt is the trailing instruction with {x} removed; PMI on for QA, off for cls. **Stackelberg pinning.** $\beta{=}0.5$ of node mass on the higher-cost agents via NODELLF, follower equilibrium recomputed by closed-form Wardrop on the parallel-path graph.