Specialists Hold, Generalists Discount: Asymmetric Equilibrium in LLM Routing Auctions

Xinyu Hou
University of Science and Technology of China
Hefei, Anhui
houxinyu123@mail.ustc.edu.cn

Yang Lu
University of Houston
Houston, TX 77204
ylu17@central.uh.edu

Rabimba Karanjai
PayPal AI, PayPal
San Jose, CA 95112
rabimba@gmail.com

Pei-Chi Pan
University of Houston
Houston, TX 77204
ppan@cougarnet.uh.edu

Sen Lin
University of Houston
Houston, TX 77204
slin50@Central.UH.EDU

Lei Xu
Kent State University
Kent, OH 44242
lxu12@kent.edu

Weidong Shi
University of Houston
Houston, TX 77204
wshi3@uh.edu

Abstract

Routing systems for large language models match queries to suppliers on cost–quality grounds, usually taking supplier prices as given. We study the complementary supply-side pricing problem as a sealed-bid first-price reverse auction over collaboration modes. We establish mixed-strategy Bayesian Nash equilibrium (BNE) existence for the finite game and a conditional monotone pure-strategy existence result for continuous games with aggregate own-price monotonicity and continuous payoff aggregates. Cache-based solvers separate routing estimation from equilibrium computation. Experiments use calibrated profiles of nine open-weight LLMs spanning 3B–32B parameters and three task categories. In the tested markets, task leaders tend to retain high own-task bids while weaker substitutes discount; a dominant generalist can weaken this differentiation. Direct audits show that the learned routing caches are not uniformly monotone. We therefore compute mixed profiles by fictitious play and certify them by exhaustive finite-grid unilateral-deviation residuals, rather than invoking the conditional pure-strategy theorem; iterated best response cycles on the audited learned games and serves only as a diagnostic. Analytical-router and finer-grid controls preserve the qualitative price ordering, while cycle periods and supports change. We report monetary consumer surplus, supplier profits, and total welfare separately from the customer's price-weighted utility. The findings characterize the tested routing markets, not a universal pricing law.

1 Introduction

LLMs in production face a substantial price spread: 3–4× across providers serving the same open-weight model and roughly 100× across model classes (as of April 2026; see Appendix E.11 for source data). Routing systems (MasRouter [Yue et al., 2025], RouteLLM [Ong et al., 2024], FrugalGPT [Chen et al., 2023a], and others) exploit this widespread by matching queries to suppliers on cost-quality grounds. Almost all prior routing work optimizes from the *demand* side: given a population

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

of suppliers with fixed prices, how should the router choose? The dual question: given a routing mechanism, how should LLM suppliers *price into* it? has received almost no formal treatment, and supplier incentives feeding into it will shape the eventual welfare properties of the routing infrastructure being built now.

We model supplier-side pricing as a sealed-bid first-price reverse auction (SBFP-RA) layered over a fixed cost–quality routing policy. A selected collaboration mode may use one supplier or several suppliers in different roles. The solver is *cache-compatible*: learned, analytical, or cascade-based routing rules can supply their own allocation tables. This computational interface does not imply that all routers satisfy the hypotheses of our existence results or induce the same equilibrium. Sealed bids represent simultaneous private offers; first-price payment means that each active supplier receives its own quoted payment. The platform implements the customer's routing policy rather than choosing an additional strategic objective.

Our contributions can be summarized as follows.

(1) We formulate the supply-side pricing game induced by query-dependent cost–quality routing over heterogeneous LLM suppliers.

(2) We establish finite-game mixed BNE existence and conditional continuous-game monotone pure BNE existence. We separate cache construction from equilibrium computation, compute candidate mixed profiles by fictitious play, and certify them by exhaustive unilateral-deviation checks.

(3) We evaluate calibrated markets using nine open-weight LLMs. Task-dependent price ordering survives analytical-router and finer-grid controls in the tested cohorts, although pure versus mixed behavior, supports, and cycle periods depend on the routing rule and grid.

(4) We report welfare in common monetary units, separating consumer surplus and supplier profits from price-weighted customer utility. Reserve factors and alternative tie-breaking are candidate design interventions; their effects require separate evaluation. The code for this work is publicly available.[1]

2 Related work

LLM routing: the demand side. Recent work trains query-to-supplier routers under cost-quality tradeoffs. MasRouter [Yue et al., 2025] learns a cascaded controller over collaboration modes, roles, and suppliers; RouteLLM [Ong et al., 2024] trains routers on human preference data to direct queries to the cheapest supplier whose quality is acceptable; FrugalGPT [Chen et al., 2023a] cascades suppliers from cheapest to most expensive; HybridLLM [Ding et al., 2024] and AutoMix [Aggarwal et al., 2024] route on predicted query difficulty and self-verification, respectively. All take supplier prices as exogenous. Guo et al. [2025] studies single-leader Stackelberg pricing; we study the multi-supplier simultaneous reverse auction.

Auction theory. Classical first-price auctions and procurement provide the starting point [Krishna, 2010, McAfee and McMillan, 1987, Milgrom and Weber, 1982, Che, 1993, Asker and Cantillon, 2008]. Asymmetric competition has well-established antecedents [Maskin and Riley, 2000, Cantillon, 2008]. We use the single-crossing framework of Athey [2001] under explicit continuity conditions; general discontinuous-game results [Reny, 1999] require additional hypotheses and are not invoked automatically for deterministic routing. Unlike homogeneous-product Bertrand competition, query-dependent capability and collaboration modes allow suppliers to be substitutes on some tasks and complements on others. Unlike Stackelberg pricing, the suppliers here move simultaneously with private costs. The contribution is this calibrated demand structure and its supply-side implications, rather than a new general auction existence theorem. Where the computed equilibria are mixed, quoted prices are dispersed, as in classical models of sales and price dispersion [Varian, 1980]; here the dispersion arises from a quality-sensitive allocator rather than from consumer search.

Vertical product differentiation. The asymmetric pattern we find — specialists holding flat markups while generalists compete on price — has a natural reading in the industrial-organization literature on vertical product differentiation [Hotelling, 1929, Salop, 1979, Tirole, 1988, Shaked and Sutton, 1983, Mussa and Rosen, 1978]. The routing-auction setting can thus be read as a particular instantiation of a general phenomenon, with our contribution being the calibration to actual LLM capabilities and the resulting implications for market design.

[1] `https://github.com/paypal/research/NeurIPS%202026/Specialists-Hold-Generalists-Discount`

LLMs as economic agents. A separate line of research studies LLMs as strategic actors in auctions or pricing games [Chen et al., 2023b, Shah et al., 2025, Fish et al., 2024]. We ask a different question: *they study what LLMs do; we study what LLM markets do*. Our bidders are not prompted language models but supplier organizations setting price functions; the LLM is the priced commodity, not the strategic actor.

Research questions. We organize the paper around three questions. **(RQ1)** Does supplier-side pricing under a fixed cost-quality optimizing router admit a Bayesian Nash equilibrium, and what does the equilibrium look like? **(RQ2)** On real LLM cohorts, when does the predicted asymmetric pattern actually arise, and when does it degenerate? **(RQ3)** How does customer price-sensitivity shape the pattern? Specifically, does intensifying price competition strengthen or weaken specialist rents? Theory (§4) answers RQ1; calibrated empirics on three trios (§6.2) answer RQ2; the λ-sweep (§6.3) answers RQ3.

3 Model formulation

Multi-agent system design space. We formalize the design space of a multi-agent system (MAS) by the finite set of *collaboration modes* $\mathcal{S} = \{S_1, \ldots, S_M\}$. Each mode specifies how a group of LLM backbones is composed: a single backbone in *Solo*, sequential composition in *Chain*, or parallel aggregation in *Ensemble*; role-level specifics within a mode are abstracted as in Yue et al. [2025].

Bayesian game with three parties. Application-level examples of role-specialized LLM collaboration include UniTranslator for code translation [Karanjai et al., 2025a] and HPCAgentTester for HPC unit-test generation [Karanjai et al., 2025b]. These motivate the collaboration abstraction; we do not model their full adaptive workflows. We model the pricing interaction as a Bayesian game with n strategic LLM providers ($i \in N = \{1, \ldots, n\}$), a customer, and a non-strategic platform implementing the routing policy. Provider i has a privately known cost $c_i \in \mathcal{C} = [c_{\min}, c_{\max}]$ drawn i.i.d. across providers from a common-knowledge prior F. The customer submits queries from a distribution $\mathcal{D}$ over a finite set of types $\mathcal{T} = \{\tau_1, \ldots, \tau_K\}$ with public reward $r > 0$ and price sensitivity $\lambda > 0$. Provider i chooses a **per-type discount vector** $\mathbf{d}_i = (d_i^\tau)_{\tau \in \mathcal{T}} \in D^K$ where $D = [d_{\min}, d_{\max}]$ with $0 < d_{\min} \le d_{\max} < \infty$. A pure strategy is $\sigma_i : \mathcal{C} \to D^K$. The **base fee** in mode S is

$$F_i(S) := \rho_i \cdot \mathrm{tok}_i(S), \quad \mathrm{tok}_i(S) := m_i(S) \cdot T_i, \tag{1}$$

where $m_i(S) \ge 0$ is the token multiplier ($= 0$ if i is inactive in S), T_i the base token count, ρ_i the public per-token price; the actual payment to provider i for a type-τ query under mode S is $d_i^\tau \cdot F_i(S)$.

Timing and information. Stage (i): each provider i observes $c_i \sim F$ privately and commits to σ_i, producing $\mathbf{d}_i = \sigma_i(c_i)$. (ii) A query Q is drawn, of type $\tau = \mathrm{type}(Q)$. (iii) The customer observes $\mathbf{d} = (\mathbf{d}_1, \ldots, \mathbf{d}_n)$ and Q, then selects mode $S \in \mathcal{S}$ via the routing policy π^*. Providers commit *before* observing the query; the customer acts after observing both bids and the query.

Customer utility and routing policy. For a mode S on query Q, let $\alpha(Q, S) \triangleq \Pr(\text{correct} \mid Q, S)$. The customer's expected utility under routing policy π is

$$U_{\mathrm{cust}}(\pi; \mathbf{d}) = \mathbb{E}_{Q \sim \mathcal{D}}\, \mathbb{E}_{S \sim \pi(\cdot \mid Q, \mathbf{d})} \left[r \cdot \alpha(Q, S) - \lambda \sum_{i=1}^{n} d_i^{\mathrm{type}(Q)} F_i(S) \right], \tag{2}$$

and an ideal optimal policy satisfies $\pi^*(\mathbf{d}) \in \arg\max_\pi U_{\mathrm{cust}}(\pi; \mathbf{d})$. We also use π^* to denote a fixed implemented policy, including the learned approximation in the experiments; exact optimality is not assumed for its cache. For the analytical control, ties between maximizing modes are broken lexicographically. Such ties need not be negligible under induced bid distributions, and continuity of the optimal value does not imply continuity of the selected allocation.

The cache interface accommodates different fixed routing policies. The continuous-game pure-strategy result additionally requires aggregate own-price monotonicity and payoff continuity, stated in Section 4. Neither property follows solely from training a policy to optimize Equation (2). Learned finite games are assessed directly using their cached payoffs.

Provider utility and equilibrium. Provider i's interim expected utility against opponents' strategies σ_{-i} is

$$U_i(\mathbf{d}_i, \sigma_{-i}; c_i) = \mathbb{E}_{c_{-i} \sim F^{n-1}}\, \mathbb{E}_{\tau \sim \mathcal{D}}\, \mathbb{E}_{S \sim \pi^*(\cdot \mid \tau, \mathbf{d})} \left[d_i^\tau F_i(S) - c_i \mathrm{tok}_i(S) \right], \tag{3}$$

where $\mathbf{d} = (\mathbf{d}_i, \sigma_{-i}(c_{-i}))$. The routing rule observes bids and query information but not hidden costs. In addition, the type-τ policy depends only on the type-τ bids; this explicit type-locality assumption justifies the cache and utility decomposition. A BNE satisfies $U_i(\sigma_i^*(c_i), \sigma_{-i}^*; c_i) \geq U_i(\mathbf{d}_i', \sigma_{-i}^*; c_i)$ for every supplier, almost every cost, and every feasible deviation.

4 Theoretical results

This section separates three statements. The finite Bayesian game always admits a mixed BNE. Continuous-game existence requires additional continuity assumptions. Aggregate own-price monotonicity then yields increasing differences and, under those additional conditions, a monotone pure BNE. These statements do not guarantee convergence of iterated best response or prove the empirical specialist–generalist ordering.

(Existence of mixed-strategy BNE). We begin with the discrete formulation, in which both the type space and action space are finite grids; this is the formulation solved numerically in Section 5.

Theorem 1 (Existence in the Discretized Game)**.** *Consider the discretized game Γ_d in which the type space is replaced by a finite grid $\mathcal{C}_d = \{c^{(1)}, \ldots, c^{(L)}\} \subset \mathcal{C}$ with prior given by a probability vector $(F_\ell)_{\ell=1}^L$, and the action space is replaced by a finite grid D_d^K with $D_d = \{d^{(1)}, \ldots, d^{(J)}\} \subset D$. Then Γ_d admits at least one mixed-strategy BNE.*

Theorem 1 concerns the finite game, not convergence of a particular solver. For continuous actions we state explicit sufficient conditions for existence; the mixed result is given in Appendix A.3. The pure result below adds aggregate allocation monotonicity.

(Type decomposition and existence of monotone pure-strategy BNE). We first make the following assumptions:

Assumption 1 (Regularity)**.** *(C1) The type space $\mathcal{C} = [c_{\min}, c_{\max}]$ and the action space $D^K = [d_{\min}, d_{\max}]^K$ are nonempty compact intervals with $0 < d_{\min} \leq d_{\max} < \infty$; (C2) the prior F admits a continuous density f on $\mathcal{C}$ bounded above and below by positive constants; (C3) for each $\tau \in \mathcal{T}$ and $S \in \mathcal{S}$, the accuracy $\alpha(\cdot, S)$ and token consumption $\mathrm{tok}_i(S)$ are well-defined and bounded; (C4) the base fees $F_i(S)$ are bounded uniformly in i, S.*

Assumption 2 (Aggregate Own-Price Monotonicity)**.** *For each supplier i and query type τ, define the expected billable token allocation*

$$X_i^\tau(\mathbf{d}) := \sum_{S \in \mathcal{S}} \pi^*(S \mid \tau, \mathbf{d}) \, \mathrm{tok}_i(S). \tag{4}$$

Holding every other bid component fixed, X_i^τ is non-increasing in d_i^τ.

Assumption 2 concerns aggregate expected tokens, not each individual mode probability. The mode-wise condition is stronger: it implies aggregate monotonicity because token quantities are non-negative. The converse need not hold, since traffic can move between modes involving the same supplier. For positive public token prices, exact maximization of the stated linear-payment objective implies aggregate token monotonicity, but not mode-wise monotonicity (Remark 1 in Appendix A.4). Approximate learned optimization does not guarantee the aggregate condition. The learned caches do not satisfy the aggregate condition at every grid comparison: expected billable tokens increase after an own-price increase in 33.27% of 3,024 adjacent comparisons. We do not attribute these aggregate violations separately to router-training variation or policy structure. The numerical profiles are therefore certified directly by finite-grid deviations (Section 5; Appendix G), and Theorem 2 is not used for them.

A defining feature of our model is that a provider's utility decomposes additively across query types, which allows us to apply one-dimensional results from Athey [2001] to the multi-dimensional game.

Lemma 1 (Additive Separability)**.** *Provider i's interim expected utility decomposes as*

$$U_i(\mathbf{d}_i, \sigma_{-i}; c_i) = \sum_{\tau \in \mathcal{T}} \Pr_{\mathcal{D}_\tau}(\tau) \cdot U_i^\tau(d_i^\tau, \sigma_{-i}^\tau; c_i), \tag{5}$$

where $U_i^\tau(d, \sigma_{-i}^\tau; c) = \Phi_i^\tau(d, c)$ depends on the strategies only through their type-τ components.

Theorem 2 (Conditional Monotone Pure-Strategy BNE)**.** *Under Assumptions 1 and 2, suppose additionally that, for every i, τ, both $X_i^\tau(\mathbf{d})$ and*

$$R_i^\tau(\mathbf{d}) := d_i^\tau \sum_{S \in \mathcal{S}} \pi^*(S \mid \tau, \mathbf{d}) F_i(S)$$

are jointly continuous in the type-τ bid profile. Under the independent private-cost specification, the continuous-action game admits a pure BNE with each $\sigma_i^{\tau *}(c_i)$ non-decreasing in c_i. Best-response conditions hold for almost every cost type.

Additive separability reduces the game to scalar-action Bayesian subgames. Within each subgame, aggregate own-price monotonicity implies increasing differences between a supplier's bid and private cost type. Together with payoff continuity and the remaining standing assumptions, this allows us to apply the existence result of Athey [2001], yielding a monotone pure-strategy equilibrium for each query type. These per-type equilibria can then be combined into an equilibrium of the original game; see Appendix A.6. Deterministic mode switches can violate continuity, and finite-grid audits do not establish continuous-domain hypotheses. Theorem 2 is not used as a certificate for the learned profiles.

(Economic implications). We distinguish cost monotonicity, which follows under the theorem's conditions, from cross-task pricing comparisons, which require additional payoff ordering.

Corollary 1 (Effective Price Monotonicity)**.** *In the equilibrium of Theorem 2, the per-token quote $\rho_i \sigma_i^{\tau *}(c_i)$ is non-decreasing in c_i. For any fixed mode S, the payment $F_i(S)\sigma_i^{\tau *}(c_i)$ is also non-decreasing. This does not order expected payments when the selected mode changes.*

Corollary 1 recovers the familiar conditional prediction that higher private costs support weakly higher quotes. It does not state that all learned-grid profiles satisfy the theorem's assumptions.

Corollary 2 (Conditional Cross-Task Pricing)**.** *Fix a supplier, a cost, and opponents' strategies. Let $v_H(d)$ and $v_L(d)$ be its interim payoffs on two query types. If both are continuous on D and $v_H(d) - v_L(d)$ is non-decreasing in d, then the largest payoff-maximizing bid on type H is no smaller than the largest payoff-maximizing bid on type L.*

This sufficient condition formalizes one way cross-task bid ordering can arise. Capability advantage or lower demand sensitivity alone does not establish the required payoff ordering. The specialist–generalist pattern in Section 6 is an observation in the tested markets, not an unconditional consequence of the existence theorem.

5 Computational method

Training is confined to constructing a fixed routing cache on finite cost and bid grids; two solvers then operate on the cached game by arithmetic. Iterated best response (IBR) is used as a diagnostic, whereas fictitious play (FP) produces the reported mixed profiles, each certified by exhaustive enumeration of unilateral grid deviations. Theorem 1 guarantees that an exact mixed BNE of each finite game exists; the certificate bounds every supplier–cost type's gain from a unilateral grid deviation from the reported profile.

(Discretization). We discretize the cost interval $\mathcal{C} = [c_{\min}, c_{\max}]$ on a uniform grid $\mathcal{C}_d = \{c^{(1)}, \ldots, c^{(L)}\}$ with L points and the discount interval $D = [d_{\min}, d_{\max}]$ on a finite grid $D_d = \{d^{(1)}, \ldots, d^{(J)}\}$. A pure strategy in the discretized game is a map $\sigma_i : \mathcal{C}_d \to D_d^K$, identified with an array of L vectors of length K. Throughout, we take $L = 5$ cost points and $J = 4$ discount levels per type ($D_d = \{0.8, 0.9, 1.0, 1.1\}$); a grid control uses the seven levels $\{0.8, 0.85, \ldots, 1.1\}$. With $K = 3$ query types, each supplier chooses among $J^K = 4^3 = 64$ discount vectors at each cost point ($7^3 = 343$ on the seven-point grid).

(Customer policy cache). The most expensive primitive is the customer's optimal routing policy $\pi^*(\cdot \mid \tau, \mathbf{d})$. Any concrete instantiation of π^* , a learned policy trained to optimize (2), a cascade-based selection rule, or an analytical decision rule, can be queried at the discrete grid points and stored. We adopt this approach: rather than recomputing π^* inside the equilibrium loop, we pre-compute it once on the discretized type$\times$action grid (Appendix B describes the specific instantiation used in our experiments). The separation property in Section 3 implies π^* depends on $\mathbf{d}$ only through the

type-conditioned components $(d_1^\tau, \ldots, d_n^\tau)_{\tau \in \mathcal{T}}$, not on private costs, which makes a one-time precomputation tractable: a cache $\mathcal{Z}[\tau, d_1, \ldots, d_n] \in \Delta(\mathcal{S})$ holding the customer's mode distribution at each $(\tau, d_1, \ldots, d_n) \in \mathcal{T} \times D_d^n$, with $K \cdot J^n$ entries total. Once $\mathcal{Z}$ is constructed, every quantity needed by best-response iteration is computable in $O(|\mathcal{S}|)$ arithmetic time per query (Proposition 2 and proof in Appendix B). With two providers and three types $\mathcal{Z}$ has $3 \cdot 16 = 48$ entries; in the three-provider calibrated games, $3 \cdot 64 = 192$ on the four-point grid and $3 \cdot 7^3 = 1{,}029$ on the seven-point grid. Each entry requires one π^* evaluation ($\sim$40–60s on a single GPU when the policy is learned), totalling $\sim$30 minutes for the two-provider game and 2–3 hours for the three-provider game on the four-point grid; the cache is computed once and reused across cost grids, priors, and robustness sweeps.

(Iterated best response as a diagnostic). IBR (Algorithm 1 in Appendix B) carries out one round per provider, enumerates candidate $\mathbf{d} \in D_d^K$ at each cost point, and selects the per-cost best response against the cache; a sweep costs $O(nLJ^K L^{n-1} K|\mathcal{S}|)$ operations, well under one second for two providers. A fixed point verified against every grid deviation is a pure BNE of the finite game. IBR is a pure best-response dynamic: it can stop only at a pure profile, so it must cycle whenever it does not reach a pure equilibrium, in particular when the finite game's equilibria are mixed. A cycle alone therefore neither proves a mixed equilibrium nor rules out pure equilibria elsewhere. IBR cycles on the learned T1–T3 games (periods $(10, 2, 3)$ on the four-point grid) but reaches pure fixed points under the analytical control.

(Fictitious play: reported profiles). FP [Brown, 1951] treats each supplier–cost pair as an agent. In each round, each agent in turn best-responds to the rivals' current empirical mixtures, taking expectations over rivals' costs and over their past best responses; its reported strategy is the empirical frequency of its own past best responses (Algorithm 2 in Appendix B). The "Mix" profiles in Table 1 are these FP outputs. FP converges for special classes such as two-player zero-sum games [Robinson, 1951] but can cycle in general [Shapley, 1964]. We therefore use FP only to find a candidate and rely on the certificate below, never on FP convergence.

(Certification). For a cost-conditional mixed profile μ, the deviation gain $g_i(c; \mu)$ is the payoff of supplier i's best grid deviation at cost c minus its expected payoff under μ, computed against the product of the rivals' cost-conditional mixtures (Appendix G.2). We report $\epsilon(\mu) = \max_{i,c} g_i(c; \mu)$ in supplier-utility units; $\epsilon = 0$ is an exact finite-game BNE. By Lemma 1, the maximization splits into K scalar searches. The certificate is relative to the cached finite game: it bounds neither discretization nor routing-estimation error.

Given a cache, initialization, update schedule, and tie-breaking rule, the IBR computation is deterministic. This is cache compatibility, not invariance to the router: changing the routing rule changes the induced game. The analytical control preserves qualitative comparative-advantage pricing, but its pure fixed points differ from the learned games' cycles (Appendix B.6).

6 Empirical findings

We now study numerical pricing behavior in synthetic and calibrated finite games. The continuous pure-existence theorem is conditional and is not an empirical certificate. We present the strategy figures as descriptive snapshots of the IBR dynamics and distinguish them from audited mixed profiles reporting their provenance. The main qualitative question is which suppliers retain high task-specific bids, not whether every solver trace converges to a pure equilibrium.

6.1 Synthetic warmup: the asymmetric pattern in its simplest form

The synthetic two-provider example uses one discount factor per supplier, $r = 1$, $\lambda = 35$, $\rho_i = 10^{-5}$ USD/token, $D_d = \{0.8, 0.9, 1.0, 1.1\}$, and costs in $[2, 6] \times 10^{-6}$ USD/token (profiles in Appendix C). Figure 1 illustrates the resulting pricing pattern: A holds at the upper bound, while B bids 0.9 at the four lower cost points and 1.1 at the highest, so its bid is weakly increasing in cost, consistent with the direction of Corollary 1. This descriptive example is separate from the calibrated audits and does not verify the hypotheses of the continuous theorem; cost ordering on the calibrated caches is tested directly in Appendix G.

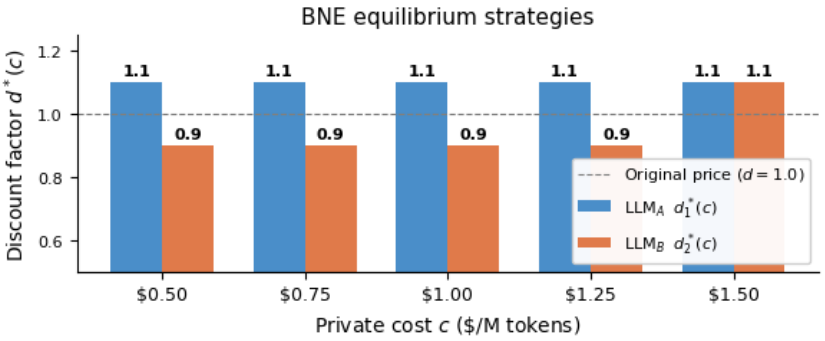


Figure 1: 2-LLM uniform-discount strategies (synthetic). Provider A holds at $d = 1.1$; B's bid is weakly increasing in its cost.

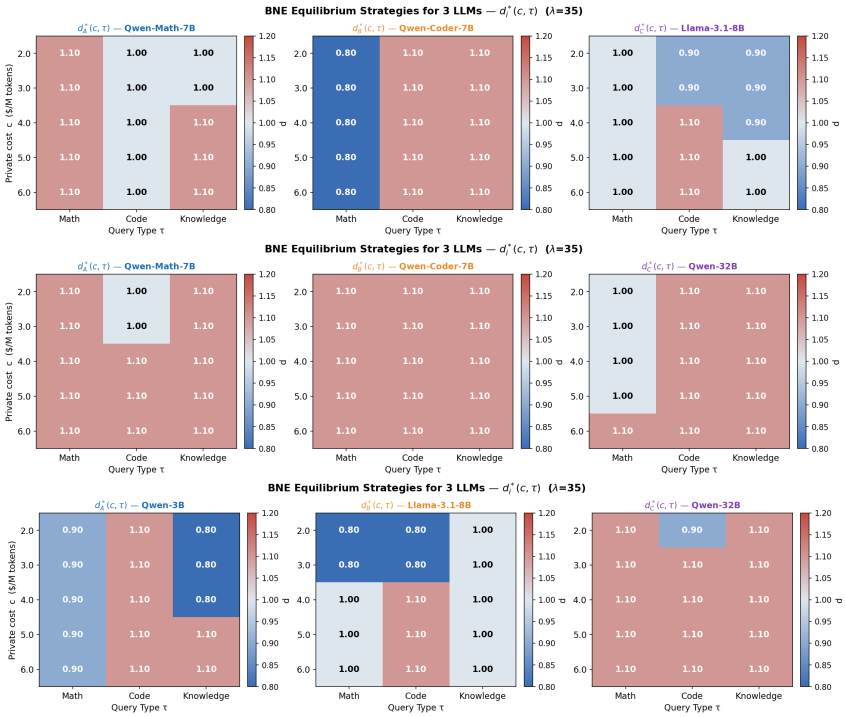


Figure 2: Calibrated IBR strategy snapshots on the four-point grid at $\lambda = 35$: T1 (top), T2 (middle), and T3 (bottom). These are IBR snapshots from one phase of each cycle that illustrate task-dependent bid patterns; they are not the FP-certified mixed profiles reported in Table 1.

6.2 Calibrated three-trio result

Calibration setup. The results above establish the asymmetric pattern under stylized capability profiles. We now ask whether it survives the move from stylized to measured.

We calibrate nine open-weight models spanning five families and 3B–32B parameters using GSM8K, MATH, HumanEval, MBPP, and MMLU, grouped into Math, Code, and Knowledge. Parameters include base accuracy, difficulty decay, chain boost, and public-price-based token costs (Appendix E; Table 4 lists which models enter which cohort). StarCoder2-7B and Mistral-Small-24B-Instruct exhibit anomalous calibration outcomes, and Gemma-2-9B scores zero on HumanEval in every bin. Three benchmark/difficulty cells (MBPP-easy, MBPP-hard, MATH-medium) read zero for every model; the two that enter the abilities (MBPP-easy, MATH-medium) enter as zeros, uniformly across models. The headline T1–T3 cohorts contain none of the three anomalous models; raw calibration entries are retained for transparency, not used to establish their intrinsic capabilities.

Three trios and headline result. We study T1 (Qwen-Math-7B, Qwen-Coder-7B, Llama-3.1-8B), T2 (Qwen-Math-7B, Qwen-Coder-7B, Qwen-32B), and T3 (Qwen-3B, Llama-3.1-8B, Qwen-32B), at $\lambda = 35$. The four-point IBR snapshots in Figure 2 illustrate leaders retaining high own-task bids and weaker suppliers discounting. Under IBR, the learned T1–T3 games exhibit cycles with periods $(10, 2, 3)$ on the four-point grid. Table 1 reports mixed profiles computed by FP on the same cached payoff tables and certified by deviation enumeration (Section 5). The IBR snapshots are therefore interpreted as descriptive solver diagnostics rather than certified equilibrium profiles.



Figure 3: IBR solver outputs for T1 as λ varies (top to bottom: $\lambda = 35, 100, 200, 300$). Task differentiation weakens as λ increases. The $\lambda = 35$ and $\lambda = 100$ panels are non-converged snapshots, whereas the $\lambda = 200$ and $\lambda = 300$ runs reach uniform upper-bound fixed points.

Cohort contrast in the IBR snapshots. The T2 snapshot has 39 of 45 displayed bid cells at $d_{\max}$, whereas T1 and T3 show stronger task-dependent differentiation. These observations motivate a comparative-advantage interpretation: a dominant generalist can absorb demand redirected by competitors' discounts. The analytical and finer-grid controls preserve the qualitative ordering, but not the same support or pure/mixed classification.

Cycling and numerical equilibrium quality. IBR cycling is a solver diagnostic and does not by itself characterize the market's equilibrium structure: it occurs in T1–T3 as well as in the follow-up cohorts T5–T8 (Appendix F). Averaging over a cycle produces a candidate distribution, not a proof of equilibrium. Evaluated by deviation enumeration, the T5–T8 cycle averages have residuals 9.77×10^{-5}, 9.92×10^{-5}, 1.10×10^{-4}, and 9.68×10^{-5}, respectively; the evaluated T8 cycle has period 4. Across six independently estimated payoff caches—one cache for each of T1 and T7 at seeds 7, 11, and 42—IBR cycle averages have residuals 6.31×10^{-5}–2.48×10^{-4}, whereas

Table 1: Absolute welfare per query in the model's common monetary units, rounded to six decimals. "Mix" denotes the mixed profile computed by FP and certified by deviation enumeration ($\epsilon = 7.44 \times 10^{-6}$, 4.67×10^{-7}, and 3.08×10^{-7} for T1–T3), rather than an IBR snapshot or cycle average. U_{cust} is reported separately because its payment weight is $\lambda = 35$. Supplier-level profits and planner values appear in Table 11.

Trio	Profile	CS_{mon}	$\sum_i PS_i$	W_{mon}	U_{cust}
T1	No discount	.588878	.007501	.596379	.163829
	Mix	.603248	.008141	.611389	.151522
T2	No discount	.640743	.007098	.647841	.238522
	Mix	.662589	.008266	.670855	.215091
T3	No discount	.638612	.007740	.646352	.200033
	Mix	.642173	.007999	.650172	.195297

FP reaches 4.47×10^{-7}–3.53×10^{-5}. Differences among solver traces alone do not demonstrate multiple equilibria or distinct equilibrium basins.

6.3 Counterintuitive sensitivity to customer price-sensitivity

The T1 λ sweep (Figure 3) shows weaker task differentiation as the customer places more weight on price, approaching uniform upper-bound bids in the displayed high-λ runs. This is an observed sensitivity of this calibration, not a universal comparative-static theorem. Changing λ changes routing thresholds and therefore the supplier's quantity response. It does not directly multiply supplier payments or supplier profit in Equation (3). We interpret the plots as consistent with threshold-driven allocation and do not infer a general "margin loss scales with λ" mechanism.

Synthesis. Synthetic examples illustrate heterogeneous bid incentives. Calibrated audits add finite-grid deviation residuals, absolute welfare accounting, and controls for routing rules, grid resolution, and cache seeds. They support task-dependent comparative-advantage pricing in the tested markets, while revealing router-dependent equilibrium structure and non-uniform learned-cache monotonicity.

6.4 Robustness

We report two complementary robustness analyses: synthetic parameter sweeps and calibrated controls. The synthetic sweeps vary the bid grid, cost prior, query mix, and λ and summarize task differentiation using

$$\text{asym} := \left(\bar{d}_A^{\text{Math}} - \bar{d}_A^{\text{Code}}\right) + \left(\bar{d}_B^{\text{Code}} - \bar{d}_B^{\text{Math}}\right), \tag{6}$$

where $\bar{d}_i^{\tau}$ averages the displayed bids over cost types. Figure 4 describes these synthetic runs (Appendix D). In the calibrated control, on the seven-point grid $\{.8, .85, .9, .95, 1, 1.05, 1.1\}$ with its own cache, learned T1–T3 cycle under IBR with periods $(30, 4, 4)$, against $(10, 2, 3)$ on the four-point grid. Certified monetary welfare changes by -1.19%, $-.34\%$, and $-.57\%$. T1 specialists' own-task bids move from $1.10/1.10$ to $1.08/1.10$, while the T2/T3 Knowledge leader remains at 1.10. Thus qualitative ordering is more stable than support, period, or pure/mixed classification. Appendix G reports the analytical and seed controls.

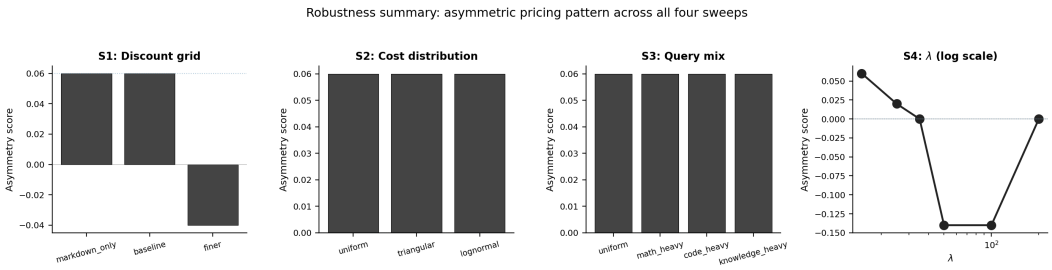


Figure 4: Summary of the four robustness sweeps. The asymmetry score (6) is positive in every configuration tested.

7 Welfare and practical implications

We evaluate the allocations induced by the pricing profiles and compare their monetary welfare with a planner benchmark. Price-weighted customer utility and monetary surplus answer different questions and must not be combined without accounting for their units.

Customer utility weights payment by λ, whereas supplier profit weights it by one. Let expectations include query types, private costs, routing randomization, and any independently mixed bids. In common monetary units per query, define

$$CS_{\text{mon}} := \mathbb{E}\left[r\alpha(Q,S) - \sum_i d_i^\tau F_i(S)\right], \tag{7}$$

$$PS_i := \mathbb{E}\left[d_i^\tau F_i(S) - c_i \text{tok}_i(S)\right], \tag{8}$$

$$W_{\text{mon}} := CS_{\text{mon}} + \sum_i PS_i = \mathbb{E}\left[r\alpha(Q,S) - \sum_i c_i \text{tok}_i(S)\right]. \tag{9}$$

Here $r\alpha$ is the model's monetized value of correctness, not measured customer willingness to pay. Transfers cancel for a fixed allocation, but bids and λ can change that allocation and hence expected monetary welfare. We report U_{cust} separately; $W_{\text{nat}} := U_{\text{cust}} + \sum_i PS_i$ is a utility index, not monetary total surplus when $\lambda \neq 1$. The planner chooses a mode maximizing $r\alpha(Q,S) - \sum_i c_i \text{tok}_i(S)$ with the same feasible modes and information on realized costs. Its monetary gap defines $\text{DWL} := W_{\text{mon}}^{\text{SP}} - W_{\text{mon}}^{\text{alloc}} \geq 0$. No discount means $d_i^\tau = 1$, distinct from the upper-bound initialization $d_{\max} = 1.1$.

Welfare results and implications. The certified FP profiles increase monetary consumer surplus relative to no discount by approximately 2.44%, 3.41%, and 0.56% in T1–T3, while price-weighted customer utility decreases. These measures answer different questions because U_{cust} applies the payment weight $\lambda = 35$, whereas monetary consumer surplus is expressed in common monetary units. Monetary welfare remains below the planner benchmark (Table 11). For T3, the certified FP profile changes U_{cust} from 0.200033 to 0.195297 (-2.37%), while the diagnostic IBR cycle-average profile has $U_{\text{cust}} = 0.193561$ and a deviation residual of 4.16×10^{-5}. Because these values are evaluated under different strategy profiles, we use the FP-certified value as the reported equilibrium result and retain the IBR cycle average only as a solver diagnostic. Reserve bid factors and alternative tie-breaking remain candidate interventions, not experimentally established remedies.

8 Conclusion and limitations

We study supply-side pricing induced by heterogeneous LLM routing, with cache-compatible solvers and fictitious-play profiles certified by direct finite-grid deviation audits. The theoretical guarantees distinguish unconditional finite-game mixed existence from conditional continuous-game existence. The tested controls support qualitative comparative-advantage pricing, not universal router invariance, exact mixed BNE from cycling, or a continuous-grid limit. Limitations include one-shot independent private costs, nine calibrated models, three broad task categories, anomalous calibration and grading outcomes, and finite bid and cost grids, with the five-point cost grid held fixed. Latency, capacity constraints, repeated pricing, platform incentives, larger supplier sets, richer task distributions, and evaluated mechanism interventions remain open. Absolute monetary welfare and price-weighted customer utility must be interpreted separately.

Acknowledgments and Disclosure of Funding

Funding. Funding in direct support of this work: the North Atlantic Treaty Organisation under SPS G7200.

Competing interests. The authors declare no competing interests.

References

Yanwei Yue, Guibin Zhang, Boyang Liu, Guancheng Wan, Kun Wang, Dawei Cheng, and Yiyan Qi. Masrouter: Learning to route llms for multi-agent systems. In *Proceedings of the 63rd*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15549–15572, 2025.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs with preference data. *arXiv preprint arXiv:2406.18665*, 2024. URL https://arxiv.org/abs/2406.18665.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023a. URL https://arxiv.org/abs/2305.05176.

Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Ruhle, Laks V. S. Lakshmanan, and Ahmed Hassan Awadallah. Hybrid LLM: Cost-efficient and quality-aware query routing. *arXiv preprint arXiv:2404.14618*, 2024. URL https://arxiv.org/abs/2404.14618.

Pranjal Aggarwal, Aman Madaan, Ankit Anand, Srividya Pranavi Potharaju, Swaroop Mishra, Pei Zhou, Aditya Gupta, Dheeraj Rajagopal, Karthik Kappaganthu, Yiming Yang, et al. AutoMix: Automatically mixing language models. *arXiv preprint arXiv:2310.12963*, 2024. URL https://arxiv.org/abs/2310.12963.

Zhendong Guo, Wenchao Bai, and Jiahui Jin. Pricing online LLM services with data-calibrated stackelberg routing game. *arXiv preprint arXiv:2511.09062*, 2025. URL https://arxiv.org/abs/2511.09062.

Vijay Krishna. *Auction Theory*. Academic Press, 2nd edition, 2010.

R. Preston McAfee and John McMillan. Auctions and bidding. *Journal of Economic Literature*, 25 (2):699–738, 1987.

Paul R. Milgrom and Robert J. Weber. A theory of auctions and competitive bidding. *Econometrica*, 50(5):1089–1122, 1982.

Yeon-Koo Che. Design competition through multidimensional auctions. *The RAND Journal of Economics*, 24(4):668–680, 1993.

John Asker and Estelle Cantillon. Properties of scoring auctions. *The RAND Journal of Economics*, 39(1):69–85, 2008.

Eric Maskin and John Riley. Asymmetric auctions. *The Review of Economic Studies*, 67(3):413–438, 2000. URL https://ideas.repec.org/a/oup/restud/v67y2000i3p413-438.html.

Estelle Cantillon. The effect of bidders' asymmetries on expected revenue in auctions. *Games and Economic Behavior*, 62(1):1–25, January 2008. URL https://ideas.repec.org/a/eee/gamebe/v62y2008i1p1-25.html.

Susan Athey. Single crossing properties and the existence of pure strategy equilibria in games of incomplete information. *Econometrica*, 69(4):861–889, 2001.

Philip J. Reny. On the existence of pure and mixed strategy Nash equilibria in discontinuous games. *Econometrica*, 67(5):1029–1056, 1999.

Hal R. Varian. A model of sales. *American Economic Review*, 70(4):651–659, 1980.

Harold Hotelling. Stability in competition. *The Economic Journal*, 39(153):41–57, 1929. doi: 10.2307/2224214.

Steven C. Salop. Monopolistic competition with outside goods. *The Bell Journal of Economics*, 10 (1):141–156, 1979. doi: 10.2307/3003323.

Jean Tirole. *The Theory of Industrial Organization*. MIT Press, Cambridge, MA, 1988. ISBN 978-0-262-20071-4.

Avner Shaked and John Sutton. Natural oligopolies. *Econometrica*, 51(5):1469–1484, 1983.

Michael Mussa and Sherwin Rosen. Monopoly and product quality. *Journal of Economic Theory*, 18 (2):301–317, 1978.

Jiangjie Chen, Siyu Yuan, Rong Ye, Bodhisattwa Prasad Majumder, and Kyle Richardson. Put your money where your mouth is: Evaluating strategic planning and execution of llm agents in an auction arena, 2023b. URL https://arxiv.org/abs/2310.05746.

Anand V. Shah, Kehang Zhu, Yanchen Jiang, Jeffrey George Wang, Arif Kerem Dayi, John Joseph Horton, and David C. Parkes. Learning from synthetic labs: Language models as auction participants, 2025. URL https://arxiv.org/abs/2507.09083.

Sara Fish, Yannai A. Gonczarowski, and Ran I. Shorrer. Algorithmic collusion by large language models, 2024. URL https://arxiv.org/abs/2404.00806.

Rabimba Karanjai, Sam Blackshear, Lei Xu, and Weidong Shi. Collaboration is all you need: Llm assisted safe code translation. In *Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering*, pages 671–675, 2025a.

Rabimba Karanjai, Lei Xu, and Weidong Shi. Hpcagenttester: A multi-agent llm approach for enhanced hpc unit test generation. In *2025 2nd IEEE/ACM International Conference on AI-powered Software (AIware)*, pages 213–222. IEEE, 2025b.

George W. Brown. Iterative solution of games by fictitious play. In Tjalling C. Koopmans, editor, *Activity Analysis of Production and Allocation*, pages 374–376. John Wiley & Sons, New York, 1951.

Julia Robinson. An iterative method of solving a game. *Annals of Mathematics*, 54(2):296–301, 1951. doi: 10.2307/1969530.

Lloyd S. Shapley. Some topics in two-person games. In Melvin Dresher, Lloyd S. Shapley, and Albert W. Tucker, editors, *Advances in Game Theory*, number 52 in Annals of Mathematics Studies, pages 1–28. Princeton University Press, Princeton, NJ, 1964.

John C. Harsanyi. Games with incomplete information played by 'Bayesian' players, I–III. *Management Science*, 14:159–182, 320–334, 486–502, 1967–1968.

John Nash. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36(1):48–49, 1950.

Paul R. Milgrom and Robert J. Weber. Distributional strategies for games with incomplete information. *Mathematics of Operations Research*, 10(4):619–632, 1985.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Llama Team, AI @ Meta. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

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

Mistral AI Team. Mistral Small: Affordable, efficient, and high-performing. https://mistral.ai/news/mistral-small-3/, 2024. Model card and release announcement.

Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, et al. StarCoder 2 and The Stack v2: The next generation. *arXiv preprint arXiv:2402.19173*, 2024. URL https://arxiv.org/abs/2402.19173.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021a. URL https://arxiv.org/abs/2103.03874.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. URL https://arxiv.org/abs/2107.03374.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021. URL https://arxiv.org/abs/2108.07732.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021b. URL https://arxiv.org/abs/2009.03300.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, 2023. URL https://arxiv.org/abs/2309.06180.

A Theoretical proofs and additional theory

This appendix collects the proofs deferred from Section 4, the continuous-game existence result, the formal single-crossing development, and several supporting remarks.

A.1 Restatement of regularity assumption

For convenience we restate the regularity assumption used throughout:

> **Assumption 1 (Regularity).** The type space $\mathcal{C}$ and action space D^K are nonempty compact intervals; the prior F admits a continuous density bounded above and below by positive constants; the per-mode accuracy and token consumption are well-defined and bounded; and the base fees are uniformly bounded.

These regularity conditions describe the continuous model. The numerical experiments replace continuous costs by finite probability distributions; a finite-grid audit is not a verification of all continuous-model hypotheses.

A.2 Proof of Theorem 1 (mixed-strategy BNE in the discretized game)

By the Harsanyi transformation [Harsanyi, 1967–1968], the Bayesian game Γ_d is equivalent to a finite strategic-form game G in which each player-type pair $(i, c^{(\ell)})$ is treated as a distinct "agent." The induced game G has nL agents, each with J^K pure actions, and payoffs that are multilinear in the agents' mixed strategies. By Nash's existence theorem [Nash, 1950], G admits at least one mixed Nash equilibrium. The standard correspondence between equilibria of G and Bayesian equilibria of Γ_d [Harsanyi, 1967–1968] then yields the result. □

A.3 Existence in the continuous game

The finite-game theorem requires no routing continuity. A sufficient condition for continuous-game mixed existence is stated separately below.

Theorem 3 (Conditional Existence in the Continuous Game)**.** *Under Assumption 1 and the independent private-cost model, suppose the expected revenue and token allocation aggregates R_i^τ and X_i^τ in Theorem 2 are jointly continuous in actions for every supplier and query type. Then the continuous game admits a BNE in distributional strategies.*

Proof. For realized cost c_i, the per-query-type payoff is $R_i^\tau(\mathbf{d}) - c_i X_i^\tau(\mathbf{d})$. It is bounded and jointly continuous in actions and c_i on compact domains; a finite query-type sum preserves these properties. In particular, action continuity is uniform in the bounded private costs. The compact metric action and type spaces and independent continuous cost densities satisfy the type-distribution and integrability requirements of the distributional-strategy existence result of Milgrom and Weber [1985]. That result supplies a BNE in distributional strategies. □

The added continuity is substantive. A deterministic utility-maximizing mode selector can jump at ties; continuity of the customer's optimal value or continuity away from bid-space tie surfaces does not establish continuity of supplier payoffs. Atomless costs do not prevent bid strategies from pooling on a switching surface. We do not apply this result to a discontinuous routing rule without a separate existence argument.

A.4 Single-crossing condition (full development)

Remark 1 (Aggregate versus mode-wise monotonicity)**.** *The weighted sum X_i^τ is the quantity required by the single-crossing proof. Non-increasing probabilities for all modes involving i imply non-increasing X_i^τ, but reallocation between such modes can violate mode-wise monotonicity without violating the aggregate condition. For the ideal exact optimizer in Equation* (2)*, write the objective with other bids fixed as $A(\pi) - \lambda\rho_i d_i^\tau X_i^\tau(\pi)$, where A is independent of d_i^τ. Optimality of π_H at d_H and π_L at d_L gives, on adding the two optimality inequalities,*

$$\lambda\rho_i(d_H - d_L)(X_i^\tau(\pi_L) - X_i^\tau(\pi_H)) \geq 0.$$

Thus $X_i^\tau(\pi_H) \leq X_i^\tau(\pi_L)$ when $d_H > d_L$ and $\lambda\rho_i > 0$. For a supplier with $\rho_i = 0$, this argument alone gives no token ordering. This exact-optimization result does not imply mode-wise probability monotonicity, action continuity, or the same property for an approximate learned policy. Finite-grid audits in Appendix G concern only tested comparisons.

Definition 1 (Single-Crossing Condition, Athey, 2001)**.** *Fix a player i, a query type $\tau \in \mathcal{T}$, and a strategy profile σ_{-i} for the opponents. The per-type interim payoff*

$$\Phi_i^\tau(d, c) \triangleq \mathbb{E}_{c_{-i}\sim F^{n-1}}\mathbb{E}_{S\sim\pi^*(\cdot|\tau,(d,\sigma_{-i}^\tau(c_{-i})))}\big[d \cdot F_i(S) - c \cdot \mathrm{tok}_i(S)\big] \tag{10}$$

satisfies the single-crossing condition (SCC) in $(d; c)$ *if for every $d_H > d_L$ in D and every $c_H > c_L$ in $\mathcal{C}$,*

$$\Phi_i^\tau(d_H, c_L) - \Phi_i^\tau(d_L, c_L) \;\geq\; 0 \implies \Phi_i^\tau(d_H, c_H) - \Phi_i^\tau(d_L, c_H) \;\geq\; 0. \tag{11}$$

A sufficient condition for the implication above is increasing differences: for every $d_H > d_L$, the incremental return $\Delta\Phi_i^\tau(d_H, d_L; c) := \Phi_i^\tau(d_H, c) - \Phi_i^\tau(d_L, c)$ is non-decreasing in c. We establish this stronger property; it is not in general equivalent to the single-crossing implication.

Proposition 1 (Single-Crossing Holds)**.** *Under Assumptions 1 and 2, the per-type interim payoff Φ_i^τ satisfies the single-crossing condition for every player i, every query type τ, and every opponent strategy profile σ_{-i}.*

Proof. Fix i, τ, σ_{-i}, and $d_H > d_L$ in D. Let $\pi_H^*(S; c_{-i}) \triangleq \pi^*(S \mid \tau, (d_H, \sigma_{-i}^\tau(c_{-i})))$ and analogously π_L^*. Decompose the incremental return:

$$\begin{aligned}
\Delta\Phi_i^\tau(d_H, d_L; c) &= \mathbb{E}_{c_{-i}} \sum_{S\in\mathcal{S}} \Big[\pi_H^*(S; c_{-i})\big(d_H F_i(S) - c\,\mathrm{tok}_i(S)\big) - \pi_L^*(S; c_{-i})\big(d_L F_i(S) - c\,\mathrm{tok}_i(S)\big)\Big] \\
&= \underbrace{\mathbb{E}_{c_{-i}} \sum_S \big[\pi_H^*(S; c_{-i})\, d_H - \pi_L^*(S; c_{-i})\, d_L\big] F_i(S)}_{\text{revenue difference; independent of } c} \\
&\quad - c \cdot \underbrace{\mathbb{E}_{c_{-i}} \sum_S \big[\pi_H^*(S; c_{-i}) - \pi_L^*(S; c_{-i})\big]\mathrm{tok}_i(S)}_{\triangleq B}\,.
\end{aligned} \tag{12}$$

Let $\mathbf{d}_H(c_{-i})$ and $\mathbf{d}_L(c_{-i})$ denote the bid profiles entering π_H^* and π_L^*; they differ only in i's type-τ bid. Then

$$B = \mathbb{E}_{c_{-i}}\left[X_i^\tau(\mathbf{d}_H(c_{-i})) - X_i^\tau(\mathbf{d}_L(c_{-i}))\right] \leq 0$$

by aggregate own-price monotonicity, pointwise in c_{-i}. The cost prior is independent of c, and the routing rule does not observe hidden costs, so B and the revenue difference are independent of c. For $c_H > c_L$,

$$\Delta\Phi_i^\tau(d_H, d_L; c_H) - \Delta\Phi_i^\tau(d_H, d_L; c_L) = -(c_H - c_L)B \geq 0.$$

Thus the payoff has increasing differences in (d, c), which implies Equation (11). Strictly positive increments also remain positive, so the strong single-crossing requirement used in the monotone-existence argument is met. □

Remark 2 (Economic Interpretation)**.** *The economic content of Proposition 1 is the classical* volume-versus-margin *trade-off. A low-cost provider (c small) finds the "volume strategy" (low d, high market share) attractive: each unit served carries a positive margin $d \cdot F_i(S) - c \cdot \mathrm{tok}_i(S)$, and there are many units to serve. A high-cost provider, by contrast, earns a small margin per unit and must rely on the "margin strategy" (high d, low market share) to remain profitable. The single-crossing condition formalizes the comparative statics: as cost rises, the marginal benefit of charging more, i.e., raising d, increases. This is the structural force driving the monotone equilibrium of Theorem 2.*

A.5 Proof of Lemma 1 (Additive separability)

The implemented type-τ routing policy is assumed to depend only on the type-τ bids. This separability is part of the cache construction, not a consequence of hidden-cost independence alone. Each payoff term then involves only the corresponding bid component, and exchanging the finite task-type and mode sums in Equation (3) yields Equation (5). □

A.6 Full proof of Theorem 2 (monotone pure-strategy BNE)

The proof proceeds in three steps.

Step 1 (Decomposition into type-τ subgames). By Lemma 1, the maximization in $\mathbf{d}_i \in D^K$ separates:

$$\arg\max_{\mathbf{d}_i \in D^K} U_i(\mathbf{d}_i, \sigma_{-i}; c_i) = \prod_{\tau \in \mathcal{T}} \arg\max_{d \in D} U_i^\tau(d, \sigma_{-i}^\tau; c_i). \tag{13}$$

Hence σ_i^* is a best response to σ_{-i}^* in the original game Γ if and only if, for every τ, the component $\sigma_i^{\tau *} : \mathcal{C} \to D$ is a best response in the one-dimensional subgame Γ^τ in which the action space is D and the payoff function is U_i^τ.

Step 2 (Conditional subgame existence). Each query-type subgame has compact interval type and action spaces and independent atomless cost densities. Its ex post payoff is $R_i^\tau(\mathbf{d}^\tau) - c_i X_i^\tau(\mathbf{d}^\tau)$, bounded and jointly continuous by the explicit hypotheses of Theorem 2. Proposition 1 gives increasing differences against every fixed opponent strategy profile, in particular against non-decreasing strategies. The continuous-payoff monotone pure-equilibrium existence result of Athey [2001] therefore applies, yielding a monotone pure BNE in each subgame, with best-response conditions for almost every cost. This uses joint payoff continuity, not continuity merely away from switching boundaries. It does not invoke the specialized discontinuous-auction Theorem 6 without checking its additional assumptions.

Step 3 (Reassembly). Define the composite strategy

$$\sigma_i^*(c_i) \triangleq \left(\sigma_i^{\tau_1 *}(c_i), \ldots, \sigma_i^{\tau_K *}(c_i)\right) \in D^K, \quad i \in N. \tag{14}$$

By Step 1, σ_i^* is a best response to σ_{-i}^* in Γ if and only if each component $\sigma_i^{\tau *}$ is a best response in Γ^τ, which holds by Step 2. Therefore $\sigma^* = (\sigma_1^*, \ldots, \sigma_n^*)$ is a pure BNE of Γ. The monotonicity of each component $\sigma_i^{\tau *}$ in c_i is preserved by the assembly. □

Remark 3 (Scope of continuous-game existence)**.** *Both continuous-game existence results impose explicit continuity conditions. The pure result also imposes aggregate monotonicity. Neither result certifies the learned finite-game profiles or guarantees convergence of IBR. Deterministic analytical routing requires a separate argument at switching boundaries if continuous-game existence is claimed.*

A.7 Proof of Corollary 1

Both ρ_i and the fixed-mode fee $F_i(S)$ are non-negative constants independent of c_i. Multiplying a non-decreasing bid by either constant preserves monotonicity. No conclusion about the distribution of selected modes is needed. □

A.8 Proof of Corollary 2

Proof. Let d_H and d_L be the greatest maximizers of v_H and v_L, which exist by continuity on compact D. Suppose $d_H < d_L$. Optimality yields $v_H(d_H) - v_H(d_L) \geq 0$ and $v_L(d_L) - v_L(d_H) \geq 0$. Monotonicity of $v_H - v_L$ gives

$$v_H(d_H) - v_L(d_H) \leq v_H(d_L) - v_L(d_L).$$

The sum of the two non-negative optimality differences is therefore at most zero. Both must be zero, so d_L is also a maximizer of v_H, contradicting the choice of d_H as its greatest maximizer. Thus $d_H \geq d_L$. □

A.9 Reduction to classical FPSBA

Remark 4 (Classical procurement comparison)**.** *For symmetric risk-neutral procurement bidders with independent uniform costs on $[0, \bar{c}]$, the textbook unrestricted-bid equilibrium is $b^*(c) = c + (\bar{c} - c)/n$ [Krishna, 2010]. This provides an economic comparison when quality is identical and allocation goes to the lowest quote. It is not a direct corollary of our continuous-payoff theorem: winner-take-all routing is discontinuous at ties, and the bounded multiplier action set need not contain all textbook bids.*

A.10 Non-uniqueness and equilibrium selection

Remark 5 (Equilibrium selection)**.** *The existence results do not assert uniqueness. If a query-type subgame has several equilibria, separability allows different combinations across types. Symmetric capabilities alone do not establish any particular asymmetric pair as an equilibrium. Likewise, different IBR cycles or snapshot welfare values do not prove multiplicity; each candidate must be evaluated by its own deviation gains.*

A.11 Iterated best-response equilibrium selection

Remark 6 (Iterated best-response selection)**.** *IBR is initialized at $d_{\max}$, an upper-bound quote rather than the $d = 1$ no-discount benchmark. Given a fixed cache, schedule, and tie-breaking, the trace is deterministic. A fixed point verified against every grid deviation is a pure BNE of that finite game. A cycle is an algorithmic outcome; neither existence nor non-existence of pure equilibria elsewhere follows. Empirical cycle distributions require a separate deviation audit. The reported profiles are therefore computed by fictitious play and certified by enumeration (Section 5, Appendix B.3); IBR serves as a diagnostic.*

B Computational method, algorithm, and implementation

This appendix collects the details deferred from Section 5: the cache-decomposition proposition and its proof, the IBR pseudocode (Algorithm 1), the fictitious-play pseudocode (Algorithm 2), a remark on IBR cycles, implementation notes, and the analytical-router control.

B.1 Cache decomposition: formal statement and proof

We formalize the cache pre-computation summarized in Section 5.

Proposition 2 (Cache Decomposition)**.** *For each query type $\tau \in \mathcal{T}$ and each discount tuple $(d_1, \ldots, d_n) \in D_d^n$, let*

$$\mathcal{Z}[\tau, d_1, \ldots, d_n] \triangleq \big(\pi^*(S \mid \tau, \mathbf{d})\big)_{S \in \mathcal{S}} \in \Delta(\mathcal{S}) \tag{15}$$

Algorithm 1 Previous-sweep IBR over cached routing policies

Require: Cache $\mathcal{Z}$, finite cost/bid grids, prior F, maximum sweeps R

Initialize $\sigma_i^{(0)}(c) = (d_{\max}, \ldots, d_{\max})$ for all i, c
for $r = 1, \ldots, R$ **do**
  for each supplier i and cost grid point c **do**
    for each candidate $\mathbf{d} \in D_d^K$ **do**
      Evaluate $u(\mathbf{d}) = U_i(\mathbf{d}, \sigma_{-i}^{(r-1)}; c)$ using the cache
      Sum over opponent costs with weight $\prod_{j \neq i} F(c_j)$
    end for
    Set $\sigma_i^{(r)}(c)$ to a maximizing candidate using fixed tie-breaking
  end for
  if the profile is unchanged after the full sweep **then**
    Enumerate every unilateral grid deviation to verify the fixed point
    return the profile and its deviation residual
  end if
  Record repeated profiles and cycle diagnostics
end for
return the trace and any candidate empirical distributions for separate auditing

denote the customer's optimal mode distribution under any discount profile $\mathbf{d}$ *whose type-*τ *components are* $(d_1, \ldots, d_n)$*. The full cache* $\mathcal{Z}$ *contains* $K \cdot J^n$ *entries; once* $\mathcal{Z}$ *is constructed, every quantity needed by the best-response iteration (provider utilities, customer utility, mode-selection frequencies) is computable in* $O(|\mathcal{S}|)$ *arithmetic time per query.*

Proof. By the separation property in Section 3, the customer's optimization for type τ depends on $\mathbf{d}$ only through its type-τ components $(d_1^\tau, \ldots, d_n^\tau)$, so $\mathcal{Z}[\tau, \cdot]$ is well-defined: any two profiles $\mathbf{d}, \mathbf{d}'$ that agree on type-τ components yield the same $\pi^*(\cdot \mid \tau, \mathbf{d}) = \pi^*(\cdot \mid \tau, \mathbf{d}')$. Substituting into Equation 3 reduces provider i's interim expected utility to a finite sum over $\mathcal{S}$ of cached probabilities times closed-form revenue and cost terms:

$$U_i(\mathbf{d}_i, \sigma_{-i}; c_i) = \mathbb{E}_{c_{-i},\tau} \sum_{S \in \mathcal{S}} \mathcal{Z}\big[\tau, d_i^\tau, \sigma_{-i}^\tau(c_{-i})\big](S) \cdot \big[d_i^\tau F_i(S) - c_i \mathrm{tok}_i(S)\big].$$

Each evaluation is therefore $O(|\mathcal{S}|)$ once the cache is loaded. The full procedure is detailed in Algorithm 1. □

B.2 Algorithm 1: iterated best response over cached customer policies

B.3 Algorithm 2: fictitious play and certification

Fictitious play [Brown, 1951] is run on the same cached finite game as IBR. Each supplier–cost pair (i, c) is an agent of the agent-normal form, whose strategy is a distribution $\mu_i(\cdot \mid c)$ over D_d^K. The reported profile is the vector of empirical best-response frequencies after the final round, and it is certified by the deviation residual of Appendix G.2. FP is a search procedure: it converges in special classes [Robinson, 1951] but can cycle in general [Shapley, 1964], so no claim rests on its convergence.

For the reported FP profiles, we use $R = 2000$ rounds and $S = 100$ Monte Carlo samples per candidate evaluation. Supplier–cost agents are updated sequentially, with no early-stopping rule. Expectations over the finite rival-cost grid are computed exactly, whereas expectations over rivals' empirical action mixtures are estimated by Monte Carlo.

B.4 Convergence and cycles

Remark 7 (Termination and cycles)**.** *After each full sweep, unchanged strategies trigger a fixed-point check against all grid deviations. If an orbit repeats, we record its period and cost-conditional action frequencies as a candidate mixed profile. Neither the last iterate nor a synchronized cycle-phase average is treated as a mixed BNE without an independent evaluation of unilateral deviations. In*

Algorithm 2 Fictitious play over cached routing policies, with certification

Require: Cache $\mathcal{Z}$, finite cost and bid grids, prior F, number of rounds R, and Monte Carlo sample size S

Initialize $n_i(c, a) \leftarrow 0$ and $\bar{\mu}_i(a \mid c) \leftarrow 1/|D_d^K|$ for every supplier i, cost point c, and action $a \in D_d^K$
for $r = 1, \ldots, R$ **do**
  for each supplier i and cost grid point c, in a fixed sequential order **do**
    Evaluate $u(\mathbf{d}) = U_i(\mathbf{d}, \bar{\mu}_{-i}; c)$ for each $\mathbf{d} \in D_d^K$ using exact summation over the rivals' finite cost grids and S Monte Carlo samples from their current empirical action mixtures
    $\mathbf{d}^\star \leftarrow$ a maximizer of u using the fixed lowest-bid tie-breaking rule
    $n_i(c, \mathbf{d}^\star) \leftarrow n_i(c, \mathbf{d}^\star) + 1$
    $\bar{\mu}_i(\cdot \mid c) \leftarrow n_i(c, \cdot) \big/ \sum_a n_i(c, a)$
  end for
end for
Set $\mu_i(\cdot \mid c) \leftarrow \bar{\mu}_i(\cdot \mid c)$ for every supplier i and cost point c
Enumerate every unilateral grid deviation against μ and compute $\epsilon(\mu)$
return μ together with $\epsilon(\mu)$

Table 2: Per-type base accuracy $\alpha(\tau)$ for the synthetic two-provider profiles. Bold entries indicate each provider's specialty type.

Setting	Provider	Math	Code	Knowledge
2-LLM (uniform-discount, per-type)	A	**0.88**	0.80	0.72
	B	0.82	**0.85**	**0.83**

the learned T1–T3 games, IBR produces cycles; the reported mixed profiles are therefore obtained using Algorithm 2. IBR cycle averages are retained only as solver diagnostics, with their deviation residuals reported in Appendix G.

B.5 Implementation and reproducibility

Our implementation is approximately 800 lines of Python. The customer policy is trained with PyTorch using the Adam optimizer, learning rate 5×10^{-3}, 25 epochs with 15 mini-batches of 32 queries each, and evaluated over 60 mini-batches. Cache files are persisted as Python pickles and loaded in milliseconds. The complete code, cached customer policies, and configuration files used to produce the figures and tables in this paper are included in the supplementary material.

B.6 Rational-baseline check

The analytical control selects a utility-maximizing mode for each query type and bid profile, breaking ties lexicographically, and supplies its own cache to the same IBR interface. The audit finds zero aggregate participation/token monotonicity violations in 3,024 four-point comparisons and 7,938 finer-grid comparisons. Only 30 of 18,144 active mode cells increase on the four-point grid, all due to within-supplier mode reallocation. Task leaders retain high own-task bids and weaker suppliers discount, but analytical IBR reaches pure fixed points while the learned T1–T3 games cycle. The control supports qualitative pricing, not uniform mode-wise monotonicity or router-independent equilibrium structure. Grid checks do not establish continuous-action continuity.

C Synthetic profiles and the two-provider illustration

This appendix gives the synthetic profiles behind Section 6.1, the full figure for the two-provider illustration (Figure 1), and the two-provider per-type game used as the baseline of the robustness sweeps (Section 6.4).

C.1 Synthetic capability profiles

All synthetic experiments use $r = 1$, $\lambda = 35$, $\rho_i = 10^{-5}$ USD/token, discount grid $D_d = \{0.8, 0.9, 1.0, 1.1\}$, and the five-point cost grid $\mathcal{C}_d = \{2.0, 3.0, 4.0, 5.0, 6.0\} \times 10^{-6}$ USD/token, with an i.i.d. uniform prior over its points.

C.2 2-LLM uniform-discount headline (full 3-panel figure)

We start with the simplest setting: two providers and a single discount factor per provider (no per-type differentiation). Both providers have similar overall capabilities ($\bar{\alpha}_A = 0.80$, $\bar{\alpha}_B = 0.83$), with A having a comparative advantage in Math and B in Code and Knowledge.

Figure 5(a) shows the synthetic strategy profiles: A holds at $d_{\max} = 1.1$ at every cost, while B bids 0.9 at the four lower cost points and 1.1 at the highest. B's bid is therefore weakly increasing in its private cost, consistent with the direction predicted by Corollary 1. This finite-grid example illustrates the resulting strategy pattern rather than verifying the hypotheses of Theorem 2; cost ordering on the calibrated caches is tested directly in Appendix G.

An informal reading: A is selected on Math queries, its strong type, for a wide range of $\mathbf{d}$, so reducing d_A does little to expand its market share and A retains its margin. B faces more elastic demand on its weaker types, where shading can shift selection. Figure 5(b)–(c) is consistent with this reading: under the computed discounts, the customer routes more queries to B-favoring modes, while A retains a higher per-query margin.

The flat-then-shade pattern observed here motivates two natural questions: Appendix C.3 lets providers price-discriminate across query types; that per-type game is the baseline of the robustness sweeps in Appendix D.

C.3 2-LLM per-type-discount extension

We now allow each provider to set a separate discount factor for each of the three query types: Math, Code, and Knowledge. The strategy space per cost realization expands from $|D_d| = 4$ to $|D_d|^K = 64$, but by Lemma 1, the BNE decomposes into three independent per-type subgames.

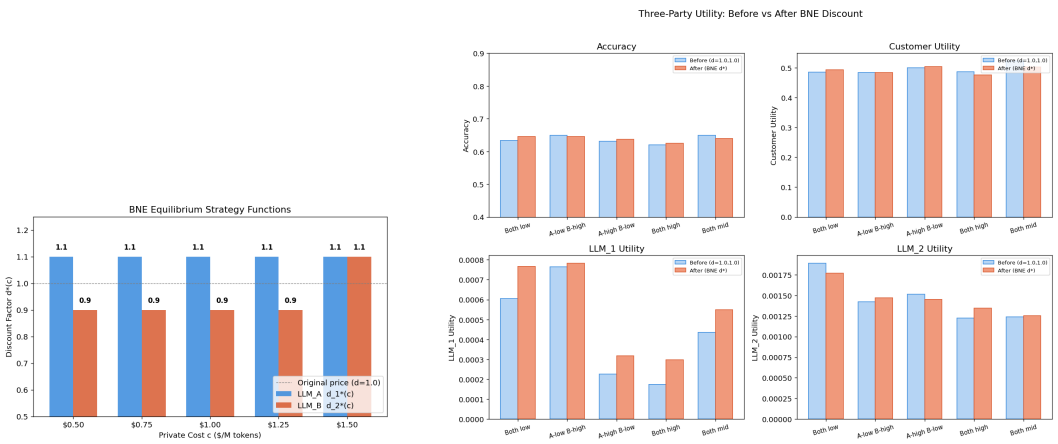
Figure 6(a) displays the equilibrium discount factor $\sigma_i^{\mathcal{T}*}(c_i)$ as a heatmap over cost and query type for each provider. The structure is striking: each provider holds at $d = 1.1$ *uniformly across all costs* on its strong type, with A on Math and B on Code and Knowledge, and shades only on its weak type. In other words, the asymmetric pattern from Appendix C.2 now manifests *per type*: each provider behaves as a "specialist" on its fortress queries and as a "competitor" on the others.

Two consequences follow. First, market share concentrates by type, as shown in Figure 6(b): Math queries route predominantly through A-favoring modes, while Code and Knowledge queries route through B-favoring modes. Second, the per-type structure does not weaken provider rent extraction; if anything, the cleaner type separation *amplifies* it relative to a uniform-discount BNE, as shown in Figure 6(c).

D Synthetic robustness sweeps

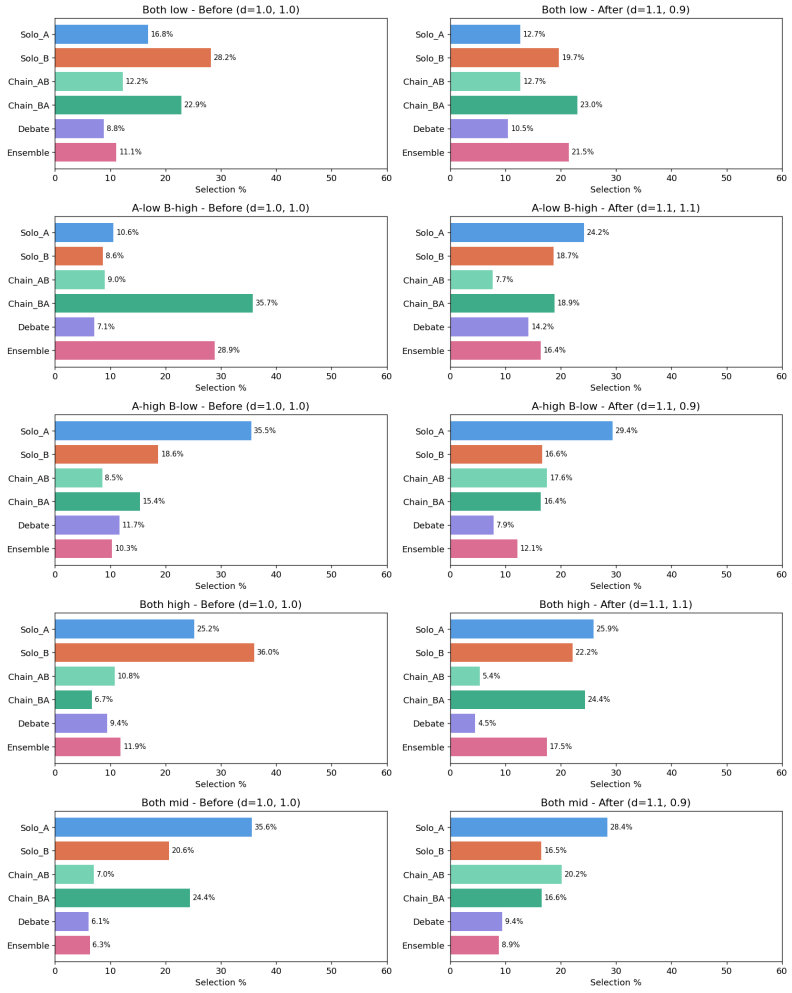
The four synthetic sweeps behind Figure 4 vary one parameter at a time around the two-provider per-type game of Appendix C.3 (profiles in Table 2; $\lambda = 35$ is chosen so that the price term in Equation (2) is comparable to a 25–35% accuracy reduction at typical fees). Table 3 lists the settings and the reported asymmetry scores (Equation (6)) of the computed IBR profiles. The scores are computed from the finite-grid profiles displayed in the corresponding sweep figures. These are descriptive synthetic results, separate from the calibrated audits.

The score is positive in most tested configurations and nonnegative in 13 of 16 configurations. It is unchanged across the tested cost priors and query mixtures, but varies with the bid grid and λ. In the price-sensitivity sweep, the score decreases from 0.06 at $\lambda = 15$ to -0.14 at $\lambda = 50$ and 100, before returning to 0.00 at $\lambda = 200$. These descriptive finite-grid results indicate stronger robustness to the cost prior and query mix than to the bid grid and price-sensitivity parameter; they are not general comparative-static results.



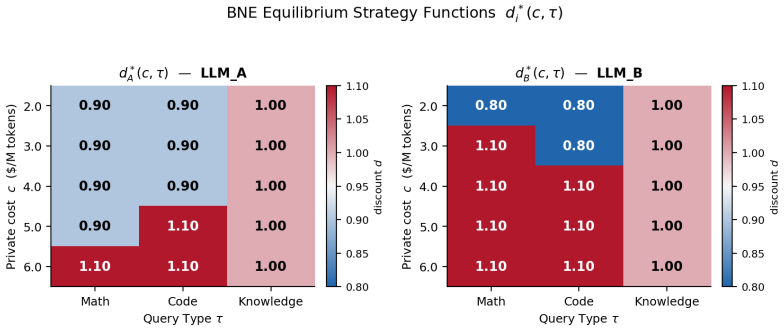
(a) Strategies $d_A^*(c)$ and $d_B^*(c)$.

(b) Three-party utility comparison.



(c) Mode distribution before and after BNE.

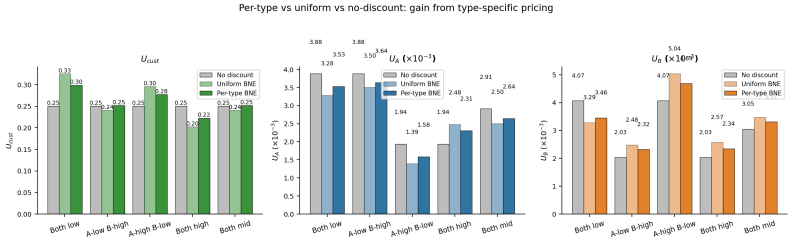
Figure 5: 2-LLM uniform-discount game. (a) Provider A holds at $d = 1.1$; provider B's bid is weakly increasing in its cost. (b) Despite lower selection probability, A retains a higher per-query margin. (c) Mode selection shifts under BNE discounts, with more queries routed through B-favoring modes.



(a) Per-type strategy heatmap $\sigma_i^{\tau *}(c)$.



(b) Mode distribution by query type.



(c) Per-type vs. uniform vs. no-discount.

Figure 6: 2-LLM per-type-discount game. (a) Each provider holds at $d_{\max} = 1.1$ on its strong type and shades on its weak types: A holds Math, while B holds Code and Knowledge. (b) Math queries flow predominantly through A; Code and Knowledge queries flow through B. (c) Per-type discounting increases each party's utility relative to uniform discounting; the asymmetric structure is preserved.

Table 3: Synthetic robustness sweeps (16 configurations) and reported asymmetry scores.

Sweep	Settings	Asymmetry score
S1 bid grid	markdown-only $\{0.7, \ldots, 1.0\}$; baseline $\{0.8, \ldots, 1.1\}$; finer $\{0.7, \ldots, 1.2\}$	0.06 (markdown-only), 0.06 (baseline), -0.04 (finer)
S2 cost prior	uniform; triangular with midpoint mode; truncated lognormal with the same support and mean	0.06 for all three priors
S3 query mix	uniform; 0.70 on Math, Code, or Knowledge, the rest split evenly	0.06 for all four query mixes
S4 λ	$\{15, 25, 35, 50, 100, 200\}$	0.06, 0.02, 0.00, -0.14, -0.14, 0.00, respectively

Table 4: The nine open-weight LLMs calibrated for the supply-side experiments. Trio column lists membership across the nine calibrated trios T1–T9 reported in Section 6.2 (main-paper trios T1–T3) and Appendix F (substitution trio T4 and follow-up cohorts T5–T9). $---$ marks models not in any reported BNE trio.

Model	**Family**	**Params**	**Role**	**Trio**
Qwen2.5-3B-Instruct	Qwen	3.0B	generalist	T3,T6,T8
Llama-3.2-3B-Instruct	Llama	3.2B	generalist	T6
Qwen2.5-Coder-7B-Instruct	Qwen	7.0B	code-tuned	T1,T2,T4,T5,T9
Qwen2.5-Math-7B-Instruct	Qwen	7.0B	math-tuned	T1,T2,T5,T7,T8
StarCoder2-7B (base)	StarCoder	7.0B	code (base)	—
Llama-3.1-8B-Instruct	Llama	8.0B	generalist	T1,T3,T4,T7,T9
Gemma-2-9B-IT	Gemma	9.0B	generalist	T5,T6,T7,T9
Mistral-Small-24B-Instruct	Mistral	24.0B	generalist	T4
Qwen2.5-32B-Instruct	Qwen	32.0B	generalist	T2,T3,T8

E Calibration details

This appendix reports the complete output of the calibration pipeline used to construct the calibrated profiles in Section 6.2. We calibrated nine open-weight models drawn from five families and a 3B–32B parameter range, each evaluated on five benchmarks (GSM8K, MATH, HUMANEVAL, MBPP, MMLU) at three difficulty bins (easy / medium / hard), with three item-sampling seeds per cell. The accuracy and summary-ability tables report three-seed means, while per-seed values are retained in the calibration file. The BNE solver reads the `accuracy` field of the aggregated calibration file, which is inherited from seed 3; the three-seed means are stored separately in `accuracy_by_bench`. The reported calibrated-game results therefore use the seed-3 accuracy profile.

The calibration pipeline produced four artifacts that the BNE solver consumes: per-task base abilities (Table 5), measured chain-boost values (Table 6), blended per-token costs (the cost column in Table 5), and the underlying per-(model, benchmark, difficulty) accuracies (Table 7). We report all four for transparency.

E.1 Cohort and roles

The cohort spans five families: Qwen2.5 [Qwen Team, 2024], Llama-3.x [Llama Team, AI @ Meta, 2024], Gemma-2 [Gemma Team, 2024], Mistral Small [Mistral AI Team, 2024], and StarCoder2 [Lozhkov et al., 2024]. All are used at released weights. The relevant Mistral release announcement is *Mistral Small 3*, dated January 30, 2025, at `https://mistral.ai/news/mistral-small-3/`.

E.2 Benchmark suite

We evaluate each model on five widely-used benchmarks chosen to cover the three task categories used in our theoretical model: mathematical reasoning, code synthesis, and broad knowledge. Each benchmark provides ground-truth answers and a deterministic grader.

- **GSM8K** [Cobbe et al., 2021]: 7,473 grade-school math word problems with numerical answers. Grading via exact-match on the extracted final number.
- **MATH** [Hendrycks et al., 2021a]: ∼5,000 competition mathematics problems across seven subject configurations (`algebra`, `counting_and_probability`, `geometry`, `intermediate_algebra`, `number_theory`, `prealgebra`, `precalculus`). Grading via symbolic equivalence on the boxed final answer. Difficulty levels 1–5 are reported in the dataset; [2]
- **HumanEval** [Chen et al., 2021]: 164 hand-written Python programming problems with function signatures and test suites. Grading via test-suite execution (PASS@1).
- **MBPP** [Austin et al., 2021]: 974 mostly basic Python programming problems with hidden test cases. Grading via test-suite execution.
- **MMLU** [Hendrycks et al., 2021b]: 14,042 multiple-choice questions across 57 subjects covering STEM, humanities, social sciences, and other fields. Grading via exact-match on the chosen answer letter (A–D).

We map benchmarks to the theoretical task categories as follows: GSM8K and MATH constitute the math category; HumanEval and MBPP constitute the code category; MMLU constitutes the knowledge category.

E.3 Inference settings

All models are served via vLLM [Kwon et al., 2023] at their released BF16 precision; no quantization, distillation, or fine-tuning is applied. StarCoder2-7B is the only base model in the cohort and is served with a minimal pass-through chat template (concatenation of message content without role markers); all other models use their published chat templates.

E.4 Decoupled grading

For each (model, problem) pair we extract a candidate answer from the model's response and compare against ground truth. Extraction patterns are benchmark-specific (boxed answer for MATH, the number after "####" (else the last number) for GSM8K, code block for HumanEval/MBPP, single letter for MMLU); the full grading code is released alongside the paper. Calibration and grading are decoupled: model responses are persisted before grading so that grader changes do not require re-inference.

E.5 Difficulty stratification

Each benchmark is stratified using an item-level difficulty proxy. MATH uses the dataset-provided level, with levels $\{1, 2\}$ assigned to easy, level 3 to medium, and levels $\{4, 5\}$ to hard. GSM8K uses the number of calculation annotations in the reference solution, with 1–3, 4–6, and at least 7 annotations defining the three bins. HumanEval uses the number of lines in the canonical solution, with 0–5, 6–12, and at least 13 lines. MBPP uses the number of provided test cases, with 1–2, 3–4, and at least 5 tests. Every MBPP item in the split used here has three tests, so only its medium bin is populated. MMLU uses fixed subject lists for the easy, medium, and hard groups. Within each populated bin, $N = 50$ items are drawn for each sampling seed.

E.6 Two notes on data quality before reading the tables

First, certain model–benchmark pairs produce anomalous accuracy patterns. StarCoder2-7B is a base code model without instruction-tuning; its measured HUMANEVAL accuracy is 0.39 in the easy bin, while its near-zero GSM8K and MATH scores are treated as anomalous calibration outcomes rather than direct evidence of intrinsic mathematical incapability. Mistral-Small-24B returns zero on both HUMANEVAL and MBPP and 0.26 on MMLU; it appears only in T4, which is used as a descriptive substitution-sensitivity trace. Gemma-2-9B-IT returns zero on HUMANEVAL in all three bins, so

[2] We use the `EleutherAI/hendrycks_math` mirror, as the original `hendrycks/competition_math` repository was removed from HuggingFace by DMCA takedown in early 2025. The mirror is bit-identical, MIT-licensed, and parquet-formatted.

Table 5: Per-task base abilities consumed by the BNE solver. Each entry is the mean of (easy, medium) accuracy over the benchmarks mapping to that task type (GSM8K, MATH for Math; HUMANEVAL, MBPP for Code; MMLU for Knowledge), including the zero-valued cells of Appendix E.6, averaged over three calibration seeds. The cost column reports the blended dollar cost per million tokens, computed from each model's published input/output pricing weighted by the input/output token mix observed during calibration. Bold marks the per-row leader.

	Math	Code	Knowledge	cost ($/M tok)	avg tokens
Qwen2.5-3B	0.63	0.40	**0.64**	0.065	346
Llama-3.2-3B	0.51	0.32	**0.61**	0.062	296
Qwen2.5-Coder-7B	0.58	0.50	**0.61**	0.149	340
Qwen2.5-Math-7B	**0.68**	0.15	0.56	0.163	624
StarCoder2-7B	0.05	**0.22**	0.19	0.147	272
Llama-3.1-8B	0.61	0.34	**0.68**	0.151	302
Gemma-2-9B-IT	0.59	0.10	**0.68**	0.160	284
Mistral-Small-24B	**0.66**	0.00	0.26	0.509	348
Qwen2.5-32B	0.70	0.42	**0.81**	0.797	322

Table 6: Chain-boost values measured via the strawman protocol: for each problem in each benchmark, we construct a deliberately-wrong predecessor answer and measure the model's correction rate when prompted to review and re-answer. Used by the simulator's ACCURACYENGINE to compute the chain-mode probability $P(\text{Chain}_{ij} \text{ correct}) = P_i + (1 - P_i) \cdot \text{boost}_j$. The reported values are estimated from up to 30 items per benchmark sampled across the difficulty bins.

	Math	Code	Knowledge
Qwen2.5-3B	0.71	0.33	0.63
Llama-3.2-3B	0.54	0.35	0.57
Qwen2.5-Coder-7B	0.75	0.71	0.63
Qwen2.5-Math-7B	0.85	0.27	0.47
StarCoder2-7B	0.08	0.17	0.10
Llama-3.1-8B	0.70	0.35	0.43
Gemma-2-9B-IT	0.69	0.18	0.70
Mistral-Small-24B	0.77	0.00	0.10
Qwen2.5-32B	0.78	0.55	0.80

its Code ability is based on MBPP alone; these entries are likewise treated as anomalous calibration outcomes. Gemma-2-9B enters T5–T7 and T9, but not the headline T1–T3 cohorts.

Second, three cells are zero for every model: MBPP-easy and MBPP-hard contain no items, while MATH-medium is treated as an anomalous calibration outcome. The two zero-valued cells that enter the base abilities (MBPP-easy and MATH-medium) are retained as 0.0 in the reported calibration; for example, Qwen2.5-Math-7B's Math ability is 0.68 with the MATH-medium entry and would be 0.90 without it. Because one zero term enters every model's four-term average, each Math and Code ability equals $3/4$ of the mean of its other three terms. Rankings within each task are unchanged, while the absolute Math and Code levels are lowered relative to Knowledge. We therefore interpret these absolute ability levels cautiously and do not treat the zero-valued cells as evidence of intrinsic model incapability. The main conclusions do not rely on the anomalous StarCoder2, Mistral, or Gemma-2 code entries.

E.7 Per-task base abilities (BNE-solver input)

E.8 Chain-boost values (BNE-solver input)

E.9 Full per-(model, benchmark, difficulty) accuracy

E.10 Implementation notes (calibration pipeline)

The calibration pipeline runs each (model, benchmark) combination at `n_per_difficulty=50` samples (typical) on a single A100 GPU via `vLLM` as the inference backend. Total compute for

Table 7: Full per-(model, benchmark, difficulty) accuracy from the calibration. Each value is the mean across three random seeds (1, 2, 3). **e/m/h** denote easy/medium/hard difficulty. Cells reported as 0.00 are either a uniformly zero outcome under our pipeline or an unpopulated (benchmark, difficulty) partition; MBPP-easy, MBPP-hard, and MATH-medium read zero for every model; MBPP-easy and MATH-medium enter the base abilities as 0.0 (Appendix E.6).

	GSM8K			MATH			HUMANEVAL			MBPP			MMLU		
Model	**e**	**m**	**h**	**e**	**m**	**h**	**e**	**m**	**h**	**e**	**m**	**h**	**e**	**m**	**h**
Qwen2.5-3B	0.89	0.78	0.62	0.83	0.00	0.45	0.63	0.56	0.38	0.00	0.41	0.00	0.67	0.62	0.34
Llama-3.2-3B	0.86	0.68	0.60	0.49	0.00	0.13	0.56	0.42	0.39	0.00	0.28	0.00	0.68	0.54	0.39
Qwen2.5-Coder-7B	0.83	0.69	0.54	0.79	0.00	0.41	0.59	0.71	0.57	0.00	0.69	0.00	0.63	0.59	0.37
Qwen2.5-Math-7B	0.91	0.92	0.89	0.87	0.00	0.62	0.22	0.20	0.06	0.00	0.17	0.00	0.59	0.53	0.39
StarCoder2-7B	0.11	0.06	0.01	0.03	0.00	0.01	0.39	0.37	0.15	0.00	0.11	0.00	0.20	0.19	0.19
Llama-3.1-8B	0.91	0.82	0.72	0.70	0.00	0.29	0.56	0.49	0.46	0.00	0.31	0.00	0.71	0.64	0.46
Gemma-2-9B-IT	0.89	0.80	0.70	0.67	0.00	0.22	0.00	0.00	0.00	0.00	0.40	0.00	0.76	0.60	0.46
Mistral-Small-24B	0.93	0.88	0.84	0.81	0.00	0.47	0.00	0.00	0.00	0.00	0.00	0.00	0.26	0.25	0.19
Qwen2.5-32B	0.97	0.94	0.90	0.89	0.00	0.65	0.40	0.57	0.56	0.00	0.69	0.00	0.87	0.75	0.77

the 9-model calibration was approximately 35 GPU-hours summed over three seeds; per-model time ranged from $\sim$1 hour for 3B-class models to $\sim$6.5 hours for the 32B model. Generation used temperature 0.7 with no per-request generation seed; the three seeds select which items are drawn. The full configuration, including chat templates, grading scripts, and per-bench prompts, is included with the code release.

E.11 Cost data provenance

The price spread cited in Section 1 is based on representative public provider quotes accessed on April 30, 2026. For Llama-3.3-70B, the listed input-token prices are \$0.23 per million tokens at DeepInfra, \$0.88 at Together AI, and \$0.90 at Fireworks, corresponding to a $3.9\times$ range; the corresponding output-token prices are \$0.40, \$0.88, and \$0.90, corresponding to a $2.25\times$ range (`https://www.aipricing.guru/meta-pricing/`). Thus, the 3–4$\times$ same-model comparison in the Introduction refers to quoted input-token rates. Across model classes, public output-token prices range from small open-weight endpoints near \$0.10–\$0.20 per million tokens to GPT-4o at \$10 and Claude 3.5 Sonnet at \$15 per million tokens, spanning roughly two orders of magnitude; see the OpenAI and Anthropic public pricing pages (`https://platform.openai.com/docs/pricing` and `https://www.anthropic.com/pricing`).

The per-token cost parameters reported in Table 5 are based on public pricing tables from three commercial inference providers—Together AI, DeepInfra, and Fireworks—accessed in late April 2026. When multiple provider listings were available for a model, we used the median quoted rate; when an exact listing was unavailable, we used a same-family model of comparable size as a pricing proxy. The resulting input- and output-token rates used in the experiments are recorded explicitly for every model in the released calibration file. We combine these rates using the input/output token mix observed during calibration:

$$\mathrm{cost}_i = \frac{p_i^{\mathrm{in}} t_i^{\mathrm{in}} + p_i^{\mathrm{out}} t_i^{\mathrm{out}}}{t_i^{\mathrm{in}} + t_i^{\mathrm{out}}},$$

where p_i^{in} and p_i^{out} are prices per million tokens and t_i^{in} and t_i^{out} are the average input and output tokens per query. The BNE solver uses a finite private-cost grid bracketing this blended rate rather than treating the blended rate itself as the supplier's private cost; see Section 5

F Follow-up calibrated cohorts (T4–T9)

This appendix defines the follow-up cohorts whose IBR residuals are reported in Section 6.2 and Appendix G. Under IBR they produce solver traces only: a cycle is not a certificate of a mixed BNE, of the absence of a pure BNE, or of equilibrium multiplicity. No welfare claim in the paper rests on these cohorts.

Table 8: Cohort overview. "IBR outcome" reports diagnostic IBR behavior on the learned four-point games and is not an equilibrium certificate. "Reported" identifies the numerical object used in this paper: FP-certified profiles for T1–T3, deviation residuals for T5–T8, and descriptive solver traces for T4 and T9.

Trio	Cohort role	IBR outcome	Reported
Main-paper trios			
T1	Specialist: Qwen-Math, Qwen-Coder, Llama-8B	Cycle, period 10	FP-certified, Table 11
T2	Dominance: Qwen-Math, Qwen-Coder, Qwen-32B	Cycle, period 2	FP-certified, Table 11
T3	Size-ladder: Qwen-3B, Llama-8B, Qwen-32B	Cycle, period 3	FP-certified, Table 11
Substitution sensitivity			
T4	Mistral-24B replaces Qwen-Math in T1	Recorded trace, not audited	Descriptive trace only
Follow-up probes			
T5	Specialist + Gemma robustness	Cycle, period 2 (audited)	Residual only
T6	Capability-symmetric small generalists	Cycle, period 2 (audited)	Residual only
T7	Knowledge tie probe	Cycle, period 2 (audited)	Residual only
T8	Code tie probe	Cycle, period 4 (audited)	Residual only
T9	Math + Knowledge double-tie probe	Near-stable trace, not audited	Descriptive trace only

F.1 Calibrated trios used in this appendix

The follow-up runs use trios constructed from the same nine open-weight LLMs as the main calibration. Table 8 summarizes all calibrated trios. The five trios beyond T1, T2, T3 introduced in this appendix are:

- **T5** (specialist + Gemma robustness): Qwen-Math-7B, Qwen-Coder-7B, Gemma-2-9B at $\lambda = 35$.
- **T6** (capability-symmetric small generalists): Llama-3.2-3B, Qwen-3B, and Gemma-2-9B at $\lambda = 35$. Qwen-3B and Gemma-2-9B have identical seed-3 Knowledge ability, 0.69.
- **T7** (Knowledge tie probe): Qwen-Math-7B, Gemma-2-9B, and Llama-3.1-8B at $\lambda = 35$. Gemma-2-9B and Llama-3.1-8B have seed-3 Knowledge abilities of 0.69 and 0.68, respectively, a gap of 0.01.
- **T8** (Code tie probe): Qwen-Math-7B, Qwen-3B, and Qwen-32B at $\lambda = 35$. Qwen-3B and Qwen-32B have seed-3 Code abilities of 0.380 and 0.405, respectively, a gap of 0.025.
- **T9** (Math-and-Knowledge double-tie probe): Gemma-2-9B, Qwen-Coder-7B, Llama-3.1-8B at $\lambda = 35$.

Each recorded trace runs IBR for 30 rounds on the four-point grid $D_d = \{0.8, 0.9, 1.0, 1.1\}$ at $\lambda = 35$. Two runs are available for each of T5–T8, and one run is available for T4 and T9. The recorded profiles represent individual phases of a cycle, or a near-stable trace in the case of T9, and are used only as solver diagnostics rather than as certified equilibrium profiles. Different traces, including the opposite-sign snapshot welfare values observed in the two T6 runs, do not establish distinct equilibria. Similarly, visually similar traces do not establish equilibrium uniqueness or a common attraction basin. T4 is retained only as a substitution-sensitivity trace and is not used for the paper's reported welfare results.

G Additional finite-game audits

This appendix reports finite-game audits for the learned routing caches. We distinguish descriptive IBR traces, cycle-average diagnostics, and FP-certified profiles. All deviation residuals are computed relative to the specified finite payoff tables and therefore bound neither discretization error nor routing-estimation error. The audited IBR calculations use sequential supplier–cost updates.

G.1 Own-price allocation monotonicity

Each adjacent-price comparison fixes all other bids and increases one supplier's bid by one grid step. At the mode-cell level, 8,569 of 18,144 learned-cache cells increase (47.23%), with a maximum probability increase of 0.377. At the supplier-comparison level, 2,737 of 3,024 comparisons have at least one increasing mode involving that supplier (90.51%). Total participation and expected billable tokens increase in 33.66% and 33.27% of the comparisons, respectively, with a maximum token increase of 302.75. We do not attribute these aggregate exceptions separately to router-training variation or policy structure. These percentages summarize different events and should not be interpreted as sharing the same denominator

The analytical control has 30 increasing mode cells out of 18,144 ($.17\%$), all reallocations, and zero aggregate participation/token violations in 3,024 comparisons. On the finer grid it again has zero aggregate violations in 7,938 comparisons. The learned cache therefore does not uniformly satisfy Assumption 2; the analytical grid tests support that condition only on tested comparisons.

G.2 Mixed-profile deviation checks

For a cost-conditional mixed strategy profile μ, write $V_i(a_i, \mu_{-i}; c_i)$ for Equation (3) evaluated by summing over opponent costs and independently drawn opponent actions. A conditional finite-grid deviation gain is

$$g_i(c_i; \mu) := \max_{a_i \in D_d^K} V_i(a_i, \mu_{-i}; c_i) - \mathbb{E}_{a_i \sim \mu_i(\cdot|c_i)} V_i(a_i, \mu_{-i}; c_i).$$

The interim criterion is $\epsilon_{\text{int}}(\mu) := \max_{i,c_i} g_i(c_i; \mu)$. An ex ante criterion is $\epsilon_{\text{ex}}(\mu) := \max_i \sum_{c_i} F(c_i) g_i(c_i; \mu)$; it need not equal the interim criterion. Exact BNE on positive-probability grid types has zero gain at every type. All residuals reported in this paper use the interim criterion ϵ_{int}, in supplier-utility units per query. Numerical residuals must retain their aggregation convention and payoff normalization when comparing runs.

A cycle supplies a marginal action distribution for each supplier conditional on its cost. Independent mixed play uses the product of supplier distributions, not a shared random cycle phase; the latter can preserve cross-supplier correlation and requires a different equilibrium concept. Within-supplier task components can be represented by their marginals when computing the additive payoff, since no term couples different query types. Cycle detection and time averaging alone are not equilibrium tests.

Table 9: Cycle-average finite-grid deviation residuals for the audited IBR runs. These values characterize the specified runs and finite payoff tables, not every recorded solver snapshot.

Cohort	Audited period	Reported residual
T5	2	9.77×10^{-5}
T6	2	9.92×10^{-5}
T7	2	1.10×10^{-4}
T8	4	9.68×10^{-5}

Across six independently estimated caches for T1 and T7 (one cache for each cohort at seeds 7, 11, and 42) IBR cycle-average deviation gains range from 6.31×10^{-5} to 2.48×10^{-4}, whereas FP on the same finite payoff tables yields residuals from 4.47×10^{-7} to 3.53×10^{-5}. This same-game comparison motivates the use of FP for the reported profiles. It does not establish convergence of IBR or imply that an IBR cycle is an exact mixed equilibrium.

G.3 Grid, seed, and private-cost checks

The seven-point set is $\{0.8, 0.85, 0.9, 0.95, 1.0, 1.05, 1.1\}$. The analytical control reaches pure fixed points on both grids and preserves the qualitative price ordering. Across six independently estimated caches for T1 and T7 (one cache for each cohort at seeds 7, 11, and 42) the mean own-price token responses are negative, but the rates of positive token responses remain between 29.40% and 34.49%. A negative mean response does not establish pointwise monotonicity.

Private-cost monotonicity is a separate test. Across these six caches, adjacent cost-type comparisons pass all 216 of 216 mean-bid tests and all 216 first-order stochastic-dominance tests. Four changes of

Table 10: Learned T1–T3 grid control. The period columns report diagnostic IBR cycles; the monetary-welfare change compares the FP-certified profile on the seven-point grid with the corresponding profile on the four-point grid.

Trio	Four-point period	Seven-point period	W_{mon} change
T1	10	30	-1.19%
T2	2	4	$-.34\%$
T3	3	4	$-.57\%$

the most frequent action occur only at modal ties. This supports cost-policy ordering on the tested grid, not own-price allocation monotonicity or an unrestricted continuous-cost theorem.

G.4 Absolute welfare and profile comparison

Table 11: Absolute welfare per query for the reported T1–T3 profiles. "Mix" denotes the FP-certified profile. Entries are rounded, so component sums may differ by one unit in the last decimal. Residuals use the interim finite-grid criterion defined in Appendix G.2.

Trio	Profile	Residual	CS_{mon}	U_{cust}	PS_A	PS_B	PS_C	W_{mon}
T1	No discount	–	.588878	.163829	.002643	.003012	.001846	.596379
	Mix	7.44×10^{-6}	.603248	.151522	.002823	.003306	.002012	.611389
	Planner	–	.776912	.157929	.001913	.004191	.004865	.787881
T2	No discount	–	.640743	.238522	.001673	.002073	.003352	.647841
	Mix	4.67×10^{-7}	.662589	.215091	.002070	.002624	.003573	.670855
	Planner	–	.845115	.226315	.001920	.004080	.004920	.856035
T3	No discount	–	.638612	.200033	.002072	.001541	.004127	.646352
	Mix	3.08×10^{-7}	.642173	.195297	.001779	.001472	.004747	.650172
	Planner	–	.788637	.177226	.003479	.001507	.005904	.799527

For T3, the FP-certified profile has $U_{\mathrm{cust}} = 0.195297$ and an interim deviation residual of 3.08×10^{-7}. The diagnostic IBR cycle-average profile instead has $U_{\mathrm{cust}} = 0.193561$ and a residual of 4.16×10^{-5}. Because these values are evaluated under different strategy profiles, the paper uses only the FP-certified value as the reported equilibrium result.

The symmetric-capability experiment is a separate two-supplier counterfactual, not the main three-supplier calibration: the symmetric BNE gives both suppliers the mean capability profile, and the planner chooses a mode maximizing $r\alpha(Q,S) - \sum_i c_i \mathrm{tok}_i(S)$, as in Section 7. Its reported common-unit accounting is

	CS_{mon}	$\sum_i PS_i$	W_{mon}
No discount	.5801	.0060	.5861
Asymmetric BNE	.5964	.0067	.6031
Symmetric-capability BNE	.5833	.0066	.5900
Planner	.7272	.0149	.7420

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and §1 state the four contributions: SBFP-RA formalization, asymmetric BNE characterization, calibrated empirics across nine open-weight LLMs, and capability-symmetric price-differentiation result; each is delivered in §4–§7.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: §8 discusses the one-shot i.i.d. private cost assumption, $n \in \{2, 3\}$ scope, discount-grid quantization, the single learned-policy router instantiation, and the cycling regime in which cohorts with tied capability on a task type produce period-2 IBR cycles requiring time-averaging to recover the mixed BNE of Theorem 1 (full discussion in Appendix F).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Assumptions are stated inline (Assumption 2 and conditions C1–C4 in §3); complete proofs of Theorem 1, Theorem 2, Lemma 1, Corollaries 1–2, and Propositions 1–3 are in Appendix A, with proof sketches in the body.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: §5 and Appendix B give the cache + IBR algorithm with all hyperparameters; Appendix E gives the calibration pipeline (benchmarks, models, difficulty bins, seeds, decoding settings) sufficient for end-to-end replication. Nine calibrated trios are reported (T1–T3 in the body §6.2; T4–T9 in Appendix F); each trio is fully specified by its cohort composition, λ value, discount grid, and cost grid, all documented in the corresponding section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Complete code, cached customer policies, and configuration files needed to reproduce every figure and table are included as an anonymized supplementary zip with this submission (with READMEs and run instructions); the five benchmarks used are public, and all open-weight models are downloadable from HuggingFace.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Hyperparameters (Adam, lr 5×10^{-3}, 25 epochs, 32-query mini-batches), inference settings (temperature 0.0 greedy + 3 seeds), discount grid, and difficulty-bin partitioning are in Appendix B and Appendix E. IBR terminates at a pure-strategy fixed point for converged trios (T1–T4) and runs to a 30-round budget with a round-30 snapshot for cycling trios (T5–T8); see Appendix F for the initialization caveat.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Calibration uses 3 random seeds per (model, benchmark, difficulty) cell with per-seed means reported; the asymmetric BNE pattern is reported across four robustness sweeps (S1–S4) covering 100+ configurations in §6.4 and Appendix D. BNE strategies are deterministic for converged trios (T1–T4); no error bars apply. For cycling trios (T5–T8) we report two IBR realizations each in Appendix F; the body cites only converged-trio welfare numbers, and quantitative cycling-trio welfare (requiring more IBR realizations and time-averaging) is deferred to follow-up work.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Single A100 GPU; cache construction $\sim$30 min for the two-provider game and 2–3 hours for the three-provider game (Appendix B); IBR converges in 3–6 rounds at well under a second per round for converged trios, with cycling trios run to a 30-round budget. All nine trios reuse the same calibration.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research conforms to the NeurIPS Code of Ethics: it uses only publicly released benchmarks and open-weight models, involves no human subjects, and analyzes a market-design question with no foreseeable misuse pathway.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: §7 discusses welfare implications: auction-based routing redistributes surplus from consumers to specialist providers, and the paper proposes mitigation (reserve discount factors, capability-blind tie-breaking) that preserve consumer welfare.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The supplementary zip releases only analysis code, pre-computed equilibrium caches, and calibration manifests: no pretrained models, generative artifacts, or scraped datasets that could pose misuse risk.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All benchmarks (GSM8K, MATH, HumanEval, MBPP, MMLU) and open-weight models (Qwen2.5, Llama-3.x, Gemma-2, Mistral-Small, StarCoder2) are cited (Appendix E) and used under their published licenses (MIT, Apache-2.0, Llama Community License, Gemma Terms of Use, BigCode OpenRAIL-M); vLLM [Kwon et al., 2023] is Apache-2.0.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The supplementary zip contains the analysis code, equilibrium caches, and calibration manifests with structured READMEs, per-experiment configuration files, and the exact commands needed to reproduce each figure and table.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper involves no crowdsourcing or human-subjects research; all empirics are model evaluations on standard public benchmarks.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: No human-subjects research; no IRB review required.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are the studied artifact (the priced commodity in the auction model), not a component of the methodology. LLM use for incidental editing assistance is outside the scope of this declaration.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.