CONTEXT-DEPENDENT AGENT EVALUATION WITH ORTHOGONAL EQUILIBRIUM LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Many applications require evaluating agents under contextual information (e.g., a prompt, task, or user group). We study how to perform such context-dependent agent evaluation from offline feedback. Existing score-based models for this purpose (e.g., Bradley–Terry) impose a transitive preference ordering, which fails to reflect collective preferences when human judgments are heterogeneous. Inspired by social choice theory, we frame evaluation as a contextual game between two players, each selecting a distribution over agents as the strategy to receive greater collective preference than the other. Then, the support of the Nash equilibrium defines a context-specific set of winners. However, learning context-specific equilibria from offline logs is difficult because each context reveals human feedback on only a subset of agents, and a naïve plug-in estimator can therefore be biased. To address these challenges, we propose NASHEVAL, a general framework for robust contextual equilibrium learning. NASHEVAL first constructs debiased estimates of the contextual payoff matrix that characterizes the game. NASHEVAL then learns the context-to-equilibrium mapping with a tailored orthogonal loss, which avoids the need to solve a separate game for each context. We show theoretically that errors in estimating the nuisance functions underlying the payoff matrix affect the risk of the learned equilibrium (i.e., exploitability) only through higher-order terms. Across various experiments, NASHEVAL improves the robustness of equilibrium learning and consistently identifies the set of top-performing agents across contexts.

1 INTRODUCTION

Many applications require agents to be evaluated together with contextual information, such as the prompt, task, language, domain, or user group (e.g., Xu et al., 2025; Frick et al., 2025). For example, in model routing, the goal is not to identify a single globally best language model, but to select an appropriate model for a particular prompt (Ong et al., 2025). More generally, agents may have different strengths across tasks and domains (Frick et al., 2025), and human preferences can also vary across users and settings (Haghtalab et al., 2026). Here, we study **context-dependent agent evaluation** from selectively observed relative feedback (e.g., pairwise preferences or partial rankings) collected from a population of voters with heterogeneous preferences to identify, for each context, a set of agents that best reflects collective preference of the population (see Figure 1).

In this paper, we propose NASHEVAL, a game-theoretic framework for contextual agent evaluation. Our framework is flexible and accommodates any form of relative feedback that can be represented as ordered partitions (e.g., pairwise preferences, partial rankings, and winner-of-subset feedback). For a given context, we formulate the evaluation as a regularized two-player zero-sum game, in which each player selects a distribution over agents and seeks to outperform the other according to the collective preference. The support of the corresponding Nash equilibrium then defines a context-specific set of winners.

To learn this equilibrium from selectively observed feedback, our framework proceeds in two stages. First, we construct a novel *debiased payoff matrix (DPM)* that is robust to nuisance misspecification. Second, we use the estimated payoff to introduce a tailored orthogonal loss to directly learn the context-to-equilibrium mapping. Thereby, we amortize equilibrium computation rather than solving a separate game for every context. We show theoretically that the robustness guarantees of the DPM

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

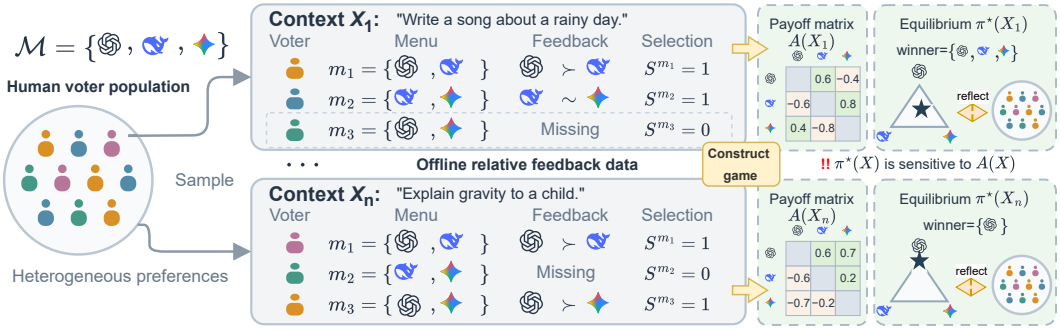


Figure 1: **Agent evaluation is subject to two challenges:** (1) preferences sampled from a large population of voters are heterogeneous and, thus, often intransitive; and (2) the capabilities of agents vary across contexts. ⇒ We propose a novel framework called NashEval for this task by framing evaluation as solving a contextual, two-player zero-sum game, constructed from the preferences.

transfer to the learned equilibrium; i.e., errors in estimating the nuisance functions underlying the payoff matrix affect the risk of the learned equilibrium (i.e., exploitability) only through higher-order terms and, under suitable conditions, the learned equilibrium can therefore achieve a quasi-oracle convergence rate. As a by-product, we derive a general equilibrium learning framework that is robust to nuisance misspecifications.

Our main **contributions**[1] are: **(i)** We propose NashEval, a novel game-theoretic framework for learning context-dependent agent evaluators from logged data with relative feedback. **(ii)** We derive strong theoretical guarantees for our equilibrium learner in the form of Neyman-orthogonality, bounds on the average exploitability, and a quasi-oracle convergence rate. **(iii)** We demonstrate the effectiveness of our NashEval framework on both simulated and real-world datasets.

2 Related work[2]

Social choice theory. This literature studies collective choice under potentially intransitive preferences (Fishburn, 1984; Brandt, 2017). Concepts like Condorcet (de Condorcet, 1785) and Copeland winners (Copeland, 1951) may fail to identify a unique or robust winner (de Condorcet, 1785; Laslier, 1997; Tideman, 1987). We build on this literature by developing a game-theoretic solution concept for context-dependent agent evaluation.

Agent evaluations. There are several approaches for related but *different* settings (see Table 1). Traditionally, evaluation methods aggregate feedback across contexts to produce a single *global* leaderboard (e.g., Chiang et al., 2024; Frauen et al., 2026), and, more recently, there are extensions to incorporate contextual information (e.g., Ong et al., 2025; Frick et al., 2025; Chiang et al., 2025). However, these approaches rely on score-based models (e.g., Bradley–Terry; Bradley & Terry, 1952), and therefore assume *transitive* preferences, meaning that if $A \succ B$ and $B \succ C$, then $A \succ C$. Importantly, this assumption can be restrictive when feedback is collected from a heterogeneous population, since aggregating individual preferences can result in intransitive collective preferences (de Condorcet, 1785; Kramer, 1973), particularly for subjective, open-ended tasks (Haghtalab et al., 2026). As a result, models that assume such transitivity can be misspecified and produce biased (or even conflicting) evaluations that do not reflect the underlying heterogeneous preferences. To the best of our knowledge, approaches for context-dependent agent evaluation with intransitive collective preferences are missing.

Game-theoretic approaches. A common approach is to frame agent evaluation as a two-player zero-sum game between two dueling evaluation policies that compete head-to-head for voter preference (Yue et al., 2012). The game is characterized by an antisymmetric *payoff matrix (PM)* constructed from the relative feedback, and its Nash equilibrium is a distribution over agents whose support is interpreted as the set of winners. Because no evaluation policy strictly beats this equilibrium, it is the best evaluation achievable under intransitive feedback and typically has favorable

[1] Code available at https://anonymous.4open.science/r/Contextual-Agent-Evaluator-1A58

[2] We provide an extended discussion of related work in Appendix A.

properties (e.g., Condorcet consistency, meaning that when a Condorcet winner exists, the equilibrium places all probability on that agent). In the dueling bandit literature, such equilibrium concepts are referred to as the *von Neumann winner* (vNw) (Dudík et al., 2015) and, in social choice theory, as the *maximal lottery* (Kreweras, 1965; Fishburn, 1984). Recently, Lanctot et al. (2023) and Khalaf et al. (2026) study maximal lotteries for *global* agent evaluation. In contrast, approaches that learn the context-specific equilibrium for agent evaluation are still missing.

Why is context-dependent agent evaluation challenging? A major difficulty of context-dependent evaluation is that, for a given context, a voter is shown only a small subset of agents, and thus the feedback is often *partially observed*. This means that the contextual PM must be *estimated* using nuisance functions[3]. When these nuisance functions are misspecified, the estimated PM can be biased, which leads to a suboptimal approximation of the equilibrium (see Proposition 1). Hence, a key challenge is to learn the contextual equilibrium in a way that is *robust to errors in the nuisance functions*.

Table 1: Key works on agent evaluation from relative feedback.

Preference \ Evaluator	Global	Contextual
Transitive	(Chiang et al., 2024; Frauen et al., 2026)	(Ong et al., 2025; Frick et al., 2025)
Intransitive	(Lanctot et al., 2023; Khalaf et al., 2026)	**NASHEVAL (ours)**

3 PRELIMINARIES

When preferences are aggregated across a population, they may become intransitive even if each individual's preferences are transitive (see Appendix E.1 for an example). In such cases, a single deterministic winner may not exist, which motivates game-theoretic solution concepts (i.e., lotteries over agents) that can represent collective preferences (Fishburn, 1984; Brandt, 2017).

Preference game. Let $A \in [-1,1]^{K\times K}$ denote the collective preference matrix with entries $A_{jk} = P(j \succ k) - P(k \succ j)$, which satisfies $A = -A^\top$. We interpret A as the payoff matrix of a symmetric two-player zero-sum game, where the two players play strategies $\pi, q \in \Delta_K$. Player π receives payoff $\mathcal{P}(\pi, q) = \pi^\top A q$, and player q receives $-\mathcal{P}(\pi, q)$. The corresponding *Nash equilibrium* is given by the following minimax formulation (von Neumann, 1928):

$$(\pi^\star, q^\star) = (\arg\max_{\pi} \min_{q} \mathcal{P}(\pi, q),\ \arg\min_{q} \max_{\pi} \mathcal{P}(\pi, q)). \tag{1}$$

By symmetry, both players share the same equilibrium strategy, and the solution is called the *von Neumann winner* (Dudík et al., 2015) or the *maximal lottery* (Kreweras, 1965; Fishburn, 1984).

The equilibrium is well suited to intransitive preferences because it depends only on pairwise comparisons and always exists by the minimax theorem. Additional properties, including Condorcet consistency and related consistency axioms, are discussed in Appendix E.

4 CONTEXT-DEPENDENT AGENT EVALUATION FROM SELECTIVE FEEDBACK

We now formulate context-dependent agent evaluation from offline, selectively observed relative feedback as a contextual game and define its equilibrium.

Setting: • *Agents*: Let $\mathcal{M} = \{1, \dots, K\}$ denote the candidate agents, which perform tasks specified by the context $X \in \mathcal{X}$ (e.g., a prompt, task, language, domain, or user group). A *menu* $m \subseteq \mathcal{M}$ is a subset of agents being compared. Let $\mathcal{Q}$ denote the collection of all possible menus. For each context, $S^m \in \{0, 1\}$ indicates whether menu m was selected to generate the feedback, Y_m.

• *Relative feedback*: The feedback Y_m is *relative*, represented by an *ordered partition* of the selected menu, i.e.,

$$Y_m = (B_1 \succ \cdots \succ B_r), \qquad B_s \neq \emptyset, \quad B_s \cap B_t = \emptyset \ \text{ for } s \neq t, \quad \bigcup_{s=1}^{r} B_s = m, \tag{2}$$

where r is the number of blocks, and where agents within the same block are tied and earlier blocks are preferred. This covers various forms of relative feedback, including pairwise preferences, partial

[3]The term "nuisance functions" refers to components that facilitate estimation of a statistical target by modeling observable distributions from data but are not themselves of direct interest.

rankings, and winner-of-subset feedback (see Table 5). For example, pairwise feedback corresponds to $|m| = 2$.

• *Logged data*: We write the logged data as $O = \big(X,\ \{S^m\ :\ m \in \mathcal{Q}\},\ \{S^m Y_m\ :\ m \in \mathcal{Q}\}\big)$ and observe i.i.d. samples $\mathcal{D}_n = \{O_i\}_{i=1}^n \sim \mathbb{P}$.

Pairwise outcomes: For each menu m, the feedback Y_m naturally induces a pairwise outcome for every $\{j, k\} \subseteq m$:

$$Z_{jk}^m = \mathbf{1}\{j \succ_{Y_m} k\} - \mathbf{1}\{k \succ_{Y_m} j\} \in \{-1, 0, 1\}, \tag{3}$$

which satisfies $Z_{jk}^m = -Z_{kj}^m$ and where $Z_{jk}^m = 0$ denotes a tie. We are then interested in the *context-dependent mean pairwise outcome*, defined as $g_{jk}^m(x) = \mathbb{E}[Z_{jk}^m \mid X = x]$, which quantifies the average collective preference of agent j over agent k under context x and menu m.

Contextual payoff: We then aggregate context-dependent mean pairwise outcomes across menus into a contextual payoff matrix (PM).

Definition 1 (Contextual payoff matrix)**.** *Let $\mathcal{Q}_{jk} = \{m \in \mathcal{Q} : \{j, k\} \subseteq m\}$ be the set of menus containing the pair $\{j, k\}$. Let $\mathcal{F} = \{f_{jk} : [-1, 1]^{\mathcal{Q}_{jk}} \to [-1, 1] \mid j, k \in \mathcal{M}, j \neq k\}$ denote the set of smooth, odd functions that serve as menu aggregators, with $f_{kj} = f_{jk}$ for every $j \neq k$. Then, the contextual payoff matrix is defined by aggregating the menu-specific mean outcomes $g_{jk}^m(x)$ via the menu aggregators $f_{jk} \in \mathcal{F}$ (with $A_{jj}(x) = 0$):*

$$A(x) = \big(A_{jk}(x)\big)_{j,k \in \mathcal{M}}, \quad \textit{where } A_{jk}(x) = f_{jk}\Big(\big\{g_{jk}^m(x)\ :\ m \in \mathcal{Q}_{jk}\big\}\Big) \quad (j \neq k). \tag{4}$$

Remark 1. *The contextual PM characterizes the expected pairwise preferences under a given context, which reflects the collective preference of the population. Since $g_{kj}^m(x) = -g_{jk}^m(x)$ and f_{jk} is odd, we have $A_{kj}(x) = -A_{jk}(x)$. Thus, $A(x)$ is skew-symmetric.*

Remark 2. *A simple instantiation of the aggregators is the uniformly weighted combination $f_{jk}(\mathbf{g}_{jk}) = \frac{1}{|\mathcal{Q}_{jk}|} \sum_{m \in \mathcal{Q}_{jk}} g_{jk}^m$. In practice, users can customize the aggregator to their needs (see Appendix F for more instantiations).*

Based on the contextual PM, we formulate a regularized contextual game and define its equilibrium.

Definition 2 (Contextual game and its equilibrium)**.** *Let Ω be continuously differentiable on a neighborhood of Δ_K and κ-strongly convex on Δ_K, where $\kappa > 0$. We consider a* regularized two-player zero-sum game*, where the expected payoff between two players with strategies $\pi, q \in \Delta_K$ is given by $F_\Omega(\pi, q; x) = \pi^\top A(x) q - [\Omega(\pi) - \Omega(q)]$. The regularized contextual equilibrium is the minimax solution of the optimization problem:*

$$\pi^\star(x) = \arg\max_{\pi \in \Delta_K} \min_{q \in \Delta_K} F_\Omega(\pi, q; x) := \pi^\top A(x) q - [\Omega(\pi) - \Omega(q)]. \tag{5}$$

Since $F_\Omega(\pi, q; x) = -F_\Omega(q, \pi; x)$, the game is symmetric and zero-sum. Strong convexity implies that the associated self-play operator is strongly monotone, which ensures both uniqueness of the equilibrium and its Lipschitz continuity w.r.t. perturbations of the PM (Dontchev & Rockafellar, 2009; Sokota et al., 2023) (see Appendix C for details).

Further, the regularization term can be used to control the equilibrium according to constraints in practice. For example, one can incorporate deployment considerations such as agent cost or encourage properties such as sparsity. A common choice is to combine a linear term for the deployment cost with an L_2 regularizer; i.e., $\Omega(\pi) = \frac{\rho}{2}\|\pi\|_2^2 + \sum_{j=1}^K c_j \pi_j$, where c_j is the cost of deploying agent j.

Target: We aim to learn the *context-to-equilibrium mapping* $x \mapsto \pi^\star(x)$ defined in Eq. (5). Here, uniqueness of the equilibrium ensures that this mapping is well-defined. We then interpret the support of the contextual equilibrium (i.e., $\mathrm{supp}(\pi^\star(x)) = \{k\ :\ \pi_k^\star(x) > 0\}$) as the set of winning agents for the given context x.

We make standard identifiability assumptions (see Assumption 1), such as positivity and missing at random (Rubin, 1974), so that we can identify the contextual PM from observed data through the context-dependent mean outcomes $\mu_{0,jk}^m(x)$:

$$g_{jk}^m(x) = \mathbb{E}[Z_{jk}^m \mid X = x] = \mathbb{E}[Z_{jk}^m \mid X = x, S^m = 1] =: \mu_{0,jk}^m(x) \quad \text{for } m \in \mathcal{Q}_{jk}. \tag{6}$$

Hence, identification of the contextual PM, and thus its equilibrium, follows from Eqs. (4) and (5) (see Appendix J.2 for details).

Nuisances: We refer to $\mu^m_{0,jk}(x)$ and $e^m_0(x)$ as *nuisance functions* and write $\eta_0 = (\boldsymbol{\mu}_0, \boldsymbol{e}_0)$, where (i) $\boldsymbol{\mu}_0 = (\mu^m_0)_{m\in\mathcal{Q}}$, $\mu^m_0 = (\mu^m_{0,jk})_{j,k\in m, j\neq k}$, is the collection of the outcome regressions and (ii) $\boldsymbol{e}_0 = (e^m_0)_{m\in\mathcal{Q}}$, with $e^m_0(x) = P(S^m = 1 \mid X = x)$, is the collection of the selection propensities across menus.

Stability: We adapt the stability result for equilibrium mappings (Dontchev & Rockafellar, 2009) to our context-dependent setting (see Appendix C.5 for the derivation and definition of $\|\cdot\|_\star$).

Proposition 1 (Contextual payoff stability)**.** *For any two contexts $x, x' \in \mathcal{X}$,*

$$\left\|\pi^\star(x) - \pi^\star(x')\right\| \leq \kappa^{-1}\left\|(A(x) - A(x'))\pi^\star(x')\right\|_\star \leq \kappa^{-1}\left\|A(x) - A(x')\right\|_\star. \tag{7}$$

Proposition 1 shows that the equilibrium is Lipschitz continuous with respect to changes in the contextual PM, which implies that errors in the contextual PM propagate directly to the equilibrium. Hence, in order to accurately learn the contextual equilibrium, we need to accurately estimate the contextual PM $A(x)$. However, this is challenging: some pairwise entries of the PM are missing, and, because estimating $A(x)$ requires estimated nuisance functions, a naïve plug-in estimator can inherit first-order bias from the nuisance estimation error. To remove this first-order bias, we propose a novel Neyman-orthogonal equilibrium learning method in the next section.

5 Orthogonal Equilibrium Learning

Overview. Our NashEval framework consists of two stages (see Figure 2): in Stage (1), we estimate the contextual payoff matrix $\widehat{A}(x)$ from the logged feedback, and in Stage (2), we use the estimated contextual payoff matrix to learn the context-to-equilibrium mapping.

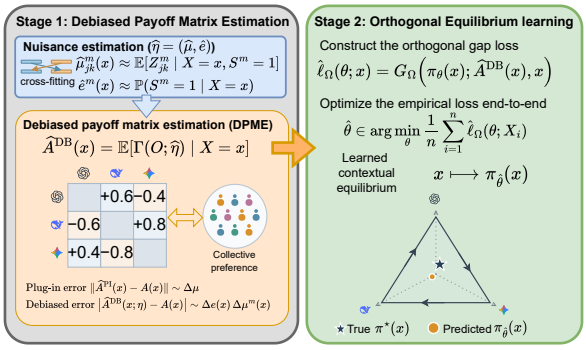


Figure 2: **Overview of our NashEval.** In Stage (1), we estimate the contextual payoff matrix using the debiased payoff matrix (DPM). In Stage (2), we use the estimated payoff matrix to learn the context-to-equilibrium mapping via an orthogonal loss.

Importantly, we show in Section 5.1 that a naïve, plug-in equilibrium learner is subject to the so-called *plug-in bias*, which can lead to a suboptimal equilibrium. As a remedy, we develop a Neyman-orthogonal equilibrium learner, and, to do so, we first propose a *debiased payoff matrix (DPM)* that is robust to nuisance misspecification (Section 5.2), which we use in Stage (1). In Section 5.3, we present a tailored orthogonal loss based on DPM for learning the contextual equilibrium in Stage (2). Finally, we show that our NashEval framework has favorable theoretical guarantees, such as Neyman-orthogonality and a quasi-oracle convergence rate (Section 5.4).

5.1 Why a naïve plug-in equilibrium learner is problematic

Plug-in equilibrium learner. A natural approach is to first estimate the outcome regressions $\widehat{\mu}^m_{jk}$ using an arbitrary machine learning method and then "plug" them into the menu aggregator:

$$\widehat{A}^{\mathrm{PI}}_{jk}(x) = \frac{1}{|\mathcal{Q}_{jk}|}\sum_{m\in\mathcal{Q}_{jk}} \widehat{\mu}^m_{jk}(x), \quad \text{where } \widehat{\mu}^m_{jk}(x) = \widehat{\mathbb{E}}[Z^m_{jk} \mid X = x, S^m = 1]. \tag{8}$$

Here, we consider the mean aggregator for simplicity. We can then learn the equilibrium $\widehat{\pi}^{\mathrm{PI}}(x)$ using the estimated contextual PM $\widehat{A}^{\mathrm{PI}}(x)$.

However, by Proposition 1, errors in the estimated contextual PM propagate directly to the learned equilibrium. In particular, the error of the equilibrium is Lipschitz-bounded by the error of the contextual PM estimator: $\|\widehat{\pi}^{\mathrm{PI}}(x) - \pi^\star(x)\| \leq \kappa^{-1}\|\widehat{A}^{\mathrm{PI}}(x) - A(x)\|_\star$, where $\widehat{A}^{\mathrm{PI}}_{jk}(x) - A_{jk}(x) =$

$\frac{1}{|\mathcal{Q}_{jk}|}\sum_{m\in\mathcal{Q}_{jk}}\Delta\mu^m_{jk}(x)$ and $\Delta\mu^m_{jk}=\widehat{\mu}^m_{jk}-\mu^m_{0,jk}$. Hence, errors in $\widehat{\mu}^m$ affect the estimated contextual PM (and therefore the learned equilibrium) at first order. In semiparametric statistics, such sensitivity to direct plug-in estimation is commonly referred to as *plug-in bias* (Kennedy, 2023). The effect can be particularly pronounced when κ is small.

Key idea: orthogonal learning. To reduce such sensitivity to nuisance estimation errors, NASHEVAL builds on orthogonal statistical learning (Foster & Syrgkanis, 2023). Let $\widehat{\eta}$ denote the estimated nuisance functions and suppose the equilibrium mapping π_θ is learned by minimizing

$$\widehat{\theta}=\arg\min_{\theta\in\Theta}\mathcal{L}(\theta;\widehat{\eta}):=\mathbb{E}\left[\ell(\pi_\theta(X),A(X;\widehat{\eta}))\right] \tag{9}$$

Our goal is thus to construct a loss for which the gradient with respect to θ is locally insensitive to errors in the nuisance functions. Formally, a loss is *Neyman-orthogonal* if

$$D_\eta D_\theta\mathcal{L}(\theta;\eta_0)[\widehat{\theta}-\theta,\widehat{\eta}-\eta_0]=0,\quad\forall\theta\in\Theta \tag{10}$$

where D_η and D_θ are the Gâteaux derivatives w.r.t. η and θ, under directions $h_\theta=\widehat{\theta}-\theta$ and $h_\eta=\widehat{\eta}-\eta_0$. Hence, orthogonality means that the gradient $D_\theta\mathcal{L}$ is robust to perturbations h_η of the nuisance functions around their true value η_0. Intuitively, orthogonality removes the first-order effect of nuisance estimation errors on the learning objective, thus allowing nuisance errors to enter the final estimation error only through higher-order terms (Foster & Syrgkanis, 2023).

In the following, we adapt this idea to NASHEVAL. However, doing so is nontrivial and requires a tailored orthogonalization for each stage, which we demonstrate in Sections 5.2 and 5.3.

5.2 STAGE ①: DEBIASED PAYOFF MATRIX

In this section, we construct the debiased payoff matrix (DPM) for Stage ① that is robust to nuisance estimation errors. We begin with the average payoff $\psi_{jk}=\mathbb{E}_X[A_{jk}(X)]=\mathbb{E}_X\left[f_{jk}(\{\mu^m_{0,jk}(X)\}_{m\in\mathcal{Q}_{jk}})\right]$. The efficient influence function (EIF) of the average payoff provides a correction (i.e., a debiasing term) to the plug-in payoff estimator, and, using this correction, we can construct a debiased PM that removes first-order sensitivity to nuisance estimation errors.

Definition 3 (Debiased payoff matrix)**.** *For generic nuisances $\eta=(\boldsymbol{\mu},\boldsymbol{e})$, we define the debiased payoff matrix as*

$$A^{\mathrm{DB}}_{jk}(x;\eta):=\mathbb{E}_0\left[\Gamma_{jk}(O;\eta)\mid X=x\right],(j\neq k)\qquad A^{\mathrm{DB}}(x;\eta):=\left(A^{\mathrm{DB}}_{jk}(x;\eta)\right)_{j,k\in\mathcal{M}}, \tag{11}$$

with diagonal entries $A^{\mathrm{DB}}_{jj}(x;\eta)$ set to 0. Here, $\Gamma_{jk}(O;\eta)$ is the EIF-based pseudo-outcome of the payoff (see Appendix J.3 for details):

$$\Gamma_{jk}(O;\eta)=\underbrace{f_{jk}\left(\{\mu^m_{jk}(X)\}_{m\in\mathcal{Q}_{jk}}\right)}_{\textit{plug-in term}}+\sum_{m\in\mathcal{Q}_{jk}}\underbrace{\frac{\partial f_{jk}}{\partial u_m}\left(\{\mu^{m'}_{jk}(X)\}_{m'\in\mathcal{Q}_{jk}}\right)\frac{S^m}{e^m(X)}\left\{Z^m_{jk}-\mu^m_{jk}(X)\right\}}_{\textit{debiasing term}}$$

Theorem 1 (Properties of the debiased payoff matrix)**.** *Under Assumptions 1–3, we have: (i) $A^{\mathrm{DB}}_{jk}(x;\eta_0)=\mathbb{E}_0\left[\Gamma_{jk}(O;\eta_0)\mid X=x\right]=A_{jk}(x)$; and (ii) for every perturbation $h=(h_\mu,h_e)$, the debiased payoff is Neyman-orthogonal w.r.t. η: $D_\eta A^{\mathrm{DB}}_{jk}(x;\eta)[h]\big|_{\eta=\eta_0}=0$.*

Proof. See Appendix J.4. □

The theorem shows that the debiased payoff matrix recovers the true contextual PM when the nuisances are correctly specified, and that it is locally insensitive to nuisance perturbations. More precisely, under a smoothness condition, we can show that the pointwise bias induced by nuisance misspecification is bounded by second-order nuisance errors:

$$\left|A^{\mathrm{DB}}_{jk}(x;\eta)-A_{jk}(x)\right|\le\frac{L_f}{2}\left\|\Delta\mu(x)\right\|_2^2+\sum_{m\in\mathcal{Q}_{jk}}\left|\frac{\partial f_{jk}}{\partial u_m}\left(\mu(x)\right)\right|\frac{|\Delta e^m(x)|}{e^m(x)}\,|\Delta\mu^m(x)|, \tag{12}$$

where ∇f_{jk} is L_f-Lipschitz. We formally derive this bound in Appendix J.5. Compared with the plug-in estimator, nuisance errors affect the debiased payoff matrix only through a squared outcome-error term and an outcome–propensity error product. In practice, given fitted nuisances $\widehat{\eta}=(\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{e}})$, we construct the DPM estimator $\widehat{A}^{\mathrm{DB}}(x;\widehat{\eta})$ by regressing $\Gamma_{jk}(O;\widehat{\eta})$ on X for each pair (j,k) with $j\neq k$, and set $\widehat{A}^{\mathrm{DB}}_{jj}:=0$.

5.3 STAGE (2): ORTHOGONAL EQUILIBRIUM LEARNING

In Stage (2), we learn the context-to-equilibrium mapping $\pi_\theta(\cdot)$ using the estimated contextual PM from Stage (1). A straightforward approach would be to solve the estimated contextual games for each context and then use the resulting equilibria as labels. However, repeatedly solving such games is computationally expensive and thus *not* scalable.

Instead, we directly learn the context-to-equilibrium mapping by minimizing a tailored, differentiable training loss with gradient-based optimization. To construct this loss, we build upon the *gap function*, which reformulates equilibrium computation as a differentiable optimization problem for a fixed payoff matrix (Auslender, 1973; Fukushima, 1992) (see Appendix C.4 for background). We extend the gap function to our contextual setting with stochastic payoffs, and, to do so, we combine it with our DPM from Stage (1) to obtain an orthogonal loss for learning $\pi^\star(\cdot)$:

Definition 4 (Orthogonal gap loss)**.** *For a generic nuisance η, let $A^{\mathrm{DB}}(x;\eta)$ be the population pseudo-outcome regression target in Eq. (11). We define the population orthogonal gap loss as*

$$\mathcal{L}(\theta;\eta) = \mathbb{E}_X\Big[G_\Omega\big(\pi_\theta(X); A^{\mathrm{DB}}(X;\eta), X\big)\Big], \tag{13}$$

where G_Ω is the regularized gap function, defined as

$$G_\Omega(\pi; A, x) = \max_{q\in\Delta_K}\big\{-F_\Omega(\pi, q; x)\big\} = \max_{q\in\Delta_K}\Big\{\big\langle -A\pi,\ \pi - q\big\rangle + \Omega(\pi) - \Omega(q)\Big\}. \tag{14}$$

Theorem 2 (Differentiability and Neyman-orthogonality)**.** *Under Assumptions 2–4, the orthogonal gap loss $\mathcal{L}(\theta;\eta)$ is differentiable with respect to θ for every fixed nuisance η satisfying Assumption 2. Under Assumption 1, the loss is also universally Neyman-orthogonal: for every $\theta\in\Theta$ and every pair of directions (h_θ, h_η) with bounded h_η, it holds that*

$$D_\eta D_\theta \mathcal{L}(\theta;\eta_0)[h_\theta, h_\eta] = 0, \qquad \forall\theta\in\Theta. \tag{15}$$

Proposition 2 (Oracle recovery under realizability)**.** *Let Assumptions 1, 3, and 4 hold, and assume that the policy class is realizable, i.e., there exists $\theta_0\in\Theta$ such that $\pi_{\theta_0}(x) = \pi^\star(x)$ for almost every x. Then, under the oracle nuisance η_0, we have*

$$\arg\min_{\theta\in\Theta}\mathcal{L}(\theta;\eta_0) = \big\{\theta\in\Theta : \pi_\theta(x) = \pi^\star(x) \text{ for almost every } x\big\}, \tag{16}$$

and the minimum population gap risk equals zero.

Proof. We prove Theorem 2 and Proposition 2 in Appendices J.7 and J.6, respectively. □

Intuitively, the gap function measures the largest advantage available to an optimal opponent against a candidate strategy π and equals zero at equilibrium. We replace the fixed contextual PM A with the debiased payoff matrix $A^{\mathrm{DB}}(X;\eta)$, which depends on the nuisance functions η, and average over contexts to obtain a population gap risk. Theorem 2 shows that our proposed loss is universally Neyman-orthogonal and therefore locally insensitive to nuisance estimation errors.

Procedure. In practice, training NASHEVAL proceeds in two stages. **(1)** We estimate the nuisance functions $\widehat{\eta}$ using K_{CF}-fold cross-fitting (Chernozhukov et al., 2018). For each fold, we fit the nuisance functions on the remaining $K_{\mathrm{CF}}-1$ folds and evaluate the pseudo-outcomes $\Gamma_{jk}(O;\widehat{\eta})$ on the held-out fold. We then regress the cross-fitted pseudo-outcomes on X to obtain $\widehat{A}^{\mathrm{DB}}(\cdot;\widehat{\eta})$. **(2)** We minimize the empirical orthogonal gap loss to obtain $\pi_{\widehat{\theta}}$:

$$\widehat{\theta}\in\arg\min_{\theta\in\Theta}\widehat{\mathcal{L}}_n(\theta), \quad \text{where } \widehat{\mathcal{L}}_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} G_\Omega\Big(\pi_\theta(X_i); \widehat{A}^{\mathrm{DB}}(X_i;\widehat{\eta}), X_i\Big). \tag{17}$$

At inference time, the learned parameter $\widehat{\theta}$ is used to predict the contextual equilibrium $\pi_{\widehat{\theta}}(x_\star)$ for a new context $x_\star\in\mathcal{X}$. In the sparsemax implementation, we approximate the equilibrium support by $S_\tau(\pi_{\widehat{\theta}}(x_\star)) = \{k : \pi_{\widehat{\theta},k}(x_\star) > \tau\}$, with $\tau = 10^{-3}$ as specified in Appendix I.2. For a downstream decision-making task with some utility function $\mathcal{U}(k,x)$, one can select a deterministic winner via $\arg\max_{k\in S_\tau(\pi_{\widehat{\theta}}(x_\star))}\mathcal{U}(k, x_\star)$.

5.4 THEORETICAL GUARANTEES

We provide theoretical guarantees for the learned equilibrium under our empirical orthogonal gap loss. To measure the quality of the equilibrium, we introduce the *regularized exploitability*, which is the regularized analogue of a standard measure of equilibrium quality in two-player zero-sum games and which is commonly used to evaluate how close a learned policy is to the true Nash equilibrium (Zinkevich et al., 2007). For a symmetric game with context x, the exploitability of the learned policy $\pi_{\widehat{\theta}}$ under the true contextual payoff matrix $A(x)$ is given by the regularized gap function $\mathrm{Expl}_{A(x)}\left(\pi_{\widehat{\theta}}(x)\right) := G_\Omega\left(\pi_{\widehat{\theta}}(x); A(x), x\right)$. The *average exploitability* is then $\mathrm{Expl}(\pi_{\widehat{\theta}}) := \mathbb{E}_X[\mathrm{Expl}_{A(X)}(\pi_{\widehat{\theta}}(X))]$, where $X \sim P_X$.

Theorem 3 (Empirical error bound for average exploitability)**.** *Let Assumptions 1–4 and 5–6 in Appendix J.1 hold. Further, let $r_\mu, r_e \to 0$ be the nuisance convergence rates and e_A the payoff regression error in Assumption 6. Let $\widehat{\mathcal{L}}$ be the population counterpart of $\widehat{\mathcal{L}}_n$. Then, the empirical minimizer $\widehat{\theta}$ in Eq. (17) satisfies*

$$\mathrm{Expl}(\pi_{\widehat{\theta}}) \le 4 \inf_{\theta\in\Theta} \mathrm{Expl}(\pi_\theta) + 4 \sup_{\theta\in\Theta} \left|\widehat{\mathcal{L}}(\theta) - \widehat{\mathcal{L}}_n(\theta)\right| + O_p\left(\frac{K^2}{\kappa}\left\{e_A^2 + r_\mu^4 + \underline{e}^{-2} r_\mu^2 r_e^2\right\}\right). \quad (18)$$

Corollary 1 (Quasi-oracle convergence rate)**.** *Under standard assumptions (see Appendix J.9), the empirical minimizer $\widehat{\theta}$ achieves the same exploitability rate as the learner that minimizes the gap loss with oracle nuisance functions $\eta_0 = (\boldsymbol{\mu}_0, \boldsymbol{e}_0)$.*

Proof. See Appendices J.8 and J.9 for the proofs. □

Theorem 3 shows that, for NASHEVAL, the nuisance estimation errors affect the exploitability rate only through higher-order terms, whereas, for the plug-in learner, the corresponding contribution of the nuisance error term would be r_μ^2. In the following, we verify the theoretical advantage of NASHEVAL over the plug-in learner through numerical experiments.

6 EXPERIMENTS

Baselines. We group the baselines into two categories: (1) Global+intransitive baselines: **Maximal lottery** (Lanctot et al., 2023), **Robust maximal lottery** (Khalaf et al., 2026), and **Pluralistic leaderboard** (Haghtalab et al., 2026); (2) Contextual+transitive baselines: Bradley–Terry (**BT-Plug-in**) (Bradley & Terry, 1952), **BT-Reg**, and **BT-debiased** (Frauen et al., 2026). We refer to our method as **NashEval-debiased**. We further compare to the plug-in learner from Section 5.1, called **NashEval-plug-in**, to demonstrate the benefits of debiasing. **NashEval-oracle** serves as an oracle-nuisance reference. We replace $\widehat{\eta}$ in NASHEVAL with the true η_0 and keep everything else unchanged. Since our framework is model-agnostic, we instantiate NASHEVAL and the baselines with the same machine learning backbone to ensure fair comparison (see Appendix I for details).

Metrics. We evaluate equilibrium quality through average unregularized exploitability and winner-set recovery using the F_1 score of the predicted equilibrium support. For RQ2, we additionally report the head-to-head game value (see Appendix I.1 for details).

RQ1 **Does NASHEVAL improve equilibrium learning?** We first verify the theoretical benefits of our method using a simulated dataset, where the true equilibrium is known. We consider three forms of relative feedback generated by N_v voters with $N_v = 1, 3, 5$. We simulate with $K = 5$ agents and $d = 5$-dimensional contexts (see Appendix G for details). **Results.** Figure 3 shows the results. (i) Across all three feedback forms, our NASHEVAL framework outperforms the baselines when $N_v \ge 3$. In particular, **NashEval-debiased** consistently performs closest to **NashEval-oracle** in terms of both exploitability and support F_1. (ii) When $N_v = 1$, preferences are transitive, and NASHEVAL performs comparably to **BT-debiased**, which is expected. (iii) **NashEval-debiased** achieves better exploitability and F_1 scores than **NashEval-plug-in**, which demonstrates the benefit of Neyman-orthogonality. (iv) Finally, **BT-Reg** and **BT-debiased** have higher exploitability and lower F_1 scores for $N_v \ge 3$. This suggests that regularization or debiasing alone cannot overcome misspecification from imposing transitivity when collective preferences are intransitive. In Appendix B.6, we show that our NASHEVAL achieves comparable equilibrium quality with much less inference runtime compared to numerical equilibrium solvers.

Table 2: Performance comparison for LLMs using RouterBench (Hu et al., 2024).

Method	Exploitability [↓]	Support F_1 [↑]	Game value [↑]
NashEval-debiased **(ours)**	$\mathbf{0.380 \pm 0.000}$	$\mathbf{0.300 \pm 0.001}$	$\mathbf{0.000 \pm 0.000}$
NashEval-plug-in	0.398 ± 0.001	0.287 ± 0.001	-0.016 ± 0.001
BT-debiased / Borda	0.395 ± 0.001	0.274 ± 0.001	-0.010 ± 0.001
BT-Plug-in	0.401 ± 0.003	0.270 ± 0.002	-0.014 ± 0.003
Maximal lottery	0.423 ± 0.000	0.246 ± 0.000	-0.030 ± 0.000
Robust maximal lottery	0.423 ± 0.000	0.246 ± 0.000	-0.030 ± 0.000
Pluralistic leaderboard	0.470 ± 0.000	0.242 ± 0.000	-0.136 ± 0.000

Table 3: Performance comparison for RL agents using the MO-Highway benchmark (Felten et al., 2023).

Method	Exploitability [↓]	Support F_1 [↑]	Game value [↑]
NashEval-debiased **(ours)**	$\mathbf{0.066 \pm 0.007}$	0.708 ± 0.023	$\mathbf{0.000 \pm 0.000}$
NashEval-plug-in	0.068 ± 0.000	$\mathbf{0.713 \pm 0.000}$	-0.027 ± 0.006
BT-debiased / Borda	0.099 ± 0.003	0.525 ± 0.014	-0.017 ± 0.004
BT-Plug-in	0.232 ± 0.000	0.438 ± 0.000	-0.072 ± 0.004
Maximal lottery	0.105 ± 0.000	0.505 ± 0.000	-0.006 ± 0.006
Robust maximal lottery	0.105 ± 0.000	0.505 ± 0.000	-0.006 ± 0.006
Pluralistic leaderboard	0.217 ± 0.000	0.556 ± 0.000	-0.147 ± 0.004

RQ2 **How does NASHEVAL perform in larger-scale, realistic agent evaluation settings?** We evaluate NASHEVAL in two large-scale real-world agent evaluation settings: **(1)** evaluating LLMs on RouterBench (Hu et al., 2024), using the model evaluations of Li et al. (2026), and **(2)** evaluating reinforcement learning (RL) agents in the multi-objective highway-driving environment of MO-Gymnasium (Felten et al., 2023), which is built on HighwayEnv (Leurent, 2018). For RouterBench, we use ten LLMs and 8,742 prompts. For the driving task, we train five RL agents on 256 configurations (i.e., contexts) with different reward models. The pairwise preference data are sampled from three external judges, each of which has internally transitive preferences (see Appendices H.2 and H.3 for details). We evaluate NASHEVAL from complete held-out pairwise comparisons against reference equilibria, which are close approximations to the true contextual equilibria. We further report the *game value* against **NashEval-debiased**, which quantifies whether each baseline has a higher or lower *collective* preference compared to NASHEVAL. In Appendix B.3, we do the analogous evaluation with the cost of the agents considered.

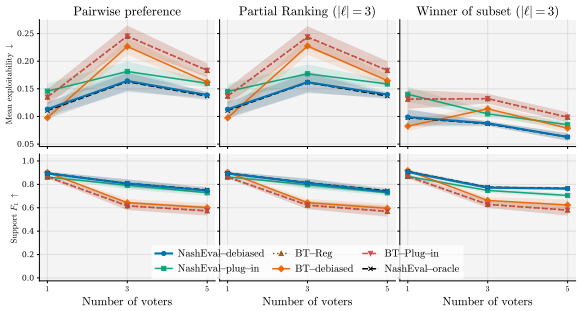


Figure 3: RQ1 **Equilibrium quality using a simulated dataset with ground truth.** Shown: mean exploitability and F_1 score on the simulated dataset for #voters$= 1, 3, 5$ over 5 random seeds.

Results. Tables 2 and 3 show the results. (1) Overall, our NashEval-debiased achieves the lowest mean exploitability on both RouterBench and the RL-agent benchmark and thus performs best. (2) On both real-world benchmarks, it also achieves lower exploitability and higher support F_1 than **BT-debiased / Borda** and the global evaluation baselines. This confirms that our NASHEVAL is more accurate in recovering the contextual evaluation under intransitive preferences. (3) The head-to-head game values favor **NashEval-debiased** over all competitors on both benchmarks. This demonstrates that **NashEval-debiased** better reflects the underlying collective preferences.

Insights. We examine how often each agent is selected as the winner in the RouterBench dataset. We use the same setup and reference evaluation as in RQ2. We divide the test data by domain and then compute the average winning rate of each agent by domain. We visualize the frequency as a heatmap in Figure 4. We observe that the estimated evaluation is largely consistent with the reference evaluation, which provides further evidence that NASHEVAL is effective in capturing context-dependent variation in agent performance. For example, frontier reasoning-based models (e.g., GPT-5, Gemini 2.5 Pro) perform better on the math domain, in line with our expectations.

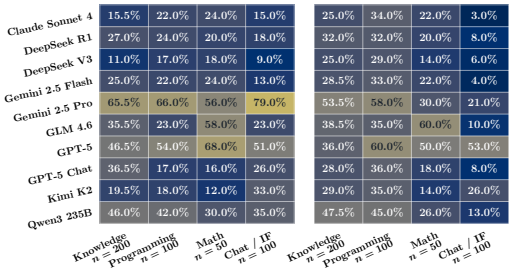


Figure 4: **Insights.** Winner frequency for each agent in the RouterBench dataset. *Left:* estimated evaluation from **NashEval-debiased**. *Right:* reference evaluation.

Conclusion. We develop, to the best of our knowledge, the first game-theoretic framework for context-dependent agent evaluation under potentially intransitive collective preferences. Our experiments confirm the theoretical properties of NASHEVAL and demonstrate the benefit of our Neyman-orthogonal equilibrium learning.

AI use statement. In this work, we used generative AI tools to generate synthetic datasets, examine and refine proofs, assist in writing proofs, provide feedback on our methodology and experiments, implement our method, process and clean data, and conduct exploratory data analysis. We also used LLMs as judges to construct pairwise preference feedback for the RouterBench experiments, as described in Appendix H.2. We independently developed the theoretical models and conceptual framework, formulated the mathematical claims, and supplied the critical ingredients for their proofs without generative AI assistance. We did not use generative AI tools for other research tasks beyond those listed above. We reviewed all the AI-assisted work and take responsibility for the final content of this work, including all AI-assisted text, code, data, and mathematical arguments.

Ethics statement. This work aims to advance the methodology of contextual agent evaluation. We only use publicly released datasets and do not collect personal data or other sensitive information. Therefore, we do not foresee any ethical issues arising from this work.

Reproducibility statement. The synthetic data-generating process and implementation details are provided in Appendices G and I. The construction of the RouterBench and reinforcement-learning datasets is described in Appendices H.2 and H.3, respectively. Proofs of the main exploitability and quasi-oracle results appear in Appendices J.8 and J.9. We publish our code at `https://anonymous.4open.science/r/Contextual-Agent-Evaluator-1A58`.

REFERENCES

Kenshi Abe and Yusuke Kaneko. Off-policy exploitability-evaluation in two-player zero-sum Markov games. In *AAMAS*, 2021.

Siavash Ameli, Siyuan Zhuang, Ion Stoica, and Michael W. Mahoney. A statistical framework for ranking LLM-based chatbots. In *ICLR*, 2025.

Enrique Areyan Viqueira, Cyrus Cousins, Eli Upfal, and Amy Greenwald. Learning equilibria of simulation-based games. *arXiv preprint arXiv:1905.13379*, 2019.

A. Auslender. Brève communication. résolution numérique d'inégalités variationnelles. *Revue française d'automatique informatique recherche opérationnelle. Mathématique*, 7(R2):67–72, 1973.

David Balduzzi, Karl Tuyls, Julien Perolat, and Thore Graepel. Re-evaluating evaluation. In *NeurIPS*, 2018.

Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Viktor Bengs, Aadirupa Saha, and Eyke Hüllermeier. Stochastic contextual dueling bandits under linear stochastic transitivity models. In *ICML*, pp. 1764–1786, 2022.

Peter J. Bickel, Chris A. J. Klaassen, Ya'acov Ritov, and Jon A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York, 1998. ISBN 978-0-387-98473-5. Reprint of the 1993 original.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Florian Brandl and Felix Brandt. Arrovian aggregation of convex preferences. *Econometrica*, 88 (2):799–844, 2020. doi: 10.3982/ECTA15749.

Florian Brandl, Felix Brandt, and Hans Georg Seedig. Consistent probabilistic social choice. *Econometrica*, 84(5):1839–1880, 2016.

Florian Brandl, Felix Brandt, and Christian Stricker. An analytical and experimental comparison of maximal lottery schemes. *Social Choice and Welfare*, 58(1):5–38, 2022. doi: 10.1007/s00355-021-01326-x.

Felix Brandt. Rolling the dice: Recent results in probabilistic social choice. In *Trends in Computational Social Choice*. 2017.

Claire Chen, Yuheng Zhang, Xinyu Liu, Zixuan Xie, Shuze Daniel Liu, and Nan Jiang. Offline two-player zero-sum Markov games with KL regularization. In *ICML*, 2026. arXiv:2605.13025.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10.1111/ectj.12097.

Wayne Chi, Valerie Chen, Anastasios Nikolas Angelopoulos, Wei-Lin Chiang, Aditya Mittal, Naman Jain, Tianjun Zhang, Ion Stoica, Chris Donahue, and Ameet Talwalkar. Copilot Arena: A platform for code LLM evaluation in the wild. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10354–10382. PMLR, 2025.

Chao-Kai Chiang, Takashi Ishida, and Masashi Sugiyama. LLM routing with dueling feedback. *arXiv preprint arXiv:2510.00841*, 2025.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot Arena: An open platform for evaluating LLMs by human preference. In *ICML*, 2024.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *NeurIPS*, 2017.

Arthur H. Copeland. A "reasonable" social welfare function. Mimeographed notes, Seminar on Applications of Mathematics to the Social Sciences, University of Michigan, 1951.

Qiwen Cui and Simon Du. When are offline two-player zero-sum Markov games solvable? *NeurIPS*, 2022.

Marquis de Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Imprimerie Royale, Paris, 1785.

Qiwei Di, Tao Jin, Yue Wu, Heyang Zhao, Farzad Farnoud, and Quanquan Gu. Variance-aware regret bounds for stochastic contextual dueling bandits. In *ICLR*, 2024.

Asen L. Dontchev and R. Tyrrell Rockafellar. *Implicit Functions and Solution Mappings: A View from Variational Analysis*. Springer Monographs in Mathematics. Springer, 2009.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *ICML*, 2011.

Miroslav Dudík, Katja Hofmann, Robert E. Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. In *COLT*, 2015.

Bhaskar Dutta and Jean-François Laslier. Comparison functions and choice correspondences. *Social Choice and Welfare*, 16(4):513–532, 1999.

Francisco Facchinei and Jong-Shi Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer Series in Operations Research. Springer, New York, 2003.

John Fearnley, Martin Gairing, Paul W. Goldberg, and Rahul Savani. Learning equilibria of games via payoff queries. *Journal of Machine Learning Research*, 16(39):1305–1344, 2015.

Dan S. Felsenthal and Moshé Machover. After two centuries, should Condorcet's voting procedure be implemented? *Behavioral Science*, 37(4):250–274, 1992.

Florian Felten, Lucas N. Alegre, Ann Nowé, Ana L. C. Bazzan, El Ghazali Talbi, Grégoire Danoy, and Bruno C. da Silva. A toolkit for reliable benchmarking and research in multi-objective reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/hash/4aa8891583f07ae200ba07843954caeb-Abstract.html`.

Peter C. Fishburn. Probabilistic social choice based on simple voting comparisons. *The Review of Economic Studies*, 51(4):683–692, 1984.

Dylan J. Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *The Annals of Statistics*, 51 (3):879–908, 2023. doi: 10.1214/23-AOS2258.

Dennis Frauen, Maresa Schröder, Konstantin Hess, and Stefan Feuerriegel. Orthogonal survival learners for estimating heterogeneous treatment effects from time-to-event data. In *NeurIPS*, 2025.

Dennis Frauen, Athiya Deviyani, Mihaela van der Schaar, and Stefan Feuerriegel. Nonparametric LLM evaluation from preference data. In *ICML*, 2026.

Evan Frick, Connor Chen, Joseph Tennyson, Tianle Li, Wei-Lin Chiang, Anastasios N. Angelopoulos, and Ion Stoica. Prompt-to-leaderboard. *arXiv preprint arXiv:2502.14855*, 2025.

Masao Fukushima. Equivalent differentiable optimization problems and descent methods for asymmetric variational inequality problems. *Mathematical Programming*, 53(1):99–110, 1992. doi: 10.1007/BF01585696.

Nika Haghtalab, Ariel D. Procaccia, Han Shao, Serena Lutong Wang, and Kunhe Yang. Pluralistic leaderboards. In *ICML*, 2026.

Daniel Halpern, Evi Micha, Ariel D. Procaccia, Benjamin Schiffer, Itai Shapira, and Shirley Zhang. AI alignment from social choice perspectives. *ACM SIGecom Exchanges*, 24(1): 47–76, June 2026. URL `https://www.sigecom.org/exchanges/volume_24/1/HALPERN.pdf`.

Frank R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974.

Qitian Jason Hu, Jacob Bieker, Xiuyu Li, Nan Jiang, Benjamin Keigwin, Gaurav Ranganath, Kurt Keutzer, and Shriyash Kaustubh Upadhyay. RouterBench: A benchmark for multi-LLM routing system. In *ICML 2024 Agentic Markets Workshop*, 2024. Poster.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *ICML*, 2016.

Patrick R. Jordan, Yevgeniy Vorobeychik, and Michael P. Wellman. Searching for approximate equilibria in empirical games. In *AAMAS*, pp. 1063–1070, 2008.

Edward H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023. doi: 10.1214/23-EJS2157.

Edward H. Kennedy. Semiparametric doubly robust targeted double machine learning: A review. In *Handbook of Statistical Methods for Precision Medicine*, pp. 207–236. Chapman and Hall/CRC, 2024.

Hadi Khalaf, Serena L. Wang, Daniel Halpern, Itai Shapira, Flavio du Pin Calmon, and Ariel D. Procaccia. Robust AI evaluation through maximal lotteries. In *ICML*, 2026.

Elon Kohlberg and Jean-François Mertens. On the strategic stability of equilibria. *Econometrica*, 54(5):1003–1037, 1986. doi: 10.2307/1912320.

Ezgi Korkmaz. The axiomatic value of regularization in AI alignment from human preferences. In *ICML*, 2026. Spotlight.

Gerald H. Kramer. On a class of equilibrium conditions for majority rule. *Econometrica*, 41(2): 285–297, 1973.

Germain Kreweras. Aggregation of preference orderings. In *Mathematics and Social Sciences I*, 1965.

Gilbert Laffond, Jean-François Laslier, and Michel Le Breton. The bipartisan set of a tournament game. *Games and Economic Behavior*, 5(1):182–201, 1993. doi: 10.1006/game.1993.1010.

Marc Lanctot, Kate Larson, Yoram Bachrach, Luke Marris, Zun Li, Avishkar Bhoopchand, Thomas Anthony, Brian Tanner, and Anna Koop. Evaluating agents using social choice theory. *arXiv preprint arXiv:2312.03121*, 2023.

Torbjörn Larsson and Michael Patriksson. A class of gap functions for variational inequalities. *Mathematical Programming*, 64(1):53–79, 1994.

Jean-François Laslier. *Tournament solutions and majority voting*. Springer-Verlag, 1997.

Edouard Leurent. An environment for autonomous driving decision-making. GitHub repository, 2018. URL `https://github.com/Farama-Foundation/HighwayEnv`.

Hao Li, Yiqun Zhang, Zhaoyan Guo, Chenxu Wang, Shengji Tang, Qiaosheng Zhang, Yang Chen, Biqing Qi, Peng Ye, Lei Bai, Zhen Wang, and Shuyue Hu. LLMRouterBench: A massive benchmark and unified framework for LLM routing. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, July 2026.

Valentyn Melnychuk, Dennis Frauen, Jonas Schweisthal, and Stefan Feuerriegel. Orthogonal representation learning for estimating causal quantities. In *AISTATS*, 2026.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015. doi: 10.1038/nature14236.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Côme Fiegel, Andrea Michi, Marco Selvi, Sertan Girgin, Nikola Momchev, Olivier Bachem, Daniel J. Mankowitz, Doina Precup, and Bilal Piot. Nash learning from human feedback. In *ICML*, 2024.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

Shayegan Omidshafiei, Christos Papadimitriou, Georgios Piliouras, Karl Tuyls, Mark Rowland, Jean-Baptiste Lespiau, Wojciech M. Czarnecki, Marc Lanctot, Julien Perolat, and Rémi Munos. α-Rank: Multi-agent evaluation by evolution. *Scientific Reports*, 9:9937, 2019.

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 from preference data. In *ICLR*, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-Baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. URL `https://jmlr.org/papers/v22/20-1364.html`.

Ronald L. Rivest and Emily Shen. An optimal single-winner preferential voting system based on game theory. In *COMSOC*, 2010.

Corby Rosset, Ching-An Cheng, Arindam Mitra, Michael Santacroce, Ahmed Awadallah, and Tengyang Xie. Direct Nash optimization: Teaching language models to self-improve with general preferences. *arXiv preprint arXiv:2404.03715*, 2024.

Mark Rowland, Shayegan Omidshafiei, Karl Tuyls, Julien Perolat, Michal Valko, Georgios Piliouras, and Rémi Munos. Multiagent evaluation under incomplete information. In *NeurIPS*, 2019.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Aadirupa Saha. Optimal algorithms for stochastic contextual preference bandits. In *NeurIPS*, 2021.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, et al. EmbeddingGemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*, 2025.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

Samuel Sokota, Ryan D'Orazio, J. Zico Kolter, Nicolas Loizou, Marc Lanctot, Ioannis Mitliagkas, Noam Brown, and Christian Kroer. A unified approach to reinforcement learning, quantal response equilibria, and two-player zero-sum games. In *International Conference on Learning Representations*, 2023.

Peter Stone. *The Luck of the Draw: The Role of Lotteries in Decision Making*. Oxford University Press, 2011. doi: 10.1093/acprof:oso/9780199756100.001.0001.

Gokul Swamy, Christoph Dann, Rahul Kidambi, Zhiwei Steven Wu, and Alekh Agarwal. A minimaximalist approach to reinforcement learning from human feedback. In *ICML*, 2024.

T. Nicolaus Tideman. Independence of clones as a criterion for voting rules. *Social Choice and Welfare*, 4(3):185–206, 1987. doi: 10.1007/BF00433944.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. doi: 10.1017/CBO9780511802256.

John von Neumann. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen*, 100(1):295–320, 1928.

Yevgeniy Vorobeychik. Probabilistic analysis of simulation-based games. *ACM Transactions on Modeling and Computer Simulation*, 20(3):16:1–16:25, 2010. doi: 10.1145/1842713.1842719.

Michael P. Wellman. Methods for empirical game-theoretic analysis. In *AAAI*, pp. 1552–1556, 2006.

Yue Wu, Tao Jin, Qiwei Di, Hao Lou, Farzad Farnoud, and Quanquan Gu. Borda regret minimization for generalized linear dueling bandits. In *ICML*, 2024.

Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. Self-play preference optimization for language model alignment. In *ICLR*, 2025.

Austin Xu, Srijan Bansal, Yifei Ming, Semih Yavuz, and Shafiq Joty. Does context matter? ContextualJudgeBench for evaluating LLM-based judges in contextual settings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9541–9564, 2025.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The K-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025a.

Yuheng Zhang, Dian Yu, Baolin Peng, Linfeng Song, Ye Tian, Mingyue Huo, Nan Jiang, Haitao Mi, and Dong Yu. Iterative Nash policy optimization: Aligning LLMs with general preferences via no-regret learning. In *ICLR*, 2025b.

Yuheng Zhang, Claire Chen, and Nan Jiang. Beyond pessimism: Offline learning in KL-regularized games. *arXiv preprint arXiv:2604.06738*, 2026a.

Ziqian Zhang, Xingjian Hu, Yue Huang, Kai Zhang, Ruoxi Chen, Yixin Liu, Qingsong Wen, Kaidi Xu, Xiangliang Zhang, Neil Zhenqiang Gong, and Lichao Sun. RankLLM: Weighted ranking of LLMs by quantifying question difficulty. In *ICLR*, 2026b.

Zesen Zhao, Shuowei Jin, and Zhuoqing Morley Mao. Eagle: Efficient training-free router for multi-LLM inference. In *Proceedings of the Workshop on Machine Learning for Systems at NeurIPS 2024*, pp. 1–8, 2024.

Han Zhong, Wei Xiong, Jiyuan Tan, Liwei Wang, Tong Zhang, Zhaoran Wang, and Zhuoran Yang. Pessimistic minimax value iteration: Provably efficient equilibrium learning from offline datasets. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27117–27142, 2022.

Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. In *NeurIPS*, 2007.

A Related Work

Agent evaluation from relative feedback. Relative feedback, particularly pairwise preferences, is widely used to evaluate generative agents. Chatbot Arena and its successor LMArena collect head-to-head judgments on LLM responses (Chiang et al., 2024), while Copilot Arena applies the same evaluation to coding agents (Chi et al., 2025). A standard approach aggregates these comparisons with a score-based model such as Bradley–Terry to produce a global leaderboard (Bradley & Terry, 1952; Chiang et al., 2024). Subsequent work proposes improved evaluation and ranking procedures by modeling ties and dependence among agents (Ameli et al., 2025), estimating nonparametric ranking functionals (Frauen et al., 2026), learning prompt-dependent Bradley–Terry scores (Frick et al., 2025), or incorporating question difficulty into rankings (Zhang et al., 2026b). Frick et al. (2025) is contextual but retains a Bradley–Terry model at each prompt. Despite these differences, these methods ultimately summarize agents by scalar scores or an ordered leaderboard, which fail to represent intransitive preferences. To accommodate potentially intransitive preferences, Rivest & Shen (2010) use the equilibrium distribution of the pairwise majority-margin game as a randomized single-winner voting rule. Recent agent-evaluation work extends this social-choice perspective to heterogeneous tasks or populations through locally stable top-k sets (Haghtalab et al., 2026) and maximal lotteries (Lanctot et al., 2023; Khalaf et al., 2026). These social-choice methods nevertheless aggregate feedback across contexts to produce a global evaluation, which can obscure context-dependent variation in agents' relative capabilities. Other related game-theoretic methods rank agents from empirical interaction payoffs using Nash averaging or α-Rank, including under incomplete payoff information (Balduzzi et al., 2018; Omidshafiei et al., 2019; Rowland et al., 2019). To our knowledge, no existing method simultaneously captures context-dependent variation in agents' relative capabilities and represents the collective preference of a heterogeneous population.

Social choice and maximal lotteries. Social choice theory studies how individual preferences can be aggregated into collective decisions, including settings where majority preferences are cyclic and no deterministic Condorcet winner exists (de Condorcet, 1785; Brandt, 2017). A central randomized rule is the maximal lottery, introduced by Kreweras (1965) and developed by Fishburn (1984): it selects a distribution over alternatives that maximizes the worst-case expected majority margin, equivalent to an equilibrium strategy of the induced two-player zero-sum game. Subsequent work has studied maximal lotteries and related tournament-based rules through game-theoretic, axiomatic, computational, and empirical analyses (Felsenthal & Machover, 1992; Laffond et al., 1993; Laslier, 1997; Dutta & Laslier, 1999; Brandl & Brandt, 2020; Brandl et al., 2022). Maximal lotteries extend the Condorcet principle and satisfy strong population- and composition-consistency properties, while robustness to cloned alternatives is a standard criterion for voting rules (Tideman, 1987; Brandl et al., 2016). More broadly, lotteries are considered suitable for decision making when a deterministic choice is undesirable (Stone, 2011). Recent work brings this perspective to AI alignment by treating heterogeneous human feedback as a preference-aggregation problem and studying the axiomatic consequences of aggregation and regularization (Halpern et al., 2026; Korkmaz, 2026). Unlike this literature, which aggregates preferences into a global majority profile without modeling context-dependent variation, we learn a context-specific maximal lottery (with convex regularization) from selectively observed offline feedback.

Double/debiased machine learning. Double/debiased machine learning (DML) builds on semiparametric efficiency theory (Bickel et al., 1998; Kennedy, 2024) to support valid inference on target parameters while using flexible machine learning models for nuisance functions. Its main tools include influence-function analysis in robust estimation (Hampel, 1974), efficient influence functions from semiparametric theory (Bickel et al., 1998), and Neyman-orthogonal moment equations (Chernozhukov et al., 2018). Foster & Syrgkanis (2023) extend this principle to general two-stage statistical learning: Neyman-orthogonality removes the first-order effect of nuisance estimation error on the target loss, allowing the learner to attain quasi-oracle rates under suitable conditions (Nie & Wager, 2021). Such an orthogonal learning paradigm is used in many scenarios, such as heterogeneous treatment-effect estimation (Kennedy, 2023), survival analysis (Frauen et al., 2025), off-policy evaluation and learning (Dudík et al., 2011; Jiang & Li, 2016), and representation learning for causal quantities (Melnychuk et al., 2026). We show that the same principle applies to contextual equilibrium learning from stochastic relative feedback by constructing a debiased payoff estimator and an orthogonal gap loss.

LLM routing from preferential feedback. LLM routing studies the selection of the best LLM(s) for a given prompt. A specific line of work is to learn a predictive router from pairwise preferential feedback (Zhao et al., 2024; Ong et al., 2025; Chiang et al., 2025; Frick et al., 2025). This can be viewed as a special case of our general formulation: the context is the prompt, and the agents are the LLMs. However, these works assume that the preference is transitive: they assign implicit contextual scores and then approximate the preference via a utility function (or link function). This leads to misspecification when preferences cycle. In contrast, our framework does not assume transitivity and instead learns a contextual equilibrium that yields better alignment.

Game-theoretic RLHF. Canonical RLHF methods learn a scalar reward model from pairwise comparisons (e.g., Christiano et al., 2017; Ouyang et al., 2022), while direct preference optimization eliminates explicit reward-model fitting but retains the Bradley–Terry preference assumption (Rafailov et al., 2023). To accommodate general, potentially intransitive preferences, recent works instead formulate language-model alignment as a two-player zero-sum game and train a Nash policy through self-play or no-regret optimization (Munos et al., 2024; Swamy et al., 2024; Rosset et al., 2024; Wu et al., 2025; Zhang et al., 2025b). Although conceptually similar, their goal is to learn the response-generation Nash policy, which is different from our agent evaluation setting with the target being the Nash policy over a set of candidates.

Contextual dueling bandits. This line of work studies how to learn a context-dependent policy when the bandit can only collect pairwise or subsetwise preferential feedback (e.g., Yue et al., 2012; Dudík et al., 2015; Saha, 2021; Bengs et al., 2022; Di et al., 2024; Wu et al., 2024). Saha (2021), Bengs et al. (2022) and Di et al. (2024) all implicitly assume stochastic transitivity, which we do not. Wu et al. (2024) use the Borda score to define a transitivity-robust regret, but still learn a single winner that is neither Condorcet-consistent nor independent of clones (Appendix E). Dudík et al. (2015) are the first to propose the concept of "von Neumann winner" in the dueling bandit setting, which allows for preference cycles and upon which we build our NASHEVAL. But they use a plug-in estimator of the preference matrix and are thus not robust to nuisance estimation errors. Finally, all of these works are targeted at an interactive setting, where the bandit iteratively selects which pair to compare, digests the feedback, and then approximates the optimal policy. In contrast, our work is targeted at an offline setting, where the feedback is collected beforehand. Thus, these works are not applicable to our setting.

Offline reinforcement learning in games. Offline RL in games studies how to learn equilibrium policies from offline multi-agent behaviors, and thus is related to our setting. Existing methods learn such policies directly from observed rewards, with or without KL regularization (Cui & Du, 2022; Zhong et al., 2022; Zhang et al., 2026a; Chen et al., 2026). Closest to our debiasing approach, Abe & Kaneko (2021) construct doubly robust off-policy estimators of the exploitability of candidate policy profiles and use them for policy selection. We instead debias the contextual PM and learn the entire context-to-equilibrium mapping with an orthogonal gap loss. Finally, RL in games has a fundamentally different observation model: the payoff is sampled directly from the reward model for joint actions, whereas we observe only the relative feedback of the agents within a menu. Hence, these methods are *not* applicable to our setting.

Equilibrium learning from stochastic/noisy games. A related literature studies equilibrium approximation when the payoff function is unavailable analytically but can be accessed through payoff queries or stochastic simulation. This line develops search procedures, probabilistic error bounds, and finite-sample guarantees for equilibria computed from empirical games (e.g., Wellman, 2006; Jordan et al., 2008; Vorobeychik, 2010; Areyan Viqueira et al., 2019; Fearnley et al., 2015). However, these equilibrium learning methods do not apply debiasing w.r.t. the nuisance functions, and thus are subject to plug-in bias. In contrast, our approach uses a debiased payoff matrix (DPM) that is robust to nuisance misspecification, and optimizes a Neyman-orthogonal loss to learn the contextual equilibria.

B ADDITIONAL EXPERIMENTS

We complement the main experiments with additional results. In Section B.1, we compare pooled and prompt-specific equilibria on RouterBench to illustrate how aggregation can hide contextual winners. In Section B.2, we examine subgroup-specific equilibria and apply NashEval-debiased to human feedback from LMArena. We use these analyses to illustrate variation in agent capabilities across contexts and the presence of multiple agents in an equilibrium's support. Because we do not know the true equilibria in LMArena, we treat comparisons with empirical subgroup lotteries as descriptive evidence rather than an accuracy assessment.

In Section B.3, we compare cost–performance trade-offs to assess whether contextual equilibrium evaluation remains useful when we account for inference costs. In Section B.4, we replace the outcome and payoff regression backbones with MLPs to examine whether the main synthetic findings persist under an alternative model instantiation. In Section B.5, we vary the regularization strength to assess the sensitivity of exploitability and winner-set recovery to this choice.

B.1 AGGREGATION HIDES REFERENCE CONTEXTUAL WINNERS

Existing maximal-lottery-based agent evaluators aggregate preferences across all or a subset of contexts to compute a global maximal lottery (Lanctot et al., 2023; Khalaf et al., 2026). Using the RouterBench dataset, we show that this aggregation can hide the reference contextual winners and lead to suboptimal evaluation.

We use MMLU-Pro and LiveCodeBench as examples. Figures 5 and 6 compare the equilibrium distribution obtained from aggregated preferences with the distributions for individual prompts (i.e., contexts). The aggregated preferences yield a Condorcet winner: Gemini 2.5 Pro for MMLU-Pro and GPT-5 for LiveCodeBench. Across individual prompts, however, the contextual equilibria are highly heterogeneous and need not contain a Condorcet winner. In some contexts, the aggregated winner is not even in the contextual equilibrium support (see Contexts 2 and 3 for LiveCodeBench). Therefore, **aggregation hides the reference contextual winners and can lead to suboptimal evaluation**. This motivates our work on a contextual evaluator that adaptively selects the best agent or agents for each context.

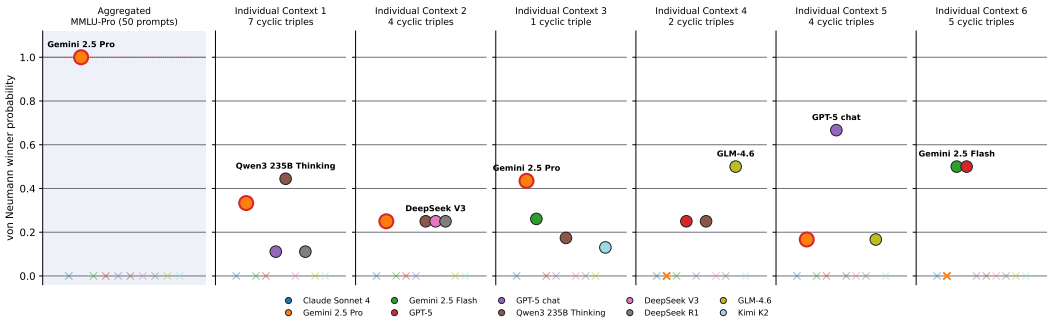


Figure 5: Equilibrium distributions for aggregated MMLU-Pro preferences and individual prompts. Aggregation yields a Condorcet winner, whereas the contextual equilibria are heterogeneous and need not contain one. **Therefore, aggregation hides the reference contextual winners and can lead to suboptimal evaluation.**

B.2 EXPERIMENT ON LMARENA

We demonstrate our **NashEval-debiased** on the Arena Human Preference 140K data. In LMArena, a user compares responses from two anonymously presented language models and supplies a preference, a tie, or a both-bad judgment (Chiang et al., 2024). The full release contains 135,634 human votes involving 53 models and 128,336 distinct first-user prompts (Appendix H.1). We restrict both training and analysis to the 17,480 comparisons between the 15 models in Table 9, which cover 17,296 distinct prompts.

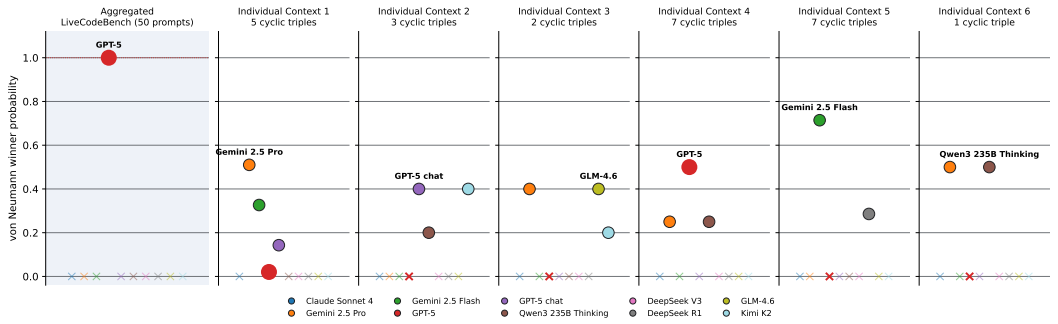
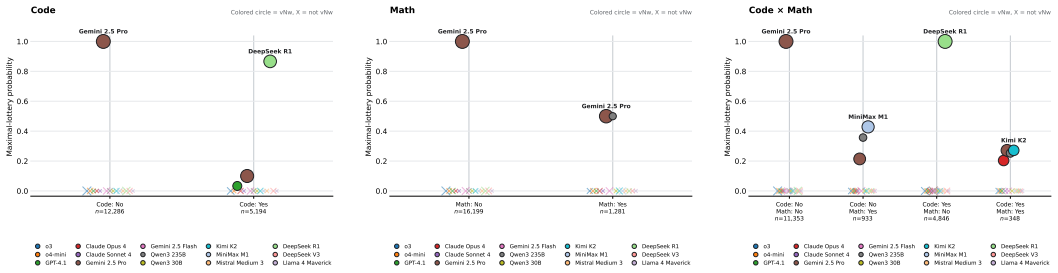


Figure 6: Equilibrium distributions for aggregated LiveCodeBench preferences and individual prompts. Aggregation yields a Condorcet winner, whereas the contextual equilibria are heterogeneous and need not contain one. Therefore, **aggregation hides the reference contextual winners and can lead to suboptimal evaluation**.

B.2.1 Heterogeneity of the agent capability across contexts

In this section, we provide evidence of the heterogeneity of agent capability across contexts in the LMArena dataset. We stratify the LMArena data by discrete context attributes (see Tables 7 and 8) and construct majority-margin matrices using win-minus-loss counts for each stratification, with unobserved model pairs replaced by ties. We then compute an equilibrium distribution for each subgroup having at least 90% pair coverage. When multiple equilibria exist, we choose the one with minimal ℓ_2 norm.

Results. We quantify heterogeneity using pairwise total-variation distances between the equilibrium distributions. For each single- or two-attribute combination, we report the mean and maximum total-variation distance in Table 4. We further visualize the subgroup-specific equilibria in Figures 7 and 8. These equilibria are highly heterogeneous across subgroups. For example, Gemini 2.5 Pro is not in the equilibrium support for $(\text{code}, \text{math}) = (\text{yes}, \text{no})$. Therefore, a single global equilibrium cannot represent how agent performance varies across contexts.



(a) Code (b) Mathematics (c) Code and mathematics

Figure 7: Subgroup-specific equilibrium distributions across code and mathematics contexts. Each plotted circle is an agent in the equilibrium support, and the y-axis shows its equilibrium probability. Further stratification can introduce new agents into the support (e.g., MiniMax M1), indicating that **a global equilibrium does not reflect context-specific agent capabilities**.

For this experiment, we define the context using only the recorded language, code, mathematics, and instruction-following labels. We display a selection of subgroups that have at least 100 comparisons and observed feedback for at least 94 of the 105 selected-model pairs. The true context-specific equilibrium is unavailable in LMArena. We therefore use the empirical maximal lottery within each subgroup as a reference, i.e., disagreement or agreement does not necessarily imply the accuracy of the method.

Results. Figure 9 shows that the evaluation of our **NashEval-debiased** is largely consistent with the empirical maximal lottery, which indicates the effectiveness of our NashEval. For English prompts without code, mathematics, or instruction following, both equilibria assign all probability to Gemini 2.5 Pro. For English coding prompts without mathematics or instruction following, the

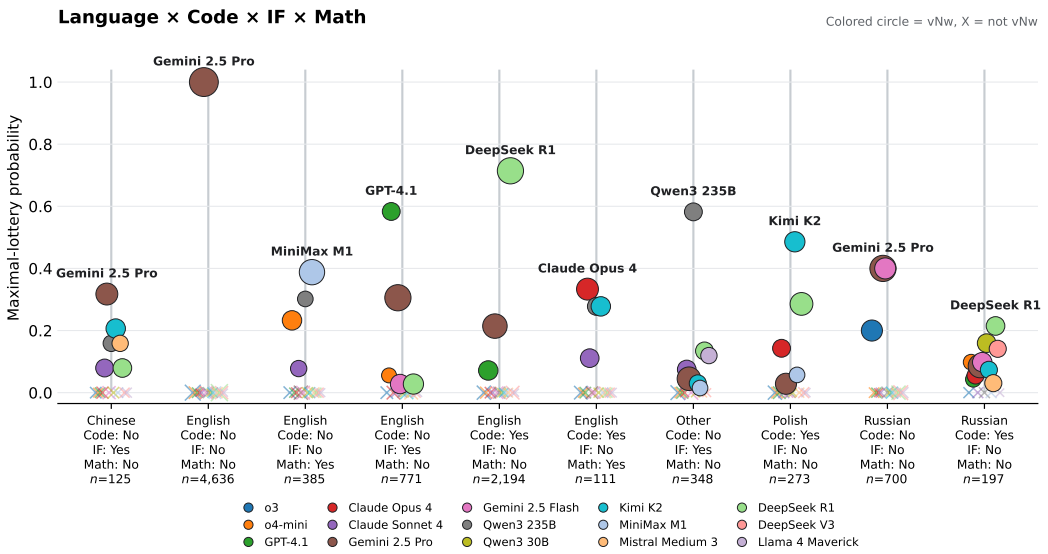


Figure 8: Subgroup-specific equilibrium distributions across language, code, instruction-following, and mathematics contexts. Each plotted circle is an agent in the equilibrium support, and the y-axis shows its equilibrium probability. Agent capabilities vary substantially across contexts.

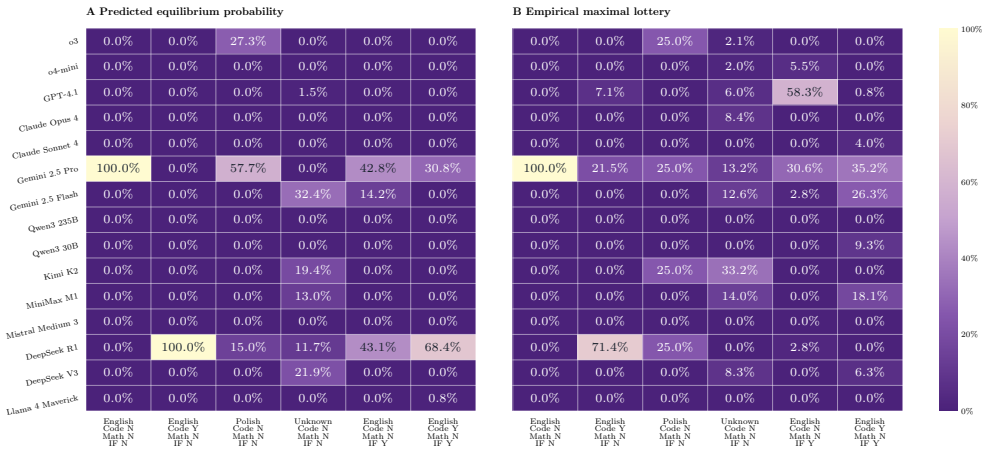


Figure 9: Attribute-specific equilibria for six LMArena subgroups defined by language, code, mathematics, and instruction-following labels. Panel A shows the NashEval-debiased equilibrium probability of each selected model. Panel B shows the regularized maximal lottery from pairwise win-minus-loss counts pooled within the same subgroup. Both panels use $\rho = 0.01$ and report percentages. Note: this is different from the heatmap in Figure 4, which reports the winner frequencies of the agents.

Table 4: Heterogeneity of subgroup-specific equilibrium distributions in LMArena illustrated by total variation within each attribute combination. Only subgroups with at least 90% pair coverage are included. **For most stratifications, subgroup equilibria differ substantially.**

Stratification	Subgroups	Mean TV	Max TV	Distinct winners
Single attribute				
Language	5	0.612	0.866	3
Code	2	0.900	0.900	2
Instruction following	2	0.000	0.000	1
Mathematics	2	0.500	0.500	1
Two attributes				
Language × code	6	0.460	0.600	1
Language × instruction following	6	0.736	1.000	4
Language × mathematics	5	0.734	1.000	4
Code × instruction following	4	0.438	0.875	2
Code × mathematics	3	1.000	1.000	3
Instruction following × mathematics	3	0.333	0.500	1

predicted equilibrium assigns all probability to DeepSeek R1, while the empirical lottery assigns it 71.4%. The predicted equilibrium changes across these contexts: DeepSeek R1 receives no probability in the first group but is the sole supported model in the second. For English instruction-following prompts without code or mathematics, the predicted equilibrium assigns 43.1% to DeepSeek R1 and 42.8% to Gemini 2.5 Pro, whereas the empirical lottery assigns 58.3% to GPT-4.1. Claude Opus 4 receives zero predicted probability in all six displayed groups, despite having the highest listed price.

B.3 Cost-constrained performance

Experimental setup. We reuse the RouterBench payoff models from RQ2 across five seeds and 500 test prompts. For NashEval-debiased, NashEval-plug-in, and Maximal lottery, we incorporate the prices in Table 11 through

$$\Omega_\beta(\pi) = \frac{\rho}{2}\|\pi\|_2^2 + \beta\widetilde{c}^\top\pi, \qquad \rho = 0.01, \qquad \widetilde{c}_j = \frac{c_j - \min_k c_k}{\max_k c_k - \min_k c_k}. \tag{19}$$

Here, c_j is the cost of $1{,}000$ input and $1{,}000$ output tokens. We vary β and compare against **BT-cost**, which maximizes its predicted pairwise score under a per-prompt expected-cost budget following Theorem 1 of Frick et al. (2025) (Appendix I.3). We compare all methods by realized mean expected cost. For the equilibrium methods, we use Eq. (19) with the price weights

$$\beta \in \{0.025t : t = 0, \dots, 12\} \cup \{0.35, 0.4, 0.5, 0.75, 1, 1.5, 2, 3, 5, 10, 20\}. \tag{20}$$

Metrics. We report unregularized exploitability against an unrestricted opponent and average strict win rate against the ten single-agent opponents with equal weight. We give ties, including self-comparisons, zero credit and retain them in the denominator.

Results. Figures 10 and 11 show that NashEval-debiased offers the strongest empirical trade-off through much of the intermediate-cost region, achieving the Pareto frontier among intermediate cost ranges for both metrics. This demonstrates that our method is able to produce the cost-constrained contextual evaluations that reflect the collective preference of the population.

B.4 Ablations on backbone

Experimental setup. We conduct an ablation study on the backbone of the outcome nuisance estimator and the second-stage payoff regressor on the simulated dataset from RQ1. We jointly replace LightGBM in these two components with multilayer perceptrons (MLPs). We retain the LightGBM propensity estimator and compute equilibria with the same numerical solver as in RQ1, without introducing a learned policy predictor. For NashEval-oracle, we use the same MLP payoff regressor as for NashEval-debiased, but construct the pseudo-outcomes with the true nuisances. NashEval-plug-in uses the fold-averaged MLP outcome predictions without a separate payoff regression. The

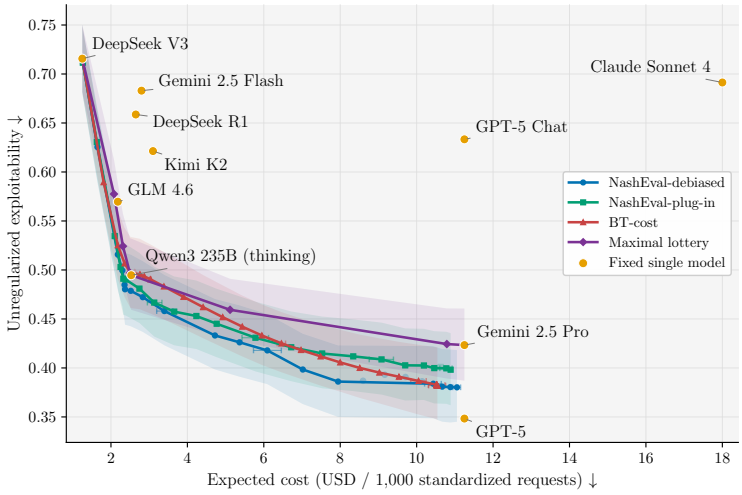


Figure 10: Cost–exploitability trade-offs on RouterBench across five seeds and 500 test prompts. Our method largely achieves the empirical Pareto frontier over costs of approximately USD 2–11 per 1,000 standardized requests among the evaluated policies. Therefore, under the same cost budget, our method produces contextual evaluations that better reflect the collective preference of the population.

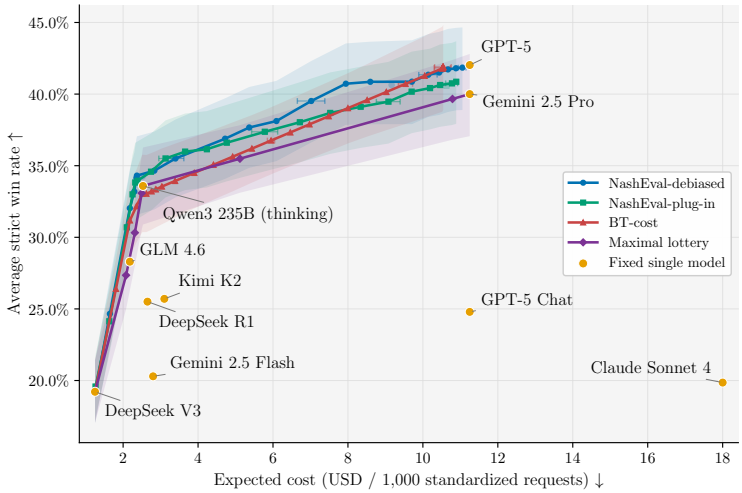


Figure 11: Cost–average-win-rate trade-offs on RouterBench. Our method largely achieves the empirical Pareto frontier at the intermediate cost range $[2, 11]$. Therefore, under the same cost budget, our method produces contextual evaluations that better reflect the collective preference of the population.

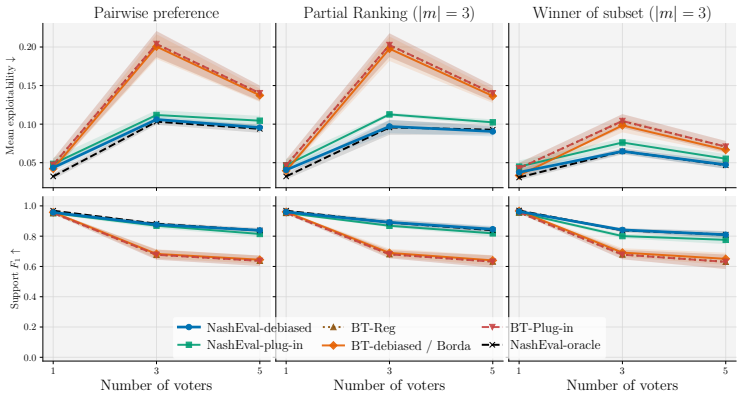


Figure 12: Ablation study on the outcome nuisance and payoff backbone on the simulated dataset. We note that our NASHEVAL outperforms the non-oracle baselines on $N_v \geq 3$ and achieves comparable performance with **BT**-based baselines on $N_v = 1$. This result suggests that the gains are not specific to the backbone.

remainder is unchanged compared with RQ1. We evaluate unregularized exploitability and support F_1, using the same support threshold of 10^{-3}, and summarize each metric by its mean and standard deviation across 5 random seeds.

We use two hidden layers with widths $(64, 64)$ for each outcome nuisance and $(128, 128)$ for each pairwise payoff regressor, with ReLU activations. We train both networks with Adam at learning rate 3×10^{-3} and an L_2 weight penalty of 10^{-4}, using batch sizes of 512 and 1,024, respectively. We allow at most 300 epochs and apply early stopping on a held-out subset of the corresponding training data with patience 12. We fix these MLP settings before evaluating the test sets. We regress the unclipped debiased pseudo-outcomes, clip predicted margins to $[-1, 1]$, and impose antisymmetry before solving for the equilibrium.

We report the mean and std of the (unregularized) exploitability and F_1 score over five random seeds in Figure 12. For $N_v \geq 3$, our NASHEVAL consistently outperforms other non-oracle baselines. For $N_v = 1$, preferences are transitive and the BT-model-based baselines are slightly better. The result is consistent with the main findings in RQ1 in Section 6. These results support the effectiveness of NASHEVAL with both the LightGBM and MLP backbones evaluated here.

B.5 PERFORMANCE OVER VARYING REGULARIZATION STRENGTHS

We investigate the effect of the regularization strength ρ on the performance of our method on the simulated dataset. The setting is similar to RQ1, but we fix the number of voters to 3. Then we vary ρ within the range $\{0.0, 0.001, 0.01, 0.1, 1.0\}$ and compare against other contextual evaluation baselines. We report the mean and std of the (unregularized) exploitability and F_1 score over five random seeds in Figure 13. We observe little change in either metric when $\rho \leq 0.01$, with performance slightly decreasing for ρ beyond that range. We also observe a slight improvement in F_1 score when ρ is increased from 0.01 to 0.1. In summary, on this grid the performance of our NASHEVAL is insensitive to $\rho \leq 0.01$.

B.6 COMPARISON WITH THE NUMERICAL EQUILIBRIUM SOLVER

A naíve approach is to only estimate the contextual PM, and then solve the equilibrium every time a new context is encountered. We compare our **NashEval-debiased** and **NashEval-plug-in** with the numerical equilibrium solver using both the debiased and plug-in PM estimators. We adopt the same settings as in RQ1, where we use unregularized exploitability and support F_1 scores to evaluate the quality of the learned equilibrium while recording the mean runtime during inference. We use a custom solver that enumerates over all possible supports $(2^K - 1)$ to search for equilibria that satisfy the KKT conditions.

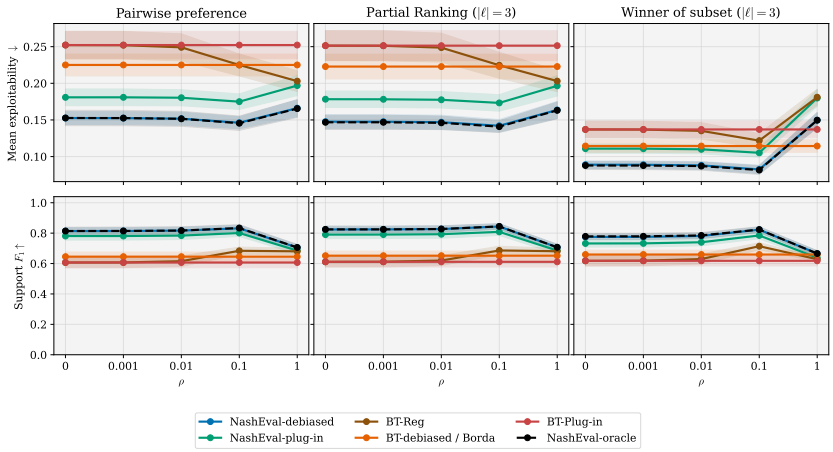


Figure 13: Mean $\pm$ std of exploitability and F_1 score on the simulated dataset for #voters$= 3$ over 5 random seeds, plotted against varying regularization strengths ρ.

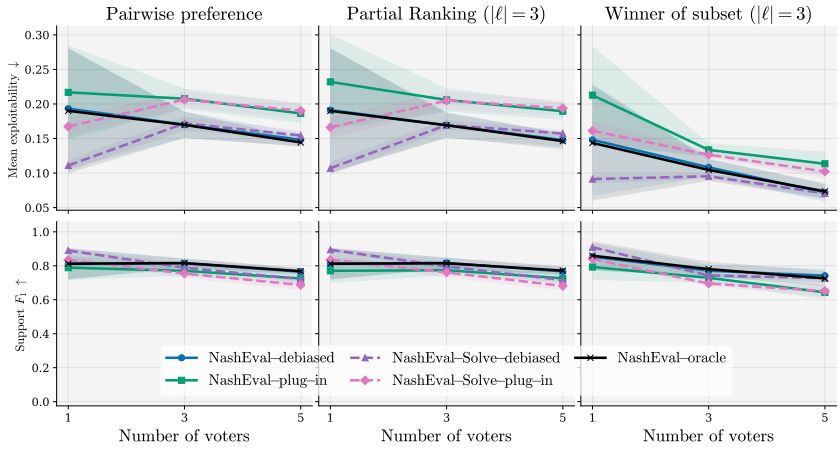


Figure 14: Performance comparison of our NASHEVAL with the numerical solver. Mean $\pm$ std of exploitability and F_1 score on the simulated dataset for #voters$= 1, 3, 5$ over 5 random seeds.

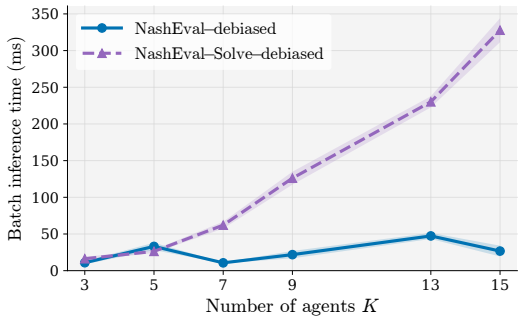


Figure 15: Inference runtime comparison between our **NashEval-debiased** and the numerical equilibrium solver. Our NASHEVAL is much more efficient at inference time.

Results. Results in Figure 14 show that our NASHEVAL performs comparably to the numerical solver when the number of voters is larger than 1. A runtime comparison is given in Figure 15. We see that our NASHEVAL is much more efficient during inference: the runtime of **NashEval-debiased** remains stable, while the inference runtime of the numerical solver increases exponentially when K increases.

C PRELIMINARIES OF GAME THEORY

In this appendix, we introduce the game-theoretic concepts needed to understand our solution concept and methods. In Section C.1, we introduce variational inequalities and monotone operators, which provide the main tools for characterizing and solving equilibria in two-player zero-sum games. In Section C.2, we introduce the self-play operator and show that its variational-inequality solutions coincide with the symmetric equilibria of our regularized game. In Section C.3, we prove existence and uniqueness of the contextual regularized equilibrium. In Section C.4, we introduce the gap function and establish its consistency, equilibrium-error control, and differentiability. In Section C.5, we prove that the equilibrium is stable under perturbations of the payoff matrix.

Notation and standing conventions. We conduct the game-theoretic analysis at a fixed context x. Unless stated otherwise, we suppress the context from the payoff notation and write $A = A(x)$. The regularizer Ω does not depend on x. Recall that $A = -A^\top$ and that Ω is continuously differentiable on a neighborhood of Δ_K and κ-strongly convex on Δ_K with respect to $\|\cdot\|$, where $\kappa > 0$. Thus, for every $p, q \in \Delta_K$,

$$\Omega(q) \geq \Omega(p) + \langle \nabla\Omega(p), q - p\rangle + \frac{\kappa}{2}\|q - p\|^2. \tag{21}$$

Game-theoretic setting. Game theory studies strategic interactions in which each player's payoff depends on the strategies chosen by all players. A game specifies the players, their feasible strategy sets, their payoff functions, and a solution concept that describes stable joint behavior. In this work, we study a contextual, regularized, two-player zero-sum game over the shared mixed-strategy simplex Δ_K. At each context x, the row player receives $F_\Omega(\pi, q; x)$ and the column player receives $-F_\Omega(\pi, q; x)$, where the skew-symmetric matrix $A(x)$ encodes collective pairwise preferences and the context-independent regularizer Ω induces strong convexity. We seek the symmetric saddle point of this game, which coincides with its Nash equilibrium and defines the contextual von Neumann winner $\pi^\star(x)$.

C.1 VARIATIONAL INEQUALITIES AND MONOTONE OPERATORS

In this subsection, we introduce variational inequalities and strong monotonicity, which provide the equilibrium characterization and uniqueness argument used throughout this appendix. A *variational inequality* is the problem of finding, for an operator $V : \Delta_K \to \mathbb{R}^K$ and the convex compact set Δ_K, a point $\pi^\star \in \Delta_K$ such that

$$\big\langle V(\pi^\star),\ \pi - \pi^\star \big\rangle \ \geq\ 0 \qquad \text{for all } \pi \in \Delta_K. \tag{22}$$

We write $\mathrm{VI}(V, \Delta_K)$ for this problem. Geometrically, Eq. (22) says that $-V(\pi^\star)$ makes a non-acute angle with every feasible direction at $\pi^\star$, so no feasible direction decreases the objective represented by V. See Facchinei & Pang (2003).

The operator property that governs uniqueness is monotonicity. An operator V is *monotone* on Δ_K if

$$\langle V(p) - V(q),\ p - q\rangle \ \geq\ 0 \qquad \text{for all } p, q \in \Delta_K,$$

and κ-*strongly monotone* with respect to $\|\cdot\|$ if the stronger inequality

$$\langle V(p) - V(q),\ p - q\rangle \ \geq\ \kappa\,\|p - q\|^2 \tag{23}$$

holds for the same quantifiers. We use two standard facts (Facchinei & Pang, 2003): the gradient of a differentiable κ-strongly convex function is κ-strongly monotone, and a strongly monotone VI has at most one solution. In our game, the linear part $\pi \mapsto -A\pi$ contributes no curvature because $\langle -Ad, d\rangle = -d^\top Ad = 0$ for $A = -A^\top$.

C.2 THE SELF-PLAY OPERATOR

In this subsection, we define the self-play operator, characterize symmetric equilibria as solutions of its variational inequality, and establish the strong monotonicity needed to prove uniqueness. For the regularized game $F_\Omega(\pi, q) = \pi^\top Aq - [\Omega(\pi) - \Omega(q)]$ of Definition 2, define the *self-play operator*

$$V_\Omega(\pi)\ =\ -A\pi + \nabla\Omega(\pi). \tag{24}$$

This is the standard self-play operator for a regularized zero-sum game (Sokota et al., 2023). The next lemma gives its exact saddle-point characterization.

Lemma 1 (Self-play variational inequality)**.** *A pair $(\pi^\star, \pi^\star)$ with $\pi^\star \in \Delta_K$ is a symmetric saddle point (i.e., a Nash equilibrium) of F_Ω if and only if*

$$\big\langle -A\pi^\star + \nabla\Omega(\pi^\star),\ \pi - \pi^\star \big\rangle \ \geq\ 0 \qquad \textit{for all } \pi \in \Delta_K. \tag{25}$$

Moreover V_Ω is κ-strongly monotone with respect to $\|\cdot\|$.

Proof. Necessity. Suppose $(\pi^\star, \pi^\star)$ is a saddle point. We fix the column strategy at $\pi^\star$ and discard the term $\Omega(\pi^\star)$, which does not depend on π. The row player then maximizes

$$f_{\pi^\star}(\pi) \ =\ \pi^\top A\pi^\star - \Omega(\pi) \tag{26}$$

over Δ_K. The function $f_{\pi^\star}$ is concave because it equals a linear function minus a convex function. The saddle-point property implies that $\pi^\star$ maximizes it over the convex set Δ_K. The first-order optimality condition for this differentiable concave problem is exactly Eq. (25).

Sufficiency. Conversely, suppose Eq. (25) holds. By convexity, $\Omega(\pi) \geq \Omega(\pi^\star) + \langle \nabla\Omega(\pi^\star), \pi - \pi^\star\rangle$ for all $\pi \in \Delta_K$. Hence, for every π,

$$f_{\pi^\star}(\pi) - f_{\pi^\star}(\pi^\star) = \langle A\pi^\star, \pi - \pi^\star\rangle - \big[\Omega(\pi) - \Omega(\pi^\star)\big] \leq \big\langle A\pi^\star - \nabla\Omega(\pi^\star),\ \pi - \pi^\star\big\rangle \leq 0,$$

where the final inequality follows from Eq. (25). We add $\Omega(\pi^\star)$ to both sides and obtain $F_\Omega(\pi, \pi^\star) \leq F_\Omega(\pi^\star, \pi^\star)$ for all $\pi \in \Delta_K$. Therefore, $\pi^\star$ is a best response for the row player. For the column player, skew-symmetry gives $F_\Omega(p, q) = -F_\Omega(q, p)$, so

$$F_\Omega(\pi^\star, q) = -F_\Omega(q, \pi^\star) \ \geq\ -F_\Omega(\pi^\star, \pi^\star) = F_\Omega(\pi^\star, \pi^\star).$$

Here we use $F_\Omega(\pi^\star, \pi^\star) = (\pi^\star)^\top A\pi^\star = 0$, which follows from $d^\top Ad = 0$ for skew-symmetric A. Thus, $\pi^\star$ is also a best response for the column player, and $(\pi^\star, \pi^\star)$ is a saddle point with value zero. This direction uses skew-symmetry twice and fails without it.

Strong monotonicity. Let $p, q \in \Delta_K$ and put $d = p - q$. Then

$$\langle V_\Omega(p) - V_\Omega(q), d\rangle = \underbrace{-\,d^\top Ad}_{=0} + \langle \nabla\Omega(p) - \nabla\Omega(q), d\rangle = \langle \nabla\Omega(p) - \nabla\Omega(q), d\rangle. \tag{27}$$

The first term vanishes because $d^\top Ad = \frac{1}{2} d^\top (A + A^\top) d = 0$ for skew-symmetric A. For the second, we write Eq. (21) twice, once at (p, q) and once at (q, p):

$$\Omega(q) \ \geq\ \Omega(p) + \langle \nabla\Omega(p), q - p\rangle + \tfrac{\kappa}{2}\|d\|^2, \qquad \Omega(p) \ \geq\ \Omega(q) + \langle \nabla\Omega(q), p - q\rangle + \tfrac{\kappa}{2}\|d\|^2.$$

We add the two inequalities and cancel $\Omega(p) + \Omega(q)$ to obtain $\langle \nabla\Omega(p) - \nabla\Omega(q), d\rangle \geq \kappa\|d\|^2$. We substitute this result into Eq. (27) and obtain Eq. (23) for V_Ω. □

C.3 Existence and uniqueness

In this subsection, we combine the minimax theorem with the self-play characterization to prove that the contextual regularized equilibrium exists, is symmetric, and is unique, which makes the target mapping $x \mapsto \pi^\star(x)$ well-defined.

Proposition 3 (Existence and uniqueness of the regularized equilibrium)**.** *Let $A = -A^\top$ and let Ω be continuously differentiable on a neighborhood of Δ_K and κ-strongly convex on Δ_K, with $\kappa > 0$. Then the game F_Ω of Definition 2 admits a saddle point, every saddle point is symmetric of the form $(\pi^\star, \pi^\star)$, the game value is zero, and $\pi^\star$ is* unique*. Consequently $\pi^\star(x)$ is a well-defined single-valued function of x.*

Proof. Existence. The strategy set Δ_K is compact and convex, and F_Ω is upper-semicontinuous and concave in π and lower-semicontinuous and convex in q. Hence Sion's minimax theorem (Sion, 1958) gives

$$\max_{\pi\in\Delta_K}\ \min_{q\in\Delta_K} F_\Omega(\pi, q) \ =\ \min_{q\in\Delta_K}\ \max_{\pi\in\Delta_K} F_\Omega(\pi, q) \ =:\ \mathrm{val}(F_\Omega),$$

with both extrema attained, so a saddle point exists.

Value zero and symmetry. Skew-symmetry of A gives $F_\Omega(p,q) = -F_\Omega(q,p)$ for all p, q, since $p^\top Aq = -q^\top Ap$ and the bracket $[\Omega(p) - \Omega(q)]$ also changes sign. The game is thus symmetric. If we exchange the roles of the players, we map saddle points to saddle points and negate the value. Hence, $\mathrm{val}(F_\Omega) = -\mathrm{val}(F_\Omega)$, so $\mathrm{val}(F_\Omega) = 0$. Let $(\pi^\star, q^\star)$ be any saddle point. Because the value is zero and $F_\Omega(p,p) = 0$ for every p, antisymmetry implies that $(q^\star, \pi^\star)$ is also a saddle point. The standard interchangeability of saddle-point components then implies that $(\pi^\star, \pi^\star)$ and $(q^\star, q^\star)$ are saddle points. It therefore suffices to establish uniqueness among symmetric saddle points.

Uniqueness. By Lemma 1, symmetric saddle points are exactly the solutions of $\mathrm{VI}(V_\Omega, \Delta_K)$, so it suffices to show that this variational inequality has at most one solution. Suppose $\pi_1, \pi_2 \in \Delta_K$ both solve it. We test the inequality for π_1 at the feasible point π_2 and test the inequality for π_2 at the feasible point π_1. This gives

$$\langle V_\Omega(\pi_1),\ \pi_2 - \pi_1\rangle\ \geq\ 0, \qquad \langle V_\Omega(\pi_2),\ \pi_1 - \pi_2\rangle\ \geq\ 0.$$

We add the two inequalities and write $d = \pi_1 - \pi_2$. We obtain

$$\langle V_\Omega(\pi_1) - V_\Omega(\pi_2),\ d\rangle\ \leq\ 0.$$

On the other hand, Lemma 1 states that V_Ω is κ-strongly monotone, so $\langle V_\Omega(\pi_1) - V_\Omega(\pi_2), d\rangle \geq \kappa\|d\|^2$. Together, these two bounds imply

$$\kappa\,\|\pi_1 - \pi_2\|^2\ \leq\ 0,$$

and since $\kappa > 0$ this forces $\|\pi_1 - \pi_2\| = 0$, i.e., $\pi_1 = \pi_2$. The symmetric saddle point is therefore unique. For any saddle point $(\pi^\star, q^\star)$, the preceding argument shows that both $(\pi^\star, \pi^\star)$ and $(q^\star, q^\star)$ are symmetric saddle points. Uniqueness gives $\pi^\star = q^\star$, so every saddle point is symmetric and $x \mapsto \pi^\star(x)$ is single-valued. □

C.4 Gap functions

In this subsection, we specialize gap functions to the self-play variational inequality and establish the properties that make the resulting gap a valid differentiable loss for learning the contextual equilibrium. Gap functions convert a variational inequality into scalar optimization by assigning each candidate solution a nonnegative measure of its violation. The classical construction appears, up to an equivalent change of sign, in Auslender (1973). Fukushima (1992) developed a differentiable regularized gap function, and Larsson & Patriksson (1994) placed this construction within a broader class of gap functions for variational inequalities. For a general operator V on a feasible set C, the classical gap has the form $\sup_{y\in C}\langle V(z), z - y\rangle$: it is nonnegative because $y = z$ is feasible, and it vanishes exactly when z solves $\mathrm{VI}(V, C)$.

In our game, Lemma 1 identifies equilibrium with the variational inequality induced by the self-play operator $V_\Omega(\pi) = -A\pi + \nabla\Omega(\pi)$. The regularized gap in Definition 4 is equivalently the maximum payoff available to an opponent against π, relative to self-play. Its inner maximizer is the unique regularized best response

$$T_\Omega(\pi; A, x) := \operatorname*{arg\,min}_{q\in\Delta_K} \left\{\pi^\top Aq + \Omega(q)\right\}. \tag{28}$$

We define the regularized gap function as

$$\begin{aligned} G_\Omega(\pi; A, x) &:= \max_{q\in\Delta_K} \Big\{\langle -A\pi, \pi - q\rangle + \Omega(\pi) - \Omega(q)\Big\} \\ &= \Omega(\pi) - \min_{q\in\Delta_K} \left\{\pi^\top Aq + \Omega(q)\right\} \\ &= \Omega(\pi) - \pi^\top A T_\Omega(\pi; A, x) - \Omega\big(T_\Omega(\pi; A, x)\big). \end{aligned} \tag{29}$$

The second equality uses $\pi^\top A\pi = 0$, which follows from the skew-symmetry of A.

Consequently, we can solve the game by minimizing $G_\Omega(\pi; A, x)$ over Δ_K. The minimum value is zero, and the equilibrium $\pi^\star(x)$ is its unique minimizer. Strong convexity makes the inner best response unique, while Danskin's theorem supplies a gradient for the outer minimization.

Proposition 4 (Properties of the gap function)**.** *(Fukushima, 1992; Larsson & Patriksson, 1994) Fix a context x and a skew-symmetric payoff A, and let $\pi^\star$ be the regularized equilibrium of the game with payoff A. Then, for every $\pi \in \Delta_K$:*

1. (Nonnegativity and consistency) *The gap is nonnegative and vanishes exactly at the equilibrium,*
$$G_\Omega(\pi; A, x) \;\geq\; 0, \qquad G_\Omega(\pi; A, x) = 0 \iff \pi = \pi^\star. \tag{30}$$
2. (Equilibrium error control) *The squared equilibrium error is bounded by the gap:*
$$\left\|\pi - \pi^\star\right\|^2 \;\leq\; \frac{2}{\kappa}\, G_\Omega(\pi; A, x). \tag{31}$$
3. (Differentiability) *The gap $G_\Omega(\cdot; A, x)$ is differentiable, with*
$$\nabla_\pi G_\Omega(\pi; A, x) = -A\, T_\Omega(\pi; A, x) + \nabla\Omega(\pi), \tag{32}$$
which involves no derivative of T_Ω with respect to π.

Proof. Nonnegativity and consistency. For fixed π, define $h_\pi(q) := \pi^\top A q + \Omega(q)$. Since $\pi^\top A \pi = 0$, Eq. (29) gives
$$G_\Omega(\pi; A, x) = h_\pi(\pi) - \min_{q \in \Delta_K} h_\pi(q) \geq 0.$$
Equality holds if and only if π minimizes h_π over Δ_K. The first-order condition for this convex problem is
$$\big\langle -A\pi + \nabla\Omega(\pi), q - \pi \big\rangle \geq 0 \qquad \text{for every } q \in \Delta_K.$$
Lemma 1 shows that this condition holds if and only if π is the unique equilibrium $\pi^\star$.

Equilibrium error control. Let $d = \pi - \pi^\star$. We evaluate the maximum in Eq. (29) at $q = \pi^\star$ and obtain
$$G_\Omega(\pi; A, x) \geq \langle -A\pi, d\rangle + \Omega(\pi) - \Omega(\pi^\star).$$
Strong convexity at $\pi^\star$ gives
$$\Omega(\pi) - \Omega(\pi^\star) \geq \langle \nabla\Omega(\pi^\star), d\rangle + \frac{\kappa}{2}\|d\|^2.$$
Skew-symmetry gives $\langle -A(\pi - \pi^\star), d\rangle = 0$, while the equilibrium variational inequality gives $\langle -A\pi^\star + \nabla\Omega(\pi^\star), d\rangle \geq 0$. Therefore,
$$G_\Omega(\pi; A, x) \geq \frac{\kappa}{2}\|\pi - \pi^\star\|^2,$$
which proves Eq. (31).

Differentiability. The maximization objective in Eq. (29) is strongly concave in q, so it has the unique maximizer $T_\Omega(\pi; A, x)$. For fixed q, skew-symmetry gives
$$\nabla_\pi \Big[\langle -A\pi, \pi - q\rangle + \Omega(\pi) - \Omega(q)\Big] = -Aq + \nabla\Omega(\pi).$$
Danskin's theorem evaluates this derivative at the unique maximizer and gives Eq. (32). □

C.5 Equilibrium stability with respect to the payoff

In this subsection, we apply variational-inequality sensitivity analysis to bound changes in the contextual equilibrium by perturbations of the payoff matrix, thereby establishing the stability result used in the main analysis.

We denote the dual norm $\|\cdot\|_\star$ as
$$\|v\|_\star := \sup_{\|u\| \leq 1} \langle v, u\rangle, \qquad \|B\|_\star := \sup_{\|u\| \leq 1} \|Bu\|_\star.$$

Proof of Proposition 1. We adapt the standard two-VI sensitivity argument (Dontchev & Rockafellar, 2009, Theorem 2F.9). We retain the perturbation term induced by the payoff matrix. We fix $x, x' \in \mathcal{X}$ and write

$$\pi = \pi^\star(x), \qquad \pi' = \pi^\star(x'), \qquad d = \pi - \pi'.$$

For brevity, we write $A = A(x)$ and $A' = A(x')$. By Lemma 1, we test the two equilibrium variational inequalities at π' and π, respectively, and obtain

$$\langle -A\pi + \nabla\Omega(\pi), \pi' - \pi\rangle \geq 0, \tag{33}$$

$$\langle -A'\pi' + \nabla\Omega(\pi'), \pi - \pi'\rangle \geq 0. \tag{34}$$

We rewrite both inequalities in terms of d and add them to obtain

$$\begin{aligned} 0 &\leq \langle A\pi - A'\pi' - \{\nabla\Omega(\pi) - \nabla\Omega(\pi')\}, d\rangle \\ &= \langle (A - A')\pi', d\rangle + \underbrace{\langle Ad, d\rangle}_{=0} - \langle \nabla\Omega(\pi) - \nabla\Omega(\pi'), d\rangle. \end{aligned} \tag{35}$$

The skew-symmetry of A gives $\langle Ad, d\rangle = 0$. Because Ω is differentiable and κ-strongly convex, its gradient is κ-strongly monotone, so

$$\langle \nabla\Omega(\pi) - \nabla\Omega(\pi'), d\rangle \geq \kappa\|d\|^2.$$

We combine this inequality with Eq. (35) and apply dual Hölder to obtain

$$\kappa\|d\|^2 \leq \langle (A - A')\pi', d\rangle \leq \|(A - A')\pi'\|_\star \, \|d\|.$$

If $d = 0$, the claim is immediate. Otherwise, we divide by $\kappa\|d\|$ and obtain

$$\|\pi^\star(x) - \pi^\star(x')\| \leq \kappa^{-1}\|\{A(x) - A(x')\}\pi^\star(x')\|_\star.$$

Finally, $\|\pi'\| \leq 1$ by the norm normalization in Appendix J, and therefore

$$\|(A - A')\pi'\|_\star \leq \|A - A'\|_\star\|\pi'\| \leq \|A - A'\|_\star.$$

This proves both inequalities in Proposition 1. □

C.5.1 Instability of the unregularized equilibrium/maximal lottery

Without regularization, a single-valued equilibrium selection is not uniformly continuous in the payoff matrix near games with multiple equilibria (Kohlberg & Mertens, 1986; Facchinei & Pang, 2003). For $\epsilon > 0$, we consider the two skew-symmetric games

$$A_\epsilon^+ = \epsilon \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \qquad A_\epsilon^- = -A_\epsilon^+.$$

The unique maximal lottery of A_ϵ^+ is e_1, while that of A_ϵ^- is e_2. Hence

$$\|A_\epsilon^+ - A_\epsilon^-\|_{\mathrm{op}} = 2\epsilon \longrightarrow 0, \qquad \|e_1 - e_2\|_2 = \sqrt{2}.$$

Thus, a vanishing payoff perturbation can produce a constant change in the selected equilibrium. The discontinuity occurs at the limiting tie game $A = 0$, whose equilibrium set is the whole simplex. We need additional regularity or strong convexity to obtain a uniformly stable single-valued target.

D Background on efficient influence functions and orthogonal statistical learning

Our work draws on efficient semiparametric inference and orthogonal statistical learning. In Section D.1, we define the efficient influence function, use the von Mises expansion to isolate plug-in bias, and specialize these concepts to the average contextual payoff. In Section D.2, we use the efficient influence function to construct a one-step estimator and then convert its uncentered form into a regression target for the contextual payoff. In Section D.3, we extend this bias correction from a scalar functional to our context-to-equilibrium mapping and explain the resulting robustness guarantee. We refer the reader to Kennedy (2024) for a detailed treatment of influence-function-based estimation and to Foster & Syrgkanis (2023) for the general theory of orthogonal statistical learning.

D.1 Efficient influence function

Let $O \sim P_0$, where the true distribution P_0 belongs to a statistical model $\mathcal{P}$, and let $\psi : \mathcal{P} \to \mathbb{R}$ be a scalar target functional. Consider any regular one-dimensional parametric submodel $\{P_t : |t| < \epsilon\} \subseteq \mathcal{P}$ through P_0, with $P_{t=0} = P_0$ and score $s(O) = \partial_t \log p_t(O)|_{t=0}$. The efficient influence function (EIF) $\varphi(O; P_0)$ is the mean-zero, finite-variance element of the tangent space that represents every pathwise derivative:

$$\frac{\mathrm{d}}{\mathrm{d}t}\psi(P_t)\bigg|_{t=0} = \mathbb{E}_0\big[\varphi(O; P_0)s(O)\big], \qquad \mathbb{E}_0\big[\varphi(O; P_0)\big] = 0. \tag{36}$$

In a nonparametric model, this representation determines the EIF uniquely. In a semiparametric model, the EIF is the influence function in the tangent space with the smallest variance, and its variance gives the semiparametric efficiency bound (Bickel et al., 1998; van der Vaart, 1998; Kennedy, 2024).

A smooth functional admits the von Mises expansion

$$\psi(\overline{P}) - \psi(P) = \int \varphi(o; \overline{P})\, \mathrm{d}(\overline{P} - P)(o) + R_2(\overline{P}, P), \tag{37}$$

where $R_2(\overline{P}, P)$ is a second-order remainder composed of products or squares of differences between components of $\overline{P}$ and P. The EIF gives the linear response of ψ to a local perturbation of the data distribution, while R_2 collects the nonlinear residual. We set $\overline{P} = \widehat{P}$ and $P = P_0$ in Eq. (37). Since $\int \varphi(o; \widehat{P})\, \mathrm{d}\widehat{P}(o) = 0$, the plug-in error satisfies

$$\psi(\widehat{P}) - \psi(P_0) = -\mathbb{E}_0\big[\varphi(O; \widehat{P})\big] + R_2(\widehat{P}, P_0). \tag{38}$$

Thus, the first term in Eq. (38) is the first-order plug-in bias, conditional on the fitted distribution $\widehat{P}$. A direct plug-in estimator retains this term, so its bias can have the same first-order magnitude as the nuisance-estimation error. An EIF-based correction estimates and cancels this linear term, leaving only the second-order remainder (Kennedy, 2024).

In our setting, we derive the EIF for the average payoff because point evaluation of a conditional regression generally need not be pathwise differentiable when X is continuous (Kennedy, 2024). For a pair (j, k), this scalar functional is

$$\psi_{jk}(P) := \mathbb{E}_P\big[f_{jk}\big(\{\mu^m_{P,jk}(X)\}_{m \in \mathcal{Q}_{jk}}\big)\big], \tag{39}$$

where $\eta_P = (\boldsymbol{\mu}_P, \boldsymbol{e}_P)$ collects the outcome regressions and menu-selection propensities induced by P. Under MAR, positivity, and the regularity conditions of Lemma 2, its EIF is

$$\varphi_{jk}(O; P) = \Gamma_{jk}(O; \eta_P) - \psi_{jk}(P), \tag{40}$$

where Γ_{jk} is defined as

$$\Gamma_{jk}(O; \eta) = f_{jk}\big(\{\mu^m_{jk}(X)\}_{m \in \mathcal{Q}_{jk}}\big) + \sum_{m \in \mathcal{Q}_{jk}} \frac{\partial f_{jk}}{\partial u_m}\Big(\{\mu^{m'}_{jk}(X)\}_{m' \in \mathcal{Q}_{jk}}\Big) \frac{S^m}{e^m(X)}\big\{Z^m_{jk} - \mu^m_{jk}(X)\big\}. \tag{41}$$

D.2 One-step estimator and pseudo-outcome

Suppose that we estimate P_0 by $\widehat{P}$ on an auxiliary sample or through cross-fitting, and let $\mathbb{P}_n$ denote the empirical measure on the evaluation observations. The generic one-step estimator is

$$\widehat{\psi}^{\mathrm{1S}} := \psi(\widehat{P}) + \mathbb{P}_n\big[\varphi(O;\widehat{P})\big]. \tag{42}$$

Equation (38) shows why this correction works. The expectation of the added EIF term equals the negative of the first-order plug-in bias, so the remaining bias is $R_2(\widehat{P}, P_0)$.

For the average payoff, Eqs. (40) and (42) reduce the one-step estimator to

$$\widehat{\psi}^{\mathrm{DB}}_{jk} = \mathbb{P}_n\big[\Gamma_{jk}(O;\widehat{\eta})\big]. \tag{43}$$

The plug-in payoff inside Γ_{jk} and the centered EIF correction combine into a single sample average. This construction explains why we call Γ_{jk} an EIF-based pseudo-outcome: it is the uncentered EIF, $\Gamma_{jk}(O;\eta_P) = \psi_{jk}(P) + \varphi_{jk}(O;P)$.

Our target is the contextual payoff $A_{jk}(x)$ rather than only its population average. The pseudo-outcome makes this conditional target accessible because

$$\mathbb{E}_0\big[\Gamma_{jk}(O;\eta_0) \mid X = x\big] = A_{jk}(x). \tag{44}$$

For a generic nuisance η, the same conditional expectation defines the nuisance-indexed surface $A^{\mathrm{DB}}_{jk}(x;\eta)$ in Eq. (11). Theorem 1 shows that this surface equals $A_{jk}(x)$ at η_0 and has zero first derivative with respect to the nuisance. Equation (12) further shows that its conditional bias depends only on $\|\Delta\mu\|_2^2$ and products of Δe^m and $\Delta\mu^m$. We therefore regress the cross-fitted values $\Gamma_{jk}(O;\widehat{\eta})$ on X to obtain the estimate $\widehat{A}^{\mathrm{DB}}_{jk}(x)$ of the full contextual surface, which we call the **debiased payoff matrix (DPM)** estimator.

D.3 Orthogonal statistical learning

Classical one-step estimation typically targets a finite-dimensional functional such as ψ_{jk}. Our target $x \mapsto \pi^\star(x)$ is a function, so we use orthogonal statistical learning (OSL) to estimate it through a two-stage risk minimization problem (Foster & Syrgkanis, 2023). Let θ index a target function and let η denote nuisance functions estimated in the first stage. For a population risk $\mathcal{R}(\theta;\eta)$, Neyman-orthogonality requires the target gradient to be locally insensitive to nuisance perturbations:

$$D_\eta D_\theta \mathcal{R}(\theta;\eta)[h_\theta, h_\eta]\big|_{\eta=\eta_0} = 0. \tag{45}$$

When Eq. (45) holds for every θ, we call the risk universally orthogonal. This property removes the first-order effect of nuisance estimation from the learning objective rather than only from a scalar estimator.

Neyman-orthogonality. Neyman-orthogonality describes the derivative of a target moment or loss gradient with respect to nuisance functions. An EIF-based correction often constructs an orthogonal moment because it cancels the first-order nuisance derivative identified by the von Mises expansion. Therefore, the EIF can often be used to construct an orthogonal loss function for OSL.

We apply the same principle twice. First, the EIF-based pseudo-outcome yields the orthogonal payoff surface $A^{\mathrm{DB}}(x;\eta)$. Second, we insert this surface into the regularized gap risk

$$\mathcal{L}(\theta;\eta) = \mathbb{E}_X\left[G_\Omega\big(\pi_\theta(X); A^{\mathrm{DB}}(X;\eta), X\big)\right]. \tag{46}$$

The derivative $D_\eta A^{\mathrm{DB}}(x;\eta)[h_\eta]$ vanishes at η_0 by Theorem 1. The chain rule therefore makes the mixed derivative of Eq. (46) vanish for every θ, as formalized in Theorem 2. In practice, we cross-fit $\widehat{\eta}$, construct $\Gamma_{jk}(O;\widehat{\eta})$, regress these pseudo-outcomes on X to obtain $\widehat{A}^{\mathrm{DB}}$, and minimize the empirical gap loss to learn $\pi_{\widehat{\theta}}$.

Orthogonality changes the leading nuisance contribution from first order to second order. In our payoff learner, Eq. (12) gives the rate $O_p(r_\mu^2 + r_\mu r_e)$ under the stated norm conditions, instead of the $O_p(r_\mu)$ plug-in bias. For a linear aggregator, the term r_μ^2 vanishes, and the remaining product term gives the usual doubly robust property. This higher-order dependence permits slower and more flexible nuisance estimation while retaining the oracle learning rate when the conditions of Corollary 1 hold. This creates favorable empirical generalization properties and allows the use of general machine learning models.

E BENEFITS OF EQUILIBRIUM-BASED EVALUATION: A SOCIAL-CHOICE PERSPECTIVE

E.1 HOW DO CYCLES ARISE?

Here we give a simple example of how the cyclic preference arises during the aggregation of voters. Consider three equally weighted raters and three alternatives a, b, c, with preferences $a \succ_1 b \succ_1 c$, $b \succ_2 c \succ_2 a$, and $c \succ_3 a \succ_3 b$. Each rater is individually transitive. In the order (a, b, c), their majority-margin matrices are

$$A^{(1)} = \begin{pmatrix} 0 & 1 & 1 \\ -1 & 0 & 1 \\ -1 & -1 & 0 \end{pmatrix}, \quad A^{(2)} = \begin{pmatrix} 0 & -1 & -1 \\ 1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix}, \quad A^{(3)} = \begin{pmatrix} 0 & 1 & -1 \\ -1 & 0 & -1 \\ 1 & 1 & 0 \end{pmatrix}.$$

Aggregating the three raters gives

$$A = \frac{1}{3} \sum_{h=1}^{3} A^{(h)} = \frac{1}{3} \begin{pmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix},$$

so that $a \succ b$, $b \succ c$, and $c \succ a$. Thus, cycles can arise from aggregation even when no individual rater is intransitive.

E.2 WHY CAN AN EQUILIBRIUM POLICY BETTER ALIGN WITH HUMAN PREFERENCES?

In the profile above, each alternative is ranked first, second, and third by exactly one rater. A policy that respects this diversity should therefore treat the alternatives symmetrically. The equilibrium does so: for this problem, it is the uniform policy $\pi^\star = (1/3, 1/3, 1/3)$. In contrast, any deterministic selection (i.e., a pure policy) would arbitrarily favor one of the three equally represented groups.

E.3 SOCIAL CHOICE AXIOMS

The maximal-lottery rule satisfies several desirable social-choice axioms (Brandl et al., 2016):

- *Condorcet consistency:* If one alternative defeats every other alternative by a positive margin, the unique maximal lottery places all mass on it.
- *Population consistency:* If a lottery is maximal for two populations of raters separately, it remains maximal after their profiles are combined.
- *Composition consistency:* A component of similar alternatives can be evaluated internally without changing its relationship to alternatives outside the component. Thus, adding variants or copies of one agent does not distort the probabilities assigned to unrelated agents.
- *Cloning consistency:* Introducing a clone of an alternative does not change the probabilities assigned to alternatives outside the cloned component.

For agent evaluation, these axioms provide concrete operational guarantees. An agent that defeats every competitor is selected with probability one; combining feedback from user groups that share the same maximal lottery preserves that lottery; and adding duplicate or closely related variants of one agent does not change the probabilities assigned to unrelated agents. Thus, the evaluation remains consistent when feedback is pooled and when the candidate set contains redundant agent variants.

Adding regularization may change which axioms can be satisfied. For the KL-regularized equilibrium, Korkmaz (2026) establishes independence of irrelevant alternatives, population consistency, a regularized form of Pareto optimality, and approximate strategy-proofness. These guarantees motivate regularization in preference-based AI systems, but they depend on the choice of regularizer and do not generally hold for arbitrary strongly convex regularizers.

Table 5: Turning comparative feedback modalities into ordered partitions, illustrated on the menu $m = \{1, \ldots, 5\}$. Note that zero verdicts may have different interpretations for different modalities.

Modality	Raw response	Ordered partition Y_m	Zero verdicts
Pairwise duel	$1 \succ 3$	$(\{1\}, \{3\})$	a declared tie
Full ranking	$1 \succ 2 \succ 3 \succ 4 \succ 5$	$(\{1\}, \{2\}, \{3\}, \{4\}, \{5\})$	none arise
Weak ranking	$\{1, 2\} \succ \{3\} \succ \{4, 5\}$	$(\{1, 2\}, \{3\}, \{4, 5\})$	revealed indifference
Top-L ($L = 2$)	$T = \{1, 2\}$	$(\{1, 2\}, \{3, 4, 5\})$	censoring: the order inside T and inside its complement is never asked
Best–worst	best 1, worst 5	$(\{1\}, \{2, 3, 4\}, \{5\})$	censoring within the middle block
Cardinal scores	$s = (5, 5, 3, 1, 1)$	$(\{1, 2\}, \{3\}, \{4, 5\})$	equal scores

F Instantiations

F.1 A summary of different comparative feedback modalities

We summarize the different relative feedback modalities and how to convert them into ordered partitions in Table 5. We further provide an example of the conversion on a menu of five agents.

Example. Take the menu $m = \{1, \ldots, 5\}$ and the reported weak ranking $Y_m = (\{1, 2\} \succ \{3\} \succ \{4, 5\})$. Let $b_{Y_m}(j)$ denote the index of the block of Y_m containing agent j, so that $j \succ_{Y_m} k$ if and only if $b_{Y_m}(j) < b_{Y_m}(k)$. Here, $b_{Y_m}(1) = b_{Y_m}(2) = 1$, $b_{Y_m}(3) = 2$, and $b_{Y_m}(4) = b_{Y_m}(5) = 3$. Applying Eq. (3) to each of the $\binom{5}{2} = 10$ pairs gives

$$Z^m = \begin{pmatrix} 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ -1 & -1 & 0 & 1 & 1 \\ -1 & -1 & -1 & 0 & 0 \\ -1 & -1 & -1 & 0 & 0 \end{pmatrix}.$$

F.2 Instantiations of the aggregator function

We provide three instantiations of the aggregator f_{jk} in Definition 1. In addition to smoothness and oddness, we typically require idempotence, $f_{jk}(u, \ldots, u) = u$. Idempotence ensures that $A_{jk}(x) = u$ whenever every contributing menu has the same payoff $g^m_{jk}(x) = u$. The choice of aggregator determines how menu-specific preferences contribute to the contextual payoff and can therefore change the equilibrium.

When aggregation is useful. An aggregator is useful when the same pair appears in multiple overlapping menus and those menus provide distinct evidence about the pairwise preference. This occurs, for example, when evaluators report full rankings, weak rankings, or top-L choices over menus of different sizes. Linear weights can represent a prespecified deployment distribution over menus, balance menu sizes, or downweight menus known to produce less reliable or more heavily censored comparisons. A nonlinear aggregator can make the influence of a menu-specific margin depend on its magnitude. The cubic-link construction below, for example, gives larger absolute margins more influence on the transformed scale. In all these cases, the aggregator forms part of the estimand: different choices can produce different payoff matrices and equilibria when preferences depend on menu composition. Therefore, researchers should specify the weights independently of the observed outcomes or treat estimated weights as additional nuisance functions. Every menu assigned positive weight must also satisfy the menu-specific coverage and positivity assumptions.

Notes on pairwise feedback. When every menu contains exactly two agents, each covered pair (j, k) has the single menu $m = \{j, k\}$, so $\mathcal{Q}_{jk} = \{\{j, k\}\}$. No cross-menu aggregation is needed, and we use the identity map

$$f_{jk}(u_{\{j,k\}}) = u_{\{j,k\}}, \qquad A_{jk}(x) = g^{\{j,k\}}_{jk}(x). \tag{47}$$

Linear aggregator. Let $w_{jk}^m \geq 0$ be fixed menu weights satisfying $\sum_{m \in \mathcal{Q}_{jk}} w_{jk}^m = 1$ and $w_{kj}^m = w_{jk}^m$. We define

$$f_{jk}\big((u_m)_{m \in \mathcal{Q}_{jk}}\big) := \sum_{m \in \mathcal{Q}_{jk}} w_{jk}^m u_m. \tag{48}$$

This aggregator is smooth, odd, idempotent, and valued in $[-1, 1]$. Uniform weights recover the mean aggregator in Section 4, while nonuniform weights enable practitioners to customize the importance of each menu.

Nonlinear aggregator. We can apply a smooth odd link before averaging. For $\alpha \geq 0$, define the strictly increasing bijection $\phi_\alpha : [-1, 1] \to [-1, 1]$ by

$$\phi_\alpha(u) := \frac{u + \alpha u^3}{1 + \alpha}, \tag{49}$$

and use the link-transformed weighted aggregator

$$f_{jk}^{(\alpha)}\big((u_m)_{m \in \mathcal{Q}_{jk}}\big) := \phi_\alpha^{-1}\left(\sum_{m \in \mathcal{Q}_{jk}} w_{jk}^m \phi_\alpha(u_m)\right). \tag{50}$$

The oddness of ϕ_α and ϕ_α^{-1} makes $f_{jk}^{(\alpha)}$ odd. The normalized weights make it idempotent. Moreover, $\phi_\alpha'(u) = (1 + 3\alpha u^2)/(1 + \alpha) > 0$, so its inverse is smooth and the weighted average remains in its range. The choice $\alpha = 0$ recovers the linear aggregator in Eq. (48). When $\alpha > 0$ and at least two menus receive positive weight, the construction generally gives a nonlinear aggregation rule. Unlike the linear case, this nonlinear aggregator can have nonzero curvature, so the quadratic Taylor remainder in Eq. (12) need not vanish.

F.3 Instantiations of the cost function

In the main text, we use the linear expected cost $\Phi(\pi) = c^\top \pi$, where $c \in \mathbb{R}_+^K$ contains the deployment costs. We next give a nonlinear soft-budget extension and an exact hard-budget formulation.

Convex cost disutility. Let $\varphi : \mathbb{R}_+ \to \mathbb{R}$ be a continuously differentiable, convex, and nondecreasing disutility of expected spend. We define

$$\Phi(\pi) := \varphi(c^\top \pi), \qquad \Omega(\pi) := \Omega_0(\pi) + \Phi(\pi), \tag{51}$$

where Ω_0 is the differentiable, κ-strongly convex base regularizer. Convexity of Φ preserves the strong-convexity modulus of Ω_0. A smooth family that interpolates between linear pricing and a soft budget is

$$\varphi(s) := \lambda_0 s + \bar{\lambda}\gamma \log\left(1 + \exp\left(\frac{s - B}{\gamma}\right)\right), \qquad \varphi'(s) = \lambda_0 + \bar{\lambda}\,\sigma\left(\frac{s - B}{\gamma}\right), \tag{52}$$

where $\lambda_0, \bar{\lambda} \geq 0$, B is the soft budget, $\gamma > 0$ controls the transition width, and $\sigma(t) = (1 + e^{-t})^{-1}$. The marginal price therefore increases smoothly from λ_0 to $\lambda_0 + \bar{\lambda}$ as expected spend crosses B. The choice $\bar{\lambda} = 0$ recovers linear pricing, while the limit $\gamma \downarrow 0$ yields the hinge disutility $\lambda_0 s + \bar{\lambda}(s - B)_+$.

Hard cost budget. Let $C \geq \min_{j \in \mathcal{M}} c_j$ and define the nonempty compact convex strategy set

$$\Delta_K(C) := \{\pi \in \Delta_K : c^\top \pi \leq C\}. \tag{53}$$

We impose the budget on both players and define the constrained contextual equilibrium as

$$\pi_C^\star(x) := \operatorname*{arg\,max}_{\pi \in \Delta_K(C)} \min_{q \in \Delta_K(C)} \left\{\pi^\top A(x) q - [\Omega_0(\pi) - \Omega_0(q)]\right\}. \tag{54}$$

Using the same feasible set for both players preserves antisymmetry and the zero game value. Compactness and convexity give existence, while strong convexity of Ω_0 gives uniqueness by the same argument as in Appendix C. Alternatively, we can also choose a large $\bar{\lambda}$ (e.g., $\bar{\lambda} = 100$) and still use the unconstrained optimization with Eq. (51). This yields an approximately hard budget formulation and usually is easier to solve.

G SIMULATED AGENT-EVALUATION DATASET

In this section, we construct a simulated agent-evaluation dataset with context-dependent agent performance and heterogeneous voter preferences. The data-generating process (DGP) generates pairwise comparisons, partial rankings, or winner-only feedback and provides the exact contextual payoff matrix for each modality, making precise evaluation possible.

G.1 CONTEXT AND VECTOR PERFORMANCE

We draw contexts independently as $X_i \sim \mathrm{Unif}([0,1]^d)$ with $d = 5$ and use $K = 5$ agents. Agent j has a performance vector $\psi_j(x) \in \mathbb{R}^p$ with $p = 12$. We partition its coordinates into three groups $\mathcal{G}_g = \{4g+1, \dots, 4g+4\}$ indexed by $g \in \{0,1,2\}$. For each coordinate $t \in \mathcal{G}_g$, we define

$$\psi_{jt}(x) = a_s s_j(x) + a_c C_{jt} + a_h \tanh\{h_{jt}(x)\}, \qquad (a_s, a_c, a_h) = (1, 2, 0.15). \tag{55}$$

Here, $s_j(x)$ represents quality shared across criteria, C_{jt} represents group specialization, and $h_{jt}(x)$ represents criterion-specific contextual variation.

We generate s_j and h_{jt} from two independent random Fourier feature fields. For a generic output a, we define the unstandardized field as

$$\begin{aligned} f_a(x) &= \sqrt{\frac{2}{D}} \sum_{b=1}^{D} V_{ba} \cos(\omega_b^\top x + \phi_b), \\ \omega_b \sim \mathcal{N}(0, \ell_f^{-2} I_d), &\qquad \phi_b \sim \mathrm{Unif}(0, 2\pi), \qquad V_{ba} \sim \mathcal{N}(0,1). \end{aligned} \tag{56}$$

We draw all frequencies, phases, and coefficients independently. Within each field, the outputs share frequencies and phases but use separate coefficient columns. The shared-quality and criterion-specific fields have K and Kp outputs, respectively. We use $D = 128$ and $\ell_f = 0.5$, then center and scale each output to unit empirical standard deviation on 4,096 independent uniform calibration contexts. The standardized outputs define $s_j(x)$ and $h_{jt}(x)$.

To specify specialization, we draw a uniform permutation $(\varpi_1, \dots, \varpi_K)$ of $\{0, \dots, K-1\}$ and set $(o_0, o_1, o_2) = (0, 2, 4)$. Then

$$C_{jt} = 2 - \frac{2}{K-1}\big[(\varpi_j - o_g) \bmod K\big], \qquad t \in \mathcal{G}_g. \tag{57}$$

We hold all latent field draws and specialization profiles fixed within each replication.

G.2 VOTERS AND POPULATION PREFERENCES

Voter v assigns utility

$$u_{vj}(x) = w_v^\top \psi_j(x), \qquad w_v \in \Delta_p. \tag{58}$$

The population contains $N_v \in \{1, 3, 5\}$ voters. We draw a pool of five voters with preferred groups $(g_1, \dots, g_5) = (0, 1, 2, 0, 1)$ and use its first N_v members. For each voter, we independently draw $\theta_v \sim \mathrm{Dirichlet}(\alpha \mathbf{1}_4)$ with $\alpha = 20$. We construct $\widetilde{\theta}_v \in \mathbb{R}^p$ by placing these four coordinates in $\mathcal{G}_{g_v}$ and setting the remaining coordinates to zero. We then set

$$w_v = \frac{\epsilon}{p}\mathbf{1}_p + (1-\epsilon)\widetilde{\theta}_v, \qquad \epsilon = 0.1. \tag{59}$$

Each member of the finite population has weight $1/N_v$, and each response samples a voter uniformly from this population.

For distinct utilities, we define the latent population margin as

$$M_{jk}(x) = \frac{1}{N_v} \sum_{v=1}^{N_v} \mathrm{sign}\{u_{vj}(x) - u_{vk}(x)\}, \qquad M_{kj}(x) = -M_{jk}(x), \quad M_{jj}(x) = 0. \tag{60}$$

The continuous field draws make exact utility ties a probability-zero event, and the implementation resolves numerical ties by agent index. With $N_v = 1$, $M(x)$ follows the voter's transitive ordering. With $N_v \in \{3, 5\}$, majority aggregation can generate cycles. The populations of sizes three and five contain preferred-group counts $(1, 1, 1)$ and $(2, 2, 1)$, respectively.

G.3 Menus, selection, and feedback

A menu m contains the agents presented to one voter. Pairwise feedback uses all ten menus of size two. Partial-ranking feedback uses the five menus

$$\mathcal{Q}_{\text{rank}} = \big\{\{1,2,3\},\{1,2,5\},\{1,4,5\},\{2,3,4\},\{3,4,5\}\big\}. \tag{61}$$

This catalog covers every pair. Winner feedback uses all $\binom{5}{3} = 10$ menus of size three. A pair belongs to one or two ranking menus and exactly three winner menus.

For each replication, we draw $b \sim \mathcal{N}(0, I_d)$ and set $\bar{b} = b/\|b\|_2$. We define $\sigma(z) = (1+e^{-z})^{-1}$ and

$$\begin{aligned} e_0^m(x) = e_0(x) = e_{\min} + (e_{\max} - e_{\min})\sigma\Big(\gamma\sqrt{12}\,\bar{b}^\top(x - \tfrac{1}{2}\mathbf{1}_d)\Big), \\ S_i^m \mid X_i \sim \text{Bernoulli}\{e_0(X_i)\}, \qquad (e_{\min}, e_{\max}, \gamma) = (0.25, 0.75, 2). \end{aligned} \tag{62}$$

We draw menu selections independently conditional on the context and independently of voter responses. Thus, the DGP satisfies MAR conditional on X. Each $e_0^m(x)$ is a menu inclusion probability, so one context can generate multiple records and the probabilities need not sum to one. The lower bound $e_{\min} = 0.25$ bounds inverse propensities by four.

For each selected (i, m), we independently sample $V_{im} \sim \text{Unif}\{1, \ldots, N_v\}$. The selected voter orders the agents in m by $u_{V_{im},j}(X_i)$. Pairwise feedback reports the preferred agent, partial-ranking feedback reports the complete order, and winner feedback reports only the highest-utility agent. All comparisons extracted from one response use the same voter. Records at the same context sample voters independently.

For a pair $j, k \in m$, the encoded margin from pairwise or ranking feedback is

$$Z_{jk}^m = \text{sign}\{u_{V_{im},j}(X_i) - u_{V_{im},k}(X_i)\}. \tag{63}$$

For winner feedback, we set $W_{im} = \arg\max_{a \in m} u_{V_{im},a}(X_i)$ and define

$$Z_{jk}^m = \mathbf{1}\{W_{im} = j\} - \mathbf{1}\{W_{im} = k\}. \tag{64}$$

Eq. (64) assigns zero when another menu member wins because winner-only feedback does not reveal the relative order of j and k.

G.4 Ground-truth game and experimental parameters

We let $\mathcal{Q}_{jk} = \{m : j, k \in m\}$ and $g_{jk}^m(x) = \mathbb{E}[Z_{jk}^m \mid X = x]$. Because the DGP satisfies MAR conditional on X and positivity ($e_0^m \geq e_{\min} > 0$), we have $g_{jk}^m(x) = \mu_{0,jk}^m(x)$. We define the ground-truth game by averaging uniformly over the menus containing each pair:

$$A_{jk}(x) = \sum_{m \in \mathcal{Q}_{jk}} \lambda(m \mid j,k) g_{jk}^m(x), \qquad \lambda(m \mid j,k) = \frac{1}{|\mathcal{Q}_{jk}|}. \tag{65}$$

For pairwise and ranking feedback, $g_{jk}^m(x) = M_{jk}(x)$ on every containing menu, so $A(x) = M(x)$. For winner feedback, we define

$$p_j^m(x) = \frac{1}{N_v}\sum_{v=1}^{N_v} \mathbf{1}\left\{j = \arg\max_{a \in m} u_{va}(x)\right\}, \qquad g_{jk}^m(x) = p_j^m(x) - p_k^m(x). \tag{66}$$

Table 6 gives the parameters used in Figure 3. Within each replication, we share the performance fields, specialization permutation, selection direction, voter pool, and training and test contexts across the nine modality–population-size settings.

Table 6: Parameters of the simulated dataset.

Parameter	Value	Meaning
K, d, p	$5, 5, 12$	Agents, context dimensions, and performance criteria.
X	$\mathrm{Unif}([0,1]^5)$	Independent training and test contexts.
D, ℓ_f	$128, 0.5$	Fourier features and field length scale.
n_{cal}	4,096	Uniform contexts for field standardization.
(a_s, a_c, a_h)	$(1, 2, 0.15)$	Shared quality, specialization, and criterion variation amplitudes.
(o_0, o_1, o_2)	$(0, 2, 4)$	Offsets of the three specialization profiles.
N_v	$1, 3, 5$	Equally weighted voters, drawn as prefixes of one pool.
$(g_1, \ldots, g_5)$	$(0, 1, 2, 0, 1)$	Preferred criterion groups in the voter pool.
ϵ, α	$0.1, 20$	Uniform weight allocation and Dirichlet concentration.
Menu sizes	$2, 3, 3$	Pairwise, partial-ranking, and winner feedback.
Menu counts	$10, 5, 10$	Catalogs for the three respective modalities.
$(e_{\min}, e_{\max}, \gamma)$	$(0.25, 0.75, 2)$	Inclusion bounds and selection logit scale.
n, n_{test}	20,000, 4,096	Training and test contexts per experiment.
Test replications	5	Seeds 70–74, each covering all nine settings.

H Details of the real-world datasets

We describe the data sources, context variables, agent populations, and feedback protocols used in our application experiments. In Section H.1, we describe the human preference data from LMArena and the context attributes we use to examine heterogeneity across user prompts and tasks. In Section H.2, we explain how we combine benchmark scores with LLM judgments to construct offline pairwise feedback and complete test reference matrices for evaluating contextual policies and their performance under domain shift. In Section H.3, we describe how we train driving agents with different reward weights and construct comparisons from simulated judges with distinct utilities. This setting extends our evaluation beyond language models to agents that trade off multiple performance objectives. We distinguish observed human judgments in LMArena from the constructed preferences in RouterBench and the driving simulator, and provide the data-construction details needed to interpret and reproduce these experiments.

H.1 LMArena dataset

LMArena is a crowdsourced platform in which a (human) user submits a prompt, receives responses from two anonymously presented language models, and indicates which response is better, whether the responses tie, or whether both are unsatisfactory (Chiang et al., 2024). The Arena Human Preference 140K dataset records these votes together with the conversation histories and associated metadata. The dataset fits our framework: the observed menus are the pairs $m = \{j, k\}$, with pairwise preference Y_m. The prompt and its metadata constitute the context X. For each context, only one menu's outcome is selected, and the others are not observed. The crowdsourced votes span 126 detected language labels and a variety of tasks, providing naturally heterogeneous evaluation contexts and human preferences.

We use the Arena Human Preference 140K snapshot containing 135,634 votes collected from April 17, 2025 to July 24, 2025, covering 53 models. Our experiments retain the 15 models listed in Table 9 to ensure minimal coverage over the pairs. For each comparison, we construct the context from the conversation and salient metadata, including language, task-category annotations (e.g., mathematics, creative writing, instruction following, and hard prompts), and whether the conversation involves code.

Tables 7 and 8 report the marginal distributions of the principal discrete context attributes. We aggregate the recorded language labels into five categories: English, Polish, Russian, Chinese (including both `zh` and `zh-Hant`), and Other. The code, instruction-following, and mathematics indicators are taken directly from the dataset annotations.

Table 7: Marginal distribution of language groups.

Language group	Count	Share
English	71,175	52.5%
Polish	13,813	10.2%
Russian	9,263	6.8%
Chinese	7,346	5.4%
Other	34,037	25.1%

Table 8: Marginal distributions of binary context attributes.

Attribute	Value	Count	Share
Code	Yes	39,363	29.0%
	No	96,271	71.0%
Instruction following	Yes	24,666	18.2%
	No	110,968	81.8%
Mathematics	Yes	10,892	8.0%
	No	124,742	92.0%

Table 9: Selected agents from the Arena Human Preference 140K dataset, with list prices in USD per 10^6 tokens.

Provider	Agent	Input	Output
OpenAI	`o3-2025-04-16`	2.00	8.00
	`o4-mini-2025-04-16`	1.10	4.40
	`gpt-4.1-2025-04-14`	2.00	8.00
Anthropic	`claude-opus-4-20250514`	15.00	75.00
	`claude-sonnet-4-20250514`	3.00	15.00
Google	`gemini-2.5-pro`	1.25	10.00
	`gemini-2.5-flash`	0.30	2.50
Alibaba (Qwen)	`qwen3-235b-a22b-instruct-2507`	0.09	0.55
	`qwen3-30b-a3b`	0.12	0.50
Moonshot AI	`kimi-k2-0711-preview`	0.57	2.30
MiniMax	`minimax-m1`	0.55	2.20
Mistral AI	`mistral-medium-2505`	0.40	2.00
DeepSeek	`deepseek-r1-0528`	0.50	2.15
	`deepseek-v3-0324`	0.25	1.00
Meta	`llama-4-maverick-17b-128e-instruct`	0.20	0.80

H.2 ROUTERBENCH DATASET

RouterBench evaluates LLM routers by running multiple models on a common collection of prompts and recording their responses, task-specific correctness scores, and inference costs (Hu et al., 2024). We use the LLMRouterBench collection (Li et al., 2026), which follows this protocol, and retain the $K = 10$ flagship models in Table 11 with complete coverage across ten subsets (see Table 10). The resulting benchmark contains $8{,}742$ prompts. The context X includes the prompt, source subset, task domain, mathematics and code indicators, language group, and token length, following the LMArena context schema in Section H.1.

RouterBench provides absolute scores rather than relative feedback. A score $s_j(x)$ induces $Z_{jk}(x) = \operatorname{sign}\{s_j(x) - s_k(x)\}$, which is therefore transitive by construction. We use this verdict when the native scores distinguish a pair and elicit a preference only when both models succeed (if both fail, then we simply treat it as a both-bad verdict). For each prompt, we sample $Q = 20$ of the $\binom{10}{2} = 45$ pairs uniformly without replacement, independently across prompts, giving the known propensity $e_0^m(x) = 20/45$ and satisfying positivity.

A sampled pair (j, k) at prompt x is then resolved as

$$Z_{jk}(x) = \begin{cases} \operatorname{sign}(s_j - s_k), & s_j \neq s_k \quad \text{(the benchmark orders the pair)}, \\ 0, & s_j = s_k = 0 \quad \text{(both agents fail)}, \\ \text{judge verdict}, & s_j = s_k > 0 \quad \text{(both agents succeed)}, \end{cases} \tag{67}$$

so a judge is used only when both models succeed. We draw judges from three of the latest open-source frontier models supplied by different providers and assign them distinct emphases—rigor, depth, or efficiency—within a common six-metric rubric. This emulates the real-world rater diversity and heterogeneity, which also might introduce intransitive feedback. Each pair is judged in both presentation orders and the verdicts are averaged to mitigate position bias.

We select 500 prompts from the benchmark as a separate test set, unseen during training. All 45 pairs on the 500 test prompts are evaluated by all three judges, yielding a complete reference payoff matrix. Each of the $8{,}242$ training prompts receives the 20 sampled pairs and one randomly assigned judge. Overall, the processing yields $232{,}340$ comparisons, including $67{,}500$ test comparisons. Table 10 shows the composition. Knowledge tasks account for 67.0% of prompts, programming 17.8%, chat and instruction following 8.6%, logic 4.6%, and mathematics 2.1%. In addition, 20.7% involve code and 4.9% mathematics. Unlike LMArena, RouterBench is 96.0% English, with some Chinese, Japanese, and Korean (CJK) prompts.

Table 10: Composition of the RouterBench panel. "Feedback modality" is the benchmark's native outcome format; the last two columns give the share of sampled pairs that the label already orders and the share on which both agents succeed and a judge is therefore consulted.

Subset	Prompts	Test	Domain	Feedback modality	Ord.	Judged
MMLU-Pro	3,000	50	knowledge	Binary (correct/incorrect)	13.1%	75.5%
HLE	2,158	50	knowledge	Binary (correct/incorrect)	17.1%	3.0%
LiveCodeBench	1,055	50	programming	Binary (pass/fail)	19.5%	60.4%
ArenaHard	750	100	chat / IF	Ternary (win/tie/loss)	48.7%	42.6%
SimpleQA	500	50	knowledge	Ternary (correct/incorrect/not attempted)	30.1%	19.5%
SWE-bench	500	50	programming	Binary (pass/fail)	24.9%	10.8%
ARC-AGI	400	50	logic	Binary (exact/non-match)	23.2%	9.8%
GPQA	198	50	knowledge	Binary (correct/incorrect)	22.3%	63.5%
LiveMathBench	121	30	math	Binary (correct/incorrect)	21.9%	54.1%
AIME	60	20	math	Binary (correct/incorrect)	28.4%	56.5%
Total	**8,742**	**500**			22.1%	41.3%

H.3 REINFORCEMENT-LEARNING AGENT DATASET

We construct a contextual agent-evaluation dataset using `mo-highway-fast-v0`, the multi-objective highway-driving environment in MO-Gymnasium (Felten et al., 2023), built on High-

Table 11: Selected agents from the RouterBench dataset, with list prices in USD per 10^6 tokens. †`gpt-5-chat` is no longer served; its price is taken from the benchmark's own records.

Provider	Agent	Input	Output
OpenAI	`gpt-5` (medium)	1.25	10.00
	`gpt-5-chat`†	1.25	10.00
Anthropic	`claude-sonnet-4`	3.00	15.00
Google	`gemini-2.5-pro`	1.25	10.00
	`gemini-2.5-flash`	0.30	2.50
Alibaba	`qwen3-235b-a22b-thinking-2507`	0.23	2.30
DeepSeek	`deepseek-v3-0324`	0.25	1.00
	`deepseek-r1-0528`	0.50	2.15
Z-AI	`glm-4.6`	0.43	1.75
Moonshot AI	`kimi-k2-0905`	0.60	2.50

wayEnv (Leurent, 2018). An agent controls a vehicle among simulated traffic and receives separate rewards for speed, right-lane occupancy, and collision avoidance. The collision component is a nonpositive penalty. The tasks are contextual, containing five configuration features: lane count, vehicle count, traffic density, initial lane, and initial ego spacing. We sample lane and vehicle counts uniformly from $\{2, 3, 4, 5\}$ and $\{20, \dots, 60\}$, density from $U[0.7, 1.5]$, the initial lane uniformly among valid lanes, and ego spacing from $U[0.5, 3]$, then fix the configuration for the episode.

The driving task is inherently multi-objective: for example, agents can trade off speed against safety. We simulate external judges through linear utilities over an episode-performance vector $z_j = (\overline{r}_{\text{speed}}, \overline{r}_{\text{lane}}, 1 - \mathbf{1}\{\text{collision}\}) \in [0, 1]^3$. Three equally weighted judges use $U_h(j) = v_h^\top z_j$, with weights approximately $(0.67, 0.31, 0.02)$, $(0.04, 0.01, 0.95)$, and $(0.27, 0.01, 0.72)$. Each offline comparison samples a pair menu $m = \{j, k\}$ uniformly from the ten unordered pairs and a judge uniformly from the three judges. Both agents are then evaluated on the same contextual task. The signed verdict Z_{jk}^m is $+1$ if the judge prefers j and -1 otherwise, with exact utility ties broken by a fair coin. Because menus are sampled uniformly, $e_0^m(x) = 1/10$ per comparison. Our target $A_{jk}(x)$ further averages over the initialization scenarios.

We train agents with the scalar reward $w_j^\top r$, where w_j specifies agent j's training weights on the reward vector $r = (r_{\text{speed}}, r_{\text{lane}}, r_{\text{collision}})$. We apply deep Q-learning (Mnih et al., 2015) through Stable-Baselines3 (Raffin et al., 2021). Each agent receives $20{,}000$ interaction steps and uses a Q-network with two hidden layers of 256 units. We adapt the official HighwayEnv DQN example,[4] using learning rate 5×10^{-4}, replay capacity $15{,}000$, batch size 32, and target updates every 50 steps. The final five agents use weights $(0.1, 0.8, 0.1)$, $(0.8, 0.1, 0.1)$, $(0.45, 0.1, 0.45)$, $(0.02, 0, 0.98)$, and $(0.2, 0, 0.8)$, respectively. The first three use discount factor 0.8. The last two use 0.99 to emphasize longer-term safety. These design choices make sure that the agents trained have a range of varying capabilities.

We use 256 training configurations, with eight initializations per configuration. We generate ten pairwise comparisons per initialization, yielding $20{,}480$ offline comparisons. We use separate validation and test sets of 32 configurations each, with 32 and 64 initializations per configuration, respectively. For each validation or test configuration, we evaluate all five agents and average all three judges' verdicts to estimate the complete payoff matrix, without assigning ties. The estimated contextual payoff matrices contain a strict three-agent preference cycle in 11 of the 32 test configurations.

[4] `https://highway-env.farama.org/quickstart/`

I IMPLEMENTATION

In Section I.1, we define the evaluation metrics used in our experiments. In Section I.2, we specify the training details of our NASHEVAL. In Section I.3, we describe the baselines.

We summarize our method (i.e., **NashEval-debiased**) in Algorithm 1.

Algorithm 1 NashEval-debiased: contextual equilibrium learning from selective feedback

Require: Training log $\mathcal{D}_n = \{O_i\}_{i=1}^n$, menu aggregators $\{f_{jk}\}_{j<k}$, context-independent strongly convex regularizer Ω, number of folds K_{CF}, and policy class $\{\pi_\theta : \theta \in \Theta\}$.

Encode each observed $Y_{i,m}$ as verdicts $Z_{i,jk}^m$ using Eq. (3).
Partition the training indices into folds $I_1, \dots, I_{K_{\mathrm{CF}}}$, keeping all records from each context together.
for $b = 1, \dots, K_{\mathrm{CF}}$ **do**
  for each menu $m \in \mathcal{Q}$ **do**
    Fit $\widehat{e}^{m,(-b)}(x)$ to S_i^m on all $i \notin I_b$ using binary log loss.
    Fit $\widehat{\mu}_{jk}^{m,(-b)}(x)$ to $Z_{i,jk}^m$ for each $j < k$ in m using squared error on $i \notin I_b$ with $S_i^m = 1$.
  end for
  for each $i \in I_b$ and pair $j < k$ **do**
    Set $\widehat{\boldsymbol{\mu}}_{i,jk}^{(-b)} = \left(\widehat{\mu}_{jk}^{m,(-b)}(X_i)\right)_{m \in \mathcal{Q}_{jk}}$.
    Construct the out-of-fold target from Eq. (41):

$$\widehat{\Gamma}_{i,jk} = f_{jk}\left(\widehat{\boldsymbol{\mu}}_{i,jk}^{(-b)}\right) + \sum_{\substack{m \in \mathcal{Q}_{jk} \\ S_i^m = 1}} \frac{\partial f_{jk}}{\partial u_m}\left(\widehat{\boldsymbol{\mu}}_{i,jk}^{(-b)}\right) \frac{Z_{i,jk}^m - \widehat{\mu}_{jk}^{m,(-b)}(X_i)}{\widehat{e}^{m,(-b)}(X_i)}.$$

  end for
end for
Pool the targets from all folds and fit a shared payoff regressor:

$$\widehat{a} \in \arg\min_a \sum_{i=1}^n \sum_{j<k} \left\{a(X_i, j, k) - \widehat{\Gamma}_{i,jk}\right\}^2.$$

Define $\widehat{A}_{jk}^{\mathrm{DB}}(x) = \mathrm{clip}_{[-1,1]}\{\widehat{a}(x, j, k)\}$ for $j < k$, and set $\widehat{A}_{kj}^{\mathrm{DB}}(x) = -\widehat{A}_{jk}^{\mathrm{DB}}(x)$ and $\widehat{A}_{jj}^{\mathrm{DB}}(x) = 0$.
Minimize the empirical gap loss to learn $\pi_{\widehat{\theta}}$:

$$\widehat{\mathcal{L}}_n(\theta) = \frac{1}{n} \sum_{i=1}^n G_\Omega\left(\pi_\theta(X_i); \widehat{A}^{\mathrm{DB}}(X_i), X_i\right).$$

return $\widehat{A}^{\mathrm{DB}}$ and $\pi_{\widehat{\theta}}$.
Inference: For a new context $x_\star$, return $\pi_{\widehat{\theta}}(x_\star) \in \Delta_K$.

We evaluate the residual correction only for observed menus, so the algorithm never requires an unobserved verdict. We retain the full pseudo-outcome values during payoff regression and clip only the fitted payoff predictions. We use LightGBM for the nuisance and payoff models and the multilayer perceptron described in Section I.2 for the policy. For experiments that compute equilibria directly, we replace policy training and prediction by a numerical equilibrium solver applied to $\widehat{A}^{\mathrm{DB}}(x)$ with the experiment's regularizer.

I.1 EVALUATION METRICS

We evaluate a policy π on n_{test} contexts x_i using reference payoff matrices A_i^{ref}, with positive entries indicating preference for the row agent over the column agent. We use the true matrices in synthetic experiments and approximated matrices in RouterBench and MO-Highway.

We compute the *reference equilibria* π_i^{ref} from these matrices. In RouterBench and MO-Highway, these reference equilibria are close approximations to the true contextual equilibria. We use *reference evaluation* for the distributions over agents given by these equilibria and *reference contextual winners* for the agents in their thresholded supports.

Average exploitability. Exploitability measures the largest advantage an opponent can obtain against the evaluated policy. We average the unregularized best-response payoff over test contexts:

$$\widehat{\mathrm{Expl}}_0(\pi) = \frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \max_{q \in \Delta_K} q^\top A_i^{\mathrm{ref}} \pi(x_i). \tag{68}$$

Lower is better, and we allow an unrestricted opponent even when we impose a cost penalty on the evaluated policy.

Support F_1. Support F_1 measures how accurately a policy recovers the reference winner set, without comparing the assigned probabilities. We define $S_\tau(p) = \{j : p_j > \tau\}$ and average the overlap with the reference equilibrium π_i^{ref}:

$$\widehat{F}_1(\pi) = \frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \frac{2|S_\tau(\pi(x_i)) \cap S_\tau(\pi_i^{\mathrm{ref}})|}{|S_\tau(\pi(x_i))| + |S_\tau(\pi_i^{\mathrm{ref}})|}. \tag{69}$$

Higher is better. We use $\tau = 10^{-3}$ for both predicted and reference supports in the synthetic, RouterBench, and MO-Highway experiments. We compute the reference equilibrium using each experiment's specified regularization.

Head-to-head game value. Game value measures the evaluated policy's net preference advantage against a specified opponent policy. We report the average payoff of π against q:

$$\widehat{V}(\pi, q) = \frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \pi(x_i)^\top A_i^{\mathrm{ref}} q(x_i). \tag{70}$$

In the RQ2 tables, we take q to be NashEval-debiased, so positive values favor the row method and negative values favor NashEval-debiased.

Average strict win rate. Strict win rate measures how often a sampled agent wins against an equally weighted pool of single-agent opponents. We let $W_{i,jk}$ denote the reference probability that agent j strictly defeats agent k at x_i and compute

$$\widehat{\mathrm{WR}}(\pi) = \frac{1}{n_{\mathrm{test}} K} \sum_{i=1}^{n_{\mathrm{test}}} \sum_{j=1}^{K} \sum_{k=1}^{K} \pi_j(x_i) W_{i,jk}. \tag{71}$$

We assign ties and self-comparisons zero credit and retain them in the denominator, so payoff margins alone do not determine this metric when ties occur.

I.2 Training details

We use LightGBM to estimate the outcome regressions $\mu_{0,jk}^m$ and selection propensities e_0^m, and to fit the debiased payoff matrix $\widehat{A}^{\mathrm{DB}}$. We split training contexts into K_{CF} folds and keep all records from the same context in one fold. For each fold, we fit the nuisance models on the remaining folds and predict the nuisances on the held-out fold to construct its pseudo-outcomes. We then pool these out-of-fold pseudo-outcomes to train the payoff regressor.

Sentence embedding. For contexts that include text, we use the fixed Qwen3-Embedding-4B encoder (Zhang et al., 2025a) to map each prompt to $f_{\mathrm{emb}}(x) \in \mathbb{R}^d$. We compare Qwen3-Embedding-0.6B, Qwen3-Embedding-4B, and EmbeddingGemma-300M (Schechter Vera et al., 2025) with embedding dimensions from 32 to 1024. In these comparisons, we observe little variation in nuisance prediction accuracy across encoders and only small gains from larger embeddings. We therefore use Qwen3-Embedding-4B with $d = 64$ to limit encoding and storage costs.

Nuisance estimation. For each menu m and pair $j < k$ in m, we fit a separate LightGBM regressor to estimate $\mu_{0,jk}^m(x)$ by minimizing squared error on records with $S^m = 1$. We impose $\widehat{\mu}_{kj}^m(x) = -\widehat{\mu}_{jk}^m(x)$ and $\widehat{\mu}_{jj}^m(x) = 0$. For each menu, we fit a LightGBM binary classifier to estimate $e_0^m(x)$ from the inclusion indicator S^m using binary log loss on all training contexts. We cross-fit both nuisance estimators as described above.

Payoff regression. We fit a shared LightGBM regressor to the cross-fitted pseudo-outcomes $\Gamma_{jk}(O;\widehat{\eta})$ using (x, j, k) as inputs and treating j and k as categorical features. We minimize squared error to estimate the conditional mean of these targets and obtain $\widehat{A}^{\mathrm{DB}}(x)$. This regression targets the debiased payoff surface rather than directly observed entries of the ground-truth matrix $A(x)$.

Hyperparameter tuning. We tune the nuisance estimators and the shared payoff regressor by grid search with three-fold cross-validation. Tables 12 and 13 give the respective search spaces. We use each grid configuration's tree count as the maximum number of boosting iterations and select the stopping iteration by early stopping. We distinguish these tuning folds from the K_{CF} folds used to construct out-of-fold nuisance predictions. For each cross-fitting fold, we tune each nuisance model using three-fold cross-validation confined to that fold's training contexts. Note that we use a slightly more complex search space for the payoff regressor than for the nuisance estimators. In the synthetic experiment, we fix the hyperparameters to the values in the last columns of Tables 12, 13, and 14.

Table 12: Hyperparameter search space for the nuisance estimators. The index $\nu \in \{\mu, e\}$ denotes the outcome or propensity model. The last column gives the fixed values for μ and e in the synthetic experiment.

Hyperparameter	Symbol	Search space	Synthetic (μ / e)
Learning rate	lr_ν	$\{0.01, 0.1\}$	0.1 / 0.1
Maximum leaves per tree	$n_{\mathrm{leaf},\nu}$	$\{7, 15, 31\}$	7 / 15
Maximum tree depth	$d_{\max,\nu}$	$\{7, 10\}$	7 / 7
Minimum observations per leaf	$n_{\min,\nu}$	$\{30, 50, 100, 200\}$	100 / 50
L_2 penalty on leaf values	$\lambda_{2,\nu}$	$\{0, 0.1, 1\}$	0 / 0
Feature fraction per tree	f_ν	$\{0.6, 0.8, 1.0\}$	1.0 / 1.0
Number of trees	T_ν	$\{50, 100, 200\}$	100 / 100

Table 13: Hyperparameter search space for the shared payoff regressor. The last column gives the fixed values in the synthetic experiment.

Hyperparameter	Symbol	Search space	Synthetic
Learning rate	lr_A	$\{0.02, 0.05, 0.1\}$	0.05
Maximum leaves per tree	$n_{\mathrm{leaf},A}$	$\{15, 31, 62\}$	31
Maximum tree depth	$d_{\max,A}$	$\{10, 15\}$	10
Minimum observations per leaf	$n_{\min,A}$	$\{50, 100, 200\}$	100
L_2 penalty on leaf values	$\lambda_{2,A}$	$\{0, 0.1, 1\}$	0
Feature fraction per tree	f_A	$\{0.6, 0.8, 1.0\}$	1.0
Number of trees	T_A	$\{50, 100, 200, 400\}$	400

Equilibrium learning. We parameterize the equilibrium predictor $\pi_\theta(x)$ as a multilayer perceptron with two hidden layers, ReLU activations, and a K-dimensional sparsemax output, which maps each context to a distribution over agents by Euclidean projection onto Δ_K. We scale the initial weights of the output layer by 0.01 and set its initial bias to zero, so training starts from nearly uniform policies. During inference or testing, the threshold for a winner is 0.001. We train the network to minimize the empirical orthogonal gap loss and approximate the target equilibrium $\pi^\star(x)$. During training, we gradually decay the regularization strength geometrically from 1 to the target value ρ over the first 150 epochs and keep it at ρ thereafter. We hold out 20% of the training contexts, evaluate the gap loss with strength ρ on them after every epoch, and apply early stopping once the decay has ended. We tune the hyperparameters of the predictor by the grid search specified in Table 14 and select the configuration with the lowest held-out gap loss.

Table 14: Hyperparameter search space for the equilibrium predictor. We fix two hidden layers and vary their common width. The last column gives the values in the synthetic experiment.

Hyperparameter	Search space or fixed setting	Synthetic
Hidden-layer widths	$\{(64,64),(128,128),(256,256)\}$	$(256,256)$
Learning rate	$\{10^{-4}, 3\times10^{-4}, 10^{-3}\}$	3×10^{-4}
Weight decay	$\{0, 10^{-5}, 10^{-4}\}$	10^{-5}
Batch size	4,096	4,096
Maximum training epochs	500	500
Early-stopping patience	30 epochs	30 epochs

I.3 BASELINES

Oracle nuisance baseline. For NashEval-oracle, we substitute the exact outcome functions $\mu^m_{0,jk}(x)$ and inclusion propensities $e^m_0(x)$ into the orthogonal training target:

$$\Gamma^{\mathrm{ON}}_{i,jk} = \frac{1}{|\mathcal{Q}_{jk}|}\sum_{m\in\mathcal{Q}_{jk}}\left[\mu^m_{0,jk}(X_i) + \frac{S^m_i}{e^m_0(X_i)}\left\{Z^m_{i,jk} - \mu^m_{0,jk}(X_i)\right\}\right]. \tag{72}$$

We use the same training log, payoff regression, and policy learning as NashEval-debiased. We clip predictions to $[-1,1]$ and impose antisymmetry and a zero diagonal to obtain $\widehat{A}^{\mathrm{ON}}(x)$.

NashEval-related policies. We compute NashEval-debiased, NashEval-plug-in, and NashEval-oracle from $\widehat{A}^{\mathrm{DB}}$, $\widehat{A}^{\mathrm{PI}}$, and $\widehat{A}^{\mathrm{ON}}$, respectively. In the synthetic experiment, we set all agent costs to zero and use the following regularizer:

$$\Omega(\pi) = \frac{\rho}{2}\left\|\pi - \frac{1}{K}\mathbf{1}_K\right\|_2^2, \qquad \rho = 0.001. \tag{73}$$

BT and Borda policies. For an estimated matrix $B(x)$, we fit BT scores separately at each context:

$$\widehat{s}^B(x) = \arg\min_{s\in\mathbb{R}^K}\frac{1}{\binom{K}{2}}\sum_{j<k}\left[\log\{1+\exp(s_j-s_k)\} - \frac{1+B_{jk}(x)}{2}(s_j-s_k)\right] + \frac{\lambda_{\mathrm{BT}}}{2}\|s\|_2^2, \tag{74}$$
$$\lambda_{\mathrm{BT}} = 10^{-4}.$$

We use Newton iterations with line search and stop when the maximum absolute score gradient is at most 10^{-9}. We set $B = \widehat{A}^{\mathrm{DB}}$ for BT-debiased and $B = \widehat{A}^{\mathrm{PI}}$ for BT-Plug-in, and select $\arg\max_j \widehat{s}^B_j(x)$. For winner feedback, we use the encoded margins in Eq. (64).

BT-Reg uses the same scores $\widehat{s}^{\mathrm{PI}} = \widehat{s}^B$ with $B = \widehat{A}^{\mathrm{PI}}$ as BT-Plug-in to form

$$\widehat{A}^{\mathrm{BT}}_{jk}(x) = \tanh\left(\frac{\widehat{s}^{\mathrm{PI}}_j(x) - \widehat{s}^{\mathrm{PI}}_k(x)}{2}\right), \tag{75}$$

and computes its equilibrium with the same policy regularizer and strength ρ as NashEval in Eq. (73).

Borda selects the largest average row margin of $\widehat{A}^{\mathrm{DB}}(x)$: $\arg\max_j (K-1)^{-1}\sum_{k\neq j}\widehat{A}^{\mathrm{DB}}_{jk}(x)$. With complete, equally weighted pairs and a common L_2 score penalty, BT scores and row sums have the same ordering. With identical tie-breaking by agent index, BT-debiased and Borda therefore yield the same policy, which we label **BT-debiased / Borda**.

BT-cost on RouterBench. We train a cost-constrained adaptation of the contextual BT model of Theorem 1 in Frick et al. (2025). For each prompt, we form $\widehat{P}^{\mathrm{BT}}_{jk}(x) = \sigma(\widehat{s}_j(x) - \widehat{s}_k(x))$ and fix the opponent to $q = \mathbf{1}/K$. Given a budget $C \geq \min_j c_j$, we solve

$$\widehat{\pi}_{\text{BT-cost}}(x;C) \in \arg\max_{\pi\in\Delta_K:\, c^\top\pi\leq C} \pi^\top \widehat{P}^{\mathrm{BT}}(x)q. \tag{76}$$

We evaluate 31 equally spaced budgets between the minimum and maximum costs, together with every distinct model cost, for 37 budgets in total.

Maximal lottery. We compute a context-independent equilibrium from pooled training payoffs $\widehat{M}_{jk}$, using the regularization and cost treatment specified for each experiment. We average the observed training verdicts for each pair across contexts. Let $\mathcal{I}_{jk}$ index the training comparisons of agents j and k, and let $Z_{i,jk}^{m_i}$ denote comparison i's signed verdict, oriented in favor of j. We define

$$\widehat{M}_{jk} = \frac{1}{|\mathcal{I}_{jk}|} \sum_{i \in \mathcal{I}_{jk}} Z_{i,jk}^{m_i}, \quad j < k, \qquad \widehat{M}_{kj} = -\widehat{M}_{jk}, \qquad \widehat{M}_{jj} = 0. \tag{77}$$

Robust maximal lottery. Let $\mathcal{G}$ be five balanced random folds of the training contexts, $\widehat{M}^{(g)}$ the pooled payoff matrix of fold g, and $\widehat{w}_0$ the empirical fold frequencies. Following Khalaf et al. (2026), we define

$$\mathcal{W}(\widehat{w}_0, \tau) = \left\{ w \in \Delta_{|\mathcal{G}|} : \frac{1}{2}\|w - \widehat{w}_0\|_1 \le \tau \right\}, \qquad \mathcal{M}_\tau = \left\{ \sum_{g \in \mathcal{G}} w_g \widehat{M}^{(g)} : w \in \mathcal{W}(\widehat{w}_0, \tau) \right\} \tag{78}$$

as the ambiguity set and choose the primary radius

$$\tau_n = \min\left\{ 1, \sqrt{\frac{|\mathcal{G}|}{n}} + \sqrt{\frac{2}{n} \log \frac{2}{\delta}} \right\}, \qquad \delta = 0.05, \tag{79}$$

where n is the number of training contexts. We also evaluate $\tau \in \{0, 0.025, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 1\}$, together with τ_n, without test-based radius selection.

The robust policy is the solution of

$$\widehat{\pi}_{\mathrm{rob}} \in \arg\max_{\pi \in \Delta_K} \min_{w \in \mathcal{W}(\widehat{w}_0, \tau)} \min_{j \in [K]} \pi^\top \left(\sum_{g \in \mathcal{G}} w_g \widehat{M}^{(g)} \right) e_j. \tag{80}$$

We then solve Eq. (80) with the exact dual linear program from Theorem 17 of Khalaf et al. (2026).

Pluralistic leaderboard. We adapt the ranking method of Haghtalab et al. (2026) to incomplete comparisons by treating each training context as one voter. We infer its scores from the observed pairs $\mathcal{O}_x$ through

$$\widehat{u}_x = \arg\min_{u \in \mathbb{R}^K} \sum_{(j,k) \in \mathcal{O}_x} \big(u_j - u_k - Z_{jk}(x) \big)^2 + \|u\|_2^2. \tag{81}$$

We sort $\widehat{u}_x$ in descending order and break ties with a permutation fixed by the data-split seed. This completion step is our adapter, so the ranking guarantees concern the inferred rankings rather than the original cyclic preferences.

We implement the iterated-rounding committee procedure and the geometric-checkpoint ranking construction of Algorithms 1 and 2 in Haghtalab et al. (2026). We use the paper's parameters $\epsilon = 0.01$, $\alpha = 1/2 + 4\epsilon$, and $\beta = 1/4 + 2\epsilon$, and use growth factor two for the checkpoint sizes. Since our candidate sets have at most $K = 10$ agents, we enumerate the committees at each checkpoint and replace the paper's approximate multiplicative-weights subroutine with an exact empirical linear program. For candidate committees $C_1, \ldots, C_L$ of size k, let

$$b_{a\ell} = \frac{1}{n} \sum_{x=1}^{n} \mathbf{1}\{ a \succ_x c \text{ for every } c \in C_\ell \}. \tag{82}$$

We compute the stable lottery over committees by solving

$$\min_{\delta \in \Delta_L,\, v} \; v \quad \text{subject to} \quad \sum_{\ell=1}^{L} \delta_\ell b_{a\ell} \le v \quad \forall a \in [K]. \tag{83}$$

We use δ to select committees in order to produce the ranking. At each rounding step, we select a committee with positive probability under δ that minimizes the fraction of unsatisfied prompts at threshold $\beta - \epsilon$, and repeat on the remaining unsatisfied prompts. We construct checkpoint committees of sizes $1, 2, 4, \ldots$ and append each model to the ranking when it first appears, then append any remaining models. Let W_3 be the set of the top three models in the ranking. We convert the ranking to a uniform lottery over W_3 for policy evaluation.

J PROOFS

We collect the assumptions in Section J.1. We then prove identification, derive the EIF and the conditional bias bound, establish Neyman-orthogonality, and prove the empirical exploitability bound and quasi-oracle rate.

Notation and conventions. We write P_0 for the observed-data law and P_X for its context marginal. We fix the finite agent and menu collections, the aggregators f_{jk}, and the context-independent regularizer Ω. All functions, fitted quantities, policy selections, and suprema below are measurable, and pointwise statements hold outside a common P_X-null set. We suppress pair and menu indices when they do not affect the argument, write $\eta = (\mu, e)$ and $\eta_0 = (\mu_0, e_0)$, and set $\Delta\mu = \mu - \mu_0$ and $\Delta e = e - e_0$ for a generic nuisance, evaluated at $\eta = \widehat{\eta}$ for fitted nuisance errors.

We use $\|\cdot\|_1$ and $\|\cdot\|_2$ for the vector ℓ_1 and Euclidean norms. For the policy norm $\|\cdot\|$, we define the dual norm and induced matrix norm by

$$\|v\|_\star := \sup_{\|u\|\le 1} \langle v, u\rangle, \qquad \|B\|_\star := \sup_{\|u\|\le 1} \|Bu\|_\star.$$

We assume that the policy norm satisfies $\|p\| \le 1$ for every $p \in \Delta_K$, as do the ℓ_1 and Euclidean norms. Finite-dimensional norm equivalence gives a constant C_{mat} such that $\|B\|_\star^2 \le C_{\mathrm{mat}} \sum_{j,k} |B_{jk}|^2$. For a scalar function h, we write $\|h\|_{L_2(P_X)} = (\mathbb{E}_X[h(X)^2])^{1/2}$, so $\|\|\Delta\mu\|_2^2\|_{L_2(P_X)} = (\mathbb{E}_X\|\Delta\mu(X)\|_2^4)^{1/2}$. We write $L_2^0(P_0)$ for the square-integrable, mean-zero functions of O. All regularity constants are uniform in n, and κ is fixed for the quasi-oracle rate.

J.1 ASSUMPTIONS

We state the full list of assumptions used in the analysis.

Assumption 1 (Identifiability)**.** In order to identify the contextual payoff $A_{jk}(x)$ from the observed data, we make the following assumptions: (i) *pairwise menu coverage*: $\mathcal{Q}_{jk} \neq \emptyset$ for every pair $\{j, k\} \subseteq \mathcal{M}$; (ii) *missing at random (MAR)*: $S^m \perp Z_{jk}^m \mid X$ for every $m \in \mathcal{Q}_{jk}$; (iii) *positivity*: $e_0^m(x) = P(S^m = 1 \mid X = x) \ge \underline{e} > 0$ almost everywhere for every $m \in \mathcal{Q}_{jk}$.

The potential-verdict formulation follows Rubin (1974), and MAR (Rubin, 1976) and positivity (Kennedy, 2024) are standard identifying conditions in missing-data and causal analysis.

Assumption 2 (Boundedness)**.** We fix $\alpha_\mu > 0$ and write $I_\alpha := (-1 - \alpha_\mu, 1 + \alpha_\mu)$. We assume (i) generic and fitted nuisances used in the bias, differentiability, and rate bounds satisfy $\mu_{jk}^m(x) \in I_\alpha$ and $e^m(x) \ge \underline{e}$, with $\underline{e}$ from Assumption 1, and (ii) the fitted payoff has finite entrywise $L_2(P_X)$ error. For the quasi-oracle result, we additionally require (iii) a uniform envelope $\sup_{h\in\mathcal{H}_n} |h(x)| \le H$, where $\mathcal{H}_n$ is the loss class in Assumption 8 and H is independent of n.

Assumption 3 (Smoothness)**.** We assume (i) each f_{jk} extends to a continuously differentiable function on $I_\alpha^{\mathcal{Q}_{jk}}$, with I_α from Assumption 2, (ii) Ω is continuously differentiable on a neighborhood of Δ_K, and (iii) $\theta \mapsto \pi_\theta(x)$ is continuously differentiable when we differentiate the gap loss. For population gap-loss differentiation, we further require a neighborhood U_θ of each θ such that

$$\mathbb{E}_X[M_\theta(X)] < \infty, \qquad M_\theta(x) := \sup_{\vartheta\in U_\theta} \|D_\theta \pi_\vartheta(x)\|_{\mathrm{op}}, \tag{84}$$

where $\|J\|_{\mathrm{op}} := \sup_{\|v\|_2\le 1} \|Jv\|$. For gap-loss differentiability, orthogonality, and the quantitative remainder and rate bounds, we further assume $\|\nabla f_{jk}(u) - \nabla f_{jk}(v)\|_2 \le L_f \|u - v\|_2$ for all $u, v \in I_\alpha^{\mathcal{Q}_{jk}}$.

Such differentiability conditions are standard in influence-function and orthogonal-learning analyses (Kennedy, 2024; Foster & Syrgkanis, 2023). The Jacobian envelope justifies differentiation of the population gap loss under expectation, as we show in Section J.7.

Assumption 4 (Strong convexity)**.** We assume that Ω is κ-strongly convex on Δ_K with respect to $\|\cdot\|$, where $\kappa > 0$.

Strongly convex regularization is common in regularized zero-sum games (Sokota et al., 2023), and it makes the equilibrium unique (Facchinei & Pang, 2003).

Assumption 5 (Cross-fitting)**.** We observe i.i.d. data $O_1, \dots, O_n \sim P_0$ and use a data-independent partition into a fixed number K_{CF} of folds. Each nuisance fit uses only observations outside its evaluation fold and satisfies the stated propensity and rate conditions. We use $A^{\mathrm{DB}}(\cdot; \widehat{\eta})$ for the average of the fold-specific regression targets.

Cross-fitting is standard in DML (Chernozhukov et al., 2018).

Assumption 6 (Nuisance and payoff-estimation rates)**.** We assume $\|\|\Delta\mu\|_2\|_{L_2(P_X)} = O_p(r_\mu)$ and $\|\Delta e\|_{L_2(P_X)} = O_p(r_e)$ for deterministic rates $r_\mu, r_e \to 0$, together with the moment bounds

$$\left\| \|\Delta\mu\|_2^2 \right\|_{L_2(P_X)} = O_p(r_\mu^2), \qquad \|\Delta e\, \Delta\mu\|_{L_2(P_X)} = O_p(r_\mu r_e). \tag{85}$$

The outcome bounds hold uniformly over agent pairs, the propensity bound over contributing menus, and the product bound over pairs and contributing menus, with the product evaluated menu-wise. We define

$$e_A := \max_{j,k} \left\| \widehat{A}^{\mathrm{DB}}_{jk}(\cdot; \widehat{\eta}) - A^{\mathrm{DB}}_{jk}(\cdot; \widehat{\eta}) \right\|_{L_2(P_X)}. \tag{86}$$

For the quasi-oracle rate, we additionally require

$$e_A = O_p(n^{-1/4}), \qquad r_\mu \lesssim n^{-1/8}, \qquad \underline{e}^{-1} r_\mu r_e \lesssim n^{-1/4}, \tag{87}$$

and omit the separate condition on r_μ for affine aggregators.

Product-rate conditions of this type are standard in DML (Chernozhukov et al., 2018) and orthogonal statistical learning (Foster & Syrgkanis, 2023). For fixed $\underline{e} > 0$, we can satisfy Eq. (87) with matching nuisance rates $r_\mu, r_e \lesssim n^{-1/8}$ and payoff-regression error $e_A = O_p(n^{-1/4})$, using flexible nuisance learners such as gradient-boosted trees (e.g., LightGBM) or neural networks (Chernozhukov et al., 2018) when they attain these rates in the stated norms.

Assumption 7 (Realizability)**.** For oracle recovery in Proposition 2, we assume that some $\theta_0 \in \Theta$ satisfies $\pi_{\theta_0}(x) = \pi^\star(x)$ for P_X-almost every x.

Assumption 8 (Loss-class complexity)**.** For the quasi-oracle result, we assume a deterministic class $\mathcal{B}_n$ containing A and, with probability tending to one, $\widehat{A}^{\mathrm{DB}}(\cdot; \widehat{\eta})$. We define

$$\mathcal{H}_n := \{x \mapsto G_\Omega(\pi_\theta(x); B(x), x) : \theta \in \Theta,\ B \in \mathcal{B}_n\}$$

and assume the expected absolute Rademacher-complexity bound

$$\mathfrak{R}_n(\mathcal{H}_n) := \mathbb{E}_{X,\sigma}\left[\sup_{h \in \mathcal{H}_n} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i h(X_i) \right| \right] \le C n^{-1/2}, \tag{88}$$

where C is independent of n, and $\sigma_1, \dots, \sigma_n$ are independent uniform signs in $\{-1, 1\}$, independent of the contexts.

Rademacher-complexity bounds are commonly used to control uniform deviations (Bartlett & Mendelson, 2002). Because the bound covers the entire class $\mathcal{H}_n$, it permits reuse of observations for payoff and policy fitting.

J.2 Identification of the contextual payoff matrix

Proof. We fix a pair $j \neq k$, a menu $m \in \mathcal{Q}_{jk}$, and a context x outside the null set on which the identification assumptions may fail. Menu-specific positivity gives $P(S^m = 1 \mid X = x) = e_0^m(x) \ge \underline{e} > 0$, so the observed-data regression $\mu^m_{0,jk}(x)$ is well-defined. Menu-specific MAR implies conditional mean independence between S^m and Z^m_{jk} given X. Therefore,

$$\begin{aligned} \mu^m_{0,jk}(x) &= \mathbb{E}[Z^m_{jk} \mid X = x, S^m = 1] \\ &= \mathbb{E}[Z^m_{jk} \mid X = x] \\ &= g^m_{jk}(x). \end{aligned}$$

This argument only requires the conditional mean equality $\mathbb{E}[Z_{jk}^m \mid X, S^m] = \mathbb{E}[Z_{jk}^m \mid X]$ almost surely. It does not require joint independence across menus. Thus, our method handles the case where only one menu is observed per context (i.e., $\sum_{m \in \mathcal{Q}} S^m = 1$).

The equality holds for every $m \in \mathcal{Q}_{jk}$. Hence, we identify each off-diagonal payoff entry by substituting the observed-data regressions into Eq. (4):

$$A_{jk}(x) = f_{jk}\big(\{\mu_{0,jk}^m(x) : m \in \mathcal{Q}_{jk}\}\big).$$

Pairwise menu coverage ensures that $\mathcal{Q}_{jk}$ is nonempty, and menu-specific positivity applies to every menu in this set. Definition 1 fixes $A_{jj}(x) = 0$, so the observed-data law identifies the entire contextual payoff matrix $A(x)$. Under the regularizer conditions in Assumptions 3(ii) and 4, Ω is fixed and the equilibrium is unique, so Eq. (5) also identifies $\pi^\star(x)$. □

J.3 Efficient influence function of the average payoff

Lemma 2 (Efficient influence function of the average payoff)**.** *Under Assumptions 1–3, fix $j \neq k$ and define the average payoff $\psi_{jk} := \mathbb{E}_0[A_{jk}(X)]$. In the nonparametric observed-data model, we obtain the efficient influence function of ψ_{jk} at P_0 by centering the pseudo-outcome in Eq. (41) at the true nuisance $\eta_0 = (\mu_0, e_0)$:*

$$\varphi_{jk}(O; P_0) = \Gamma_{jk}(O; \eta_0) - \psi_{jk}. \tag{89}$$

In particular, $\mathbb{E}_0[\Gamma_{jk}(O; \eta_0)] = \psi_{jk}$ and $\mathbb{E}_0[\varphi_{jk}(O; P_0)] = 0$.

Proof of Lemma 2. We consider a regular one-dimensional parametric submodel $\{P_t : |t| < \epsilon\}$ of the observed-data law, with $P_{t=0} = P_0$. Let the densities be p_t. We write $O = (X, \mathbf{S}, \widetilde{\mathbf{Y}})$, where $\mathbf{S} = (S^m)_{m \in \mathcal{Q}}$ and $\widetilde{\mathbf{Y}} = (S^m Y_m)_{m \in \mathcal{Q}}$. We factor the joint density as

$$p_t(o) = p_t(x)\, p_t(\mathbf{s} \mid x)\, p_t(\widetilde{\mathbf{y}} \mid x, \mathbf{s}).$$

We define the score as the derivative of the log density at $t = 0$:

$$\begin{aligned} s(O) &:= \frac{\mathrm{d}}{\mathrm{d}t} \log p_t(O)\Big|_{t=0} \\ &= s_X(X) + s_{\mathbf{S}|X}(\mathbf{S} \mid X) + s_{\widetilde{\mathbf{Y}}|X,\mathbf{S}}(\widetilde{\mathbf{Y}} \mid X, \mathbf{S}), \end{aligned}$$

where each component is the log derivative of the corresponding density factor. The score has finite second moment, and its components satisfy

$$\begin{aligned} &\mathbb{E}_0[s_X(X)] = 0, \\ &\mathbb{E}_0[s_{\mathbf{S}|X}(\mathbf{S} \mid X) \mid X] = 0, \\ &\mathbb{E}_0[s_{\widetilde{\mathbf{Y}}|X,\mathbf{S}}(\widetilde{\mathbf{Y}} \mid X, \mathbf{S}) \mid X, \mathbf{S}] = 0. \end{aligned}$$

Thus, $\mathbb{E}_0[s(O)] = 0$, $s_X(X) = \mathbb{E}_0[s(O) \mid X]$, and $s - s_X$ is the conditional score of O given X.

We fix $j \neq k$ and abbreviate $f = f_{jk}$. Under P_t, we define $\mu_{t,jk}^m(x) = \mathbb{E}_{P_t}[Z_{jk}^m \mid X = x, S^m = 1]$ and collect these regressions in $\mu_t(x) = (\mu_{t,jk}^m(x))_{m \in \mathcal{Q}_{jk}}$. We write $\psi_t = \mathbb{E}_{P_t}[f(\mu_t(X))]$ and $\psi_0 = \psi_{jk}$. The product and chain rules give

$$\begin{aligned} \dot{\psi}_0 &:= \frac{\mathrm{d}}{\mathrm{d}t}\psi_t\Big|_{t=0} \\ &= \mathbb{E}_0[\{f(\mu_0(X)) - \psi_0\} s_X(X)] + \sum_{m \in \mathcal{Q}_{jk}} \mathbb{E}_0[\partial_m f(\mu_0(X)) \dot{\mu}_0^m(X)], \end{aligned} \tag{90}$$

where $\partial_m f = \partial f / \partial u_m$ and $\dot{\mu}_0^m = \mathrm{d}\mu_{t,jk}^m / \mathrm{d}t\Big|_{t=0}$.

In order to get $\dot{\mu}_0^m$, we compute the regression derivative using the observed-data ratio

$$\mu_{t,jk}^m(x) = \frac{\mathbb{E}_{P_t}[S^m Z_{jk}^m \mid X = x]}{\mathbb{E}_{P_t}[S^m \mid X = x]}.$$

The quotient rule and the conditional score identity yield

$$\begin{aligned}\dot{\mu}_0^m(x) &= \frac{1}{e_0^m(x)}\mathbb{E}_0\big[S^m\{Z_{jk}^m - \mu_{0,jk}^m(x)\}\{s(O) - s_X(x)\} \mid X = x\big] \\ &= \mathbb{E}_0\left[\frac{S^m}{e_0^m(x)}\{Z_{jk}^m - \mu_{0,jk}^m(x)\}s(O) \mid X = x\right].\end{aligned} \tag{91}$$

For the second equality, we use $\mathbb{E}_0[S^m\{Z_{jk}^m - \mu_{0,jk}^m(x)\} \mid X = x] = 0$. Then, we substitute Eq. (91) into Eq. (90) and apply iterated expectations. Since $f(\mu_0(X))$ is a function of X, we obtain

$$\begin{aligned}\dot{\psi}_0 &= \mathbb{E}_0[\{f(\mu_0(X)) - \psi_0\}s(O)] \\ &\quad + \sum_{m\in\mathcal{Q}_{jk}} \mathbb{E}_0\left[\partial_m f(\mu_0(X))\frac{S^m}{e_0^m(X)}\{Z_{jk}^m - \mu_{0,jk}^m(X)\}s(O)\right] \\ &= \mathbb{E}_0[\{\Gamma_{jk}(O;\eta_0) - \psi_{jk}\}s(O)]\,.\end{aligned} \tag{92}$$

We now verify the EIF definition in Section D.1 for $\varphi_{jk}(O;P_0) := \Gamma_{jk}(O;\eta_0) - \psi_{jk}$. Each residual correction has conditional mean zero given X, and $\mathbb{E}_0[f(\mu_0(X))] = \psi_{jk}$, so $\mathbb{E}_0[\varphi_{jk}] = 0$. Bounded verdicts, bounded derivatives, and positivity give $\mathbb{E}_0[\varphi_{jk}^2] < \infty$. Thus, φ_{jk} belongs to $L_2^0(P_0)$, the space of square-integrable mean-zero functions, which is the closed tangent space of the nonparametric observed-data model. Since the submodel was arbitrary, Eq. (92) represents every pathwise derivative exactly as required by Eq. (36), with $\psi = \psi_{jk}$ and $\varphi = \varphi_{jk}$. These properties establish that φ_{jk} is the unique canonical gradient and hence the EIF.

At a generic law P, we obtain $\varphi_{jk}(O;P) = \Gamma_{jk}(O;\eta_P) - \psi_{jk}(P)$, as stated in Eq. (40). We add back $\psi_{jk}(P)$ and replace η_P by a generic nuisance η to obtain the uncentered pseudo-outcome in Eq. (41). □

J.4 Neyman-orthogonality of Γ_{jk}

We prove conditional unbiasedness and Neyman-orthogonality under Assumptions 1–3.

Proof. We fix $j \neq k$, a direction h, and a context x outside the null sets in the assumptions and the conditional expectation identities below. We write $\eta_t = \eta_0 + th$ for $|t| \le t_0$. Since $\mu_{0,jk}^m(x) \in [-1,1]$ and $e_0^m(x) \ge \underline{e}$, we can choose $t_0 = t_0(x,h) > 0$ small enough that $\mu_t^m(x) \in I_\alpha$ and $e_t^m(x) > 0$ for every $m \in \mathcal{Q}_{jk}$. We write $f = f_{jk}$ and collect the outcome components into $u_t(x) = \{\mu_t^m(x)\}_{m\in\mathcal{Q}_{jk}}$, so $u_t(x) = u_0(x) + th_\mu(x)$. Assumption 1 gives $g_{jk}^m(x) = \mu_{0,jk}^m(x)$ and hence $A_{jk}(x) = f(u_0(x))$. Define

$$G(t) \;:=\; \mathbb{E}_0\big[\Gamma_{jk}(O;\eta_t)\;\big|\;X = x\big], \qquad |t| \le t_0, \tag{93}$$

where the expectation remains under the fixed law P_0 as the nuisance argument varies. We will show $G(0) = A_{jk}(x)$ and $G'(0) = 0$.

Bounded verdicts, continuity of the aggregator gradient, and positive propensities give conditional integrability along the path. Conditional on $X = x$, the nuisance factors in Eq. (41) are deterministic, so

$$G(t) = f\big(u_t(x)\big) + \sum_{m\in\mathcal{Q}_{jk}} \frac{\partial f}{\partial u_m}\big(u_t(x)\big)\,\frac{1}{e_t^m(x)}\,\underbrace{\mathbb{E}_0\Big[S^m\big\{Z_{jk}^m - \mu_t^m(x)\big\}\;\Big|\;X = x\Big]}_{=:\,T_m(t)}. \tag{94}$$

The selected residual has an explicit conditional mean. By the definition of the observed-data regression $\mu_{0,jk}^m$,

$$\mathbb{E}_0[S^m Z_{jk}^m \mid X = x] = e_0^m(x)\mathbb{E}_0[Z_{jk}^m \mid X = x, S^m = 1] = e_0^m(x)\mu_{0,jk}^m(x). \tag{95}$$

This identity does not require MAR. We use Assumption 1 only to identify the regression with the full-data margin. Since $\mathbb{E}_0[S^m\mu_t^m(x) \mid X = x] = e_0^m(x)\mu_t^m(x)$, subtracting gives

$$T_m(t) = e_0^m(x)\big\{\mu_{0,jk}^m(x) - \mu_t^m(x)\big\} = -\,t\,e_0^m(x)\,h_\mu^m(x), \tag{96}$$

where the last equality uses $\mu_t = \mu_0 + t h_\mu$. Substituting this expression into Eq. (94), we obtain

$$G(t) = f\big(u_t(x)\big) \;-\; t\,B(t), \qquad B(t) := \sum_{m\in\mathcal{Q}_{jk}} \frac{\partial f}{\partial u_m}\big(u_t(x)\big)\,\frac{e_0^m(x)}{e_t^m(x)}\,h_\mu^m(x). \tag{97}$$

We use this expression as a version of the conditional expectation for all $|t| \le t_0$. At $t = 0$ it gives $G(0) = f(u_0(x)) = A_{jk}(x)$, proving conditional unbiasedness. Continuity of the gradient and positivity of the denominators imply that $B(t)$ is continuous at zero.

We can now differentiate the explicit expression for G, without interchanging a derivative and an expectation. For $t \neq 0$, Eq. (97) gives

$$\frac{G(t) - G(0)}{t} = \frac{f\big(u_t(x)\big) - f\big(u_0(x)\big)}{t} - B(t). \tag{98}$$

By Assumption 3, the chain rule along the affine path $u_t(x) = u_0(x) + t h_\mu(x)$ gives

$$\frac{\mathrm{d}}{\mathrm{d}t} f\big(u_t(x)\big)\Big|_{t=0} = \sum_{m\in\mathcal{Q}_{jk}} \frac{\partial f}{\partial u_m}\big(u_0(x)\big)\,h_\mu^m(x) = \big\langle \nabla f\big(u_0(x)\big),\; h_\mu(x)\big\rangle. \tag{99}$$

Since $e_t^m(x) = e_0^m(x)$ at $t = 0$, the limit of the correction term is

$$B(0) = \sum_{m\in\mathcal{Q}_{jk}} \frac{\partial f}{\partial u_m}\big(u_0(x)\big)\,h_\mu^m(x) = \big\langle \nabla f\big(u_0(x)\big),\; h_\mu(x)\big\rangle. \tag{100}$$

The limits in Eqs. (99) and (100) cancel in Eq. (98), giving

$$G'(0) = \big\langle \nabla f(u_0(x)), h_\mu(x)\big\rangle - \big\langle \nabla f(u_0(x)), h_\mu(x)\big\rangle = 0.$$

Thus, for every direction h, Neyman-orthogonality holds for P_X-almost every x. □

J.5 Exact conditional bias and second-order remainder

We derive the exact conditional bias and prove Eq. (12) under the identification, boundedness, and smoothness conditions in Assumptions 1–3.

Lemma 3 (Exact conditional bias)**.** *Under these conditions, we fix $j \neq k$, write $f = f_{jk}$ and $\mu(x) = (\mu_{jk}^m(x))_{m\in\mathcal{Q}_{jk}}$, and consider a generic nuisance $\eta = (\mu, e)$ with $\mu_{jk}^m(x) \in I_\alpha$ and $e^m(x) > 0$. We define $\Delta\mu = \mu - \mu_0$ and $\Delta e^m = e^m - e_0^m$. Then, for almost every x,*

$$\mathbb{E}_0\big[\Gamma_{jk}(O;\eta) \mid X = x\big] - A_{jk}(x) = -R_f(x) + \sum_{m\in\mathcal{Q}_{jk}} \frac{\partial f}{\partial u_m}\big(\mu(x)\big)\,\frac{\Delta e^m(x)}{e^m(x)}\,\Delta\mu^m(x), \tag{101}$$

where $R_f(x) = f(\mu_0(x)) - f(\mu(x)) - \big\langle \nabla f(\mu(x)),\, \mu_0(x) - \mu(x)\big\rangle$ is the first-order Taylor remainder of f at $\mu(x)$. Consequently, Eq. (12) holds.

Proof. We fix x and suppress it below. The selected residual has conditional mean $\mathbb{E}_0[S^m(Z_{jk}^m - \mu_{jk}^m) \mid X = x] = -e_0^m \Delta\mu^m$. We substitute this identity into Eq. (41) and use $A_{jk}(x) = f(\mu_0)$ to obtain

$$\begin{aligned} A_{jk}^{\mathrm{DB}}(x;\eta) - A_{jk}(x) &= f(\mu) - f(\mu_0) - \sum_{m\in\mathcal{Q}_{jk}} \partial_m f(\mu)\frac{e_0^m}{e^m}\Delta\mu^m \\ &= \underbrace{f(\mu) - f(\mu_0) - \langle \nabla f(\mu), \Delta\mu\rangle}_{-R_f(x)} + \sum_{m\in\mathcal{Q}_{jk}} \partial_m f(\mu)\left(1 - \frac{e_0^m}{e^m}\right)\Delta\mu^m. \end{aligned}$$

The identity $1 - e_0^m/e^m = \Delta e^m/e^m$ gives Eq. (101). Since $\mu_0(x) \in [-1,1]^{\mathcal{Q}_{jk}}$ and $I_\alpha^{\mathcal{Q}_{jk}}$ is convex, the segment between $\mu(x)$ and $\mu_0(x)$ lies in $I_\alpha^{\mathcal{Q}_{jk}}$. The Lipschitz gradient bound therefore gives $|R_f(x)| \le (L_f/2)\|\Delta\mu(x)\|_2^2$, and the triangle inequality proves Eq. (12). The Lipschitz gradient on the bounded cube $I_\alpha^{\mathcal{Q}_{jk}}$ also gives $G_f := \max_{j\neq k} \sup_{u\in I_\alpha^{\mathcal{Q}_{jk}}} \|\nabla f_{jk}(u)\|_1 < \infty$. If $e^m(x) \ge \underline{e}$, as in Assumption 2, then

$$\Big|\mathbb{E}_0\big[\Gamma_{jk}(O;\eta) \mid X = x\big] - A_{jk}(x)\Big| \;\le\; \tfrac{L_f}{2}\big\|\Delta\mu(x)\big\|_2^2 + \frac{G_f}{\underline{e}} \max_{m\in\mathcal{Q}_{jk}} \big|\Delta e^m(x)\Delta\mu^m(x)\big|. \tag{102}$$

□

Under the remainder-component conditions in Eq. (85), this bound yields conditional bias of order $O_p(r_\mu^2 + \underline{e}^{-1} r_\mu r_e)$ in $L_2(P_X)$. For an affine aggregator, $R_f \equiv 0$, and Eq. (101) vanishes if either $\mu = \mu_0$ or $e = e_0$. For a nonlinear aggregator, correct propensities alone leave the outcome-curvature term $-R_f(x)$.

J.6 Proof of Proposition 2

Proof. Theorem 1 gives $A^{\mathrm{DB}}(x; \eta_0) = A(x)$, and Proposition 4 gives $G_\Omega(\pi; A(x), x) \geq 0$, with equality exactly when $\pi = \pi^\star(x)$. Under Assumption 7, we choose θ_0 with $\pi_{\theta_0}(X) = \pi^\star(X)$ almost surely, so $\mathcal{L}(\theta_0; \eta_0) = 0$ is the minimum risk. Every minimizer therefore has zero expected nonnegative gap, which implies $\pi_\theta(X) = \pi^\star(X)$ almost surely. Conversely, every such policy has zero risk and attains the minimum. □

J.7 Proof of Theorem 2

We first prove differentiability under Assumptions 2–4. We then use identification and the payoff remainder to prove universal Neyman-orthogonality.

Proof. **Differentiability.** We fix a nuisance η satisfying Assumption 2 and write $B(x) = A^{\mathrm{DB}}(x; \eta)$. The Lipschitz aggregator gradients on the bounded domain $I_\alpha^{\mathcal{Q}_{jk}}$ and the propensity lower bound give a uniform bound on $B(x)$ through Eq. (41). For any matrix B, strong convexity gives the unique maximizer $q_B(p) := \arg\min_{q \in \Delta_K} \{\Omega(q) - \langle Bp, q \rangle\}$ in the variational gap. Danskin's theorem gives

$$g_B(p) := \nabla_p G_\Omega(p; B, x) = \nabla\Omega(p) - (B + B^\top)p + B^\top q_B(p). \tag{103}$$

This formula holds without skew-symmetry of B and reduces to Eq. (32) when B is skew-symmetric. We fix a direction h_θ and define $\ell_t(x) := G_\Omega(\pi_{\theta + t h_\theta}(x); B(x), x)$. The chain rule gives

$$\frac{\mathrm{d}}{\mathrm{d}t} \ell_t(x) = \left\langle g_{B(x)}(\pi_{\theta + t h_\theta}(x)), D_\theta \pi_{\theta + t h_\theta}(x)[h_\theta] \right\rangle.$$

The uniform bound on $B(x)$ and continuity of $\nabla\Omega$ on the compact simplex give $\|g_{B(x)}(p)\|_\star \leq C_g$ uniformly in x and $p \in \Delta_K$. For sufficiently small $|t|$, the segment from θ to $\theta + t h_\theta$ lies in U_θ. The mean value theorem and duality therefore give

$$\left| \frac{\ell_t(X) - \ell_0(X)}{t} \right| \leq C_g M_\theta(X) \|h_\theta\|_2.$$

Eq. (84) makes the bound integrable, so dominated convergence justifies differentiation under expectation. The chain rule at $t = 0$ then gives

$$D_\theta \mathcal{L}(\theta; \eta)[h_\theta] = \mathbb{E}_X \left[\left\langle g_{B(X)}(\pi_\theta(X)), D_\theta \pi_\theta(X)[h_\theta] \right\rangle \right]. \tag{104}$$

Therefore, the loss is differentiable with respect to θ.

Neyman-orthogonality. We fix θ, a policy direction h_θ, and a bounded nuisance direction h_η. We write $B_t(x) = A^{\mathrm{DB}}(x; \eta_0 + t h_\eta)$, so $B_0(x) = A(x)$. Positivity and boundedness of h_η give $e_t^m(x) \geq \underline{e}/2$ for sufficiently small $|t|$. Since $\mu_{0,jk}^m(x) \in [-1, 1]$, boundedness of h_η also gives $\mu_{0,jk}^m(x) + t h_\mu^m(x) \in I_\alpha$ when $|t| \|h_\mu\|_\infty < \alpha_\mu$. Eq. (12), the Lipschitz aggregator gradients, and the finite pair and menu collections therefore give

$$\operatorname*{ess\,sup}_x \|B_t(x) - A(x)\|_\star \leq C_h t^2, \tag{105}$$

where C_h is independent of x and t.

We next bound the change in the policy gradient induced by a payoff perturbation. We fix x, suppress it below, and use the maximizer $q_B(p)$ defined above. We add the first-order optimality inequalities for $q_B(p)$ and $q_A(p)$ and use strong convexity to obtain

$$\kappa \|q_B(p) - q_A(p)\|^2 \leq \langle (B - A)p, q_B(p) - q_A(p) \rangle,$$
$$\|q_B(p) - q_A(p)\| \leq \kappa^{-1} \|B - A\|_\star.$$

The second inequality uses duality and the norm normalization in Appendix J. We apply Eq. (103) and $\|B^\top\|_\star = \|B\|_\star$ to obtain

$$\begin{aligned}\|g_B(p) - g_A(p)\|_\star &\le 3\|B - A\|_\star + \|A\|_\star \|q_B(p) - q_A(p)\| \\ &\le (3 + \kappa^{-1}\|A\|_\star)\|B - A\|_\star. \end{aligned} \tag{106}$$

The true payoff is uniformly bounded, so $C_A := 3 + \kappa^{-1} \operatorname{ess\,sup}_x \|A(x)\|_\star < \infty$ bounds the factor on the right uniformly in x and p.

We now combine the payoff remainder with the policy-gradient bound. Along the path $\eta_0 + th_\eta$, the same domination argument applies with propensity lower bound $\underline{e}/2$. We apply Eq. (104) and the bounds in Eqs. (105) and (106) to obtain

$$\begin{aligned}&|D_\theta \mathcal{L}(\theta; \eta_0 + th_\eta)[h_\theta] - D_\theta \mathcal{L}(\theta; \eta_0)[h_\theta]| \\ &\quad \le C_A C_h t^2 \|h_\theta\|_2 \mathbb{E}_X[M_\theta(X)]. \end{aligned} \tag{107}$$

We divide by $|t|$ and let $t \to 0$. Hence,

$$D_\eta D_\theta \mathcal{L}(\theta; \eta_0)[h_\theta, h_\eta] = \lim_{t \to 0} \frac{D_\theta \mathcal{L}(\theta; \eta_0 + th_\eta)[h_\theta] - D_\theta \mathcal{L}(\theta; \eta_0)[h_\theta]}{t} = 0.$$

Since θ was arbitrary, this proves the universal Neyman-orthogonality in Eq. (15). □

J.8 Empirical error bound

We first establish the deterministic oracle comparison underlying Theorem 3 using boundedness, smoothness, and strong convexity in Assumptions 2–4. We then use identification, cross-fitting, and the estimation rates in Assumptions 1, 5, and 6 to obtain the rate form. The comparison does not require realizability or a specified generalization rate.

Proof. Let $\widehat{\theta}$ be an empirical minimizer of $\widehat{\mathcal{L}}_n$ as defined in Eq. (17). We define the population loss and abbreviate the deviation and payoff error as

$$\begin{aligned}\widehat{\mathcal{L}}(\theta) &:= \mathbb{E}_X[G_\Omega(\pi_\theta(X); \widehat{A}^{\mathrm{DB}}(X; \widehat{\eta}), X)], \\ \varepsilon_{\mathrm{gen}} &:= \sup_{\theta \in \Theta} |\widehat{\mathcal{L}}(\theta) - \widehat{\mathcal{L}}_n(\theta)|, \qquad \mathcal{E}_A := \kappa^{-1} \mathbb{E}_X \|\widehat{A}^{\mathrm{DB}}(X; \widehat{\eta}) - A(X)\|_\star^2. \end{aligned}$$

Fix a context x at which the assumptions hold, write $A = A(x)$ and $B = \widehat{A}^{\mathrm{DB}}(x; \widehat{\eta})$, and let $p \in \Delta_K$ be arbitrary. For either payoff $C \in \{A, B\}$, define

$$\varphi_C(p, q) := \langle -Cp, p - q \rangle + \Omega(p) - \Omega(q), \qquad q \in \Delta_K.$$

By the variational expression in the gap definition in Eq. (14), $G_\Omega(p; C, x) = \max_{q \in \Delta_K} \varphi_C(p, q)$. By Assumption 3, Ω is continuous, so $q \mapsto \varphi_C(p, q)$ is continuous on the compact simplex and attains a finite maximum. Taking $q = p$ gives $\varphi_C(p, p) = 0$, so both gaps are nonnegative. The candidate p need not be an equilibrium of either game.

To compare the gaps, we use strong concavity in the opponent strategy. For $q_0, q_1 \in \Delta_K$ and $\alpha \in [0, 1]$, put $q_\alpha = (1 - \alpha)q_0 + \alpha q_1$. Assumption 4 gives

$$\Omega(q_\alpha) \le (1 - \alpha)\Omega(q_0) + \alpha\Omega(q_1) - \frac{\kappa}{2}\alpha(1 - \alpha)\|q_1 - q_0\|^2.$$

The term $\langle -Bp, p - q \rangle$ is affine in q, and $\Omega(p)$ is constant in q. Negating the displayed inequality and adding these terms therefore yields

$$\begin{aligned}\varphi_B(p, q_\alpha) &\ge (1 - \alpha)\varphi_B(p, q_0) + \alpha\varphi_B(p, q_1) \\ &\quad + \frac{\kappa}{2}\alpha(1 - \alpha)\|q_1 - q_0\|^2. \end{aligned} \tag{108}$$

Thus $q \mapsto \varphi_B(p, q)$ is κ-strongly concave with respect to the same norm used in the theorem.

Fix $\alpha \in (0,1)$. For any $q \in \Delta_K$, apply Eq. (108) with $q_0 = p$ and $q_1 = q$. Since $(1-\alpha)p+\alpha q \in \Delta_K$ and $\varphi_B(p,p) = 0$,

$$\begin{aligned} G_\Omega(p; B, x) &\geq \varphi_B\big(p, (1-\alpha)p + \alpha q\big) \\ &\geq \alpha \varphi_B(p,q) + \frac{\kappa}{2}\alpha(1-\alpha)\|p-q\|^2. \end{aligned}$$

Dividing by α and rearranging gives

$$\varphi_B(p,q) \leq \frac{1}{\alpha} G_\Omega(p; B, x) - \frac{\kappa(1-\alpha)}{2}\|p-q\|^2 \qquad (q \in \Delta_K). \tag{109}$$

The midpoint choice $\alpha = 1/2$ gives the coefficient 2 on the gap. Retaining α allows that coefficient to approach 1.

We now relate the two payoff matrices. Write $E = B - A$. Subtracting the two variational gap objectives gives the exact identity

$$\varphi_A(p,q) = \varphi_B(p,q) + \langle Ep, p-q\rangle.$$

This identity holds for arbitrary real matrices A and B. Let $a = \|Ep\|_\star \geq 0$. The dual-norm inequality gives $\langle Ep, p-q\rangle \leq a\|p-q\|$. Combining this with Eq. (109) yields

$$\begin{aligned} G_\Omega(p; A, x) &\leq \frac{1}{\alpha} G_\Omega(p; B, x) + \max_{q \in \Delta_K} \left\{ a\|p-q\| - \frac{\kappa(1-\alpha)}{2}\|p-q\|^2 \right\} \\ &\leq \frac{1}{\alpha} G_\Omega(p; B, x) + \sup_{t \geq 0} \left\{ at - \frac{\kappa(1-\alpha)}{2} t^2 \right\}. \end{aligned}$$

The second inequality enlarges the set of feasible distances to all nonnegative real numbers. Completing the square,

$$at - \frac{\kappa(1-\alpha)}{2}t^2 = \frac{a^2}{2\kappa(1-\alpha)} - \frac{\kappa(1-\alpha)}{2}\left(t - \frac{a}{\kappa(1-\alpha)}\right)^2.$$

The supremum is $a^2/[2\kappa(1-\alpha)]$, attained at $t = a/[\kappa(1-\alpha)]$. Taking $\alpha = 1/2$ gives

$$G_\Omega(p; A, x) \leq 2 G_\Omega(p; B, x) + \frac{1}{\kappa}\|(B-A)p\|_\star^2. \tag{110}$$

This argument applies to every $p \in \Delta_K$, including boundary policies. Interchanging A and B gives the reverse comparison with the same squared error, since $\|(A-B)p\|_\star = \|(B-A)p\|_\star$.

Apply Eq. (110) with $p = \pi_{\widehat{\theta}}(X)$ and $B = \widehat{A}^{\mathrm{DB}}(X; \widehat{\eta})$. Under our measurability conventions, the expectation over contexts is well defined, and

$$\mathrm{Expl}(\pi_{\widehat{\theta}}) \leq 2\widehat{\mathcal{L}}(\widehat{\theta}) + \frac{1}{\kappa}\mathbb{E}_X\left[\left\|(\widehat{A}^{\mathrm{DB}}(X; \widehat{\eta}) - A(X))\pi_{\widehat{\theta}}(X)\right\|_\star^2\right].$$

All integrands are nonnegative, so the inequality also holds with an infinite right-hand side if the payoff error has no finite second moment. The $L_2(P_X)$ payoff error in Assumption 2 and compactness of the strategy domain ensure a finite population gap. By the definition of the uniform deviation,

$$\widehat{\mathcal{L}}(\widehat{\theta}) \leq \widehat{\mathcal{L}}_n(\widehat{\theta}) + \varepsilon_{\mathrm{gen}}.$$

This inequality applies to the data-dependent $\widehat{\theta}$ because the supremum defining $\varepsilon_{\mathrm{gen}}$ is over all $\theta \in \Theta$. Substituting it gives the intermediate bound in Eq. (111). Empirical minimization further allows $\widehat{\mathcal{L}}_n(\widehat{\theta})$ to be written as $\inf_{\theta\in\Theta} \widehat{\mathcal{L}}_n(\theta)$, although the bound itself holds for any fitted policy in the class.

Thus the resulting bound is

$$\mathrm{Expl}(\pi_{\widehat{\theta}}) \leq 2\big\{\widehat{\mathcal{L}}_n(\widehat{\theta}) + \varepsilon_{\mathrm{gen}}\big\} + \frac{1}{\kappa}\mathbb{E}_X\left[\left\|(\widehat{A}^{\mathrm{DB}}(X; \widehat{\eta}) - A(X))\pi_{\widehat{\theta}}(X)\right\|_\star^2\right]. \tag{111}$$

Write $\mathcal{L}^A(\theta) := \mathbb{E}_X[G_\Omega(\pi_\theta(X); A(X), X)] = \mathrm{Expl}(\pi_\theta)$, and let

$$\delta(\theta) := \frac{1}{2\kappa}\mathbb{E}_X\left[\left\|(\widehat{A}^{\mathrm{DB}}(X;\widehat{\eta}) - A(X))\pi_\theta(X)\right\|_\star^2\right].$$

Applying Eq. (110) in both directions and taking expectations over X gives

$$\mathcal{L}^A(\theta) \le 2\widehat{\mathcal{L}}(\theta) + 2\delta(\theta), \qquad \widehat{\mathcal{L}}(\theta) \le 2\mathcal{L}^A(\theta) + 2\delta(\theta).$$

For any comparator $\theta \in \Theta$, the definition of $\varepsilon_{\mathrm{gen}}$ and empirical optimality yield

$$\begin{aligned}\widehat{\mathcal{L}}(\widehat{\theta}) &\le \widehat{\mathcal{L}}_n(\widehat{\theta}) + \varepsilon_{\mathrm{gen}} \\ &\le \widehat{\mathcal{L}}_n(\theta) + \varepsilon_{\mathrm{gen}} \le \widehat{\mathcal{L}}(\theta) + 2\varepsilon_{\mathrm{gen}}.\end{aligned} \tag{112}$$

Using the forward comparison at $\widehat{\theta}$ and the reverse comparison at θ therefore gives

$$\mathrm{Expl}(\pi_{\widehat{\theta}}) \le 4\mathcal{L}^A(\theta) + 4\varepsilon_{\mathrm{gen}} + 4\delta(\theta) + 2\delta(\widehat{\theta}). \tag{113}$$

The norm normalization in Appendix J gives $\delta(\theta) \le \mathcal{E}_A/2$ uniformly over θ. Taking the infimum gives the deterministic oracle inequality

$$\mathrm{Expl}(\pi_{\widehat{\theta}}) \le 4 \inf_{\theta\in\Theta} \mathrm{Expl}(\pi_\theta) + 4\varepsilon_{\mathrm{gen}} + 3\mathcal{E}_A. \tag{114}$$

No population minimizer is required. If $\mathcal{E}_A$ or $\varepsilon_{\mathrm{gen}}$ is infinite, the inequality holds with an infinite right-hand side.

To obtain the rate form in the theorem, we use the remaining assumptions. The norm normalization in Appendix J gives $\|(\widehat{A}^{\mathrm{DB}} - A)\pi_{\widehat{\theta}}\|_\star \le \|\widehat{A}^{\mathrm{DB}} - A\|_\star$. We separate the error of the second-stage regression from the bias induced by the fitted nuisances:

$$\widehat{A}^{\mathrm{DB}}(\cdot;\widehat{\eta}) - A = \left\{\widehat{A}^{\mathrm{DB}}(\cdot;\widehat{\eta}) - A^{\mathrm{DB}}(\cdot;\widehat{\eta})\right\} + \left\{A^{\mathrm{DB}}(\cdot;\widehat{\eta}) - A\right\}.$$

The first term has maximum entrywise $L_2(P_X)$ error e_A by Eq. (86). Let

$$b_n := \max_{j,k} \|A_{jk}^{\mathrm{DB}}(\cdot;\widehat{\eta}) - A_{jk}\|_{L_2(P_X)}.$$

The positivity and smoothness conditions in Assumptions 2 and 3 allow us to apply Eq. (102). Since $\|\max_m |v_m|\|_{L_2}^2 \le \sum_m \|v_m\|_{L_2}^2$ and the menu collections are fixed, the remainder-component rates in Assumption 6 give $b_n = O_p(r_\mu^2 + \underline{e}^{-1} r_\mu r_e)$. Under Assumption 5, the triangle inequality gives the same bound for the average of the fold-specific targets. The triangle inequality also gives $\max_{j,k} \|\widehat{A}_{jk}^{\mathrm{DB}} - A_{jk}\|_{L_2(P_X)} \le e_A + b_n$. Thus, writing C_{mat} for the norm-comparison constant defined in the notation paragraph,

$$\mathcal{E}_A \le \frac{C_{\mathrm{mat}}}{\kappa} \sum_{j,k} \|\widehat{A}_{jk}^{\mathrm{DB}} - A_{jk}\|_{L_2(P_X)}^2 \le \frac{2C_{\mathrm{mat}}K^2}{\kappa}(e_A^2 + b_n^2).$$

Squaring the rate for b_n yields

$$\mathcal{E}_A = \frac{1}{\kappa}\mathbb{E}_X\|\widehat{A}^{\mathrm{DB}} - A\|_\star^2 = O_p\left(\frac{K^2}{\kappa}\left(e_A^2 + r_\mu^4 + \underline{e}^{-2} r_\mu^2 r_e^2\right)\right). \tag{115}$$

The O_p constant depends on C_{mat}, L_f, and G_f. For non-Euclidean policy norms, C_{mat} can itself depend on K, whereas the Euclidean norm gives $C_{\mathrm{mat}} = 1$. For the affine aggregators of Section 4, $L_f = 0$ and $R_f \equiv 0$, so the r_μ^4 term is absent and the nuisance contribution is $\underline{e}^{-2} r_\mu^2 r_e^2$.

Substituting Eq. (115) into Eq. (114), expanding $\varepsilon_{\mathrm{gen}}$ as its defining supremum, and absorbing the factor 3 into $O_p(\cdot)$ proves exactly Eq. (18). The stochastic remainder is nonnegative and is bounded by a stochastically bounded multiplier of the displayed scale, which includes the random regression error e_A^2. Neither realizability nor a rate for the generalization deviation is used in this argument. □

J.9 Proof of Corollary 1

We specialize Theorem 3 under realizability and loss-class complexity in Assumptions 7 and 8, the bounded loss envelope in Assumption 2, and the additional rates in Eq. (87).

We prove a stronger comparison: the learner with estimated nuisances attains the same exploitability rate as an oracle that knows the true payoff matrix A. This oracle's error bound contains only the generalization term, whereas the oracle-nuisance learner in the corollary also incurs payoff-regression error.

Proof. Write $\mathcal{L}^A(\theta) := \mathrm{Expl}(\pi_\theta)$. Assumption 7 implies $\inf_{\theta\in\Theta} \mathcal{L}^A(\theta) = 0$. We first establish the true-payoff oracle rate, which does not use the nuisance or payoff convergence rates in Eq. (87).

We use the i.i.d. sampling in Assumption 5 and the complexity bound in Assumption 8 to control generalization. For a measurable function h, write $Ph := \mathbb{E}_X h(X)$ and $P_n h := n^{-1}\sum_i h(X_i)$, and set $Z_n := \sup_{h\in\mathcal{H}_n} |Ph - P_n h|$. The envelope H in Assumption 2 ensures that Z_n is finite and measurable. Let $X'_1, \ldots, X'_n$ be an independent copy of the contexts, with empirical average P'_n. Jensen's inequality gives the first bound below. Independently swapping each pair (X_i, X'_i) leaves its law unchanged, which gives the equality after introducing the signs. The triangle inequality then gives

$$\begin{aligned}
\mathbb{E}Z_n &\le \mathbb{E}_{X,X'} \sup_{h\in\mathcal{H}_n} |P'_n h - P_n h| \\
&= \mathbb{E}_{X,X',\sigma} \sup_{h\in\mathcal{H}_n} \left| \frac{1}{n}\sum_{i=1}^n \sigma_i\{h(X'_i) - h(X_i)\} \right| \\
&\le 2\mathfrak{R}_n(\mathcal{H}_n) \le 2Cn^{-1/2}.
\end{aligned}$$

Markov's inequality implies $Z_n = O_p(n^{-1/2})$.

To define the true-payoff oracle benchmark, let

$$\mathcal{L}_{n,\mathrm{or}}(\theta) := \frac{1}{n}\sum_{i=1}^n G_\Omega(\pi_\theta(X_i); A(X_i), X_i), \qquad \widehat{\theta}_{\mathrm{or}} \in \arg\min_{\theta\in\Theta} \mathcal{L}_{n,\mathrm{or}}(\theta).$$

Because $A \in \mathcal{B}_n$, its deviation $\varepsilon_{\mathrm{gen,or}} := \sup_{\theta\in\Theta} |\mathcal{L}^A(\theta) - \mathcal{L}_{n,\mathrm{or}}(\theta)|$ also satisfies $\varepsilon_{\mathrm{gen,or}} \le Z_n = O_p(n^{-1/2})$. For every $\theta \in \Theta$,

$$\mathcal{L}^A(\widehat{\theta}_{\mathrm{or}}) \le \mathcal{L}_{n,\mathrm{or}}(\widehat{\theta}_{\mathrm{or}}) + \varepsilon_{\mathrm{gen,or}} \le \mathcal{L}_{n,\mathrm{or}}(\theta) + \varepsilon_{\mathrm{gen,or}} \le \mathcal{L}^A(\theta) + 2\varepsilon_{\mathrm{gen,or}}.$$

Taking the infimum and using realizability proves $\mathrm{Expl}(\pi_{\widehat{\theta}_{\mathrm{or}}}) = O_p(n^{-1/2})$.

For the learner using the estimated payoff, retain $\varepsilon_{\mathrm{gen}}$ and $\mathcal{E}_A$ as defined in the theorem proof. On the event that the fitted payoff belongs to $\mathcal{B}_n$, we have $\varepsilon_{\mathrm{gen}} \le Z_n$. This event has probability tending to one, so $\varepsilon_{\mathrm{gen}} = O_p(n^{-1/2})$ even when the fitted payoff uses the policy-training observations. The deterministic comparison in Eq. (114) and realizability give

$$\mathrm{Expl}(\pi_{\widehat{\theta}}) \le 4\varepsilon_{\mathrm{gen}} + 3\mathcal{E}_A.$$

The matrix-norm comparison in the notation paragraph and the payoff-error decomposition used in Eq. (115) give

$$\mathcal{E}_A \lesssim \frac{1}{\kappa}\sum_{j,k} \|\widehat{A}^{\mathrm{DB}}_{jk} - A_{jk}\|^2_{L_2(P_X)} = O_p\left(\frac{K^2}{\kappa}\{e_A^2 + r_\mu^4 + \underline{e}^{-2} r_\mu^2 r_e^2\}\right). \tag{116}$$

The conditions in Eq. (87) imply $e_A^2 = O_p(n^{-1/2})$, $r_\mu^4 \lesssim n^{-1/2}$, and $\underline{e}^{-2} r_\mu^2 r_e^2 \lesssim n^{-1/2}$. Since K and the regularity constants are fixed, $\mathcal{E}_A = O_p(n^{-1/2})$. Substitution into Eq. (114) proves the same exploitability rate for $\widehat{\theta}$, which minimizes the estimated-payoff loss. For affine aggregators, the r_μ^4 term vanishes, so the separate condition on r_μ is unnecessary. Replacing the payoff and nuisance rate conditions by the corresponding little-o conditions makes $\mathcal{E}_A = o_p(n^{-1/2})$.

Finally, let $\widehat{A}^{\mathrm{ON}}$ be the payoff regression fitted using oracle nuisances, and let $\widehat{\theta}_{\mathrm{ON}}$ minimize its empirical gap loss. Since $A^{\mathrm{DB}}(\cdot;\eta_0) = A$, the nuisance remainder vanishes. Under the same loss-class conditions and payoff-regression rate $\max_{j,k} \|\widehat{A}^{\mathrm{ON}}_{jk} - A_{jk}\|_{L_2(P_X)} = O_p(n^{-1/4})$, Eq. (114) gives

$$\mathrm{Expl}(\pi_{\widehat{\theta}_{\mathrm{ON}}}) \le 4\varepsilon_{\mathrm{gen,ON}} + \frac{3}{\kappa}\mathbb{E}_X\|\widehat{A}^{\mathrm{ON}}(X) - A(X)\|_\star^2 = O_p(n^{-1/2}),$$

where $\varepsilon_{\mathrm{gen,ON}}$ is the uniform generalization deviation for the oracle-nuisance learner's fitted-payoff loss. Thus the learner with estimated nuisances attains the same convergence rate as this oracle-nuisance learner, as claimed in the corollary. □