RE-EVALUATING OPEN-ENDED EVALUATION OF LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluation has traditionally focused on ranking candidates for a specific skill. Modern generalist models, such as Large Language Models (LLMs), decidedly outpace this paradigm. Open-ended evaluation systems, where candidate models are compared on user-submitted prompts, have emerged as a popular solution. Despite their many advantages, we show that the current Elo-based rating systems can be susceptible to and even reinforce biases in data, intentional or accidental, due to their sensitivity to redundancies. To address this issue, we propose evaluation as a 3-player game, and introduce novel game-theoretic solution concepts to ensure robustness to redundancy. We show that our method leads to intuitive ratings and provide insights into the competitive landscape of LLM development.

1 INTRODUCTION

We can only improve what we measure, yet measuring the performance of Large Language Models (LLMs) has become an elusive endeavor owing to their breadth and depth of capabilities. Real-world benchmarks are costly to curate, increasingly requiring feedback from human domain experts (Hendrycks et al., 2021; Rein et al., 2023). Synthetic benchmarks can help, but their relevance to real-world performance is less clear (Zhang et al., 2024; Hsieh et al., 2024). An even more vexing challenge of static benchmarks is that of test set contamination, a phenomenon difficult to prevent despite efforts (Golchin & Surdeanu, 2024; Balloccu et al., 2024; Palavalli et al., 2024). Enumerating skills of interests with narrowly defined static benchmarks seems to be an uphill battle from the outset, as frontier models become generally capable.

An emerging trend in LLM evaluation is therefore to rely on open-ended evaluation systems, a notable example being the LMSYS Chatbot Arena (Chiang et al., 2024). In such a system, users submit prompts of interest, with each model assigned an Elo score (Elo, 1978) based on how they compare to each other on all prompts. In contrast to static benchmarks, this open-ended approach enjoys liveness, diversity and scale, lending itself to become an important reference in LLM development. Despite an intuitive sense of progress, issues around redundancy, bias and quality of crowdsourced data have been raised (Chiang et al., 2024; Ahuja et al., 2023; Li et al., 2024b). Several recent studies reverted back to centralized curation for quality (Taori et al., 2023; Lee et al., 2024; White et al., 2024). Increasing commercial efforts have been invested in private and proprietary evaluation too.

Perhaps this tension between quality and open-endedness is to be expected in LLM evaluation. Biases, redundancies and quality issues in the prompt distribution can affect Elo ratings, as they reflect performance *on average*. This along with other identified deficiencies of the Elo system (Balduzzi et al., 2018; Bertrand et al., 2023; Lanctot et al., 2023) raise crucial questions for LLM development: how does an Elo-based open-ended evaluation system affect model development today, and how can we mitigate its drawbacks, if any, in the future? In this paper, we provide an empirical simulated-based investigation of the former and lean on game theory for a solution to the latter.

The connection between evaluation and game theory needs unpacking. Consider a set of agents and a set of tasks, a naive approach to evaluation would rank agents by their average performance over tasks, propagating biases and redundancies in the task set. A game-theoretic approach (Balduzzi et al., 2018), would be to consider evaluation as an *agent-vs-task* game where the *agent* (*task*) player chooses one of its agents (tasks) and is rewarded (penalized) by the agent's performance on the task. This game-theoretic perspective accomplishes two goals simultaneously. First, it lets the evaluation system designers express their goals in players' objectives: here, Balduzzi et al. (2018) evaluates

agents under *adversarial* task selection. Second, a game-theoretic equilibrium decides which actions are played during evaluation: quality and redundancies in players' action sets do not matter. It is in this sense that game theory complements open-ended evaluation at a fundamental level.

Applying game theory to LLM evaluation however has its own challenges. Indeed, the decision of Balduzzi et al. (2018) in comparing agents under *adversarial* task selection was not a choice but a necessity. In 2-player zero-sum games, approximating a Nash equilibrium (NE, Nash et al. (1950)) is computationally tractable. NEs are also interchangeable in this setting as playing any NE guarantees zero exploitability. Beyond this setting, both benefits are lost: approximating NEs is computationally hard in the worst case (Daskalakis et al., 2006) and despite recent progress important challenges remain (Gemp et al., 2022; 2024). Equilibrium selection in this generalised setting remains a long-standing challenge too (Harsanyi & Selten, 1988; Rinott & Scarsini, 2000). For instance, driving on either side of the road is an equilibrium, but it is unclear which equilibrium should be used for evaluation. Past attempts at game-theoretic evaluation have therefore been restricted to the 2-player zero-sum settings when LLM evaluation calls for at least 3 players (e.g., *model-vs-model-vs-prompt*).

In this paper, we make several contributions that lead up to our equilibrium rating framework:

1. We show via a simulated example (Section 1.1) that the risk of models specializing in a few skills, at the expense of others, as they maximise their Elo ratings. Similarly, popular practice in prompt selection further reinforces this trend;
2. We introduce novel equilibrium solution concepts for N-player general-sum games that are unique and clone-invariant, a pre-requisite for our equilibrium rating method (Section 3);
3. We show our method scales to a real-world LLM evaluation dataset (Section 4.2) and provide ratings that are invariant to redundancy and correspond to our intuition in the sense of risk-dominance (Harsanyi & Selten, 1988), with empirical evidence (Appendix F.4);
4. We provide examples of analyzing these equilibrium structures of the game, drawing insights into the competitive landscape of LLM evaluation (Section 4.3).

1.1 Elo rating improvement path: a simulated example

With models continually improving their Elo ratings in systems such as LMSYS Chatbot Arena (Li et al., 2024a), it is worth asking if higher Elo scores translate to meaningful progress across skills of interest. This is difficult to answer from real-world data: we cannot replicate LLM development at scale nor can we disentangle factors driving model development besides maximizing leaderboard ratings. A synthetic example can provide insights in a controlled setting.

Consider S orthogonal skills of interests, M models and P prompts with each prompt a probability vector $\boldsymbol{p} \in \Delta^S$ over the skills and each model a vector $\boldsymbol{m} \in \mathbb{R}_+^S$, representing its competencies in each skill. We can then define the utility of selecting model $\boldsymbol{m}_i$ when compared to model $\boldsymbol{m}_j$ on prompt $\boldsymbol{p}_k$, as $u_m(\boldsymbol{p}_k, \boldsymbol{m}_i, \boldsymbol{m}_j) = \boldsymbol{p}_k^T(\boldsymbol{m}_i - \boldsymbol{m}_j)$ with $i, j \in [M]$ and $k \in [K]$. A less common but equally valid question is what should be the utility, if any, for selecting a prompt. We follow a similar definition as Li et al. (2024b) and define the utility in choosing prompt $\boldsymbol{p}_k$ as $u_p(\boldsymbol{p}_k, \boldsymbol{m}_i, \boldsymbol{m}_j) = |u_m(\boldsymbol{p}_k, \boldsymbol{m}_i, \boldsymbol{m}_j)|$. The *separability* of a prompt is then $\frac{1}{M^2}\sum_{ij} u_p(\boldsymbol{p}_k, \boldsymbol{m}_i, \boldsymbol{m}_j)$, consistent with the prompt selection criterion used in offline benchmarks such as `arena-hard-v0.1`.

We now observe how this system evolves with rating-maximizing players. Consider two settings: a) the "initial prompts" setting where the set of prompts is fixed but the set of models expands; and b) the "additional prompts" setting where prompt and model players alternate to introduce new prompts and models. We use a simple evolutionary process for our simulation (see Appendix F.1 for pseudocode). Let P_t and M_t be the number of prompts and models at iteration t and P_0, M_0 the number of initial prompts and models sampled from $\text{Dirichlet}(\mathbf{1}_{1:S})$. We introduce a model at each iteration which is a sum of improvement vectors sampled from $\text{Dirichlet}(\mathbf{1}_{1:S})$, such that the new addition receives the highest rating according the rating method used (i.e. Elo or our equilibrium-based method). In the "additional prompts" setting, a best-of-64 prompt is added at each iteration, selected by their separability when Elo ratings are used, and by their equilibrium ratings otherwise.

Figure 1 *(Center)* shows our findings. Let $H(\bar{\boldsymbol{p}}_t)$, $H(\bar{\boldsymbol{m}}_t)$ be the prompt and model skill entropy at iteration t with $\bar{\boldsymbol{p}}_t = \frac{1}{P_t}\sum_i^{P_t} \boldsymbol{p}_i$ and $\bar{\boldsymbol{m}}_t = \frac{1}{M_t}\sum_i^{M_t} \boldsymbol{m}_i$ and H the Shannon entropy. The Elo rating method leads to a consistent decline in skill entropy: the sequence of models improve

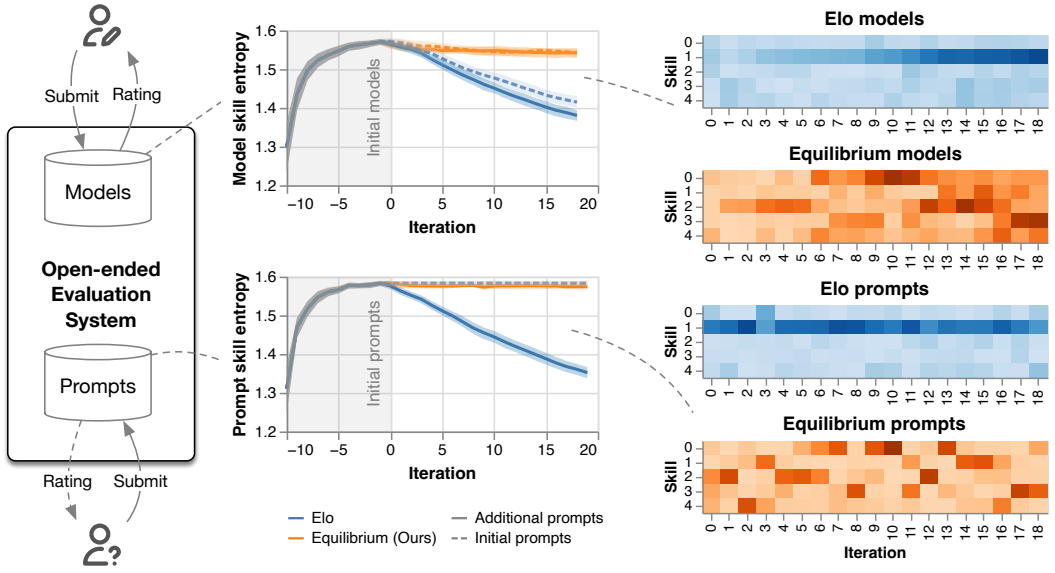

Figure 1: *(Left)* We simulate the effect of the rating method on model development with users submitting highly rated models (and prompts) iteratively. *(Center)* We show how model and prompt skill entropy evolves under different rating methods over 32 trials. *(Right)* We show an example sequence of models and prompts maximising their respective ratings. Darker indicates higher value.

along specific skill dimensions that is over-represented in the fixed set of initial prompts (dashed). Adding prompts with high separability further reinforces this trend in both model and prompt skill entropy (solid). We offer an intuitive explanation. Improvement on the Elo ratings or the separability metric reflects improvements against the *average*. At iteration t, the expected utility to model $\boldsymbol{m}_i$ is given by $u_m(\bar{\boldsymbol{p}}_t, \boldsymbol{m}_i, \bar{\boldsymbol{m}}_t)$ with its gradient defined by $\bar{\boldsymbol{p}}_t$. Improving on the most prevalent skill in $\bar{\boldsymbol{p}}_t$ therefore leads to the steepest ascent in utility. Similarly, the gradient for a prompt vector $\boldsymbol{p}_k$ is defined by the absolute deviation of the model vectors along each skill dimension $\frac{1}{M_t}\sum_i^t |\boldsymbol{m}_i - \bar{\boldsymbol{m}}_t|$. Prompts that target the skill dimension with the highest "spread" averaged across all model pairs is therefore the most highly rated. Figure 1 *(Right)* illustrates this phenomenon from a single trial. The Elo models become specialists on skill 1 due to prompt redundancy while new prompts become highly concentrated in search for higher separability.

The underlying challenge, one that we address, is to propose a practical rating method that compares models and prompts in a way that is intuitive and robust to redundancies. Figure 1 *(Center)* suggests that our NE ratings maintain skill entropy. Indeed, Figure 1 *(Right)* shows models focusing on different skills across iterations. As prompts are no longer measured against the *average* model pairs, they also remain diverse. In both cases, ratings are derived from a game-theoretic equilibrium, instead of an undifferentiated average. We now present our equilibrium-based evaluation framework.

2 Background

Normal-form game A normal-form game is a tuple $(N, \mathcal{A}, u)$ where N is a finite set of players $N = \{1, \ldots, n\}$ indexed by i, a tuple of strategy (action) sets $\mathcal{A} = (\mathcal{A}_1, \ldots, \mathcal{A}_n)$, and a tuple of utility functions $u = (u_1, \ldots, u_n)$ with $\mathcal{A}_i$ and $u_i : \mathcal{A} \to \mathbb{R}$ being player i's strategy set and utility respectively. Let $\boldsymbol{a} \in \mathcal{A} = (a_1, \ldots, a_n)$ with $a_i \in \mathcal{A}_i$ for all i denote a strategy *profile*. We allow strategy profiles to be selected randomly according to a distribution $\boldsymbol{x} \in \Delta(\mathcal{A})$ over *joint* actions. Let x_i denote the marginal distribution over player i's strategy set $\mathcal{A}_i$, i.e., $\boldsymbol{x}$ with all players $j \neq i$ marginalized out. Likewise, let x_{-i} denote the distribution with player i marginalized out. We call $\boldsymbol{x}$ a *pure* strategy if it places all mass on a single action profile and *mixed* otherwise. Each player's utility function is naturally extended to randomized strategy profiles by considering its expected value $u_i(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{x}}[u_i(\boldsymbol{a})]$. Similarly, let $u_i(x_i, x_{-i}) = \mathbb{E}_{\substack{a_i\sim x_i \\ a_{-i}\sim x_{-i}}}[u_i(\boldsymbol{a})]$.

Coarse Correlated Equilibrium (CCE) and Nash Equilibrium (NE) An equilibrium is a strategy profile $\boldsymbol{x}$ from which no player has an incentive to unilaterally deviate. Define player i's incentive to deviate to $x_i' \in \Delta(\mathcal{A}_i)$ unilaterally as $\text{regret}_i(x_i', \boldsymbol{x}) = u_i(x_i', x_{-i}) - u_i(\boldsymbol{x})$, where $\Delta(\mathcal{A}_i)$ is the simplex over $\mathcal{A}_i$. Then, player i's maximum regret for deviating from $\boldsymbol{x}$, is defined as:

$$\text{regret}_i(\boldsymbol{x}) = \max_{x_i' \in \Delta(\mathcal{A}_i)} \Big[\text{regret}_i(x_i', \boldsymbol{x})\Big] = \max_{x_i' \in \Delta(\mathcal{A}_i)} \Big[u_i(x_i', x_{-i})\Big] - u_i(\boldsymbol{x}). \tag{1}$$

The profile $\boldsymbol{x}$ is an approximate Coarse Correlated Equilibrium (Aumann, 1974; 1987) (ϵ-CCE) iff $\forall i, \text{regret}_i(\boldsymbol{x}) \leqslant \epsilon$. If $\boldsymbol{x}$ can be factorized into player marginals such that players cannot correlate, i.e., $\boldsymbol{x} = \times_{i=1}^{N} x_i$, then $\boldsymbol{x}$ is also an ϵ-NE. NEs are a subset of CCEs.

Equilibrium Selection Games can have many equilibria. Additional criteria are often introduced to make their selection unique. The set of CCEs is always convex, and so any strictly convex objective function such as negative Shannon entropy can used to select a unique equilibrium.

In contrast, the set of NEs need not be convex, however, several solutions have been proposed to solve for unique Nash equilibria in general-sum games (Harsanyi & Selten, 1988). The LLE was originally defined by McKelvey & Palfrey (1995) along with their introduction of quantal response (logit) equilibria (QREs). QREs are defined by a temperature parameter τ and can be interpreted as the Nash equilibria of a game with payoffs perturbed by Gumbel$(0, \tau)$ noise. Computing the LLE involves tracing a continuum of QREs, starting at temperature $\tau = \infty$ (corresponding to the uniform strategy profile) and ending at the LLE in the limit of $\tau = 0$. The LLE is unique in all games except a 0-measure set (McKelvey & Palfrey, 1995; Goeree et al., 2003). Another reason to solve for an LLE is that it falls into the family of homotopy methods (Herings & Peeters, 2010), which were shown to select *risk-dominant* equilibria in some general settings, a Nobel prize winning result of Harsanyi & Selten (1988). Empirically, LLEs have also been shown to approximate human play in games (McKelvey & Palfrey, 1995; Goeree et al., 2003).

3 Method

We now describe our rating method in terms of gamification, equilibrium solving and its selection. In gamification, we endow prompt and model players with utility functions, partly inspired by prior works, such that actions played at an equilibrium reflect our intuition. We note that our specific gamification defines an N-player general-sum game where equilibrium solving and selection requires more careful consideration. For equilibrium solving, we build on existing methods for approximating NEs and CCEs, reformulated to accommodate entropy-based techniques that select unique equilibria and explain why ratings derived from these equilibria remain vulnerable to manipulation in the face of redundant actions. We then propose a family of algorithms based on a novel kernelized entropy that select unique equilibria yet are also robust to redundant actions. Finally, for a given equilibrium solution $\boldsymbol{x}$, we define the rating of an action a_i to be $\text{regret}_i(a_i, \boldsymbol{x})$.

3.1 Gamification: Evaluation via a Game Between Models and Prompts

We study a 3-player general-sum game in our experiments. Consider a *prompt* player with $a_p \in \mathcal{A}_p$ the set of prompts, a *king* player and a *rebel* player each with actions $a_m \in \mathcal{A}_m$ the set of models. Let $u_k(a_p, a_m, a_m') \in \{-1, -1/2, 0, +1/2, +1\}$ be the utility function to the king player representing a preference towards king model response a_m over the rebel model response a_m' on a prompt a_p. The prompt player is rewarded for separating the models, with $u_p(a_p, a_m, a_m') = |u_k(a_p, a_m, a_m')|$. The rebel player receives $u_r(a_p, a_m, a_m') = -u_k(a_p, a_m, a_m')$ except for when $a_m = a_m'$ in which case $u_r(a_p, a_m, a_m') = -1$. This asymmetry discourages the same model being played by both model players deterministically with a prompt player indifferent over its actions. We refer to this game as `king-of-the-hill` as it favours the king player, leaving the rebel player to mount its best resistance without relying on some of the best models that the king player may choose. We refer to the king player ratings as the model ratings in our results.

Given a collection of prompts and models, the utility function can be tabulated with $|\mathcal{A}_p| \times |\mathcal{A}_m|^2$ pairwise preference ratings. In our experiments we query a `gemini-1.5-pro-api-0514` judge LLM for preference ratings similar to Zheng et al. (2023); Verga et al. (2024). We caveat that our

empirical results could therefore suffer from *self-preference* (Panickssery et al., 2024) and should *not* be viewed as an objective assessment of frontier LLMs.

3.2 Equilibrium Solving

For an instance of the evaluation game, we can compute different equilibrium solutions $\boldsymbol{x}$ which then define ratings. Here we present two options as they are unique, scalable and lead to intuitive, invariant ratings when combined with a selection criteria that we describe in Section 3.3.

Nash Equilibrium (NE) While LLE computation is typically formulated as solving a differential equation that evolves the temperature τ towards 0 while obeying the logit constraint $x_i = \mathrm{softmax}(\frac{1}{\tau}\nabla_{x_i} u_i)$ for all i (Turocy, 2005), this is also equivalent to satisfying the constraint $x_i = \arg\max_{z_i \in \Delta} u_i(z_i, x_{-i}) + \tau S(x_i)$ where $S(x_i)$ is the Shannon entropy of x_i. In this work, we choose another condition

$$x_i = \arg\max_{z_i \in \Delta} \left\{ u_i^\tau(z_i, x_{-i}) \stackrel{\text{def}}{=} u_i(z_i, x_{-i}) - \tau D_{\mathrm{KL}}(z_i || t_i) \right\} \tag{2}$$

which is equivalent in the case where the target strategy t_i is set to player i's uniform strategy. Using this definition of $u_i^\tau(z_i, x_{-i})$, we can define a loss function as in Gemp et al. (2022) such that $\arg\min_{\boldsymbol{x}} \mathcal{L}^\tau(\boldsymbol{x})$ is a QRE at temperature τ:

$$\mathcal{L}^\tau(\boldsymbol{x}) = \sum_i u_i^\tau(\mathrm{BR}_i, x_{-i}) - u_i^\tau(x_i, x_{-i}) \tag{3}$$

where player i's best response $\mathrm{BR}_i = \mathrm{softmax}(\frac{1}{\tau}\nabla_{x_i} u_i + \log(t_i))$.[1] By annealing τ from a high value and successively re-solving for the global minimum of $\mathcal{L}^\tau$, we can approximately trace the QRE continuum to the LLE. In Section 3.3, we explore non-uniform t_i to achieve clone-invariance.

Coarse Correlated Equilibrium (CCE) Solving for a unique CCE is computationally easier than NE as the problem is convex (Equation (1)). Therefore any strictly convex function can be used to uniquely select an equilibrium. For example, maximum entropy would be a suitable default criterion following the principle of maximum entropy. However, as we show in Section 3.3, a different target formulation is necessary for clone-invariance. As such, we opt for maximum relative entropy to a target joint $t = \times_{i=1}^n t_i$ to allow for non-uniform target joint distributions. A number of off-the-shelf solvers (Domahidi et al., 2013) and frameworks (Diamond & Boyd, 2016) can be used to compute solutions to this problem. We used a particularly efficient dual space gradient based algorithm described in Appendix A for scaling.

3.3 Invariant Equilibrium Selection

There may be many NEs and CCEs in N-player general-sum normal-form games of even moderate sizes (McLennan & Park, 1999; Sturmfels, 2002; McLennan, 2005). Many equilibria rely on sparse or heavily skewed strategy profiles (see examples in Appendix F.4). Intuitively, these equilibria are *risky* in the sense of *risk dominance*: playing one such equilibrium when other players do not would be a costly mistake. Our goal is to propose a selection procedure that in conjunction with our equilibrium solving algorithms, approximate a clone-invariant equilibrium.

Shannon entropy plays a key role in several equilibrium selection approaches, however, its definition is vulnerable to redundancy in games. Consider a game with 2 distinct actions A and B per player and introduce $b-1$ clones of B into player 1's action set. The maximum entropy strategy for player 1 in the new game is uniform across their actions with mass $\frac{1}{1+b}$ on each, but this induces a distribution that places $\frac{b}{1+b}$ cumulative mass on the cloned action B. From Section 3.2, the maximum Shannon entropy profile defines the precise starting point for tracing the path of QREs towards the LLE. This starting point is sensitive to clones. Hence, if we compute the LLE using the uniform distribution in this new game, we will effectively start from the (A, B) mixed-strategy $(\frac{1}{1+b}, \frac{b}{1+b})$ rather than the desired mixed-strategy $(1/2, 1/2)$; hence, will not necessarily arrive at the LLE of the original game.

Desired properties. A clone-invariant entropy definition should be:

[1] If t_i is the uniform distribution, $\log(t_i)$ is a constant vector and hence has no impact on the softmax.

P1. Real-valued, finite, and non-negative for any distribution x;

P2. Have a well-defined gradient for any x in the interior of the simplex;

P3. Its maximizers should form a convex set. In the case of duplicate strategies (clones), the maximizers should form precisely the set of distributions which arbitrarily distribute a mass of $\frac{1}{c}$ across each of the c sets of clones. In addition, they should achieve an entropy value which is equal to the entropy of the system with clones removed;

P4. Amenable to efficient estimation and flexible to re-interpretation of redundancy.

Note **P3.** resolves the issue with Shannon entropy that we highlighted above. **P1** is necessary for a reasonable measure of information content. **P2** is necessary for gradient-based optimization, and **P4** is practically helpful for efficient implementation and adaptation to bespoke game settings. We now introduce *affinity* entropy $H_a^p : \Delta \to \mathbb{R}$, a generalized Tsallis entropy (Tsallis, 1988) that recognizes similar or redundant strategies. Its derivation from the above axioms can be found in Appendix B.

Definition 1 (Affinity Entropy H_a^p)**.**

$$H_a^p(\boldsymbol{x}) = \frac{1}{p}\left[1 - \mathbf{1}^\top (U^{(p)}\boldsymbol{x})^{p+1}\right] \tag{4}$$

with entropic-index parameter $p \in (0, 1]$, $U^{(p)} = K\Lambda_p^{-1}$, and K a similarity kernel with entries in $[0, 1]$ with 1 indicating two strategies are clones, and Λ_p a diagonal matrix containing the $(p + 1)$-norms of the columns of K on its diagonal.

Theorem 1. *Affinity entropy H_a^p satisfies all desiderata **P1-P4**.*

In experiments, we define a similarity kernel $K^{(i)}$ for each player i with entries $K^{(i)}_{\alpha\beta}$ with

$$D^{(i)}_{\alpha\beta} = \mathbb{E}_{\boldsymbol{a}\sim U(\mathcal{A})}\left[\left(u_i(\alpha, a_{-i}) - u_i(\beta, a_{-i})\right)^2\right] \tag{5}$$

$$K^{(i)}_{\alpha\beta} = \exp(-D^{(i)}_{\alpha\beta}/(2\sigma)^2) \tag{6}$$

where D measures the strategic *dis*-similarity between player i's strategies α and β and K is simply a radial basis function (RBF) kernel under the metric D. Note $D^{(i)}_{\alpha\beta}$ is zero iff two strategies α and β achieve exactly the same utility for player i irrespective of the actions chosen by other players in the game. It should also be clear from the definition how one might Monte-Carlo estimate D. To select for an NE or a CCE, we set $t = \arg\max H_a^{p=1}(x)$ in Equation (2) and Equation (7) respectively.

4 Results

We use the same hyper-parameters for equilibrium solving in all results (see Appendix F.2). For evaluation on real-world prompts, we consider the `arena-hard-v0.1` dataset with 500 prompts, selected to separate frontier LLMs, as well as responses from many candidate LLMs. We consider responses from 17 LLMs in particular and queried `gemini-1.5-pro-api-0514` for 8 pairwise preference ratings on each prompt for each model pair. See Appendix F.3 for more details.

4.1 Equilibrium rating improvement path: a simulated example

Recall from Figure 1 that contrary to the Elo improvement path, maximizing equilibrium ratings led to models (and prompts) improving across skills. We inspect the equilibrium improvement path and offer our interpretation. Figure 2 *(Right)* shows that the shifts in focus between skills by the model player coincides with transitions in the NE prompts, or prompts weighted by their NE strategies (shown in Figure 2 *(Center)*). Similarly, to gain support under an NE, new prompts must highlight a skill dimension along which equilibrium models are better differentiated (Figure 2 *(Left)*). In sum, equilibrium prompts separate equilibrium models. This dynamic encourages exploration of new skill dimensions and incentivizes models to be well-rounded across skills.

4.2 Invariant Evaluation

We now turn to `arena-hard-v0.1` and show that candidate LLMs' equilibrium ratings are invariant to redundancies when their Elo ratings are not. In this experiment, we will introduce prompts

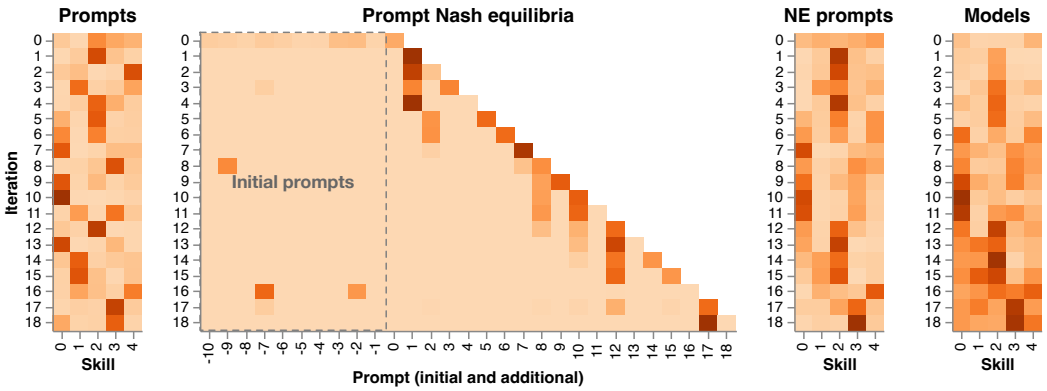

Figure 2: We inspect the model improvement path induced by NE ratings as shown in Figure 1 *(Right)*. *(Left)* shows the sequence of additional prompts added at each iteration. Each prompt is the best-of-64 samples according to their NE ratings. *(Center)* shows the sequence of prompt player NEs. Each row defines a distribution over prompts. *(Right)* shows the equilibrium-weighted prompt skills and the sequence of king player models. Recall prompts and models are non-negative vectors over skills, darker indicates higher focus or capability in each skill.

targeted at bringing down the rating of a certain action (in this case, model). Specifically, let $\bar{\boldsymbol{u}}_k(a_k) = \frac{1}{|\mathcal{A}_m|}\sum_{a_r} u_k(\cdot, a_k, a_r)$ be the vector of expected king player payoffs when playing action a_k against a randomly chosen rebel model on each prompt. We can then sample prompts adversarial to a_k from $\mathrm{softmax}(-\lambda\bar{\boldsymbol{u}}_k(a_k))$ and add them to the prompt set. Figure 3 reports the king model rankings under different methods with $a_k = \texttt{gemini-1.5-pro-api-0514}$ and $\lambda = 10$.

Our first observation is that without redundant adversarial prompts, our proposed equilibrium rankings of LLMs are fairly consistent with their Elo rankings, with a few models moving up or down one or two positions. This deserves attention. Out of a multiplicity of equilibria, the NE and CCE we selected led to rankings that correspond to our intuition. Indeed, we show in Appendix F.4 that the NE we select is risk-dominant among 128 mixed-strategy NEs of this game. Second, the Elo ratings can be arbitrarily influenced by redundancy, with the top-ranked model falling through the ranks. Equilibrium rankings remain invariant. In fact, while we lose the invariance guarantee with *near* redundant prompts, we show models' equilibrium rankings to degrade gracefully in Appendix F.5. Third, the CCE ratings show the top-3 models to tie for the first place: correlating models with prompts affects the competitive landscape which we inspect in Section 4.3. Lastly, solving for a unique equilibrium is not sufficient for invariant ratings. We show in Figure 3 *(Right)* that using Shannon's entropy for tracing the QRE continuum or for selecting a max-entropy CCE would not lead to invariant ratings. For completeness, we provide a detailed breakdown of our equilibrium ratings in terms of action ratings and marginals for each player in Appendix F.5.

4.3 Interpreting Equilibrium Solutions

Besides rankings, the equilibrium solutions can surface insights into the evaluation game dynamics. We share two examples using NE and CCE solutions respectively from ratings shown in Section 4.2.

Nash Equilibrium Prompts We have shown that equilibrium ratings are intuitive and invariant to redundancy. A follow-up question is which actions are highly-rated and which actions affect other players' ratings (i.e. with positive support at the NE).

Recall that the prompt player utility $u_p(a_p, a_k, a_r) = |u_k(a_p, a_k, a_r)|$ reflects the extent to which a prompt separates the pair of responses from models a_k and a_r. The prompt player's equilibrium rating is then $\mathrm{regret}(a_p, \boldsymbol{x}) = \mathbb{E}_{\substack{a_k\sim x_k \\ a_r\sim x_r}} u_p(a_p, a_k, a_r)$ with x_k, x_r the NE strategies of the king and rebel player respectively. By definition, prompts that are highly rated under NE ratings separate

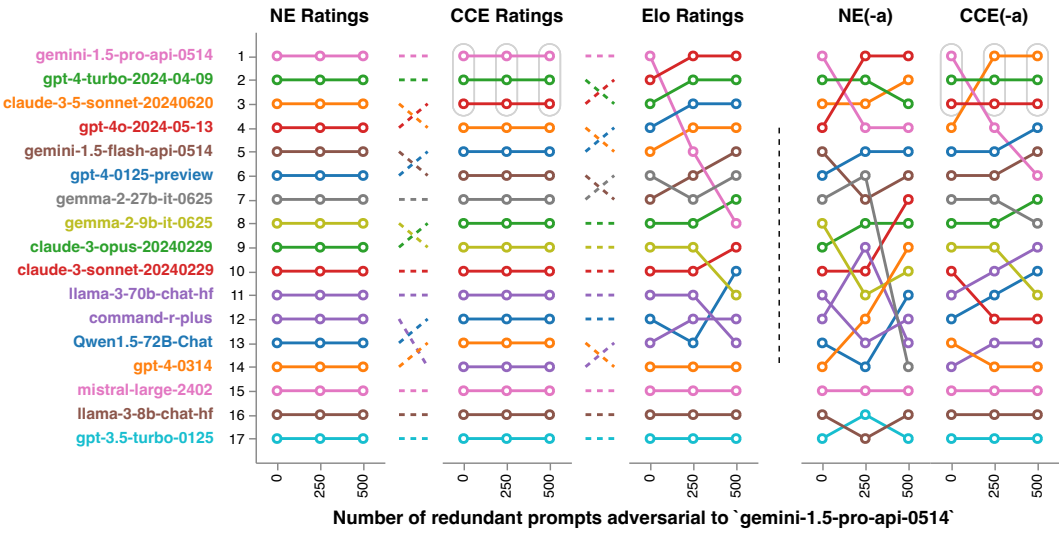

Figure 3: We introduce an increasing number of redundant copies of prompts adversarial to `gemini-1.5-pro-api-0514` and show model rankings under each method. Models at the same rank are grouped in grey and ordered alphabetically. *(Right)* We show equilibrium rankings under NE(-a) and CCE(-a) selected using Shannon's entropy instead of the affinity entropy. Dotted lines connecting different rating panels indicate continuity in the labeling. For instance, `gemini-1.5-pro-api-0514` consistently ranks first under our NE and CCE ratings, despite the introduction of up to 500 redundant adversarial prompts. However, its ranking suffered significantly under the Elo ratings as soon as 250 adversarial prompts have been introduced.

models played at the NE. In other words, while the Elo ratings reflect the strength of an action on average, equilibrium ratings reflect the strength of actions at the selected equilibrium.

We can now illustrates these phenomena using the same game investigated in the second columns of Figure 3, with 250 redundant prompts added to the game. First, we show in Figure 4 *(Top)* the *king-vs-rebel* payoff matrices induced by 6 sample prompts, with increasing equilibrium prompt ratings. Prompts with low ratings tend to fail to differentiate performant models (i.e. top-left block of each heatmap). Second, we can ask which prompts should we expect to have support at an equilibrium. Figure 4 *(Bottom)* shows that empirically, highly rated prompts are played more often at the equilibrium we select. This implies that the model ratings are heavily influenced by a small subset of prompts that separate frontier models. We note that this correlation is not guaranteed, following our discussion in Section 4.2 on redundant actions. Indeed, our final observation is that prompts that are redundant with other prompts tend to receive lower probability mass than their ratings would have required. In fact, since we have introduced 250 redundant prompts explicitly, we can highlight in gray prompts that are indeed redundant — many of these prompts enjoy high ratings, but significantly lower support. In other words, equilibrium ratings reflect quality of an action in isolation while equilibrium support further takes into account redundancy of an action with respect to other actions available to a player. This observation is even clearer in research games studied in Appendix D-E.

Marginal rating contribution by co-player action With ratings derived from underlying equilibria, we can decompose the rating of each action into a sum of marginal contributions from each co-player's actions. Recall from Equation (1) that the rating of an action a_i is its $\text{regret}_i(a_i, \boldsymbol{x}) = u_i(a_i, x_{-i}) - u_i(\boldsymbol{x})$. We can decompose the rating of player i's action a_i into a weighted sum of each of player j's contributions, with $\delta(a_i, a_j, \boldsymbol{x}) = x_j(a_j)\left[u_i(a_j, x_{-j}) - u_i(a_i, a_j, x_{-i,-j})\right]$ the marginal contribution of a_j to a_i's equilibrium rating. Note that $\text{regret}(a_i, \boldsymbol{x}) = \sum_{a_j} \delta(a_i, a_j, \boldsymbol{x})$. The rating of a_i describes the *cost* incurred by player i if it were to deviate to play a_i from an equilibrium $\boldsymbol{x}$. $\delta(a_i, a_j, \boldsymbol{x})$ therefore explains a_j's contribution in player i's decision in not deviating.

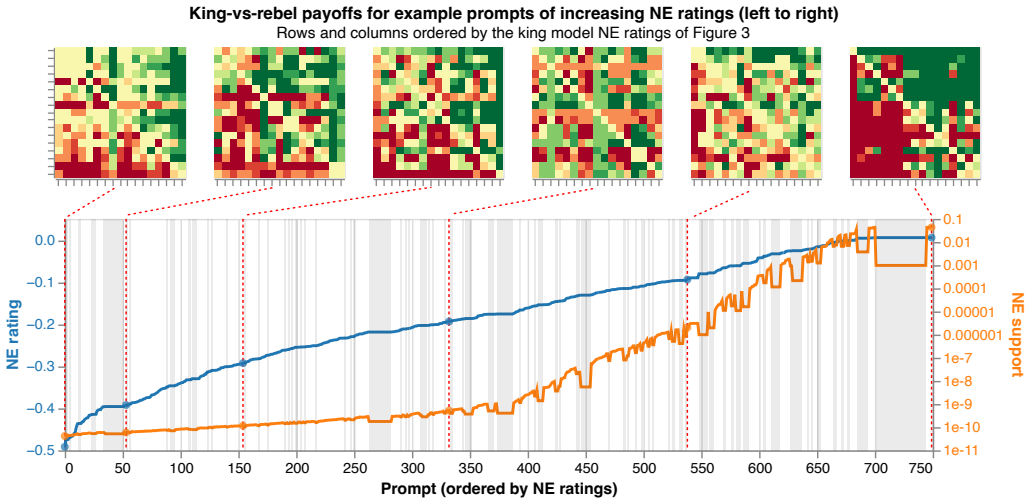

Figure 4: We show that highly rated prompts generally have high support under the NE. Redundant prompts (highlighted by the gray bands) would receive identical ratings but notably lower support. In sum, equilibrium ratings reflect separability of each prompt with respect to the model equilibrium strategies in isolation, whereas equilibrium support of each prompt further accounts for its redundancy with respect to other prompts. *(Top)* We show the *king-vs-rebel* payoffs induced by example prompts. Green indicates king-player winning and red losing. Highly rated prompts tend to discriminate between strong models (top-left corners). *(Bottom)* We show the NE supports and ratings of all prompts, ordered by their NE ratings.

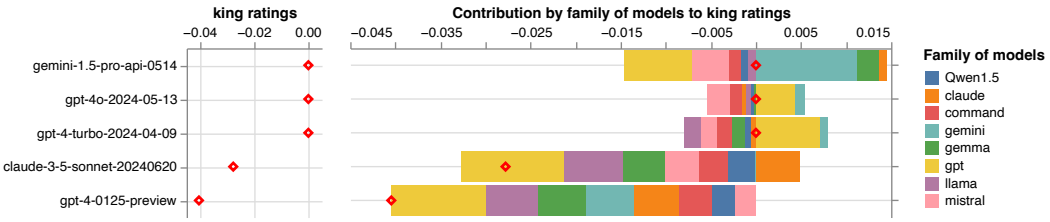

Figure 5: The CCE joint distribution surfaces interesting game dynamics in the underlying comparison data which we can leverage for more insights. Here, the width of each bar representing a model family $\mathcal{F}$ corresponds to $\sum_{a_j \in \mathcal{F}} \delta(a_i, a_j, \boldsymbol{x})$ with a_i a king player model choice and a_j a rebel model belonging to the family $\mathcal{F}$. A model's family is determined by its model name prefix. For brevity, we show the king model rating breakdown for the top 5 out of 17 models. Interestingly, we observe that the 3 top-ranked king models tend to do well against other models in its own family. The Gemma family of models, improves the rating of `gemini-1.5-pro-api-0514` but detract from the ratings of GPT models. Mistral and Llama family of models, on the other hand, surface weaknesses in all 3 top-ranked models, contributing negatively to top-ranked king models' ratings.

Recall from Figure 3 where several models tied for the first place under the CCE dynamics but are fully differentiated under NE. We can now leverage the marginal contribution analysis to understand the mechanism underlying this phenomenon. Figure 5 shows the CCE king model ratings decomposed from the perspective of the rebel player. In other words, we ask which rebel models contribute most positively or negatively to each king model's CCE rating. For clarity of presentation, we focus on the top 5 models and we group rebel models into families of models if they share the same naming prefix. The contribution of each family of model is therefore the sum of the contribution by models within each family $\mathcal{F}$ or $\sum_{a_r \in \mathcal{F}} \delta(a_k, a_r, \boldsymbol{x})$ with a_k a king model and a_r a rebel model.

We make several remarks. First, all 3 top-ranked king models benefit the most when compared against rebel models in their own model family: the GPT family (Achiam et al., 2023) of models contribute positively to the ratings of `gpt-4o-2024-05-13` and `gpt-4-turbo-2024-04-09`. Similarly, `gemini-1.5-flash-api-0514`, the only other model in the Gemini family (Team et al., 2023), improves `gemini-1.5-pro-api-0514`'s rating the most. We speculate that this can be a result of model developers selecting models to release based on favourable comparisons to their earlier or smaller models. Second, all top-ranked models remain vulnerable to open-weight models such as the Mistral (Jiang et al., 2023) and Llama (Dubey et al., 2024) families of models. More fine-grained analysis may shed light on the prompts on which these losses tend to occur.

We caveat that our results are in part derived from the preference ratings of a `gemini-1.5-pro-api-0514` model and may not reflect the true dynamics of real-world LLM development. Nevertheless, the interpretability offered by the game-theoretic equilibria further distinguishes game-theoretic evaluation from prior works to be discussed in the Section 5.

To further help build intuition in interpreting equilibrium ratings, we refer readers to Appendix D-E where we show examples of similar analysis in the context of traditional research games of *rock-paper-scissors* and *chicken*.

5 Related Works

There is a rich body of literature studying rating methods with applications in Chess, Go, Tennis and video games. One family of probabilistic methods follows the Bradley-Terry model and predict pairwise win probabilities from ratings. A widely used example is Elo (Elo, 1978) with extensions Bayes-Elo, mElo and Elo-MMR (Coulom, 2008; Balduzzi et al., 2018; Ebtekar & Liu, 2021; Vadori & Savani, 2024) capturing temporal variation, cyclicality and ordinal ranks in data. A separate line of works draws from Social Choice Theory (SCT, Sen (1977); Lanctot et al. (2023)), traditionally studied in the context of voting systems driven by axioms. One relevant axiom here is the *independence of clones*: rankings should be invariant to redundant candidates being added[2]. However, redundancy in votes is out of scope. Finally, game-theoretic evaluation has been previously studied in Balduzzi et al. (2018) and Marris et al. (2022b). Our method generalises to N-player general-sum settings, with practical equilibrium solving and selection algorithms.

6 Conclusions

We studied the effect of maximizing Elo ratings in the context of open-ended evaluation and showed that its sensitivity to redundancy could bias model (and prompt) selection. We then proposed an equilibrium rating framework, with practical equilibrium solving and selection algorithms that can scale to real-world LLM evaluation. We show our method to provide intuitive and robust rankings of models (and prompts), with interpretable structures.

We see several exciting future directions. First, although our methods can scale to tens of thousands of prompts and tens of models on commodity hardware, scaling further would be challenging. Tabulating the evaluation payoff tensor with pairwise preference ratings can be costly too. Stochastic equilibrium solving (Gemp et al., 2022) or payoff prediction (Liu et al., 2024) might help. Finally, research into alternative solution concepts, or how we could leverage their equilibrium structure for analysis (e.g. prompt and model pruning) is also promising. Finally, while we target LLM evaluation in specific, our methodology can be applied more generally to other domains.

[2]One way to satisfy this axiom is game-theoretic in nature: the *voter margin matrix* (Lanctot et al., 2023) can be viewed as a 2-player zero-sum game, such that invariance properties of max-entropy NE ratings applies.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, et al. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*, 2023.

Robert J Aumann. Subjectivity and correlation in randomized strategies. *Journal of mathematical Economics*, 1(1):67–96, 1974.

Robert J Aumann. Correlated equilibrium as an expression of bayesian rationality. *Econometrica: Journal of the Econometric Society*, pp. 1–18, 1987.

David Balduzzi, Karl Tuyls, Julien Perolat, and Thore Graepel. Re-evaluating evaluation. *Advances in Neural Information Processing Systems*, 31, 2018.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondřej Dušek. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source llms. *arXiv preprint arXiv:2402.03927*, 2024.

Quentin Bertrand, Wojciech Marian Czarnecki, and Gauthier Gidel. On the limitations of the elo, real-world games are transitive, not additive. In *International Conference on Artificial Intelligence and Statistics*, pp. 2905–2921. PMLR, 2023.

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

Rémi Coulom. Whole-history rating: A bayesian rating system for players of time-varying strength. In *International conference on computers and games*, pp. 113–124. Springer, 2008.

Constantinos Daskalakis, Aranyak Mehta, and Christos Papadimitriou. A note on approximate nash equilibria. In *International Workshop on Internet and Network Economics*, pp. 297–306. Springer, 2006.

Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

A. Domahidi, E. Chu, and S. Boyd. ECOS: An SOCP solver for embedded systems. In *European Control Conference (ECC)*, pp. 3071–3076, 2013.

Konstantinos Drakakis, UCD CASL, and Barak A Pearlmutter. On the calculation of the l2→ l1 induced matrix norm. *International Journal of Algebra*, 3(5):231–240, 2009.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak,

Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li,

Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Aram Ebtekar and Paul Liu. Elo-mmr: A rating system for massive multiplayer competitions. In *Proceedings of the Web Conference 2021*, pp. 1772–1784, 2021.

Arpad E. Elo. *The Rating of Chessplayers, Past and Present*. Arco Pub., New York, 1978. ISBN 0668047216 9780668047210. URL http://www.amazon.com/Rating-Chess-Players-Past-Present/dp/0668047216.

Ian Gemp, Rahul Savani, Marc Lanctot, Yoram Bachrach, Thomas Anthony, Richard Everett, Andrea Tacchetti, Tom Eccles, and János Kramár. Sample-based approximation of nash in large many-player games via gradient descent. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pp. 507–515, 2022.

Ian Gemp, Luke Marris, and Georgios Piliouras. Approximating nash equilibria in normal-form games via stochastic optimization. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=cc8h3I3V4E.

Jacob K Goeree, Charles A Holt, and Thomas R Palfrey. Risk averse behavior in generalized matching pennies games. *Games and Economic Behavior*, 45(1):97–113, 2003.

Shahriar Golchin and Mihai Surdeanu. Time travel in LLMs: Tracing data contamination in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2Rwq6c3tvr.

John C Harsanyi and Reinhard Selten. A general theory of equilibrium selection in games. *MIT Press Books*, 1, 1988.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

P Jean-Jacques Herings and Ronald Peeters. Homotopy methods to compute equilibria in game theory. *Economic Theory*, 42(1):119–156, 2010.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36, 2024.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline, 2024a.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From live data to high-quality benchmarks: The arena-hard pipeline, April 2024b. URL https://lmsys.org/blog/2024-04-19-arena-hard/.

Siqi Liu, Luke Marris, Georgios Piliouras, Ian Gemp, and Nicolas Heess. Nfgtransformer: Equivariant representation learning for normal-form games. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=4YESQqIys7.

Luke Marris, Ian Gemp, Thomas Anthony, Andrea Tacchetti, Siqi Liu, and Karl Tuyls. Turbocharging solution concepts: Solving NEs, CEs and CCEs with neural equilibrium solvers. *CoRR*, abs/2210.09257, 2022a. doi: 10.48550/ARXIV.2210.09257. URL https://arxiv.org/abs/2210.09257.

Luke Marris, Marc Lanctot, Ian Gemp, Shayegan Omidshafiei, Stephen McAleer, Jerome Connor, Karl Tuyls, and Thore Graepel. Game theoretic rating in n-player general-sum games with equilibria, 2022b. URL https://arxiv.org/abs/2210.02205.

Richard D McKelvey and Thomas R Palfrey. Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38, 1995.

Andrew McLennan. The expected number of nash equilibria of a normal form game. *Econometrica*, 73(1):141–174, 2005.

Andrew McLennan and In-Uck Park. Generic $4\times$ 4 two person games have at most 15 nash equilibria. *Games and Economic Behavior*, 26(1):111–130, 1999.

John F Nash et al. Non-cooperative games. *Annals of Mathematics*, 1950.

Medha Palavalli, Amanda Bertsch, and Matthew R. Gormley. A taxonomy for data contamination in large language models, 2024. URL https://arxiv.org/abs/2407.08716.

Arjun Panickssery, Samuel R Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*, 2024.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level Google-proof Q&A benchmark, 2023.

Yosef Rinott and Marco Scarsini. On the number of pure strategy nash equilibria in random games. *Games and Economic Behavior*, 33(2):274–293, 2000.

Amartya Sen. Social choice theory: A re-examination. *Econometrica: journal of the Econometric Society*, pp. 53–89, 1977.

Bernd Sturmfels. *Solving systems of polynomial equations*. Number 97. American Mathematical Soc., 2002.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Constantino Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics*, 52:479–487, 1988.

Theodore L Turocy. A dynamic homotopy interpretation of the logistic quantal response equilibrium correspondence. *Games and Economic Behavior*, 51(2):243–263, 2005.

Nelson Vadori and Rahul Savani. Ordinal potential-based player rating. In *International Conference on Artificial Intelligence and Statistics*, pp. 118–126. PMLR, 2024.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*, 2024.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, et al. Livebench: A challenging, contamination-free llm benchmark. *arXiv preprint arXiv:2406.19314*, 2024.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. ∞Bench: Extending long context evaluation beyond 100K tokens. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15262–15277, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.814.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.