A COMPUTING THE MAXIMUM RELATIVE ENTROPY CCE

A maximum relative entropy CCE, that minimises the distance of the log-joint, $\log(x(a))$, to a target log-joint, $t(a) \in \mathbb{R}^{|\mathcal{A}|}$, can be computed using gradient descent. We formulate the problem in dual space (Marris et al., 2022a) with dual parameters, $\alpha_i(a'_i) \in \mathbb{R}_+^{|\mathcal{A}_i|}\ \forall i$, defined as functions, $\alpha_i(a'_i) = \text{softplus}(\theta_i(a'_i))\ \forall i$, of learned parameters, $\theta(a'_i) \in \mathbb{R}^{|\mathcal{A}_i|}\ \forall i$. Let $l_\theta(a)$ be a logit term used to construct the loss function.

$$l_\theta(a) = -\sum_i \sum_{a'_i} \alpha_p(a'_i)\left[u_i(a'_i, a_{-i}) - u_i(a)\right] + t(a) \tag{8}$$

Minimizing a loss function, $\min_\theta L_\theta$, converges to optimal dual variables, $\alpha_i^*(a'_i) = \text{softplus}(\theta_i^*(a'_i))\ \forall i$ with $L_\theta = \log\left[\sum_a \exp\left[l_\theta(a)\right]\right]$. The loss is convex, deterministic, and unconstrained. Therefore many optimization algorithms are suitable. The primal joint can be simply recovered from the optimal logit term $x_\theta(a) = \text{softmax}\left[l_{\theta*}(a)\right]$.

B AFFINITY ENTROPY

Consider defining a modified Tsallis entropy H_a^p with temperature parameter $p \in (0, 1]$ as:

$$H_a^p(\boldsymbol{x}) = \frac{1}{p}\Big[1 - \boldsymbol{z}^\top \boldsymbol{z}\Big] = \frac{1}{p}\Big[1 - \sum_i (U_i^{(p)}\boldsymbol{x})^{p+1}\Big] \tag{9}$$

where $\boldsymbol{z} = (U^{(p)}\boldsymbol{x})^{\frac{p+1}{2}}$. Note that this definition recovers the standard definition of Tsallis entropy when $U^{(p)}$ is the identity matrix.

Remark. *$U_{ij}^{(p)} \geqslant 0$ for all entries for H_a^p to be real-valued.*

$U_{ij}^{(p)}$ must be non-negative for every i, j, otherwise, there exists $\boldsymbol{x} = \boldsymbol{e}_j$ where $\boldsymbol{e}_j$ is a standard-basis vector such that $U_i^{(p)}\boldsymbol{x} < 0$ and $(U_i^{(p)}\boldsymbol{x})^{p+1}$ is not real for $p \in (0, 1)$.

Remark. *The $(p+1)$-norm of each column of $U^{(p)}$ must be less than or equal to 1 for H_a^p to be non-negative for any $\boldsymbol{x} \in \Delta$.*

We need $\boldsymbol{z}^\top \boldsymbol{z} \leqslant 1$ for $p \in (0, 1]$ and any $\boldsymbol{x} \in \Delta$. Equivalently, we require $(\boldsymbol{z}^\top \boldsymbol{z})^{\frac{1}{p+1}} \leqslant 1$ for $p \in (0, 1]$.

Note $(\boldsymbol{z}^\top \boldsymbol{z})^{\frac{1}{p+1}} = \left(\sum_i (U_i^{(p)}\boldsymbol{x})^{p+1}\right)^{\frac{1}{p+1}} = ||U^{(p)}\boldsymbol{x}||_{p+1}$. Therefore, we require

$$1 \geqslant \sup_{\boldsymbol{x}\in\Delta} ||U^{(p)}\boldsymbol{x}||_{p+1} \tag{10}$$

$$= \sup_{||\boldsymbol{x}||_1=1} ||U^{(p)}\boldsymbol{x}||_{p+1} \quad \text{for} \quad U^{(p)} \geqslant 0 \tag{11}$$

$$= ||U^{(p)}||_{1,p+1} \tag{12}$$

$$= \max_{j=1} ||U_{\cdot,j}^{(p)}||_{p+1} \quad \text{by Drakakis et al. (2009).} \tag{13}$$

Remark. *Among all admissible $U^{(p)}$, defining $U^{(p)}$ such that its columns have exactly unit $(p+1)$-norm achieves $\min_{U^{(p)}} \min_{\boldsymbol{x}\in\Delta} H_p^a(\boldsymbol{x})$.*

This follows from the previous remark and is desireable for the sake of defining a "tight" definition of entropy. Intuitively, by the conditions set thus far, $U^{(p)} = \mathbf{0}$ is admissible. Yet, this gives a loose definition of entropy where $H_a^p = 1/p$. It turns out that this intuition is required in the limit as $p \to 0$.

Remark. *$U^{(p)}$ must be precisely column stochastic for H_a^p to remain finite in the limit of $p \to 0$.*

In the limit $p \to 0$, the denominator of H_a^p goes to zero, therefore, by L'Hôpital's rule, the numerator must as well. The numerator goes to $z^\top z = \sum_i U_i^{(p)} x = \mathbf{1}^\top U^{(p)} x$. Therefore,

$$\forall x \in \Delta^{d-1} \quad 1 - \mathbf{1}^\top U^{(p)} x = 0. \tag{14}$$

Finite distributions only obey a single equality constraint, that is $x^\top \mathbf{1} = 1$, therefore it must be the case that $\mathbf{1}^\top U^{(p)} = \mathbf{1}^\top$, i.e., $U^{(p)}$ is column stochastic.

Remark. *H_a^p is concave in $\boldsymbol{x}$.*

Let $y_i = U_i^{(p)}\boldsymbol{x}$. Then each element of the sum, y_i^{p+1} is a convex function in y_i, which itself is a linear transformation on $\boldsymbol{x}$. Therefore, $\sum_i (U_i^{(p)}\boldsymbol{x})^{p+1}$ is convex in $\boldsymbol{x}$. Hence H_a^p is concave in $\boldsymbol{x}$.

Remark. *The gradients $\nabla_{\boldsymbol{x}} H_a^p$ are well-defined.*

Recall (9), then:

$$\frac{\partial H_a^p}{\partial x_j} = -\frac{p+1}{p}\sum_i (U_i^{(p)}\boldsymbol{x})^p U_{ij}^{(p)} \tag{15}$$

$$\nabla_{\boldsymbol{x}} H_a^p = -\frac{p+1}{p}(U^{(p)})^\top (U^{(p)}\boldsymbol{x})^p \tag{16}$$

which is well-defined for any choice of $U_{ij}^{(p)} \geqslant 0$ for all i, j.

Remark. *H_a^p is well-defined in the limit as $p \to 0$, i.e., Shannon* affinity *entropy is well-defined.*

It is known that Shannon entropy can be recovered from Tsallis entropy in the limit as $p \to 0$. We repeat that derivation here and use L'Hôpital's rule. The derivative of the denominator is 1, hence we find the limit is given by the finite derivative of the numerator:

$$\frac{d[pH_a^p]}{dp} = -\frac{d}{dp}\Big[\sum_i y_i^{p+1}\Big] \tag{17}$$

$$= -\frac{d}{dp}\Big[\sum_i e^{(p+1)\log(y_i)}\Big] \tag{18}$$

$$= -\sum_i \big(\log(y_i) + (p+1)\frac{1}{y_i}\frac{dy_i}{dp}\big)e^{(p+1)\log(y_i)}. \tag{19}$$

In the limit $p \to 0$, the derivative evaluates to

$$\frac{d[pH_a^p]}{dp} = -\sum_i \Big[e^{(p+1)\log(y_i)}\log(y_i)\Big]\Big|_{p=0} - (p+1)\sum_i \Big[\frac{1}{y_i}\frac{dy_i}{dp}e^{(p+1)\log(y_i)}\Big]\Big|_{p=0} \tag{20}$$

$$= -\sum_i y_i \log(y_i) - \sum_i \frac{dy_i}{dp}\Big|_{p=0} \tag{21}$$

$$= S(y) - \sum_i \frac{dy_i}{dp}\Big|_{p=0}. \tag{22}$$

Remark. *Let K be a similarity matrix between actions with non-negative entries with positive column-sums. Then $U^{(p)} = K\,\mathtt{diag}\big(1/(\mathbf{1}^\top K^{p+1})^{1/(p+1)}\big)$ satisfies the conditions stated above for $U^{(p)}$.*

Remark. *Under the above choice of $U^{(p)}$, Shannon* affinity *entropy $S_a = H_a^{p\to 0}$ can be derived as:*

$$S_a(\boldsymbol{x}) = S(U^{(0)}\boldsymbol{x}) - \sum_j \Big[\log(\sum_i K_{ij}) - \sum_i U_{ij}^{(0)}\log(K_{ij})\Big]x_j. \tag{23}$$

The necessary y_i term can be rewritten and its derivative (evaluated at $p = 0$) can be derived as follows:

$$y_i = U_i^{(p)}\boldsymbol{x} = \sum_j \frac{K_{ij}}{(\sum_{i'} K_{i'j}^{p+1})^{\frac{1}{p+1}}} x_j \tag{24}$$

$$= \sum_j K_{ij}x_j (\sum_{i'} K_{i'j}^{p+1})^{-\frac{1}{p+1}} \tag{25}$$

$$= \sum_j K_{ij}x_j e^{-\frac{1}{p+1}\log(\sum_{i'} K_{i'j}^{p+1})} \tag{26}$$

$$= \sum_j K_{ij}x_j e^{-\frac{1}{p+1}\log\left(\sum_{i'} e^{(p+1)\log(K_{i'j})}\right)} \tag{27}$$

$$\frac{dy_i}{dp} = \sum_j K_{ij}x_j e^{-\frac{1}{p+1}\log\left(\sum_{i'} e^{(p+1)\log(K_{i'j})}\right)} \Big[\frac{1}{(p+1)^2}\log\left(\sum_{i'} e^{(p+1)\log(K_{i'j})}\right) \tag{28}$$

$$- \frac{1}{p+1}\frac{1}{\sum_{i'} e^{(p+1)\log(K_{i'j})}} \sum_{i'} \log(K_{i'j}) e^{(p+1)\log(K_{i'j})}\Big] \tag{29}$$

$$= \sum_j K_{ij}x_j (\sum_{i'} K_{i'j}^{p+1})^{-\frac{1}{p+1}} \Big[\frac{1}{(p+1)^2}\log(\sum_{i'} K_{i'j}^{p+1}) \tag{30}$$

$$- \frac{1}{p+1}\frac{1}{\sum_{i'} K_{i'j}^{p+1}} \sum_{i'} \log(K_{i'j}) K_{i'j}^{p+1}\Big] \tag{31}$$

$$= \sum_j \Big[\frac{1}{(p+1)^2}\log(\sum_{i'} K_{i'j}^{p+1}) - \frac{1}{p+1}\sum_{i'} (U_{i'j}^{(p)})^{p+1}\log(K_{i'j})\Big] U_{ij}^{(p)} x_j \tag{32}$$

$$\frac{dy_i}{dp}\Big|_{p=0} = \sum_j \Big[\log(\sum_{i'} K_{i'j}) - \sum_{i'} U_{i'j}^{(0)}\log(K_{i'j})\Big] U_{ij}^{(0)} x_j \tag{33}$$

where we define $K_{ij}\log(K_{ij}) = 0$ if $K_{ij} = 0$ (which implies $(U_{ij}^{(p)})^{p+1}\log(K_{ij}) = 0$ if $K_{ij} = 0$. Plugging this back into the second term in the formula for Shannon *affinity* entropy, we find

$$\sum_i \frac{dy_i}{dp}\Big|_{p=0} = \sum_i \sum_j \Big[\log(\sum_{i'} K_{i'j}) - \sum_{i'} U_{i'j}^{(0)}\log(K_{i'j})\Big] U_{ij}^{(0)} x_j \tag{34}$$

$$= \sum_j \Big[\log(\sum_{i'} K_{i'j}) - \sum_{i'} U_{i'j}^{(0)}\log(K_{i'j})\Big] x_j \sum_i U_{ij}^{(0)} \tag{35}$$

$$= \sum_j \Big[\log(\sum_{i'} K_{i'j}) - \sum_{i'} U_{i'j}^{(0)}\log(K_{i'j})\Big] x_j \tag{36}$$

completing the claim.

Remark. *In the case of duplicate strategies (clones), the maximizers of H_a^p form precisely the set of distributions which arbitrarily distribute a mass of $\frac{1}{C}$ across each of the C sets of clones.*

Consider the case of exact clones, i.e., K is block diagonal (w.l.o.g.) with blocks of ones. Let there be C clone groups each of size n_c for $c \in \{1, \ldots, C\}$. Let $c(i)$ map an action i to its clone set. In this case, it can be shown that $U_{ij}^{(p)} = n_{c(i)}^{-\frac{1}{p+1}}$ if $c(i) = c(j)$, otherwise $U_{ij}^{(p)} = 0$. Note that the gradient of entropy w.r.t. $\boldsymbol{x}$ must be proportional to the ones vector for $\boldsymbol{x}$ to be a maximizer in the interior of the simplex. Let $\boldsymbol{x} = [\frac{1}{C}\boldsymbol{x}_1, \ldots, \frac{1}{C}\boldsymbol{x}_C]$ with each $\boldsymbol{x}_c \in \mathbb{R}^{n_c}$ w.l.o.g. We will show that the set of maximizers of H_a^p is necessarily the set of $\boldsymbol{x}$ where each $\boldsymbol{x}_c \in \Delta^{n_c - 1}$. For $\boldsymbol{x}$ to be a maximizer, the gradient must be equal to the ones vector multiplied by a scalar $-d \in \mathbb{R}$:

$$\forall j \ \frac{\partial H_a^p(\boldsymbol{x})}{\partial x_j} = -\frac{p+1}{p}\sum_i (U_i^{(p)}\boldsymbol{x})^p U_{ij}^{(p)} \tag{37}$$

$$= -\frac{p+1}{p}\sum_i (\sum_k U_{ik}^{(p)} x_k)^p U_{ij}^{(p)} \tag{38}$$

$$= -\frac{p+1}{p}\sum_i \big(\frac{1}{C} n_{c(i)}^{-\frac{1}{p+1}} \mathbf{1}^\top \boldsymbol{x}_{c(i)}\big)^p U_{ij}^{(p)} \tag{39}$$

$$= -\frac{p+1}{p} n_{c(j)} \big(\frac{1}{C} n_{c(j)}^{-\frac{1}{p+1}} \mathbf{1}^\top \boldsymbol{x}_{c(j)}\big)^p n_{c(j)}^{-\frac{1}{p+1}} \tag{40}$$

$$= -\frac{p+1}{p} n_{c(j)} n_{c(j)}^{-\frac{p+1}{p+1}} \big(\frac{1}{C} \mathbf{1}^\top \boldsymbol{x}_{c(j)}\big)^p \tag{41}$$

$$= -\frac{p+1}{p} \big(\frac{1}{C} \mathbf{1}^\top \boldsymbol{x}_{c(j)}\big)^p = -d. \tag{42}$$

We also require $\boldsymbol{x} \in \Delta$, which implies

$$x_j \geqslant 0 \implies x_{c(j)} \geqslant \mathbf{0} \tag{43}$$

$$1 = \sum_j x_j = \sum_c \frac{1}{C} \mathbf{1}_{n_c}^\top \boldsymbol{x}_c \tag{44}$$

$$= C\Big(\frac{dp}{p+1}\Big)^{1/p} = d^{1/p} C \Big(\frac{p}{p+1}\Big)^{1/p} \implies d = C^{-p}\Big(\frac{p+1}{p}\Big). \tag{45}$$

Finally, we know from (42)

$$(\frac{1}{C}\mathbf{1}^\top \boldsymbol{x}_{c(j)})^p = \frac{dp}{p+1} = C^{-p} \tag{46}$$

$$\implies \mathbf{1}^\top \boldsymbol{x}_{c(j)} = 1 \tag{47}$$

proving the claim.

Remark. *In the case of duplicate strategies (clones), the maximizers of H_a^p achieve an entropy value which is equal to the Tsallis entropy of the system with clones removed.*

If we evaluate the max entropy distribution we find

$$H_a^p(\boldsymbol{x}) = \frac{1}{p}\Big[1 - \sum_i (U_i^{(p)}\boldsymbol{x})^{p+1}\Big] \tag{48}$$

$$= \frac{1}{p}\Big[1 - \sum_i \big(\frac{1}{C} n_{c(i)}^{-\frac{1}{p+1}} \mathbf{1}^\top \boldsymbol{x}_{c(i)}\big)^{p+1}\Big] \tag{49}$$

$$= \frac{1}{p}\Big[1 - \sum_c n_c \big(\frac{1}{C} n_c^{-\frac{1}{p+1}} \mathbf{1}^\top \boldsymbol{x}_c\big)^{p+1}\Big] \tag{50}$$

$$= \frac{1}{p}\Big[1 - \sum_c n_c \big(\frac{1}{C} n_c^{-\frac{1}{p+1}}\big)^{p+1}\Big] \tag{51}$$

$$= \frac{1}{p}\Big[1 - \sum_c n_c n_c^{-1} \big(\frac{1}{C}\big)^{p+1}\Big] \tag{52}$$

$$= \frac{1}{p}\Big[1 - \sum_c \big(\frac{1}{C}\big)^{p+1}\Big] \tag{53}$$

which is precisely the Tsallis entropy of the uniform distribution over C distinct clones.

C Integrals over Simplex

It is possible to derive a closed-form result for the dis-similarity kernel in (6) by appealing to known results of integrals of polynomial functions over the simplex.

Let $T^d = \{(x_1, \ldots, x_d) : x_i \geqslant 0, \sum_{i=1}^d x_i \leqslant 1\}$ be the standard simplex in $\mathbb{R}^d$. Let $\nu_i > 0$ for all i, then

$$\int_{T^d} x_1^{\nu_1 - 1} \ldots x_d^{\nu_d - 1}(1 - x_1 - \ldots - x_d)^{\nu_0 - 1} = \frac{\prod_{i=0}^d \Gamma(\nu_i)}{\Gamma(\sum_{i=0}^d \nu_i)}. \tag{54}$$

Proposition C.1. *From player i's perspective, the expected dis-similarity between two actions p and q under a uniform distribution over all opponent joint strategy profiles x_{-i} is equal to*

$$D_{pq}^{(i)} = \frac{1}{(d_i + 1)(d_i + 2)}\Big[||U_p^{(i)} - U_q^{(i)}||^2 + \big(1^\top (U_p^{(i)} - U_q^{(i)})\big)^2\Big] \tag{55}$$

where $U^{(i)}$ is a $|\mathcal{A}_i| \times |\mathcal{A}_{-i}|$ matrix where each entry $U^{(i)}_{a_i, a_{-i}}$ is the expected utility for player i playing action a_i against the background joint action a_{-i}. $U^{(i)}_{a_i}$ indicates an entire row of the matrix. The integer $d_i = \prod_{j \neq i} |\mathcal{A}_j|$.

Proof. Recall (54) and $\Gamma(n) = (n-1)!$ for $n \in \mathbb{N}$. Let $r_p = \sum_w U_{pw} x_w$ be the rating for the pth action under an opponent strategy profile x_{-i}.

Then we want to compute $\mathbb{E}_{x_{-i} \sim Dir(\mathbf{1})}[(r_p - r_q)^2]$. Recall the volume of the simplex is $\frac{1}{d!}$. Then

$$\begin{aligned}
\mathbb{E}_{x_{-i} \sim Dir(\mathbf{1})}[(r_p - r_q)^2] &= \frac{\int_{T^d}(r_p - r_q)^2 dx_{-i}}{\int_{T^d} dx_{-i}} && (56)\\
&= d! \int_{T^d} (r_p - r_q)^2 dx_{-i} && (57)\\
&= d! \int_{T^d} (\sum_w U^{(i)}_{pw} x_w - \sum_w U^{(i)}_{qw} x_w)^2 dx_{-i} && (58)\\
&= d! \int_{T^d} \Big[(\sum_w \sum_y U^{(i)}_{pw} U^{(i)}_{py} x_w x_y) + (\sum_w \sum_y U^{(i)}_{qw} U^{(i)}_{qy} x_w x_y) && (59)\\
&\quad - 2(\sum_w \sum_y U^{(i)}_{pw} U^{(i)}_{qy} x_w x_y)\Big] dx_{-i} && (60)\\
&= d! \sum_w \sum_y \Big[\big(U^{(i)}_{pw} U^{(i)}_{py} \underbrace{\int_{T^d} x_w x_y dx_{-i}}_{\frac{2}{(d+2)!} \text{ if } w=y \text{ else } \frac{1}{(d+2)!}}\big) && (61)\\
&\quad + \big(U^{(i)}_{qw} U^{(i)}_{qy} \int_{T^d} x_w x_y dx_{-i}\big) - 2\big(U^{(i)}_{pw} U^{(i)}_{qy} \int_{T^d} x_w x_y dx_{-i}\big)\Big] && (62)\\
&= \frac{d!}{(d+2)!} \sum_w \Big[(U^{(i)2}_{pw} + U^{(i)2}_{qw} - 2U^{(i)}_{pw} U^{(i)}_{qw}) && (63)\\
&\quad + \sum_y \big(U^{(i)}_{pw} U^{(i)}_{py} + U^{(i)}_{qw} U^{(i)}_{qy} - 2U^{(i)}_{pw} U^{(i)}_{qy}\big)\Big] && (64)\\
&= \frac{1}{(d+1)(d+2)}\Big[\sum_w (U^{(i)}_{pw} - U^{(i)}_{qw})^2 + (\sum_w U^{(i)}_{pw} - \sum_w U^{(i)}_{qw})^2\Big] && (65)\\
&= \frac{1}{(d+1)(d+2)}\Big[||U_p^{(i)} - U_q^{(i)}||^2 + \big(1^\top (U_p^{(i)} - U_q^{(i)})\big)^2\Big]. && (66)
\end{aligned}$$

□

Proposition C.2. *From player i's perspective, the expected dis-similarity between two actions p and q under a uniform distribution over all factorize-able opponent strategy profiles $x_{-i} = \prod_{j \neq i} x_j$ is equal to*

$$\begin{aligned}
D_{pq}^{(i)} = \prod_{j \neq i} \frac{1}{(d_j + 1)(d_j + 2)} \Big(& && (67)\\
\sum_{a_{-i} \in \mathcal{A}_{-i}} \sum_{a'_{-i} \in \mathcal{A}_{-i}} & \big(u_i(p, a_{-i}) - u_i(q, a_{-i})\big)\big(u_i(p, a'_{-i}) - u_i(q, a'_{-i})\big)\big(2^{\#a=a'}\big)\Big) && (68)
\end{aligned}$$

where the integer $d_i = |\mathcal{A}_i|$ *and "*$\#a{=}a'$*"* $= \sum_{j\neq i} \mathbf{1}[a_j = a'_j]$ *indicates the number of action matches between two opponent profiles.*

Proof. Let $r_p = \sum_{a_{-i}\in\mathcal{A}_{-i}} u_i(p, a_{-i}) \prod_{j\neq i} x_{j,a_j}$ be the rating for the pth action under an opponent profile $x_{-i} = \prod_{j\neq i} x_j$. Let dx_{-i} be a shorthand for dx_{-i}. Likewise, let $\int_{T^{d_{-i}}}$ be a shorthand for $\int_{T^{d_1}} \cdots \int_{T^{d_{i-1}}} \int_{T^{d_{i+1}}} \cdots \int_{T^{d_n}}$.

Then we want to compute $\mathbb{E}_{x_j\sim Dir(\mathbf{1})\forall j\neq i}[(r_p - r_q)^2]$. Recall the volume of a simplex is $\frac{1}{d!}$. Then

$$\mathbb{E}_{x_j\sim Dir(\mathbf{1})\forall j\neq i}[(r_p - r_q)^2] \tag{69}$$

$$= \frac{\int_{T^{d_{-i}}} (r_i - r'_i)^2 dx_{-i}}{\int_{T^{d_{-i}}} dx_{-i}} \tag{70}$$

$$= \Big(\prod_{j\neq i} d_j!\Big) \int_{T^{d_{-i}}} (r_i - r'_i)^2 dx_{-i} \tag{71}$$

$$= \Big(\prod_{j\neq i} d_j!\Big) \int_{T^{d_{-i}}} \Big(\sum_{a_{-i}\in\mathcal{A}_{-i}} u_i(p, a_{-i}) \prod_{j\neq i} x_{j,a_j} - \sum_{a_{-i}\in\mathcal{A}_{-i}} u_i(q, a_{-i}) \prod_{j\neq i} x_{j,a_j} \Big)^2 dx_{-i} \tag{72}$$

$$= \Big(\prod_{j\neq i} d_j!\Big) \int_{T^{d_{-i}}} \Big(\sum_{a_{-i}\in\mathcal{A}_{-i}} \prod_{j\neq i} x_{j,a_j} \big(u_i(p, a_{-i}) - u_i(q, a_{-i})\big) \Big)^2 dx_{-i} \tag{73}$$

$$= \Big(\prod_{j\neq i} d_j!\Big) \int_{T^{d_{-i}}} \Big(\tag{74}$$

$$\sum_{a_{-i}\in\mathcal{A}_{-i}} \sum_{a'_{-i}\in\mathcal{A}_{-i}} \big(\prod_{j\neq i} x_{j,a_j}\big)\big(\prod_{j\neq i} x_{j,a'_j}\big) \big(u_i(p, a_{-i}) - u_i(q, a_{-i})\big) \big(u_i(p, a'_{-i}) - u_i(q, a'_{-i})\big) \Big) dx_{-i} \tag{75}$$

$$= \Big(\prod_{j\neq i} d_j!\Big) \int_{T^{d_{-i}}} \Big(\tag{76}$$

$$\sum_{a_{-i}\in\mathcal{A}_{-i}} \sum_{a'_{-i}\in\mathcal{A}_{-i}} \big(u_i(p, a_{-i}) - u_i(q, a_{-i})\big) \big(u_i(p, a'_{-i}) - u_i(q, a'_{-i})\big) \big(\prod_{j\neq i} x_{j,a_j} x_{j,a'_j}\big) \Big) dx_{-i} \tag{77}$$

$$= \Big(\prod_{j\neq i} d_j!\Big) \Big(\tag{78}$$

$$\sum_{a_{-i}\in\mathcal{A}_{-i}} \sum_{a'_{-i}\in\mathcal{A}_{-i}} \big(u_i(p, a_{-i}) - u_i(q, a_{-i})\big) \big(u_i(p, a'_{-i}) - u_i(q, a'_{-i})\big) \big(\prod_{j\neq i} \underbrace{\int_{T^{d_j}} x_{j,a_j} x_{j,a'_j} dx_j}_{\frac{2}{(d_j+2)!} \text{ if } a_j = a'_j \text{ else } \frac{1}{(d_j+2)!}} \big) \Big) \tag{79}$$

$$= \Big(\prod_{j\neq i} d_j!\Big) / \Big(\prod_{j\neq i} (d_j+2)!\Big) \Big(\tag{80}$$

$$\sum_{a_{-i}\in\mathcal{A}_{-i}} \sum_{a'_{-i}\in\mathcal{A}_{-i}} \big(u_i(p, a_{-i}) - u_i(q, a_{-i})\big) \big(u_i(p, a'_{-i}) - u_i(q, a'_{-i})\big) \big(2^{\#a=a'}\big) \Big) \tag{81}$$

$$= \prod_{j\neq i} \frac{1}{(d_j+1)(d_j+2)} \Big(\tag{82}$$

$$\sum_{a_{-i}\in\mathcal{A}_{-i}} \sum_{a'_{-i}\in\mathcal{A}_{-i}} \big(u_i(p, a_{-i}) - u_i(q, a_{-i})\big) \big(u_i(p, a'_{-i}) - u_i(q, a'_{-i})\big) \big(2^{\#a=a'}\big) \Big). \tag{83}$$

□

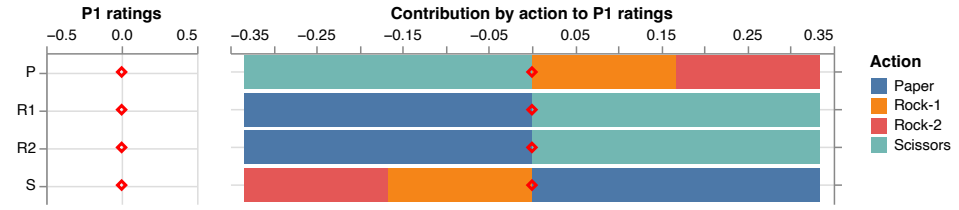

Figure 6: We visualise the marginal NE rating contributions of each player 2 action to each player 1 action. We show that a) all actions receive zero ratings and b) the rating of each action is interpretable and corresponds to our intuition.

D WARMUP: GAME-THEORETIC RANKING OF *rock-paper-scissors*

We provide a demonstration of game-theoretic ranking on the classic 2-player, 3-action zero-sum Rock-Paper-Scissors game. Balduzzi et al. (2018) proposed rating actions under the max-entropy Nash equilibrium of the game. In that case, each action receives a rating of zero. If we duplicate the Rock action, for example, the ratings remain zero under the max-entropy NE. Our proposed LLE based approach returns the same ratings.

	Rock	Paper	Scissors
Rock	$0, 0$	$-1, +1$	$+1, -1$
Paper	$+1, -1$	$0, 0$	$-1, +1$
Scissors	$-1, +1$	$+1, -1$	$0, 0$

	Rock1	Rock2	Paper	Scissors
Rock1	$0, 0$	$0, 0$	$-1, +1$	$+1, -1$
Rock2	$0, 0$	$0, 0$	$-1, +1$	$+1, -1$
Paper	$+1, -1$	$+1, -1$	$0, 0$	$-1, +1$
Scissors	$-1, +1$	$-1, +1$	$+1, -1$	$0, 0$

Figure 7: Rock-Paper-Scissors (RPS) Game and RPS Game with Duplicate Rock Action.

In Figure 6, we show that the equilibrium underlying the scalar ratings reflects incentive structure of the game — player 1 does not wish to deviate to the *Paper* action precisely because doing so would lead to losses against the *Scissors* action despite wins against the two *Rock* actions.

E VULNERABILITY OF STANDARD SHANNON ENTROPY

Prior work has shown max-entropy Nash equilibrium (equivalently max-entropy (C)CE) to be invariant to clones in 2-player zero-sum games (Balduzzi et al., 2018). We include a simple experiment here to illustrate why max-entropy Nash equilibrium becomes vulnerable to redundancy in the N-player general-sum setting.

Chicken Game Consider the 2-player 2-action general-sum *Chicken* game. Let players receive 0 if they both *swerve*. If one player swerves while the other goes straight, the one who swerves receives -1 and the other $+1$. If both go straight, then they both receive -12. This game has three NEs. Two are pure in which one player goes straight and the other swerves. The third is symmetric and the max-entropy NE of this game; each player swerves with probability $^{11}/_{12}$. Both *straight* and *swerve* have an expected payoff of $-^{1}/_{12}$ under this NE. If we duplicate the *straight* action, the original max-entropy NE becomes the *min*-entropy NE! The other two NEs representing each player swerving while the other goes straight now have higher entropy. The player that swerves rates their swerve and straight actions as -1 and -12 respectively. The player that goes straight rates their swerve and straight actions as 0 and 1 respectively, demonstrating that the max-entropy NE solution concept is not invariant to clones in the general-sum setting.

The story in the max-entropy CCE setting is more nuanced. We find that although the CCE ratings change under the addition of clones, the ratio of the ratings of the two actions remains stable. Further investigation is necessary to understand whether max-entropy CCE ratings are equivariant (robust up to affine transformations of the ratings) to cloned actions.

By contrast, we show in Figure 9 that all actions would receive zero ratings under our proposed equilibrium ratings. In other words, our equilibrium selection procedure continues to select the

	Swerve	Straight
Swerve	$0, 0$	$-1, +1$
Straight	$+1, -1$	$-12, -12$

	Swerve	Straight	Straight
Swerve	$0, 0$	$-1, +1$	$-1, +1$
Straight	$+1, -1$	$-12, -12$	$-12, -12$
Straight	$+1, -1$	$-12, -12$	$-12, -12$

Figure 8: Chicken Game and Chicken Game with Duplicate Straight Actions.

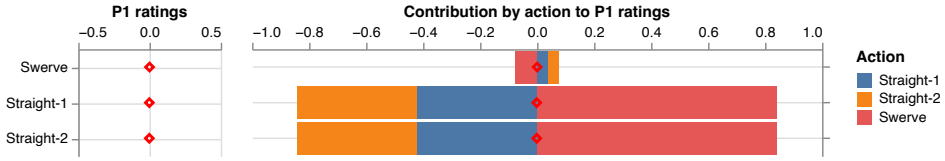

Figure 9: We visualise the marginal NE rating contributions of each player 2 action to each player 1 action. We show that a) all actions receive zero ratings and b) the rating of each action is interpretable and corresponds to our intuition.

mixed-strategy NE in the original game, unaffected by the additional redundant "straight" action. Further, the widths of the bars are interpretable: suggesting that deviating to the *Swerve* action is a safe option without major risk or reward. Deviating to one of the *Straight* actions however, can lead to high rewards but also catastrophic losses.

F Experiments

F.1 Simulated model and prompt improvement path

Algorithm 1 describes our simulated model and prompt improvement procedure. At each iteration, we add a new prompt and a model following an evolutionary procedure. We require all prompts to be probability distributions over skill dimensions. We model for a transitive dimension for models by representing each model vector as a sum of probability vectors over skills. A new model is added to the set of models $\mathcal{A}_m$ if and only if it becomes top-ranked according to the rating function r. A new prompt is added as long as it is the best-of-P' sampled prompts and does not have to be top-ranked.

F.2 Equilibrium-solving hyper-parameters

We use the same set of hyper-parameters for all our experiments. For affinity-entropy $H_a^p(x)$, we use $p = 1$ and set kernel variance to $1\text{e}{-6}$. To solve for a max affinity-entropy distribution we use gradient descent. The max affinity-entropy distribution is then used in NE and CCE solving.

For NE solving using LLE approximation, we initialize temperature $\tau = 1.0$ which is annealed exponentially with a decay rate of 0.95 every 250 gradient updates if and only if the exploitability in the annealed game $\mathcal{L}^\tau(\boldsymbol{x})$ (Equation (4)) is at most $1\text{e}{-5}$. We set the terminal temperature to $\tau = 1\text{e}{-2}$. We early terminate the equilibrium solving if we have found an $\epsilon-$NE with $\epsilon = 1\text{e}{-3}$. For CCE solving, the optimization problem is convex and we minimize Equation 8 directly. For gradient descent, we use an Adam optimizer Kingma (2014) with a fixed learning rate $1\text{e}{-2}$ for all steps (maximizing affinity-entropy and equilibrium solving).

F.3 The `arena-hard-v0.1` evaluation data

We evaluate our method on the `arena-hard-v0.1` dataset (Li et al., 2024b) with 500 prompts and 17 competing models. The set of prompts as well as model responses are downloaded from LMSYS data repository (`https://huggingface.co/spaces/lmsys/arena-hard-browser`), with the exception of `gemini-1.5-pro-api-0514` and `gemini-1.5-flash-api-0514`. As we need to tabulate the payoff tensor for all model pairs, we sampled 8 preference ratings using `gemini-1.5-pro-api-0514` for each model pair, with 4 samples for each permutation to account for potential position bias of the LLM rater. Pairwise model utility is averaged over all ratings samples.

Algorithm 1 Evolutionary model and prompt selection procedure

Let K be the number of orthogonal skill dimensions.
Let $r : \mathcal{A}_p \times \mathcal{A}_m \to \boldsymbol{r}_p, \boldsymbol{r}_m$ be a rating function assigning a scalar rating to each action.
Let P_0, M_0 be the number of initial prompts and models.
Let P', M' be the number of sampled candidate prompts and models at each iteration.

$\mathcal{A}_p^0 \sim \text{Dirichlet}(\mathbf{1}_{1:K}, P_0)$ ▷ P_0 sampled initial prompts.
$\mathcal{A}_m^0 \sim \text{Dirichlet}(\mathbf{1}_{1:K}, M_0)$ ▷ M_0 sampled initial models.

for $t \in [1, \dots]$ **do**
  if additional prompts **then** ▷ If adding new prompts.
    $\mathcal{A}'_p \sim \text{Dirichlet}(\mathbf{1}_{1:K}, P')$ ▷ Sampling P' candidate prompts.
    $\boldsymbol{r}_p, \_ \leftarrow r(A'_p \cup A_p, A_m)$
    $\mathcal{A}_p \leftarrow \mathcal{A}_p \cup \{A'_p[\arg\max \boldsymbol{r}_p[:P']]\}$ ▷ Add best-of-P' prompt.
  end if
  $\mathcal{A}'_m \leftarrow \mathbf{0}$
  while true **do**
    $\Delta_m \leftarrow \text{Dirichlet}(\mathbf{1}_{1:K}, M')$ ▷ Sampling M' model improvement vectors.
    $\_, \boldsymbol{r}_m \leftarrow r(A_p, \{\mathcal{A}'_m + \Delta_m\} \cup A_m)$ ▷ Evaluate improved candidate models.
    $\mathcal{A}'_m \leftarrow \mathcal{A}'_m + \Delta_m[\arg\max \boldsymbol{r}_m[:M']]$
    if $\arg\max \boldsymbol{r}_m[:M'] = \arg\max \boldsymbol{r}_m$ **then**
      $\mathcal{A}_m = \{\mathcal{A}'_m[\arg\max \boldsymbol{r}_m]\} \cup \mathcal{A}_m$ ▷ Add a new top-ranked model.
      break
    end if
  end while
end for

Table 1: Prompt and king actions that each define 16 pure-strategy Nash equilibria — any rebel action except the model played by the king player is a pure-strategy NE.

Prompt	King
"Can you implement a python tool that is intended to ru..."	gemini-1.5-pro-api-0514
"Hi. I have this URL which I can paste in my Microsoft ..."	gemini-1.5-pro-api-0514
"Please provide a simple RESPONSE to the following PROM..."	claude-3-5-sonnet-20240620
"Take on the rol eof an Gherkin expert. Can you improve..."	claude-3-5-sonnet-20240620
"Write a small python function that get all the links o..."	gemini-1.5-flash-api-0514

F.4 RISK-DOMINANT EQUILIBRIA

Our `king-of-the-hill` evaluation game admits a multitude of Nash equilibria, among them 80 are pure-strategy NEs (see Table 1). Additionally, we computed 128 mixed-strategy NEs with exploitability at most $1e{-}2$ that each derives a distinct set of ratings. In particular, one of the 128 mixed-strategy NEs is pre-computed by our NE solving and selection procedure by tracing the QRE continuum, which we refer to as the 0-th equilibrium, or $\boldsymbol{x}^0$.

A longstanding challenge in game theory is that of equilibrium selection. Suppose that every player knows that there are many equilibria in the game, each player must confront the following question during play: out of all equilibria, which equilibrium strategy should I play and relatedly, which equilibrium would each of my co-players play? This is critical, as miscoordinating could lead to arbitrarily bad outcome, despite each player playing one of its equilibrium strategies. For instance, everyone driving on the right or left hand side of the road are two valid equilibria, but miscoordinating would be devastating.

It is for this reason that the notion of risk-dominance of Harsanyi & Selten (1988) is critically important: the Nobel-prize winning theorem suggests that players would each iterate on their prior beliefs over which equilibria its co-players would play and choose the one that is the least *risky* when players *miscoordinate* under such priors. Here, we show empirically that our solution concept leads to risk-dominant equilibria as suggested by Herings & Peeters (2010). To do so, we simultaneously minimize the exploitability of several profiles in parallel with a regularizer that maximizes the L_2

rating differences between any two profiles by gradient descent as in Liu et al. (2024). This yields an additional 127 NEs with exploitability at most $1\mathrm{e}{-2}$ that we analyze in Figure 10.

Figure 10 *(Top)* shows the 128 mixed-strategy NEs with distinct model ratings. Figure 10 *(Center)* shows the expected payoffs to player i when it plays its p-th equilibrium strategy x_i^p when other players uniformly choose one of theirs, or $\mathbb{E}_{q\sim\pi_u}\left[u_i(x_i^p, x_{-i}^q)\right]$ with π_u a uniform distribution over 128 equilibria. In yellow, we show the sum of per-player expected payoffs. We confirm that many NEs are indeed *risky*, as their stability relies heavily on all players coordinating on the same equilibrium. Figure 10 *(Bottom)* takes things one step further and follows the intuition of risk dominance more closely. Starting from a uniform prior belief over player i's choice of equilibria, $\pi_i^0 = \pi_u$, each player iterate their believes over other players' choices of equilibrium based on the expected payoff of them playing each equilibrium.

Specifically, we let

$$\pi_i^{t+1} = \mathrm{softmax}\Big(\log \pi_i^t + \eta \mathbb{E}_{\substack{\forall j\neq i \\ t(j)\sim\pi_j}}\left[\boldsymbol{u}_i(\dots, x_{i-1}^{t(i-1)}, x_{i+1}^{t(i+1)}, \dots)\right]\Big) \tag{84}$$

with $\eta = 1\mathrm{e}{-2}$ the step-size and we compute the expected payoffs to player i when playing its k-th equilibrium at $T = 10,000$ as

$$\mathbb{E}_{\substack{\forall j\neq i \\ e(j)\sim\pi_j^T}}\left[u_i(\dots, x_{i-1}^{e(i-1)}, x_i^k, x_{i+1}^{e(i+1)}, \dots)\right] \tag{85}$$

Ordered by the sum of expected payoffs for all players, we observe that the Nash equilibrium our procedure selects (equilibrium $\boldsymbol{x}^0$) is the least risky among 128 mixed-strategy NEs of the game, without any player being particularly worse off than others even when players miscoordinate.

F.5 Invariant Evaluation

We show in Figure 11 the effect of introducing *near* redundant adversarial prompts on the equilibrium ratings. While our invariant property is limited to exact clones, our results show that our approach results in rankings that degrade gracefully in this approximate case, even with 1,000 adversarial prompts. The Elo rating system suffers from such bias in data similarly as in the exact case Figure 3.

In Figure 12 we provide a detailed breakdown of our NE and CCE ratings results (without redundant adversarial prompts). We show the actions of each player ranked by their equilibrium ratings and by their support under the equilibrium marginal distribution.

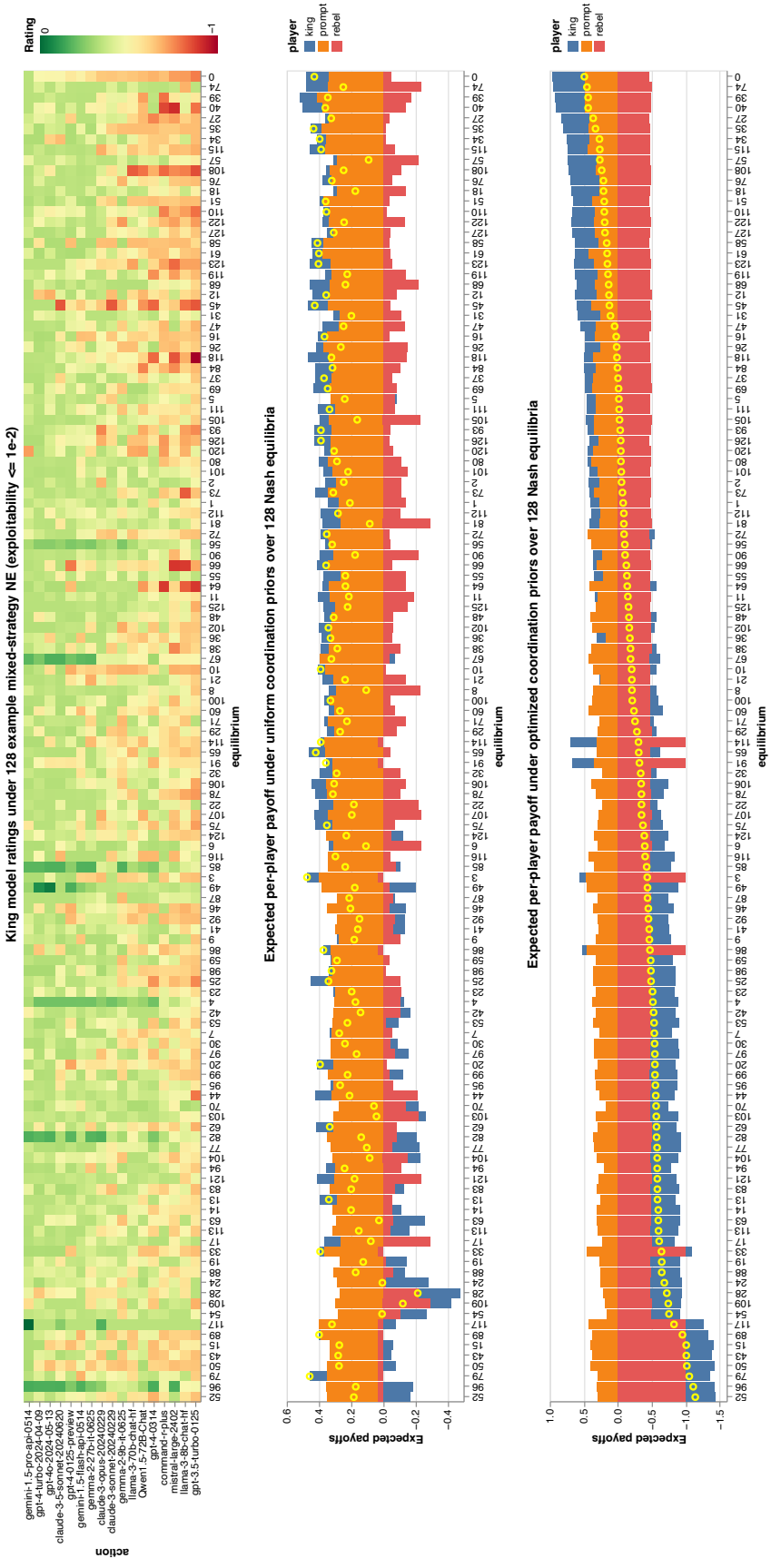

Figure 10: From top to bottom: a) we show the distinct king player action ratings derived from 128 mixed-strategy NEs of the `king-of-the-hill` game. All NEs have exploitability at most $\epsilon \leqslant 1e{-}2$; b) we show the expected payoff to each player under uniform priors over their 128 equilibria; yellow circles show the sum of expected per-player payoffs; c) we show the same analysis as in b) but the expectation is taken under optimized equilibrium priors. Equilibrium 0 (rightmost) is the LLE our NE solving procedure select.

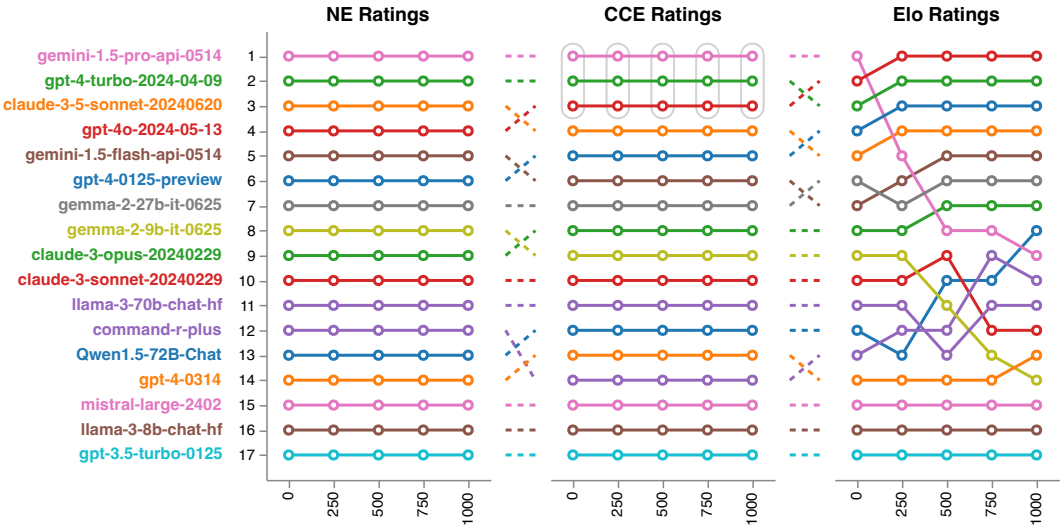

Figure 11: We introduce an increasing number of redundant copies of prompts adversarial to `gemini-1.5-pro-api-0514` with noise sampled from Uniform$(-0.01, 0.01)$ applied to their payoffs. Equilibrium ratings with a clone invariant selection procedure degrades gracefully to noisy redundancy while the Elo ratings become incrementally skewed. Models at the same rank (with an absolute rating difference at most $1e{-4}$) are grouped in grey and ordered alphabetically. We caveat that the specific rankings reported are subject to the LLM preference model used which in this case may exhibit a self-preference to the Gemini family of models.

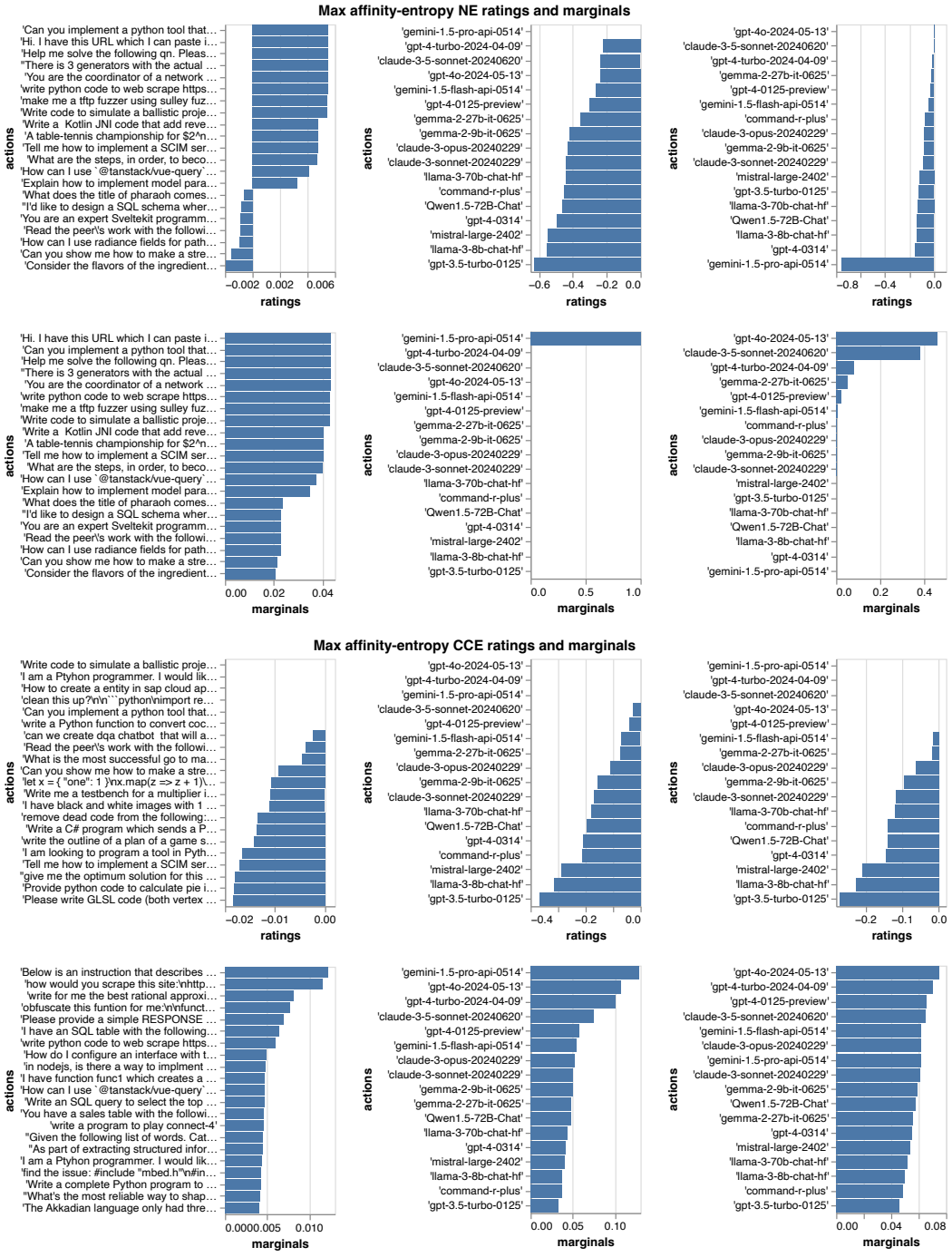

Figure 12: We show actions of each player ranked by their rating and equilibrium support under NE *(Top)* and CCE *(Bottom)* profiles respectively.