ASYMPTOTIC UNIVERSAL ALIGNMENT: A NEW ALIGNMENT FRAMEWORK VIA TEST-TIME SCALING

Yang Cai[*], **Weiqiang Zheng**[*]
Department of Computer Science
Yale University
New Haven, CT 06511, USA
{yang.cai, weiqiang.zheng}@yale.edu

ABSTRACT

Aligning large language models (LLMs) to serve users with heterogeneous and potentially conflicting preferences is a central challenge for personalized and trustworthy AI. We formalize an ideal notion of *universal alignment* through *test-time scaling*: for each prompt, the model produces $k \geq 1$ candidate responses and a user selects their preferred one. We introduce $(k, f(k))$-*robust alignment*, which requires the k-output model to have win rate $f(k)$ against any other single-output model, and *asymptotic universal alignment (U-alignment)*, which requires $f(k) \to 1$ as $k \to \infty$. Our main result characterizes the optimal convergence rate: there exists a family of *single-output* policies whose k-sample product policies achieve U-alignment at rate $f(k) = \frac{k}{k+1}$, and no method can achieve a faster rate in general. We show that popular post-training methods, including Nash learning from human feedback (NLHF), can fundamentally underutilize the benefits of test-time scaling. Even though NLHF is optimal for $k = 1$, sampling from the resulting (often deterministic) policy cannot guarantee win rates above $\frac{1}{2}$ except for an arbitrarily small slack. This stems from a lack of output diversity: existing alignment methods can collapse to a single majority-preferred response, making additional samples redundant. In contrast, our approach preserves output diversity and achieves the optimal test-time scaling rate. In particular, we propose a family of symmetric *multi-player alignment games* and prove that any symmetric Nash equilibrium policy of the $(k+1)$-player alignment game achieves the optimal $(k, \frac{k}{k+1})$-robust alignment. Finally, we provide theoretical convergence guarantees for self-play learning dynamics in these games and extend the framework to opponents that also generate multiple responses.

1 INTRODUCTION

Large language models (LLMs) have demonstrated remarkable versatility across domains such as text generation, code synthesis, information retrieval, and mathematical reasoning. Owing to their broad generalization capabilities, LLMs are increasingly integrated into users' daily workflows to assist with information seeking, solution generation, and decision-making. However, users vary widely in their preferences—when given the same prompt, they may favor responses that differ in content, tone, style, values, or level of detail. Consequently, aligning LLMs to serve people with diverse values and perspectives (Sorensen et al., 2024) has been recognized as a central challenge in building more personalized and trustworthy AI systems.

We formalize an ideal notion of alignment across diverse preferences, in which an AI system is aligned with every possible user—a property we term *universal alignment*. While one could, in principle, achieve universal alignment by training or fine-tuning a bespoke model for each individual user, such an approach is prohibitively expensive given the substantial computational and human resources required. In this paper, we ask to what extent universal alignment can be approached using

[*]Authors are alphabetically ordered.

a *single model* with *test-time scaling*, i.e., sampling multiple responses from the model. In particular,

*Can we efficiently approach **universal alignment** using a single model and test-time scaling?*

At first glance, this may seem impossible: a single model must reconcile conflicting preferences, and satisfying one group may inevitably come at the expense of another. We show that, with appropriate post-training alignment and test-time scaling, such reconciliation is, in fact, possible. Moreover, we provide theoretical guarantees characterizing how the amount of test-time scaling required relates to the desired level of alignment.

1.1 Our Model and Results

In this section, we introduce a formal framework for studying universal alignment under test-time scaling and establish theoretical guarantees and fundamental limits within this framework. To facilitate the discussion, we assume that each user has an individual ranking over responses y for every prompt x. This simplified model captures the essential ideas while keeping the presentation clear; however, our results extend to other preference structures such as Plackett–Luce and mixtures of Plackett–Luce (see Section 2 for details). Intuitively, we allow the model to generate $k \geq 1$ candidate responses for each prompt and let the user select their preferred one.[1]

1.1.1 Robust Alignment and Universal Alignment

We begin by defining the notion of *robust alignment*.

Definition 1.1 ($(k, f(k))$-Robust Alignment)**.** Let $f : \mathbb{N}_{>0} \to [0, 1]$ be a function, and let $k \in \mathbb{N}_{>0}$ denote the number of responses generated at test time. A policy $\pi : \mathcal{X} \to \Delta(\mathcal{Y}^k)$ is said to achieve $(k, f(k))$-robust alignment if, for every prompt $x \in \mathcal{X}$, its win rate against any other policy $\pi' : \mathcal{X} \to \Delta(\mathcal{Y})$ satisfies $\Pr[\pi \succcurlyeq \pi' \mid x] \geq f(k)$ [2].

An immediate implication of $(k, f(k))$-robust alignment is that, for any prompt x, a robustly aligned policy π must include every response y that is the favorite of more than a $(1 - f(k))$-fraction of the population. Otherwise, if such a response y is omitted by π, a policy π' that always outputs y on input x would violate the $(k, f(k))$-robust alignment condition. Equipped with the definition of robust alignment, we now introduce the notion of *asymptotic universal alignment*, which we may refer to simply as U-alignment throughout the paper.

Definition 1.2 (Asymptotic Universal Alignment (U-Alignment))**.** Let $f : \mathbb{N}_{>0} \to [0, 1]$ be a rate function satisfying $\lim_{k\to\infty} f(k) = 1$. We say that a family of policies $\{\pi_k\}_{k\in\mathbb{N}_{>0}}$ achieves *asymptotic universal alignment* with rate f if, for every $k \in \mathbb{N}_{>0}$, the policy π_k achieves $(k, f(k))$-robust alignment. In other words, as the number of generated responses k increases, the alignment rate $f(k)$ approaches 1, corresponding to alignment with almost all users in the population.

Note that any policy π that assigns nonzero probability to every possible response under any prompt—e.g., one that samples a response uniformly at random—technically satisfies our notion of U-alignment by sampling k times independently from π for each prompt. However, the convergence rate in this case is extremely slow: unless k is on the order of $|\mathcal{Y}|$, we have $f(k) = o(1)$. Such a convergence rate is clearly undesirable, both theoretically and practically. Can we approach universal alignment more efficiently?

1.1.2 Our Result

We give a positive answer to this question by characterizing the optimal convergence rate.

Informal Theorem. There exists a family of single-output policies $\{\pi_k\}_{k\in\mathbb{N}_{>0}}$ such that the corresponding test-time scaled policies $\{\pi_k^{\otimes k}\}_{k\in\mathbb{N}_{>0}}$ achieve U-alignment with rate $k/k+1$. Here, each π_k maps any prompt in $\mathcal{X}$ to a distribution over $\mathcal{Y}$, and $\pi_k^{\otimes k}$ denotes the policy that independently

[1] For clarity, we assume that the user selects their favorite response. In practice, the model provider may employ a lightweight auxiliary model to predict or select the user's preferred response based on their features and past interactions.

[2] We formally define the probability $\Pr[\pi \succcurlyeq \pi' \mid x]$ in Section 2. For now, it can be interpreted as the probability that a randomly selected user from the population prefers at least one of the k responses generated by π to the response generated by π'.

samples from π_k k times for each prompt. Moreover, no family of policies can achieve U-alignment at a faster rate.

This statement provides the theoretical foundation for U-alignment by characterizing the optimal convergence rate achievable through test-time scaling. The formal version of this informal theorem is given in Theorem 4.4. In the remainder of this section, we examine the features, extensions, and implications of this result from four complementary perspectives.

We begin by highlighting the importance of achieving this optimal rate using single-output policies, which are more consistent with current model architectures and training–inference pipelines than multi-output policies. Next, we address computational considerations, showing how optimal policies can be characterized via self-play in symmetric multi-player alignment games. We then analyze the limitations of existing post-training alignment methods under test-time scaling. Finally, we explore the broader connection between U-alignment and diversity preservation in post-training alignment.

Optimal U-alignment via Single-Output Policies. By drawing a connection to the notion of the Condorcet winning set in social choice theory, we can leverage existing results to establish the existence of a family of policies that achieve U-alignment at rate $k/k+1$ for any $k \geq 1$ (Theorem 3.1). However, this existence result applies only to a *multi-output* policy that maps a prompt $x \in \mathcal{X}$ directly to a distribution over $\mathcal{Y}^k$, which is less desirable than achieving the same guarantee by scaling a single-output policy at test time. In particular:

- training a multi-output policy is nonstandard and requires optimization over a response space whose size grows exponentially with k; and
- inference from a multi-output policy is slower, since parallel computation cannot be exploited for acceleration.

Our main result shows that, for any $k \in \mathbb{N}_{>0}$, a single-output policy suffices to achieve the same optimal convergence rate, thereby avoiding these drawbacks.

Computing the optimal U-alignment policies via self-play. Our main result establishes the existence of a family of single-output policies that achieve the optimal convergence rate to U-alignment. We next characterize this family by relating it to a class of symmetric $(k+1)$-player alignment games defined in Definition 4.2. For any $k \in \mathbb{N}_{>0}$, we show that the symmetric Nash equilibrium policy of such a game corresponds to π_k in our main result (see Theorem 4.3). Finally, we extend self-play methods from the two-player setting to the multi-player setting and provide convergence guarantees for these dynamics (Proposition D.2 and Proposition D.3).

Limits of Test-Time Scaling of NLHF and RLHF. We briefly review two popular alignment approaches and discuss their limitations under test-time scaling. Formal definitions can be found in Section 2.2.

- **RLHF.** Reinforcement learning from human feedback (RLHF) (Christiano et al., 2017; Bai et al., 2022; Ouyang et al., 2022) assumes that aggregated human preferences follow the Bradley–Terry (BT) model (Bradley & Terry, 1952). RLHF first trains a scalar reward model from human preference data, and then optimizes the base model with respect to this learned reward via reinforcement learning. However, any scalar reward model fails to capture the full diversity of human preferences—particularly when preferences are non-transitive—introducing systematic bias and potential misalignment in RLHF (Munos et al., 2024).
- **NLHF.** Nash learning from human feedback (NLHF) (Munos et al., 2024; Swamy et al., 2024) relaxes the scalar reward assumption and directly models general preferences. NLHF formulates alignment as a two-player zero-sum game and seeks its Nash equilibrium policy. The resulting Nash policy satisfies a desirable guarantee: it achieves at least a 50% win rate against any other policy, corresponding to $(1, \frac{1}{2})$-robust alignment in our framework.

In summary, RLHF builds on the BT model and fails to capture diverse preferences, whereas NLHF achieves the optimal robust-alignment guarantee for $k = 1$. Given NLHF's prevalence, one may wonder whether test-time scaling NLHF yields nontrivial convergence to U-alignment. We provide a negative answer in Proposition 4.1: for any $k \geq 1$, there exist instances in which test-time scaling of NLHF cannot guarantee $(k, \frac{1}{2} + \varepsilon)$-robust alignment beyond an arbitrarily small slack. The intuition is that NLHF can lack response diversity and may collapse to (nearly) deterministic policies, which makes additional samples redundant. We elaborate on this point in the next paragraph.

U-alignment and Diversity Preservation during Post-Training. A necessary condition for U-alignment is the ability to generate a diverse set of responses that reflect heterogeneous human preferences. However, post-training alignment methods such as RLHF and NLHF often induce *mode collapse*, reducing output diversity (Kirk et al., 2024). When a majority of users prefer a particular response, RLHF and NLHF tend to converge toward producing it almost deterministically. Although aligning with the majority's preference may seem optimal, disregarding minority preferences prevents RLHF and NLHF from achieving U-alignment. The resulting lack of output diversity also limits the effectiveness of LLMs in applications that require creativity and variation, such as creative writing and synthetic data generation.

Our work demonstrates that it is possible to post-train a model to improve alignment while preserving diversity, challenging the common view that post-training alignment inevitably leads to mode collapse and reduced diversity.

We discuss additional related works in Section A.

2 PRELIMINARIES

Let the universe of prompts be $\mathcal{X}$ and the universe of responses be $\mathcal{Y}$, both assumed to be finite. A single-output language model (policy) is a mapping $\pi : \mathcal{X} \to \Delta(\mathcal{Y})$, and $\pi(y \mid x) \in [0, 1]$ the probability of outputting response $y \in \mathcal{Y}$ given a prompt $x \in \mathcal{X}$. A multi-output language policy $\pi : \mathcal{X} \to \Delta(\mathcal{Y}^k)$ specifies the probability $\pi(S \mid x)$ of outputting a multiset of responses $S \subseteq \mathcal{Y}^k$ given a prompt $x \in \mathcal{X}$.[3] We sometimes write a multiset S as a vector $m_S \in \mathbb{Z}_+^{\mathcal{Y}}$ where $m_S(y)$ is the number of copies of y in S. Given a vector $m \in \mathbb{Z}_+^{\mathcal{Y}}$, we denote $S(m)$ the corresponding multiset. Given a single-output policy π and $k \geq 1$, the product policy $\pi^{\otimes k}$ is the k-output policy that outputs k independent samples from π for any given prompt. For two multisets S and S', we define $S + S'$ to be the multiset union of S and S', i.e., the multiplicity of each element in $S + S'$ is the sum of its multiplicities in S and in S'.

2.1 PREFERENCES

Our results hold for a general class of preferences. We first introduce the Plackett-Luce (PL) preference model (Luce et al., 1959), a generalization of the Bradley-Terry (BT) model (Bradley & Terry, 1952) from comparing two responses to comparing two sets of responses, which is widely used in LLM alignment. We derive a simple and useful property of the aggregated population preference when each person has a PL preference. We then consider the standard social choice setting in which each person has a complete ranking of the responses.

Mixtures of Plackett-Luce Models There are n different types of *Plackett-Luce (PL)* preferences. Each preference $i \in [n]$ is associated with a reward function $r_i(x, y)$. For any pair of multisets of responses S, S', the preference of S over S' is defined as

$$\mathbb{P}_i[S \succcurlyeq S' \mid x] = \mathbb{P}_i[S \succ S' \mid x] := \frac{\sum_{y \in S} \exp(r_i(x, y))}{\sum_{y \in S} \exp(r_i(x, y)) + \sum_{y \in S'} \exp(r_i(x, y))}, \forall x, S, S'.$$

Given a distribution $\mathcal{D}$ over $[n]$, the aggregated *population preference* $\mathbb{P}_{\mathcal{D}}$ is a mixture of the n PL preferences. Specifically, for any $k, k' \geq 1$ and a k-output policy π and k'-output policy π', the population preference of $\pi \succcurlyeq \pi'$ is defined as

$$\mathbb{P}_{\mathcal{D}}[\pi \succcurlyeq \pi' \mid x] = \mathbb{P}_{\mathcal{D}}[\pi \succ \pi' \mid x] := \mathbb{E}_{i \sim \mathcal{D}} \mathbb{E}_{S \sim \pi(\cdot|x), S' \sim \pi'(\cdot|x)}[\mathbb{P}_i[S \succ S' \mid x]], \forall x, \pi, \pi'.$$

We identify and summarize useful properties that the population preference $\mathbb{P}_{\mathcal{D}}$ has.

Property 1. A mixture of PL preferences $\mathbb{P}_{\mathcal{D}}$ satisfies the following properties

(1) Antisymmetry: $\mathbb{P}_{\mathcal{D}}[\pi \succcurlyeq \pi' \mid x] + \mathbb{P}_{\mathcal{D}}[\pi' \succ \pi \mid x] = 1$ for any two multi-output policies π, π' and prompt x.

(2) Multi v.s. Single Copy: $\mathbb{P}_{\mathcal{D}}[\pi^{\otimes k} \succcurlyeq \pi \mid x] \geq 1 - \frac{1}{k+1}$ for any single-output policy π and x.

[3] Here we slightly abuse notations to denote the universe of multisets of size k as $\mathcal{Y}^k$, which usually denotes tuples of size k.

(3) Subadditivity: $\mathbb{P}_{\mathcal{D}}[S_1 + S_2 \succ S \mid x] \leq \mathbb{P}_{\mathcal{D}}[S_1 \succ S \mid x] + \mathbb{P}_{\mathcal{D}}[S_2 \succ S \mid x]$ for any multisets S_1, S_2, S and prompt x.

(4) Submodularity: For $m \in \mathbb{Z}_+^{\mathcal{Y}}$, let $F_{x,y}(m) := \mathbb{P}_{\mathcal{D}}[S(m) \succcurlyeq y \mid x]$. For any prompt x and response y, the function $F_{x,y}(\cdot)$ satisfies $F_{x,y}(0) = 0$, is monotone and DR-submodular.[4]

Remark 2.1. All of our results hold for any population preference satisfying Property 1(1–3); Theorem 4.5 further assumes condition (4). This covers a broad class of preference models. In particular, the following two generalizations of the PL model have Property 1 by definition: (1) each user's preference is a mixture of PL preference, and the population preference is a distribution over the possible mixtures; (2) based on (1), we further allow the population distribution for different prompts $x \in \mathcal{X}$ to be different.

Population of Rankings Now, we consider the setting where each user has a total ranking over responses for each prompt. We model each possible preference $\succ_i$ as orders $\{\succ_{i,x}\}_{x\in\mathcal{X}}$ over $\mathcal{Y}$ such that $y \succ_{i,x} y'$ means y is preferred than y' under prompt x. To incorporate ties between identical responses, we extend the preference to weak preferences $\succcurlyeq_i$ and define the following function

$$\mathbf{1}[y \succcurlyeq_{i,x} y'] := \begin{cases} 1, & y \succ_{i,x} y' \text{ or } y = y' \\ 0, & y' \succ_{i,x} y \end{cases}, \quad \mathbf{1}[y' \succ_{i,x} y] := 1 - \mathbf{1}[y \succcurlyeq_{i,x} y'].$$

Thus $\mathbf{1}[y \succcurlyeq_{i,x} y'] + \mathbf{1}[y' \succ_{i,x} y] = 1$ for all pairs of (y, y'). We extend the above function naturally to compare sets of responses $S, S' \subseteq \mathcal{Y}$ for $\succ_{i,x}$ by comparing the most preferred response (the maximal element according to $\succ_{i,x}$) in each set. Formally, let $y_* \in S$ be the most preferred response in S, i.e., there is no other $y \in S$ such that $y \succ_{i,x} y_*$. Similarly, let $y'_* \in S'$ be the most preferred responses in S'. Then we define

$$\mathbf{1}[S \succcurlyeq_{i,x} S'] = \mathbf{1}[y_* \succcurlyeq_{i,x} y'_*], \quad \mathbf{1}[S \succ_{i,x} S'] = \mathbf{1}[y_* \succ_{i,x} y'_*].$$

Given a population of users, a distribution $\mathcal{D}$ over $\{\succ_i\}_{i\in[n]}$, the aggregated population preference $\mathbb{P}_{\mathcal{D}}$ is: for any π, π' and x,

$$\mathbb{P}_{\mathcal{D}}[\pi \succcurlyeq \pi' \mid x] := \mathbb{E}_{i\sim\mathcal{D}}\mathbb{E}_{S\sim\pi(\cdot|x), S'\sim\pi'(\cdot|x)}[\mathbf{1}[S \succcurlyeq_{i,x} S']],$$
$$\mathbb{P}_{\mathcal{D}}[\pi \succ \pi' \mid x] := \mathbb{E}_{i\sim\mathcal{D}}\mathbb{E}_{S\sim\pi(\cdot|x), S'\sim\pi'(\cdot|x)}[\mathbf{1}[S \succ_{i,x} S']].$$

We show in the following proposition that an aggregated preference $\mathbb{P}_{\mathcal{D}}$ from a population of rankings also has Property 1.

Proposition 2.2. *$\mathbb{P}_{\mathcal{D}}$ satisfies Property 1.*

Remark 2.3. Since Property 1 applies to every prompt x, it also holds for the generalized preference model in which the population distribution $\mathcal{D}$ may depend on x.

Throughout the paper, we assume a general population preference $\mathbb{P}_D$ satisfying Property 1. This includes, in particular, mixtures of Plackett–Luce (PL) preferences, the standard social choice setting in which each user has a ranking over responses for each prompt, as well as generalizations in Remark 2.1 and Remark 2.3.

2.2 Alignment Methods

We briefly introduce two existing LLM alignment methods. **Reinforcement Learning From Human Feedback (RLHF)** RLHF (Christiano et al., 2017; Bai et al., 2022; Ouyang et al., 2022) first learns a reward function $r : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ according to the Bradley-Terry model from a preference dataset. Given a reference policy π_{ref}, the RLHF policy is a solution that maximizes the reward subject to KL constraints: $\forall x \in \mathcal{X}$, $\pi_{\text{RLHF}}(\cdot \mid x)$ is

$$\max_{\pi:\mathcal{X}\to\Delta(\mathcal{Y})} \mathbb{E}_{y\sim\pi(\cdot|x)}\mathbb{E}[r(x, y) - \eta \cdot \text{KL}(\pi(\cdot \mid x), \pi_{\text{ref}}(\cdot \mid x))].$$

A critical limitation of RLHF is that the BT model fails to capture diverse, possibly nontransitive, human preferences (Munos et al., 2024).

[4]We provide definitions and proofs in Section B.

Nash Learning From Human Feedback (NLHF) Given a population preference $\mathbb{P}_{\mathcal{D}}$, NLHF (Munos et al., 2024; Swamy et al., 2024) aims to find a Nash equilibrium policy $\pi_{\text{NLHF}} : \mathcal{X} \to \Delta(\mathcal{Y})$ of the following two-player *symmetric* constant-sum game: $\forall x \in \mathcal{X}$,

$$\max_{\pi(\cdot|x)\in\Delta(\mathcal{Y})} \min_{\pi'(\cdot|x)\in\Delta(\mathcal{Y})} \mathbb{P}_{\mathcal{D}}[\pi(\cdot \mid x) \succcurlyeq \pi'(\cdot \mid x)] \tag{1}$$

The NLHF policy π_{NLHF} achieves a $\frac{1}{2}$ win rate against any other single-out model under $\mathcal{P}_{\mathcal{D}}$.

3 U-Alignment via Multi-Output Policy: Possibility and Limitations

We note that by adapting existing results for the Condorcet winning set from the social choice theory literature, we can prove the possibility of U-alignment with rate $k/k+1$ using *multi-output policies*. A k-output policy $\pi : \mathcal{X} \to \Delta(\mathcal{Y}^k)$ is randomized mapping that returns a k-set of responses for each prompt x. The following theorem is adapted from (Cheng et al., 2020; Jiang et al., 2020; Charikar et al., 2025). Missing proofs in this section are in Section C

Theorem 3.1 ($(k, k/k+1)$-Robust Alignment Using a Multi-Output Policy)**.** *For any $k \geq 1$ and any population preference $P_{\mathcal{D}}$ satisfying Property 1, there exists a k-output policy $\pi : \mathcal{X} \to \Delta(\mathcal{Y}^k)$ such that for every prompt $x \in \mathcal{X}$ and every response $y \in \mathcal{Y}$, $P_{\mathcal{D}}[\pi \succcurlyeq y \mid x] \geq k/k+1$.*

We remark that Theorem 3.1 is based on the following construction: any Nash equilibrium policy of the following two-player constant-sum game (a min-max optimization problem) gives a $(k, k/k+1)$-robust aligned multi-output policy: $\forall x \in \mathcal{X}$,

$$\max_{\pi(\cdot|x)\in\Delta(\mathcal{Y}^k)} \min_{\pi'\in\Delta(\mathcal{Y})} \mathbb{P}_{\mathcal{D}}[\pi \succcurlyeq \pi' \mid x]. \tag{2}$$

This two-player game objective (2) generalizes Nash Learning from Human Feedback (NLHF) (Munos et al., 2024) in (1), which is the special case of (2) with $k = 1$.

3.1 Limitations of Multi-Output Policies

Although Theorem 3.1 guarantees the existence of a $(k, k/k+1)$-robust aligned multi-output policy, it does so by optimizing over general multi-output policies rather than standard single-output policies. This is unsatisfactory and impractical from both optimization and efficiency standpoints:

- The space of k-output policies, i.e., distributions in $\Delta(\mathcal{Y}^k)$, is vastly larger—and typically harder to optimize over—than the space of single-output policies $\Delta(\mathcal{Y})$. When $|\mathcal{Y}| = m$, a k-output policy is a distribution over $\mathcal{Y}^k$, whose support can be as large as m^k (exponential in k), whereas a single-output policy is a distribution over $\mathcal{Y}$ with support size at most m.
- Training a multi-output policy departs from standard practice, since mainstream LLM training pipelines are built for single-output policies. Although one could plausibly modify existing architectures to produce k responses, such modifications would entail significant changes to the training pipeline, making them impractical in most settings.
- The objective in (2) is *asymmetric*. As a result, self-play approaches to finding a Nash equilibrium must maintain and update two separate models, one for each player. This is less practical than maintaining and training a single model.

In light of the optimization and efficiency concerns for multi-output policies, together with the asymmetry of (2), existing results from social choice theory do not immediately carry over to the LLM alignment setting, where one typically trains a single-output policy.

3.2 Tightness of the Rate $k/k+1$

We show that the rate $k/k+1$ is tight with two lower bounds. The first lower bound assumes a Plackett-Luce (PL) preference model. We show that for any $k \geq 1$, there exists a PL preference such that it is impossible to achieve $(k, f(k))$-robust alignment with $f(k) > k/k+1$.

Proposition 3.2 (Lower Bound for Plackett-Luce Model)**.** *For any $k \geq 1$, there exists a Plackett-Luce preference $\mathbb{P}$ such that for any policy $\pi : \mathcal{X} \to \Delta(\mathcal{Y}^k)$, there exists a prompt $x \in \mathcal{X}$ and a response $y \in \mathcal{Y}$ such that $\mathbb{P}[\pi \succcurlyeq y|x] \leq k/k+1$.*

The second lower bound is for a population of rankings. We show that for any $k \geq 1$, there exists a population preference over rankings such that the best-possible policy achieves $(k, f(k))$-robust alignment with $f(k) \leq k/k+1 \cdot (1 + 1/|\mathcal{Y}|) \to k/k+1$ when the total number of responses $|\mathcal{Y}| \to \infty$.

Proposition 3.3 (Lower Bound for Rankings). *For any $k \geq 1$, there exists a population preference over rankings $\mathbb{P}_{\mathcal{D}}$ such that for any policy $\pi : \mathcal{X} \to \Delta(\mathcal{Y}^k)$, there exists a prompt $x \in \mathcal{X}$ and a response $y \in \mathcal{Y}$ such that $\mathbb{P}_{\mathcal{D}}[\pi \succcurlyeq y \mid x] \leq k/k+1 \cdot (1 + 1/|\mathcal{Y}|)$.*

4 U-Alignment via Single-Output Policy: Power of Test-Time Scaling

In this section, we explore the power of standard single-output policies. In the context of alignment, the Nash learning from human feedback (NLHF) policy gives $\frac{1}{2}$ win rate against any other single-output policy, but the win rate can not be further improved. However, the barrier of $\frac{1}{2}$ for single-output policies is for sampling and evaluating *one* response, but not for sampling k responses in test time. A natural question is: *can test-time scaling of a **single-output policy** achieve U-alignment with rate $k/k+1$?*

More formally, for any $k \geq 1$, does there exist a single-output policy π such that the test-time scaled policy $\pi^{\otimes k}$ achieves $(k, k/k+1)$-robust alignment? We remark that even proving the existence of such a policy is nontrivial. Although test-time scaling is clearly no worse than sampling a single response, it is unclear whether—and to what extent—a policy exhibits strong test-time performance. In particular, even if a model is optimal for $k = 1$, test-time scaling of this model need not yield improved performance.

4.1 Limits of Test-Time Scaling of NLHF

We illustrate this phenomenon by showing that test-time scaling the exact NLHF policy, even though it is optimal for $k = 1$, need not improve alignment beyond $\frac{1}{2}$, even when the number of samples is very large. In particular, we prove a fundamental limitation for NLHF: for any $k > 1$, the k-sample policy obtained by drawing k i.i.d. responses from the NLHF policy cannot guarantee a win rate above $\frac{1}{2}$ except for an arbitrarily small slack. Our lower bound is established in a simple non-contextual setting, thereby strengthening the impossibility result.

Proposition 4.1 (Limits of Test-Time Scaling of NLHF). *For any $\varepsilon \in (0, \frac{1}{2})$, there exist a response set $\mathcal{Y}$, a reward function $r : \mathcal{Y} \to \mathbb{R}$, and a population preference model $\mathbb{P}_{\mathcal{D}}$ such that the following statements hold the population preference $\mathbb{P}_{\mathcal{D}}$ admits a unique NLHF policy π_{NLHF}. Moreover, for every $k \geq 1$, $\min_{y \in \mathcal{Y}} \ \mathbb{P}_{\mathcal{D}}\big[\pi_{\mathrm{NLHF}}^{\otimes k} \succcurlyeq y\big] \ \leq \ \frac{1}{2} + \varepsilon$.*

We note that the limitation of NLHF stems from its lack of diversity. In particular, when there exists a response y that is preferred by a strict majority of the population (that is, y is a Condorcet winner), the NLHF policy collapses to outputting y with probability 1. This is optimal when $k = 1$, since it guarantees a win rate of at least $\frac{1}{2}$ against any fixed alternative. However, always outputting the majority-preferred response ignores minority's preferences and therefore provides no benefit from test-time scaling.

4.2 Multi-Player Preference Optimization for U-Alignment

Although NLHF does not exhibit the desired test-time scaling and, under single-sample evaluation, is limited by a $\frac{1}{2}$ barrier, there nonetheless exists a single-output policy π whose test-time scaled policy $\pi^{\otimes k}$ achieves $(k, k/k+1)$-robust alignment. Moreover, we characterize such a policy as the symmetric Nash equilibrium of a simple $(k + 1)$-player game.

We focus on the non-contextual setting; the contextual extension follows by applying the construction pointwise to each context. To this end, we introduce a *multi-player alignment game*, which generalizes the two-player NLHF game.

Definition 4.2 (Multi-Player Alignment Game and Nash equilibrium). Fix $k \geq 1$ and a population preference model $\mathbb{P}_{\mathcal{D}}$ over $\mathcal{Y}$. The *$(k + 1)$-player alignment game* has player set $[k + 1]$, and each player $j \in [k + 1]$ chooses an action $\pi_j \in \Delta(\mathcal{Y})$. Given a strategy profile (π_j, π_{-j}), where

$\pi_{-j} = \{\pi_\ell\}_{\ell \in [k+1]\setminus\{j\}}$, the utility of player j is

$$u_j(\pi_j, \pi_{-j}) := \mathbb{P}_{\mathcal{D}}\left[\pi_j \succ \bigotimes_{\ell \neq j} \pi_\ell\right]. \tag{3}$$

A *Nash equilibrium* is a strategy profile $(\pi_1^*, \ldots, \pi_{k+1}^*)$ such that no player can improve her utility by deviating unilaterally. That is, for every $j \in [k+1]$ and every $\pi \in \Delta(\mathcal{Y})$,

$$u_j(\pi_j^*, \pi_{-j}^*) \;\geq\; u_j(\pi, \pi_{-j}^*).$$

We show that the symmetric Nash equilibrium of this game, which we call the *Multi-Player Nash Equilibrium (MPNE)* policy, enjoys the desired test-time scaling property. All proofs in this section are deferred to Section D.

Theorem 4.3 (Properties of MPNE policy of the Multi-Player Alignment Game)**.** *For any population preference $\mathbb{P}_{\mathcal{D}}$ satisfying Property 1, the $(k+1)$-player alignment game (Definition 4.2) admits a symmetric Nash equilibrium in which every player uses the same policy π^*. Moreover, the test-time scaled policy $(\pi^*)^{\otimes k}$ achieves $(k, k/k+1)$-robust alignment: $\min_\pi \; \mathbb{P}_{\mathcal{D}}[(\pi^*)^{\otimes k} \succcurlyeq \pi] \;\geq\; k/k+1$.*

We are now ready to state the paper's main theorem.

Theorem 4.4. *For any population preference $\mathbb{P}_{\mathcal{D}}$ satisfying Property 1, there exists a family of single-output policies $\{\pi_k\}_{k\in\mathbb{N}_{>0}}$ such that their test-time scaled policies $\{\pi_k^{\otimes k}\}_{k\in\mathbb{N}_{>0}}$ achieve U-alignment at the optimal rate $k/k+1$. In particular, for every $k \geq 1$, there exists a single-output policy $\pi_k : \mathcal{X} \to \Delta(\mathcal{Y})$ satisfying $\min_{\pi'} \; \mathbb{P}_{\mathcal{D}}[\pi_k^{\otimes k} \succcurlyeq \pi'] \;\geq\; k/k+1$.*

Proof. Optimality follows from Propositions 3.2 and 3.3, which show that the rate $k/k+1$ cannot be improved, even when allowing multi-output policies. The existence of the stated family of single-output policies $\{\pi_k\}_{k\in\mathbb{N}_{>0}}$ follows directly from Theorem 4.3. □

A few remarks are in order. Compared with existing alignment frameworks such as RLHF and NLHF, the new alignment framework formulates the LLM alignment problem as a multi-player alignment game (Definition 4.2) and aims to find the multi-player Nash equilibrium (MPNE) policy. We briefly discuss the advantages of MPNE over RLHF and NLHF.

1. A simple test-time sampling MPNE policy achieves an optimal $(k, k/k+1)$-robust alignment guarantee. In contrast, test-time scaling of NLHF only yields a $(k, 1/2 + \varepsilon)$-robust alignment guarantee.
2. As a consequence of its strong test-time scaling property, the MPNE policy can generate diverse responses that accommodate both the preferences of the majority and the minority. In contrast, the NLHF/RLHF policy may collapse into a near-deterministic policy that favors the majority.

4.3 Test-Time Performance of a Fixed Robust Policy

Suppose π_k achieves $(k, k/k+1)$-robust alignment. What is its performance when, at test time, we evaluate using $\ell \neq k$ samples? We show that the test-time performance of the product policy $\pi_k^{\otimes \ell}$ degrades gracefully as ℓ decreases, and maintains the win-rate of $k/k+1$ once $\ell \geq k$.

Theorem 4.5. *For any population preference $\mathbb{P}_{\mathcal{D}}$ satisfying Property 1 and any policy π_k achieving $(k, k/k+1)$-robust alignment, we have for all $\ell \geq 1$, $\min_{\pi:\mathcal{X}\to\Delta(\mathcal{Y})} \mathbb{P}_{\mathcal{D}}[\pi_k^{\otimes \ell} \succcurlyeq \pi] \geq \frac{\min\{k,\ell\}}{k+1}$.*

This result shows that robustness confers additional test-time stability: a policy optimized for a sample size k continues to guarantee meaningful performance even when evaluated with fewer samples $\ell < k$, and achieves its full robust-alignment guarantee whenever $\ell \geq k$.

Due to space limitations, we defer results on self-play no-regret learning dynamics in the multi-player alignment game to Section D.6, while we show that gradient dynamics's fixed points are the desirable multi-player Nash equilibrium (MPNE) policy (Proposition D.2). Moreover, we show that by no-regret learning dynamics, we can obtain a policy with $(k, k/k+1 - \varepsilon)$-robust alignment within $T = O(1/\varepsilon^2)$ iterations (Proposition D.3).

In Section D.7, we also extend the definitions of robust alignment and U-alignment to allow the opponent policy to generate $\ell \geq 1$ responses. We show that the MPNE policy remains robust against multi-output opponents.

REFERENCES

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

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

Yang Cai, Argyris Oikonomou, and Weiqiang Zheng. Finite-time last-iterate convergence for learning in multi-player games. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Yang Cai, Gabriele Farina, Julien Grand-Clément, Christian Kroer, Chung-Wei Lee, Haipeng Luo, and Weiqiang Zheng. Fast last-iterate convergence of learning in games requires forgetful algorithms. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=hK7XTpCtBi.

Daniele Calandriello, Daniel Guo, Remi Munos, Mark Rowland, Yunhao Tang, Bernardo Avila Pires, Pierre Harvey Richemond, Charline Le Lan, Michal Valko, Tianqi Liu, et al. Human alignment of large language models through online preference optimisation. *arXiv preprint arXiv:2403.08635*, 2024.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Moses Charikar, Alexandra Lassota, Prasanna Ramakrishnan, Adrian Vetta, and Kangning Wang. Six candidates suffice to win a voter majority. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pp. 1590–1601, 2025.

Feng Chen, Allan Raventos, Nan Cheng, Surya Ganguli, and Shaul Druckmann. Rethinking fine-tuning when scaling test-time compute: Limiting confidence improves mathematical reasoning. In *Workshop on Reasoning and Planning for Large Language Models*, 2025. URL https://openreview.net/forum?id=9L5t04WYAs.

Yu Cheng, Zhihao Jiang, Kamesh Munagala, and Kangning Wang. Group fairness in committee selection. *ACM Transactions on Economics and Computation (TEAC)*, 8(4):1–18, 2020.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In *the 32nd Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

Edith Elkind, Jérôme Lang, and Abdallah Saffidine. Choosing collectively optimal sets of alternatives based on the condorcet criterion. In *IJCAI*, volume 11, pp. 186–191, 2011.

Edith Elkind, Jérôme Lang, and Abdallah Saffidine. Condorcet winning sets. *Social Choice and Welfare*, 44(3):493–517, 2015.

Sara Fish, Paul Gölz, David C Parkes, Ariel D Procaccia, Gili Rusak, Itai Shapira, and Manuel Wüthrich. Generative social choice. In *Proceedings of the 25th ACM Conference on Economics and Computation*, pp. 985–985, 2024.

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

Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Zhihao Jiang, Kamesh Munagala, and Kangning Wang. Approximately stable committee selection. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 463–472, 2020.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=PXD3FAVHJT.

Germain Kreweras. Aggregation of preference orderings. In *Mathematics and Social Sciences I: Proceedings of the seminars of Menthon-Saint-Bernard, France (1–27 July 1960) and of Gösing, Austria (3–27 July 1962)*, pp. 73–79, 1965.

Kaizhao Liu, Qi Long, Zhekun Shi, Weijie J Su, and Jiancong Xiao. Statistical impossibility and possibility of aligning llms with human preferences: From condorcet paradox to nash equilibrium. *arXiv preprint arXiv:2503.10990*, 2025.

Yixin Liu, Argyris Oikonomou, Weiqiang Zheng, Yang Cai, and Arman Cohan. Comal: A convergent meta-algorithm for aligning llms with general preferences. In *NeurIPS 2024 Workshop on Fine-Tuning in Modern Machine Learning: Principles and Scalability*, 2024.

R Duncan Luce et al. *Individual choice behavior*, volume 4. Wiley New York, 1959.

Roberto-Rafael Maura-Rivero, Marc Lanctot, Francesco Visin, and Kate Larson. Jackpot! alignment as a maximal lottery. *arXiv preprint arXiv:2501.19266*, 2025.

Remi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Côme Fiegel, et al. Nash learning from human feedback. In *Forty-first International Conference on Machine Learning*, 2024.

Thanh Nguyen, Haoyu Song, and Young-San Lin. A few good choices. *arXiv preprint arXiv:2506.22133*, 2025.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. *Advances in Neural Information Processing Systems*, 2013.

Zhekun Shi, Kaizhao Liu, Qi Long, Weijie J Su, and Jiancong Xiao. Fundamental limits of game-theoretic llm alignment: Smith consistency and preference matching. *arXiv preprint arXiv:2505.20627*, 2025.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position: A roadmap to pluralistic alignment. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=gQpBnRHwxM.

Gokul Swamy, Christoph Dann, Rahul Kidambi, Steven Wu, and Alekh Agarwal. A minimaximalist approach to reinforcement learning from human feedback. In *Forty-first International Conference on Machine Learning*, 2024.

Daniil Tiapkin, Daniele Calandriello, Denis Belomestny, Eric Moulines, Alexey Naumov, Kashif Rasul, Michal Valko, and Pierre Menard. Accelerating nash learning from human feedback via mirror prox. *arXiv preprint arXiv:2505.19731*, 2025.

Mingzhi Wang, Chengdong Ma, Qizhi Chen, Linjian Meng, Yang Han, Jiancong Xiao, Zhaowei Zhang, Jing Huo, Weijie J Su, and Yaodong Yang. Magnetic preference optimization: Achieving last-iterate convergence for language model alignment. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=PDnEDS244P.

Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Linear last-iterate convergence in constrained saddle-point optimization. In *International Conference on Learning Representations (ICLR)*, 2021.

Fang Wu, Xu Huang, Weihao Xuan, Zhiwei Zhang, Yijia Xiao, Guancheng Wan, Xiaomin Li, Bing Hu, Peng Xia, Jure Leskovec, et al. Multiplayer nash preference optimization. *arXiv preprint arXiv:2509.23102*, 2025a.

Yongtao Wu, Luca Viano, Yihang Chen, Zhenyu Zhu, Kimon Antonakopoulos, Quanquan Gu, and Volkan Cevher. Multi-step alignment as markov games: An optimistic online gradient descent approach with convergence guarantees. *arXiv preprint arXiv:2502.12678*, 2025b.

Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. Self-play preference optimization for language model alignment. In *The Thirteenth International Conference on Learning Representations*, 2025c. URL https://openreview.net/forum?id=a3PmRgAB5T.

Jiancong Xiao, Zhekun Shi, Kaizhao Liu, Qi Long, and Weijie J Su. Theoretical tensions in rlhf: Reconciling empirical success with inconsistencies in social choice theory. *arXiv preprint arXiv:2506.12350*, 2025.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025.

Yuheng Zhang, Dian Yu, Baolin Peng, Linfeng Song, Ye Tian, Mingyue Huo, Nan Jiang, Haitao Mi, and Dong Yu. Iterative nash policy optimization: Aligning LLMs with general preferences via no-regret learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Pujt3ADZgI.

Runlong Zhou, Maryam Fazel, and Simon S Du. Extragradient preference optimization (egpo): Beyond last-iterate convergence for nash learning from human feedback. *arXiv preprint arXiv:2503.08942*, 2025.

CONTENTS

A Related works

We discuss related works from social choice theory.

Social Choice Theory Social choice theory focuses on the problem of choosing a winner or a committee from a set of alternatives according to voters' preferences over the alternatives, which finds application in LLMs recently (see e.g., (Fish et al., 2024; Shi et al., 2025; Xiao et al., 2025)). We review the notions in social choice theory that are closely related to our work. A well-known impossibility result is Condorcet's paradox. This paradox shows that a *Condorcet winner*, an alternative that is preferred by at least half of the voters against any other alternative, may not exist. However, if we allow randomization, a *maximal lottery* (Kreweras, 1965; Fishburn, 1984) exists, which is a distribution over alternatives and guarantees $1/2$ expected win rate against any other alternative. Maximal lottery is the foundation for Nash learning from human feedback (NLHF) (Munos et al., 2024; Swamy et al., 2024). Our work focuses on the test-time scaling version of NLHF and is related to another relaxed notion of Condorcet winner called *Condorcet winning set* (Elkind et al., 2011; 2015), a committee of alternatives such that there is no other alternative that has a win rate $> 1/2$ against all alternatives in the committee. A generalized notion of Condorcet winning set is α-*undominated committee*. A committee S is α-undominated if for all $a \notin S$, less than α-fraction of voters prefer a over each member of S. There is a line of work on finding the smallest α-undominated set (Cheng et al., 2020; Jiang et al., 2020; Charikar et al., 2025; Nguyen et al., 2025). Our alignment framework is closely related to the α-undominated set, as achieving $(k, f(k))$-robust alignment is equivalent to finding a randomized committee of size k that is $(1 - f(k))$-undominated. We show in Section 3 that we can adapt existing results from social choice to show the existence of a multi-output policy that achieves U-alignment with rate $k/k+1$. However, multi-output policies are unsatisfactory in the LLM setting (see Section 3.1 for a detailed discussion). Achieving U-alignment with test-time scaling of single-output policy requires new insights from game theory, as we presented in Section 4.

Game-Theoretic Approaches in Alignment As we have discussed, the Nash learning from human feedback (NLHF) (Munos et al., 2024; Swamy et al., 2024) approach formulates the LLM alignment problem as a two-player zero-sum game and aims to find the Nash equilibrium (NE) policy. The NE

policy coincides with the maximal lottery in social choice and achieves a $1/2$ win rate against any other policies. There are many recent developments on NLHF (see e.g., (Calandriello et al., 2024; Wu et al., 2025c; Zhang et al., 2025; Maura-Rivero et al., 2025; Liu et al., 2025)). In particular, initiated by Liu et al. (2024); Wang et al. (2025), there is a line of works on last-iterate convergent methods for NLHF (Zhou et al., 2025; Wu et al., 2025b; Tiapkin et al., 2025), that applies recent advances in online learning and game theory (Rakhlin & Sridharan, 2013; Daskalakis & Panageas, 2018; Wei et al., 2021; Cai et al., 2022; 2024) to LLM alignment.

Post-Training and Test-Time Scaling One of our work's broader implications is that post-training and test-time scaling may not be aligned, and we need to carefully design the post-training objective in order to achieve good test-time performance. We discuss several related works on the misalignment between post-training and test-time scaling. The work by Yue et al. (2025) shows that the model after post-training via reinforcement learning performs worse than the base model on pass@K (the model is considered correct if at least one of the K samples is correct) when K is large, indicating that post-training may even hurt the reasoning capabilities of the model. Relatedly, Chen et al. (2025) show that in math reasoning settings, supervised finetuning (SFT) with the cross entropy loss can be misaligned with the test-time scaling approach of pass@K. Chen et al. (2025) then suggest a new SFT loss that is more aligned with pass@K. Our work focuses on post-training with general preferences, and we show that existing methods like RLHF and NLHF do not align well with test-time scaling (see Proposition 4.1 for a formal argument). Our new objective, U-alignment, leads to a provable test-time scaling guarantee.

Comparison with (Wu et al., 2025a) A concurrent and independent work (Wu et al., 2025a) that also propose a multiplayer-game formulation for LLM alignment. However, the game objectives in their work and our Definition 4.2 are very different:

- In the game defined in (Wu et al., 2025a), each player i's utility is the *averaged* win rate $\frac{1}{k}\sum_{j\neq i}\mathbb{P}[\pi_i \succ \pi_j]$. This game is symmetric, and we note that the NLHF policy is a symmetric Nash equilibrium of the game. Therefore, solving this multi-player game does not offer a theoretical advantage over the two-player game in NLHF. Moreover, test-time scaling of the Nash equilibrium policy does not achieve U-alignment (Proposition 4.1).
- In our multi-player alignment game (Definition 4.2), each player i's utility is the win rate against the product policy of other players' policies $\mathbb{P}[\pi_i \succ \otimes_{j\neq i}\pi_j]$. Our game is symmetric, and we prove that a symmetric Nash equilibrium achieves U-alignment with optimal rate (Theorem 4.3).

We remark that (Wu et al., 2025a) shows that with diverse opponents, training with their objective exhibits better empirical performance than NLHF. However, using different opponent models breaks the symmetry and no longer solves the original game.

B Missing Details in the Preliminaries

For $m, n \in \mathbb{Z}_+^{\mathcal{Y}}$, we say $m \leq n$ if $m(y) \leq n(y)$ for all $y \in \mathcal{Y}$. For $z \in \mathcal{Y}$, we denote e_z the unit vector with $e_z(z) = 1$ and $e_z(y) = 0$ for $y \neq z$.

We introduce the definition of monotonicity and submodularity with diminishing returns (DR-submodular) below.

Definition B.1 (DR-submodularity on multisets)**.** Let $F : \mathbb{Z}_+^{\mathcal{Y}} \to \mathbb{R}$. We say F is *DR-submodular* if for all $m \leq n$ and all $z \in Y$,

$$F(m + e_z) - F(m) \;\geq\; F(n + e_z) - F(n).$$

We say F is *monotone* if $m \leq n$ implies $F(m) \leq F(n)$, and *normalized* if $F(0) = 0$.

B.1 Proof of Mixture of PL Preferences Satisfying Property 1

The first three properties hold immediately by definition. We prove that the function $F_{x,y} : \mathbb{Z}_+^{\mathcal{Y}} \to [0, 1]$ defined by

$$F_{x,y}(m) := \mathbb{P}_{\mathcal{D}}[S(m) \succcurlyeq y \mid x]$$

is normalized, monotone, and DR-submodular.

It is clear that $F_{x,y}(0) = 0$, hence the function is normalized.

Fix user i, its reward function r_i, and prompt x. Define weights $w_i(x,z) = \exp(r_i(x,z))$. Then for any opponent response $y \in \mathcal{Y}$, we have

$$F^{(i)}_{x,y}(m) := \mathbb{P}_i[S(m) \succcurlyeq y \mid x] = \frac{\sum_{z\in\mathcal{Y}} m(z)\cdot w_i(x,z)}{w_i(x,y) + \sum_{z\in\mathcal{Y}} m(z)\cdot w_i(x,z)} = g\left(\sum_{z\in\mathcal{Y}} m(z)\cdot w_i(x,z)\right),$$

where $g(s) = \frac{s}{a+s}$ with $a = w_i(x,y)$. Since the map g is increasing and concave over $\mathbb{R}_+$ and the function $m \to \sum_{z\in\mathcal{Y}} m(z)\cdot w_i(x,z)$ is increasing, we have $F^{(i)}_{x,y}$ is monotone and DR-submodular. Taking expectation over $i \sim \mathcal{D}$ concludes that $F_{x,y}$ is monotone and DR-submodular. This concludes the proof.

B.2 Proof of Proposition 2.2

Proof. **For the antisymmetry property**, let us fix $\succ_i$. When we draw $S \sim \pi(\cdot \mid x)$, $S' \sim \pi'(\cdot \mid x)$, we have the point-wise identity

$$\mathbf{1}[S \succcurlyeq_{i,x} S'] + \mathbf{1}[S' \succ_{i,x} S] = 1.$$

Taking the expectation over $\succ_i \sim \mathcal{D}$ and the draws of S, S' yields the antisymmetry property.

For the second property, let us fix any single-output policy π, any ranking $\succ_i$, and any prompt x. Draw $k+1$ i.i.d. samples $Z_1, \ldots, Z_k, Z_{k+1} \sim \pi(\cdot \mid x)$. Let us also sample a permutation σ over the index set $[k+1]$ uniformly at random and define $Y_i = Z_{\sigma(i)}$ for $i \in [k+1]$. It is clear that $\{Y_i\}_{i\in[k+1]}$ are also $k+1$ i.i.d. samples from $\pi(\cdot \mid x)$. Now fix $Z = \{Z_i\}_{i\in[k+1]}$, let $y = Y_{k+1}$ and $S = \{Y_1, \ldots, Y_k\}$. Over the randomness of the permutation σ, we have

$$\mathbb{E}\big[\mathbf{1}[y \succ_{i,x} S]\big] = \frac{1}{k+1}\sum_{j=1}^{k+1} \mathbf{1}[Z_j \succ Z \setminus \{Z_j\}] \le \frac{1}{k+1},$$

$$\mathbb{E}\big[\mathbf{1}[S \succcurlyeq_{i,x} y]\big] = 1 - \mathbb{E}\big[\mathbf{1}[y \succ_{i,x} S]\big] \ge 1 - \frac{1}{k+1}.$$

Since the above holds for all realizations of Z and all $\succ_i$, we can take the expectation over Z and $\succ_i \sim \mathcal{D}$ to conclude

$$\mathbb{P}_{\mathcal{D}}[\pi^{\otimes k} \succcurlyeq \pi \mid x] \ge 1 - \frac{1}{k+1}.$$

For the third property, we note that the maximal element in $S_1 \cup S_2$ either lies in S_1 or lies in S_2. Therefore, we have

$$\mathbf{1}[S_1 + S_2 \succ_i S \mid x] \le \mathbf{1}[S_1 \succ_i S \mid x] + \mathbf{1}[S_2 \succ_i S \mid x]$$

Taking expectation over $\succ_i \sim \mathcal{D}$ completes the proof.

For the fourth property, fix user i, prompt x, and the opponent response y. Let the set of responses that is preferred over y given prompt x be

$$U_{i,x}(y) := \{z \in \mathcal{Y} : z \succcurlyeq_{i,x} y\}.$$

Then for user i, the win rate function is an indicator function:

$$F^{(i)}_{x,y}(m) = \mathbf{1}\left[\sum_{z\in U_{i,x}(y)} m(z) \ge 1\right].$$

Clearly $F^{(i)}_{x,y}(0) = 0$ and $F^{(i)}_{x,y}(\cdot)$ is increasing. Moreover, its marginal is

$$F^{(i)}_{x,y}(m + e_z) - F^{(i)}_{x,y}(m) = \begin{cases} 1 & \text{if } z \in U_{i,x}(y) \text{ and } \sum_{z\in U_{i,x}(y)} m(z) = 0 \\ 0 & \text{otherwise} \end{cases}$$

It is clear that as m increases, the marginal can only decrease. So $F^{(i)}_{x,y}(\cdot)$ is DR-submodular. Taking expectation over $i \sim \mathcal{D}$ concludes that $F_{x,y}$ is monotone and DR-submodular. This concludes the proof. □

C Missing Proofs in Section 3

C.1 Proof of Theorem 3.1

Proof. Fix any $x \in \mathcal{X}$. It suffices to prove that

$$\max_{\pi(\cdot|x)\in\Delta(\mathcal{Y}^k)} \min_{y\in\mathcal{Y}} \mathbb{P}_{\mathcal{D}}[\pi \succcurlyeq y \mid x] = \max_{\pi(\cdot|x)\in\Delta(\mathcal{Y}^k)} \min_{\pi'\in\Delta(\mathcal{Y})} \mathbb{P}_{\mathcal{D}}[\pi \succcurlyeq \pi' \mid x] \geq \frac{k}{k+1}$$

By von Neumann's minimax theorem, we have

$$\begin{aligned}
\max_{\pi(\cdot|x)\in\Delta(\mathcal{Y}^k)} \min_{\pi'\in\Delta(\mathcal{Y})} \mathbb{P}_{\mathcal{D}}[\pi \succcurlyeq \pi' \mid x] &= \min_{\pi'\in\Delta(\mathcal{Y})} \max_{\pi(\cdot|x)\in\Delta(\mathcal{Y}^k)} \mathbb{P}_{\mathcal{D}}[\pi \succcurlyeq \pi' \mid x] \\
&\geq \min_{\pi'\in\Delta(\mathcal{Y})} \mathbb{P}_{\mathcal{D}}[(\pi')^{\otimes k} \succcurlyeq \pi' \mid x] \\
&\geq \frac{k}{k+1},
\end{aligned}$$

where the first inequality holds since the k-product distribution $(\pi')^{\otimes k}$ is a valid k-output policy, and the last inequality holds by Property 1. □

C.2 Proof of Proposition 3.2

Proof. Consider a simple case where $\mathcal{X} = \{x\}$ and $\mathcal{Y} = \{y_1, y_2, \ldots, y_{k+1}\}$. Consider a Plackett-Luce preference with a uniform reward function $r_i(x, y) = 1$ for all $y \in \mathcal{Y}$. For any (multi)-set S of size k, we have $\mathbb{P}[S \succcurlyeq y_1 \mid x] = \frac{k}{k+1}$. Thus for any policy $\pi : \mathcal{X} \to \Delta(\mathcal{Y}^k)$, we have $\mathbb{P}[\pi \succcurlyeq y_1] = \mathbb{E}_{S\sim\pi(\cdot|x)}[P[S \succcurlyeq y_1 \mid x]] = \frac{k}{k+1}$. This completes the proof. □

C.3 Proof of Proposition 3.3

Proof. Let $\mathcal{X} = \{x\}$ and $\mathcal{Y} = \{y_1, \ldots, y_m\}$. We consider the population preference $\mathbb{P}_{\mathcal{D}}$ generated from the uniform distribution over all the permutations of m responses, i.e., $m!$ rankings $\{\succ_i\}_{i\in[m!]}$. Consider any multiset $S \in \mathcal{Y}^k$ and $y \in \mathcal{Y}$, we have

$$\mathbb{P}_{\mathcal{D}}[S \succcurlyeq y \mid x] \leq \mathbf{1}[y \in S] + \mathbf{1}[y \notin S] \cdot \frac{k}{k+1} = \mathbf{1}[y \in S] \cdot \frac{1}{k+1} + \frac{k}{k+1},$$

where the inequality holds since (1) $y \in S \implies S \succcurlyeq_i y$ for all i; (2) $y \notin S \implies \mathbb{P}_{\mathcal{D}}[y \succ S] \geq \frac{1}{k+1}$ as the win rate only depends on the relative ordering between $\{y\} \cup S$ and the population is the uniform distribution over all permutations of m responses.

Now let us fix any policy $\pi : \mathcal{X} \to \Delta(\mathcal{Y}^k)$. Since the policy π samples k responses, we have

$$\sum_{j=1}^{m} \Pr_{S\sim\pi(\cdot|x)}[y_j \in S] \leq k.$$

Therefore, there exists $j \in [m]$ such that the response y_j satisfies

$$\Pr_{S\sim\pi(\cdot|x)}[y_j \in S] \leq \frac{k}{m}.$$

Consequently, we have

$$\mathbb{P}_{\mathcal{D}}[\pi \succcurlyeq y_j \mid x] \leq \mathbb{E}_{S\sim\pi(\cdot|x)}\left[\mathbf{1}[y_j \in S] \cdot \frac{1}{k+1} + \frac{k}{k+1}\right] \leq \frac{k}{m \cdot (k+1)} + \frac{k}{k+1}.$$

This completes the proof. □

D MISSING PROOFS IN SECTION 4

D.1 PROOF OF PROPOSITION 4.1

Proof of Proposition 4.1. Let $\mathcal{Y} = \{y_1, y_2\}$ and fix any $\varepsilon \in (0, \frac{1}{2})$. Consider the population distribution $\mathcal{D}$ over two rankings:

$$\mathcal{D} = \begin{cases} y_1 \succ y_2 & \text{with probability } \frac{1}{2} + \varepsilon, \\ y_2 \succ y_1 & \text{with probability } \frac{1}{2} - \varepsilon. \end{cases}$$

Then $\mathbb{P}_{\mathcal{D}}[y_1 \succcurlyeq y_2] = \frac{1}{2} + \varepsilon$.

It is clear that the unique Nash equilibrium policy under $\min_{\pi_1} \max_{\pi_2} \mathbb{P}_{\mathcal{D}}[\pi_1 \succcurlyeq \pi_2]$ is the deterministic policy π_{NLHF} that outputs y_1. Since π_{NLHF} is deterministic, test-time scaling does not change the output. Therefore, for any $k \geq 1$,

$$\min_{y \in \mathcal{Y}} \mathbb{P}_{\mathcal{D}}[\pi_{\mathrm{NLHF}}^{\otimes k} \succcurlyeq y] = \mathbb{P}_{\mathcal{D}}[y_1 \succcurlyeq y_2] = \frac{1}{2} + \varepsilon.$$

This completes the proof. □

D.2 PROOF OF THEOREM 4.3

Proof. The $(k+1)$-player alignment game is symmetric, and each player's strategy space $\Delta(\mathcal{Y})$ is compact and convex, with utilities continuous and linear in each argument. Hence, by standard existence results for symmetric games, the game admits a symmetric Nash equilibrium in which every player uses the same policy π^*.

By the definition of Nash equilibrium, no player can improve her utility by deviating unilaterally. In particular, for any player j,

$$\max_{\pi} u_j(\pi, (\pi^*)^{\otimes k}) \;\leq\; u_j(\pi^*, (\pi^*)^{\otimes k}).$$

By Property 1, the win rate of π^* against k independent copies of itself is at most $\frac{1}{k+1}$, so

$$u_j(\pi^*, (\pi^*)^{\otimes k}) \leq \frac{1}{k+1}.$$

Therefore,

$$\begin{aligned} \min_{\pi} \mathbb{P}_{\mathcal{D}}\big[(\pi^*)^{\otimes k} \succcurlyeq \pi\big] &= 1 - \max_{\pi} \mathbb{P}_{\mathcal{D}}\big[\pi \succ (\pi^*)^{\otimes k}\big] \\ &\geq 1 - u_j(\pi^*, (\pi^*)^{\otimes k}) \\ &\geq 1 - \frac{1}{k+1}. \end{aligned}$$

This proves that $(\pi^*)^{\otimes k}$ achieves the claimed robust alignment guarantee. □

D.3 PROOF OF THEOREM 4.5

We first establish a useful lemma based on (4) in Property 1.

Lemma D.1 (DR-submodularity ⇒ concavity in sample size)**.** *Assume $F : \mathbb{Z}_+^Y \to \mathbb{R}$ is normalized, monotone, and DR-submodular. Let $M_\ell = \sum_{t=1}^{\ell} e_{Z_t}$ for $Z_t \overset{iid}{\sim} \pi$ and define $G(\ell) = \mathbb{E}[F(M_\ell)]$. Then:*

1. *$G(\ell)$ is nondecreasing in ℓ.*
2. *$G(\ell)$ is* discrete concave*: $G(\ell+2) - G(\ell+1) \leq G(\ell+1) - G(\ell)$ for all $\ell \geq 0$.*
3. *Consequently, for any integer $k \geq 1$ and all $0 \leq \ell \leq k$,*

$$G(\ell) \;\geq\; \frac{\ell}{k}\, G(k).$$

Proof. (1) By monotonicity, $F(M_{\ell+1}) = F(M_\ell + e_{Z_{\ell+1}}) \geq F(M_\ell)$, so taking expectation gives $G(\ell+1) \geq G(\ell)$.

(2) Let $\Delta(z \mid m) := F(m + e_z) - F(m)$ be the marginal gain of adding one copy of z. Then

$$G(\ell+1) - G(\ell) = \mathbb{E}\big[F(M_\ell + e_{Z_{\ell+1}}) - F(M_\ell)\big] = \mathbb{E}\big[\Delta(Z_{\ell+1} \mid M_\ell)\big].$$

DR-submodularity guarantees that for every fixed z, the function $m \mapsto \Delta(z \mid m)$ is nonincreasing w.r.t. the partial order. Under the natural coupling $M_{\ell+1} = M_\ell + e_{Z_{\ell+1}} \geq M_\ell$, and since $\Delta(z \mid \cdot)$ is nonincreasing,

$$\mathbb{E}\big[\Delta(Z_{\ell+2} \mid M_{\ell+1})\big] \;\leq\; \mathbb{E}\big[\Delta(Z_{\ell+2} \mid M_\ell)\big].$$

Using independence of $Z_{\ell+2}$ from $(M_\ell, M_{\ell+1})$, the RHS equals $\mathbb{E}[\Delta(Z_{\ell+1} \mid M_\ell)]$. Hence $G(\ell+2) - G(\ell+1) \leq G(\ell+1) - G(\ell)$, proving discrete concavity.

(3) Discrete concavity implies the forward differences $d_\ell := G(\ell) - G(\ell-1)$ are nonincreasing in ℓ. Thus for $1 \leq \ell \leq k$,

$$\frac{G(\ell)}{\ell} = \frac{1}{\ell}\sum_{t=1}^{\ell} d_t \;\geq\; \frac{1}{k}\sum_{t=1}^{k} d_t = \frac{G(k)}{k},$$

which rearranges to $G(\ell) \geq \frac{\ell}{k} G(k)$. This completes the proof. □

Proof of Theorem 4.5. We know π_k achieves $(k, {}^{k}/_{k+1})$ robust alignment. Fix any prompt x and opponent response y. Recall the function $F_{x,y} : \mathbb{Z}_+^{\mathcal{Y}} \to \mathbb{R}$ defined as $F_{x,y}(m) = \mathbb{P}_{\mathcal{D}}[S(m) \succcurlyeq y]$ is monotone, normalized, and DR-submodular (Property 1 (4)). Then using the notations in Lemma D.1, we have

$$G(k) = \mathbb{E}_{S \sim \pi_k^{\otimes k}}[F_{x,y}[m_S]] = \mathbb{P}_{\mathcal{D}}[\pi_k^{\otimes k} \succcurlyeq y \mid x] \geq \frac{k}{k+1}.$$

By Lemma D.1, we have for any $0 \leq \ell \leq k$,

$$\mathbb{P}_{\mathcal{D}}[\pi_k^{\otimes \ell} \succcurlyeq y \mid x] = G(\ell) \geq \frac{\ell}{k} G(k) \geq \frac{\ell}{k+1}.$$

For $k \geq \ell$, monotonicity of $F_{x,y}(\cdot)$ implies $\mathbb{P}_{\mathcal{D}}[\pi_k^{\otimes \ell} \succcurlyeq y \mid x] = G(\ell) \geq G(k) = {}^{k}/_{k+1}$. This completes the proof. □

D.4 Proof of Proposition D.2

Proof. It is not hard to see that the map $\pi \mapsto \nabla_{\pi_1} u_1(\pi, \pi^{\otimes k})$ is continuous. The projection operator $\Pi_{\Delta(\mathcal{Y})}$ is continuous, and $\Delta(\mathcal{Y})$ is compact and convex. Therefore, F_η is a continuous map from $\Delta(\mathcal{Y})$ to itself. By Brouwer's fixed-point theorem, F_η has at least one fixed point.

Now let π be any fixed point of F_η, and write

$$g^\pi := \nabla_{\pi_1} u_1(\pi, \pi^{\otimes k}).$$

The fixed-point condition $\pi = \Pi_{\Delta(\mathcal{Y})}(\pi + \eta \cdot g^\pi)$ implies, by the first-order optimality condition for Euclidean projection onto a convex set, that

$$\langle g^\pi,\ \pi' - \pi \rangle \leq 0, \qquad \forall \pi' \in \Delta(\mathcal{Y}).$$

Equivalently,

$$\langle g^\pi,\ \pi' \rangle \leq \langle g^\pi,\ \pi \rangle, \qquad \forall \pi' \in \Delta(\mathcal{Y}).$$

Since $u_1(\cdot, \pi^{\otimes k})$ is linear in its first argument, the gradient g^π represents the linear functional defining player 1's payoff against $\pi^{\otimes k}$, and the inequality above is exactly the best-response condition:

$$u_1(\pi', \pi^{\otimes k}) \leq u_1(\pi, \pi^{\otimes k}), \qquad \forall \pi' \in \Delta(\mathcal{Y}).$$

By symmetry, the same holds for every player. Hence $(\pi, \ldots, \pi)$ is a symmetric Nash equilibrium of the multi-player alignment game. □

D.5 Proof of Proposition D.3

Proof. By the definition of regret, for any $\pi \in \Delta(\mathcal{Y})$,

$$\frac{1}{T}\sum_{t=1}^{T} u(\pi, (\pi^t)^{\otimes k}) \le \frac{1}{T}\sum_{t=1}^{T} u(\pi^t, (\pi^t)^{\otimes k}) + \frac{\mathrm{Reg}^T}{T}.$$

Since σ^T uniformly samples from $\{(\pi^t)^{\otimes k}\}_{t\in[T]}$, the left-hand side equals $u(\pi, \sigma^T)$. Moreover, by Property 1, for each t we have $u(\pi^t, (\pi^t)^{\otimes k}) \le \frac{1}{k+1}$, and hence the first term on the right-hand side is also at most $\frac{1}{k+1}$. Therefore,

$$\max_{\pi\in\Delta(\mathcal{Y})} u(\pi, \sigma^T) \le \frac{1}{k+1} + \frac{\mathrm{Reg}^T}{T}.$$

Applying Property 1 once more, we obtain

$$\min_{\pi\in\Delta(\mathcal{Y})} \mathbb{P}_{\mathcal{D}}[\sigma^T \succcurlyeq \pi] \ge \frac{k}{k+1} - \frac{\mathrm{Reg}^T}{T},$$

which proves the claim. □

D.6 Theoretical Guarantees on Convergence of Self-Play Learning Dynamics

We study the convergence properties of self-play no-regret learning dynamics in multi-player alignment games (Definition 4.2) to the desirable multi-player Nash equilibrium (MPNE) policy. When $k = 1$, the resulting two-player alignment game is a constant-sum game and coincides with Nash learning from human feedback (NLHF). In this case, a Nash equilibrium policy can be learned efficiently via self-play using no-regret algorithms (Munos et al., 2024; Swamy et al., 2024), and algorithms with last-iterate convergence guarantees are known (Liu et al., 2024; Wang et al., 2025). Moreover, these no-regret methods are practical to implement and have been successfully applied to large-scale LLM alignment with strong empirical performance (Wu et al., 2025c; Zhang et al., 2025; Liu et al., 2024).

When $k > 1$, the $(k+1)$-player alignment game becomes substantially more challenging to solve. In this regime, there are currently no general theoretical guarantees for the convergence of no-regret self-play dynamics to Nash equilibria. In the remainder of this section, we establish two theoretical guarantees for self-play learning dynamics in multi-player alignment games.

Our first guarantee shows that any symmetric Nash equilibrium policy is a fixed point of gradient-based learning dynamics. Since the game is symmetric, all players apply the same update rule. Accordingly, we write π^t for the common policy at iteration $t \ge 1$. This result implies that if the gradient-ascent dynamic converges, then its limit point must be a symmetric Nash equilibrium of the alignment game.

Proposition D.2 (Fixed Point of Projected Gradient Ascent)**.** *Let $u_1(\pi_1, \pi_{-1})$ be the utility of player 1 in the $(k+1)$-player alignment game. For any step size $\eta > 0$, define the projected gradient-ascent operator $F_\eta : \Delta(\mathcal{Y}) \to \Delta(\mathcal{Y})$ by $F(\pi) := \Pi_{\Delta(\mathcal{Y})}\Big(\pi + \eta \cdot \nabla_{\pi_1} u_1(\pi, \pi^{\otimes k})\Big)$, where $\Pi_{\Delta(\mathcal{Y})}$ denotes Euclidean projection onto the simplex. Then F_η has at least one fixed point. Moreover, any fixed point π of F_η induces a symmetric Nash equilibrium $(\pi, \ldots, \pi)$ of the multi-player alignment game.*

Our second result provides finite-time guarantees for no-regret self-play learning dynamics (Cesa-Bianchi & Lugosi, 2006), at the cost of storing a sequence of past policies. Let each player employ the same online learning algorithm. Since the alignment game is symmetric, each player produces the same sequence of strategies $\{\pi^t\}_{t\in[T]}$. Recall that each player's utility function is $u(\cdot, (\pi^t)^{\otimes k}) = \mathbb{P}_{\mathcal{D}}(\cdot \succ \sigma)$, which is linear. The regret of each player is

$$\mathrm{Reg}^T := \max_{\pi\in\Delta(\mathcal{Y})} \sum_{t=1}^{T} u(\pi, (\pi^t)^{\otimes k}) - \sum_{t=1}^{T} u(\pi^t, (\pi^t)^{\otimes k}).$$

Consider any no-regret learning algorithm with regret Reg^T. We claim that the policy which samples $t \sim \mathrm{Unif}([T])$ and then executes $(\pi^t)^{\otimes k}$ achieves $(k, k/k+1 - \mathrm{Reg}^T/T)$-robust alignment.

Proposition D.3 (No-Regret Self-Play Dynamics)**.** *Suppose the players self-play using a no-regret learning algorithm and generate iterates $\{\pi^t\}_{t\in[T]}$ with regret Reg^T. Let σ^T denote the policy that samples $t \sim \mathrm{Unif}([T])$ and then executes $(\pi^t)^{\otimes k}$. Then σ^T achieves $(k, k/k+1 - \mathrm{Reg}^T/T)$-robust alignment.*

We note that many no-regret algorithms, such as online gradient ascent and multiplicative weights update, achieve $\mathrm{Reg}^T = O(\sqrt{T})$ (Hazan et al., 2016; Orabona, 2019). Consequently, taking $T = O(1/\varepsilon^2)$ iterations suffices to obtain a policy that achieves $(k, f(k))$-robust alignment with $f(k) \geq k/k+1 - \varepsilon$.

D.7 Extensions to Multi-Output Opponents

We extend the definitions of robust alignment and U-alignment to allow the opponent policy to generate $\ell \geq 1$ responses. This strictly generalizes the standard case $\ell = 1$.

Definition D.4 ($(k, \ell, f_\ell(k))$-Robust Alignment)**.** Let $k, \ell \in \mathbb{N}_{>0}$ denote the numbers of responses generated at test time by the model and the opponent, respectively. Let $f_\ell : \mathbb{N}_{>0} \to [0, 1]$ be a function. A policy $\pi : \mathcal{X} \to \Delta(\mathcal{Y}^k)$ is said to achieve $(k, \ell, f_\ell(k))$-robust alignment if, for every prompt $x \in \mathcal{X}$, its win rate against any opponent policy $\pi' : \mathcal{X} \to \Delta(\mathcal{Y}^\ell)$ satisfies $\min_{\pi'} \Pr[\pi \succcurlyeq \pi' \mid x] \geq f_\ell(k)$.

We note that $(k, f(k))$-robust alignment is the special case $\ell = 1$ of the above definition, that is, $(k, 1, f_1(k))$-robust alignment. Equipped with this extension, we now introduce the notion of *ℓ-U-alignment*, which reduces to U-alignment when $\ell = 1$.

Definition D.5 (ℓ-Asymptotic U-alignment)**.** Fix $\ell \in \mathbb{N}_{>0}$ as the number of responses generated at test time by the opponent model. Let $f_\ell : \mathbb{N}_{>0} \to [0, 1]$ be a rate function satisfying $\lim_{k\to\infty} f_\ell(k) = 1$. We say that a family of policies $\{\pi_k\}_{k\in\mathbb{N}_{>0}}$ achieves *ℓ-U-alignment with rate f_ℓ* if, for every $k \in \mathbb{N}_{>0}$, the policy π_k achieves $(k, \ell, f_\ell(k))$-robust alignment.

In Theorem 4.4, we show that when $\ell = 1$ we can achieve 1-U-alignment with rate $f_1(k) = k/k+1$. A natural question is: what is the optimal rate for ℓ-U-alignment? On the lower bound side, we can adapt the proof of Proposition 3.2 to obtain a lower bound of $k/k+\ell$ (for all $k \geq 1$) under the Plackett–Luce model. On the upper bound side, combining Theorem 4.4 with a union bound shows that the Nash equilibrium policy of the multi-player alignment game (Definition 4.2) already achieves ℓ-U-alignment with rate $k+1-\ell/k+1$. In particular, if the opponent outputs ℓ responses and we seek a $\frac{1}{2}$ win rate guarantee, then it suffices for the model to output 2ℓ responses. The upper and lower bounds coincide for $\ell = 1$, but a gap remains for $\ell > 1$. Closing this gap is an interesting direction for future work.

Theorem D.6. *There exists a family of single-output policies $\{\pi_k\}_{k\in\mathbb{N}_{>0}}$ such that the corresponding test-time scaled policies $\{\pi_k^{\otimes k}\}_{k\in\mathbb{N}_{>0}}$ achieve ℓ-U-alignment with rate $k+1-\ell/k+1$ for any $\ell \in \mathbb{N}_{>0}$. In particular, π_k can be chosen as a symmetric Nash equilibrium of the $(k+1)$-player alignment game (Definition 4.2).*

D.8 Proof of Theorem D.6

Proof. Fix any $k \geq 1$ and let π^* be a symmetric Nash equilibrium policy of the $(k+1)$-player alignment game. By Theorem 4.3,

$$\max_{y\in\mathcal{Y}} \mathbb{P}_{\mathcal{D}}\left[y \succ (\pi^*)^{\otimes k}\right] \leq \frac{1}{k+1}.$$

By the subadditivity property in Property 1, for any $S = (y^{(1)}, \ldots, y^{(\ell)}) \in \mathcal{Y}^\ell$,

$$\mathbb{P}_{\mathcal{D}}\left[S \succ (\pi^*)^{\otimes k}\right] \leq \sum_{i=1}^{\ell} \mathbb{P}_{\mathcal{D}}\left[y^{(i)} \succ (\pi^*)^{\otimes k}\right] \leq \frac{\ell}{k+1},$$

and hence

$$\max_{S\in\mathcal{Y}^\ell} \mathbb{P}_{\mathcal{D}}\left[S \succ (\pi^*)^{\otimes k}\right] \leq \frac{\ell}{k+1}.$$

Using antisymmetry from Property 1, we obtain

$$\begin{aligned}\min_{\pi\in\Delta(\mathcal{Y}^\ell)} \mathbb{P}_{\mathcal{D}}\big[(\pi^*)^{\otimes k} \succcurlyeq \pi\big] &= 1 - \max_{\pi\in\Delta(\mathcal{Y}^\ell)} \mathbb{P}_{\mathcal{D}}\big[\pi \succ (\pi^*)^{\otimes k}\big] \\ &= 1 - \max_{S\in\mathcal{Y}^\ell} \mathbb{P}_{\mathcal{D}}\big[S \succ (\pi^*)^{\otimes k}\big] \\ &\geq 1 - \frac{\ell}{k+1} = \frac{k+1-\ell}{k+1}.\end{aligned}$$

This completes the proof. □