CURATRON: Complete and Robust Preference Data for Rigorous Alignment of Large Language Models

Son The Nguyen
University of Illinois Chicago
Chicago, Illinois, USA
snguye65@uic.edu

Niranjan Uma Naresh
Independent
Kirkland, Washington, USA
un.niranjan@gmail.com

Theja Tulabandhula
University of Illinois Chicago
Chicago, Illinois, USA
theja@uic.edu

1 Introduction

Large Language Models (LLMs) are highly advanced Artificial Intelligence (AI) systems capable of understanding, interpreting, and generating languages. The integration of AI chatbots like ChatGPT into our daily lives and businesses has had a profound impact on both society and industries (Eloundou et al., 2023). However, the success of GPTs/LLMs depends not only on their ability to generate responses and perform tasks well but also on their alignment with human values and expectations.

The prevalent method for aligning AI/LLMs currently involves preference learning (PL) through human feedback. However, gathering human feedback is slow and expensive and often results in incomplete or imperfect data (Bai et al., 2022; Lee et al., 2023). Furthermore, participants may intentionally provide inaccurate or harmful feedback due to malicious intentions, as pointed out by (Casper et al., 2023). These factors can lead to unintended consequences in estimating rankings from preference datasets from models such as BTL. They pose a considerable challenge in ensuring the integrity and reliability of the preference datasets used for aligning LLMs, especially when scaling up the alignment process with large-scale responses and participants.

Approaching the issues, we consider the following learning problem: Suppose there are n responses we wish to order based on a notion of comparison, between every pair of responses, with probabilistic outcomes. Further, we are given a set, $\aleph = \{(i, j, \{y_{ij}^k\})\}$, consisting of K independent pairwise comparison outcomes, denoted by $\{y_{ij}^k\} \in \{0, 1\}, k \in [K]$, between pairs of responses $(i, j) \subseteq [n] \times [n]$, a significant proportion of which might be corrupted by an adversary.

In this passive learning setting, our contributions are as follows. We give a generic definition of (additive) adversarial noise and show that if it is not accounted for, the quality of the estimated ranking can be quite poor. To address this, we develop an efficient and correct ranking method called Robust Preference Data for Rigorous Alignment (RORATRON), which is robust against adversarial noise. Under certain assumptions, we prove that our method guarantees high-probability learnability with a small margin of error. We also devise a method called Complete Robust Preference Data for Rigorous Alignment (CURATRON) to handle the scenario where not all pairs are compared, and the observed pairwise data is adversarially corrupted.

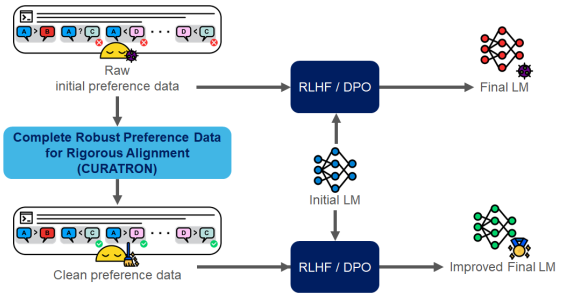

Figure 1: CURATRON corrects incomplete and adversarially corrupted preference data to improve RLHF/DPO alignment results compared to using the raw initial preference data.

2 Related Work

LLM Alignment with PL from human feedback: PL was initially developed to train agents in simulated environments to perform nuanced behaviors that are hard to define but easy to observe and recognize (Christiano et al., 2017). It has recently been found successful in aligning LLMs to human intentions and values such as harmfulness, helpfulness, factuality, and safety. Some of the methods of PL in LLMs are RLHF (Ouyang et al., 2022), RLAIF (Bai et al., 2022; Lee et al., 2023), DPO/ψPO (Rafailov et al., 2023; Tunstall et al., 2023; Zhao et al., 2023),

Proceedings of the Fifth Workshop on Data Science with Human-in-the-Loop (Language Advances), pages 31–39

and SLiC-HF (Zhao et al., 2023).

Ranking Models: In the BTL model, item i has an associated score w_i; then, the probability that item i is preferred over j is given by $P_{ij} = e^{-w_i}/(e^{-w_i} + e^{-w_j})$ where $\mathbf{w} \in \mathbb{R}^n$ is the BTL parameter vector to be estimated from data; here, $\mathbf{P} \in \mathbb{R}^{n \times n}$ is called the 'preference matrix'. A closely related model, in the non-active setting, is the recently proposed LR model (Rajkumar and Agarwal, 2016) wherein a generic class of preference matrices is characterized to be those having low rank under transformations using certain functions; specifically, for BTL-like models, the logit function defined as $\psi(x) = \log(x/(1-x))$ turns out to right choice as shown in their paper. However, while their model accounts for missing information, they do not consider the harder problem of handling adversarial noise.

Robust Subspace Recovery: The Robust PCA (RPCA) problem (Netrapalli et al., 2014) addresses the following question: suppose we are given a data matrix $\mathbf{M}$ which is the sum of an unknown low-rank matrix $\mathbf{L}$ and an unknown sparse matrix $\mathbf{S}$, can we recover each of the component matrices? While several works (Yi et al., 2016; Hsu et al., 2011) analyze this problem, it is shown in (Netrapalli et al., 2014) that, under information-theoretically tight assumptions, a simple iterative algorithm based on non-convex alternating projections of appropriate residuals provably yields an ϵ-accurate solution in $O(\log(1/\epsilon))$ iterations with an overall computational complexity of $O(n^2 r^2 \log(1/\epsilon))$ where r is the rank of $\mathbf{L}$. We will use this result, in particular, to derive guarantees for our ranking problem.

3 Problem Setup

3.1 Notation

We denote the set of all permutations of n LLM responses/items as $\mathcal{S}_n$. If not specifically defined, we use lower-case letters for scalars, upper-case letters for global constants, lower-case bold-face letters for vectors and upper-case bold-face letters for matrices; specifically, $\mathbf{P}$ denotes a preference matrix. Let $\mathcal{P}_n := \{\mathbf{P} \in [0,1]^{n \times n} | P_{ij} + P_{ji} = 1\}$ denote the set of all pairwise preference matrices over n responses. Let the set of stochastic-transitive matrices be $\mathcal{P}_n^{ST} := \{\mathbf{P} \in \mathcal{P}_n | P_{ij} > 1/2, P_{jk} > 1/2 \implies P_{ik} > 1/2\}$. Let the set preference matrices described by the BTL model be $\mathcal{P}_n^{BTL} := \{\mathbf{P} \in \mathcal{P}_n | \exists \mathbf{w} \in \mathbb{R}^n \text{ s.t. } e^{-w_i}/(e^{-w_i} + e^{-w_j})\}$. Let $\psi : [0,1] \mapsto \mathbb{R}$ be a strictly increasing bijective L-Lipschitz function and define the class of low-rank preference matrices with respect to ψ as $\mathcal{P}_n^{LR(\psi,r)} = \{\mathbf{P} \in \mathcal{P}_n | \text{rank}(\psi(\mathbf{P})) \leq r\}$ where $r \in [n]$; when we apply such a transformation to a matrix, it is applied entry-wise. In this paper, we take ψ to be the logit function.

For any matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$, let the infinity norm be denoted by $\|\mathbf{M}\|_\infty = \max_{i,j} |M_{ij}|$, the Frobenius norm be denoted by $\|\mathbf{M}\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^n M_{ij}^2}$, the spectral norm be denoted by $\|\mathbf{M}\|_2 = \max_{\mathbf{x},\mathbf{y} \in \mathbb{R}^n} \mathbf{x}^\top \mathbf{M} \mathbf{y}$. Denoting the indicator function by $\mathbb{1}$, define the zero norm of a matrix to be the maximum number of non-zero elements in any row/column, ie, $\|\mathbf{M}\|_0 = \max(\max_j \sum_{i=1}^n \mathbb{1}(M_{ij} \neq 0), \max_i \sum_{j=1}^n \mathbb{1}(M_{ij} \neq 0))$. Let the Singular Value Decomposition (SVD) of a square matrix be given by $\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ where $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{n \times r}$ are orthonormal matrices (whose columns are singular vectors) and $\mathbf{\Sigma} \in \mathbb{R}^{r \times r}$ is the diagonal matrix of singular values. Now, $\mathbf{M}$ is said to be μ-incoherent if $\max\left(\max_i \left\|\mathbf{e}_i^\top \mathbf{U}\right\|_2, \max_i \left\|\mathbf{e}_i^\top \mathbf{V}\right\|_2\right) \leq \mu\sqrt{r/n}$ where $\mathbf{e}_i$ denotes the i^{th} basis vector in $\mathbb{R}^n$. Also, let $\sigma_{\max} := \max_i \Sigma_{ii}$ and $\sigma_{\min} := \min_i \Sigma_{ii}$.

We define the distance between a permutation $\sigma \in \mathcal{S}_n$ and a preference matrix $\mathbf{P} \in \mathcal{P}_n$ as:

$$\text{dist}(\sigma, \mathbf{P}) := \binom{n}{2}^{-1} \sum_{i<j} \mathbb{1}\left((P_{ij} > 1/2) \wedge (\sigma(i) \succ \sigma(j))\right) + \binom{n}{2}^{-1} \sum_{i<j} \mathbb{1}\left((P_{ji} > 1/2) \wedge (\sigma(j) \succ \sigma(i))\right)$$

Note that the above loss function basically is the number of pairs on which the ordering with respect σ and $\mathbf{P}$ differ divided by the number of ways to choose two out of n responses. Finally, let $P_{\min} = \min_{i \neq j} P_{ij}$ and $\Delta = \min_{i \neq j} |\psi(P_{ij}) - \psi(1/2)|$.

3.2 Characterization of the Adversary

The following (weak) assumption characterizes the properties of the adversary.

Assumption 1. *The (additive) adversarial noise which corrupts a μ-incoherent preference matrix $\mathbf{P} \in \mathcal{P}_n^{LR(\psi,r)}$ is modeled by a skew-symmetric sparse matrix $\mathbf{S}$ so that the corrupted preference matrix $\mathbf{P^c} \in \mathcal{P}_n$ is given by $\mathbf{P^c} = \mathbf{P} + \mathbf{S}$. We assume the (deterministic) bounded degree condition that $\|\mathbf{S}\|_0 \leq d < n$ such $d < n/512\mu^2 r$ where $r \leq n$.*

So, why do existing non-robust algorithms not recover the true response ordering in the presence

of an adversarial noise source? This question is answered by the following proposition, which precisely quantifies how bad a ranking could be when an algorithm uses the corrupted pairwise preference matrix. The key idea is to construct an adversary that intentionally flips true comparison results.

Claim 1 (Upper bound on estimation error). *Under Assumption 1 it is possible that* $\mathrm{dist}(\widehat{\sigma}, \mathbf{P^c}) = O(1)$.

Proof. Assume that we are exactly given the entries of the preference matrix as opposed to sampling them. Note that in order to estimate a ranking from a given preference matrix, we still need to use a pairwise ranking procedure. Let $\widehat{\sigma} \in \mathcal{S}_n$ be the output of any Pairwise Ranking (PR) procedure with respect to an underlying preference matrix $\mathbf{Q} \in \mathcal{P}_n$. For a constant $\gamma > 1$, $\widehat{\sigma}$ is said to be γ-approximate if $\mathrm{dist}(\widehat{\sigma}, \mathbf{Q}) \leq \gamma \min_{\sigma \in \mathcal{S}_n} \mathrm{dist}(\sigma, \mathbf{Q})$. Define the following distance which measures the fraction of response pairs over which two preference matrices $\{\mathbf{Q}, \mathbf{R}\} \in \mathcal{P}_n$ disagree.

$$\mathrm{dist}(\mathbf{Q},\mathbf{R}) := \binom{n}{2}^{-1} \sum_{i<j} \mathbb{1}\left((Q_{ij} > 1/2) \wedge (R_{ij} < 1/2)\right) + \binom{n}{2}^{-1} \sum_{i<j} \mathbb{1}\left((Q_{ij} < 1/2) \wedge (R_{ij} > 1/2)\right)$$

By Lemma 20 of (Rajkumar and Agarwal, 2016), for $\mathbf{Q} \in \mathcal{P}_n^{ST}$ and $\mathbf{R} \in \mathcal{P}_n$, we have $\mathrm{dist}(\widehat{\sigma}, \mathbf{Q}) \leq (1+\gamma)\,\mathrm{dist}(\mathbf{Q}, \mathbf{R})$. But note that it is possible that $\mathrm{dist}(\mathbf{Q}, \mathbf{R}) = 1$ as it is easy to construct by $\mathbf{R}$ that disagrees with $\mathbf{Q}$ in every entry by simply setting $\mathbf{R} = \mathbf{Q}^\top$. Now, we may set $\mathbf{Q} = \mathbf{P}$ and $\mathbf{R} = \mathbf{P^c}$ for any algorithm that uses $\mathbf{P^c}$ for ranking; specifically, for the adversary satisfying Assumption 1, we can see by a direct counting argument that $\mathrm{dist}(\mathbf{Q}, \mathbf{R}) \leq \frac{d(2n-1-d)}{n(n-1)}$ which proves the claim. □

4 Fully Observed Adversarial Setting

4.1 Algorithm

We present our main algorithm for robust passive ranking from pairwise comparisons in the presence of adversarial noise in Algorithm 1. The input data consist of the set of pairwise comparison results $\aleph = \{(i, j, \{y_{ij}^k\})\}$, $(i, j) \in [n] \times [n]$, $k \in [K]$, $y_{ij}^k \in \{0, 1\}$. The algorithm assumes the true rank of $\psi(\mathbf{P})$ as an input parameter; specifically, for the BTL model, we set $r = 2$. Algorithm 1 calls the Robust PCA and γ-approximate pairwise ranking procedures.

Algorithm 1 RORATRON: Robust Preference Data for Rigorous Alignment

Input: Comparison dataset $\aleph = \{(i, j, \{y_{ij}^k\})\}$, true rank r.
Output: Ranking of n responses, $\hat{\sigma} \in \mathcal{S}_n$.
1: Estimate entries of $\widehat{\mathbf{P}}$ for $i \leq j$ as:

$$\widehat{P}_{ij} = \begin{cases} \frac{1}{K}\sum_{k=1}^{K} y_{ij}^k & \text{if } i < j \\ 1/2 & \text{if } i = j \end{cases}$$

2: Set $\widehat{P}_{ij} = 1 - \widehat{P}_{ji}$ for all $i > j$.
3: Perform robust PCA: $\{\psi(\overline{\mathbf{P}}), \widehat{\mathbf{S}}\} \leftarrow \mathrm{RPCA}(\psi(\widehat{\mathbf{P}}), r)$.
4: Using a pairwise ranking procedure after taking the inverse transform: $\widehat{\sigma} \leftarrow \mathrm{PR}(\overline{\mathbf{P}})$.
5: **return** $\widehat{\sigma}$.

4.2 Analysis

We begin with a useful short result followed by the statement and the proof of our main result that, with high probability, we achieve ϵ–accurate ranking in polynomial time using polynomial number of samples, despite the presence of adversarial noise. In this context, it is noteworthy that we present the result for LR models which strictly contain the BTL model while being much more general (Rajkumar and Agarwal, 2016); upon proving this result, we specialize it to the classic BTL model as well (Corollary 1).

Lemma 1 (Some properties of the logit function). *Let* $a, b, c \in (0, 1)$ *such that* $c = a + b$. *Then, we have,*

1. $\psi(c) = \psi(a) + \psi(a + b) + \psi(1 - a)$
2. $\psi(a) + \psi(1 - a) = 0$.

Proof. Both follow by using the definition of the logit function that $\psi(a) = \log(a/(1-a))$ and using the property that $\log(ab) = \log(a) + \log(b)$. □

Theorem 1 (Provably good estimation of ranking in LR models in the presence of adversarial noise). *Let* $\mathbf{P} \in \mathcal{P}_n^{LR(\psi,r)}$ *be the true preference matrix according to which the pairwise comparison dataset* $\aleph = \{(i, j, \{y_{ij}^k\})\}$ *is generated for all responses pairs* (i, j) *such that* $k \in [K]$. *Let* $\widehat{\mathbf{P}}$ *be the empirical preference matrix computed using* $\aleph$. *Let* $\mathbf{S} \in [0, 1]^{n \times n}$ *be the adversarial matrix that additively corrupts* $\widehat{\mathbf{P}}$. *Let* ψ *be L-Lipschitz in* $[\frac{P_{\min}}{2}, 1 - \frac{P_{\min}}{2}]$ *and* $\psi(\mathbf{P})$ *be* μ*-incoherent. Let each pair be compared independently* $K \geq 16384\mu^2(1+\gamma)L^2n^2\log^2(n)/\epsilon\Delta^2$ *times where* $\Delta = \min_{i \neq j} \left|\psi(P_{ij}) - \psi(1/2)\right|$. *Then, with probability atleast* $1 - 1/n^3$, *Algorithm 1 returns an estimated permutation* $\widehat{\sigma}$ *such that* $\mathrm{dist}(\widehat{\sigma}, \mathbf{P}) \leq \epsilon$.

Remark 1 (Computational complexity). *In Algorithm 1, Step 1 takes* $O(n^2K) = O(n^4 \log^2 n/\epsilon)$

time, Step 3 takes $O(n^2r^2\log(1/\epsilon))$, and Step 4 takes $O(n^2 + n\log n)$ time. Thus, putting together the cost of these main steps, the overall computational complexity of our robust ranking algorithm for $\mathbf{P} \in \mathcal{P}_n^{LR(\psi,r)}$ is $O(n^4\log^2 n/\epsilon)$.

Remark 2 (Identifying adversarially corrupted pairwise comparisons). *From Step 3 of Algorithm 1, using Theorem 2 of (Netrapalli et al., 2014), we also have* $\mathrm{Supp}(\widehat{\mathbf{S}}) \subseteq \mathrm{Supp}(\mathbf{S})$ *and thus we can identify the corrupted pairwise comparison results.*

Proof. Let $\widetilde{P}_{ij}$ be the empirical probability estimate of P_{ij}. Note that we compute $\widetilde{P}_{ij} = \frac{1}{K}\sum_{k=1}^{K} y_{ij}^k$ from the given pairwise comparison dataset, $\aleph = \{(i,j,\{y_{ij}^k\})\}$. Now, $\widehat{\mathbf{P}} = \widetilde{\mathbf{P}} + \mathbf{S}$. By Lemma 1, we may write the adversarially corrupted empirical probability estimate as $\psi(\widehat{\mathbf{P}}) = \psi(\widetilde{\mathbf{P}}) + \widetilde{\mathbf{S}}$ where $\widetilde{\mathbf{S}} = \psi(\widetilde{\mathbf{P}}+\mathbf{S}) + \psi(1-\widetilde{\mathbf{P}})$. We have $\psi(\widetilde{\mathbf{P}}) = \psi(\mathbf{P}) + \widetilde{\mathbf{N}}$ where $\widetilde{\mathbf{N}} = \psi(\widetilde{\mathbf{P}}) - \psi(\mathbf{P})$. Now, this noise, $\widetilde{\mathbf{N}}$, is purely due to finite-sample effects which can be controlled (using concentration arguments given in the inequality ξ_3 below) by driving it down to as small a value as we want by ensuring large enough number of comparisons for each pair. Note that we input $\psi(\widehat{\mathbf{P}}) = \psi(\mathbf{P}) + \widetilde{\mathbf{S}} + \widetilde{\mathbf{N}}$ to Subroutine **??** and obtain $\psi(\overline{\mathbf{P}})$ as the output in Step 3 of Algorithm 1. Hence, using Theorem 2 from (Netrapalli et al., 2014), if $\left\|\widetilde{\mathbf{N}}\right\|_\infty \le \sigma_{\min}(\psi(\mathbf{P}))/100n$, we have,

$$\left\|\psi(\overline{\mathbf{P}}) - \psi(\mathbf{P})\right\|_F \le \epsilon' + 2\mu^2 r\left(7\left\|\widetilde{\mathbf{N}}\right\|_2 + \frac{8n}{r}\left\|\widetilde{\mathbf{N}}\right\|_\infty\right)$$

after $T \ge 10\log(3\mu^2 r\sigma_1/\epsilon')$ iterations associated with Subroutine RPCA. Next, we have, with probability at least $1 - 1/n^3$,

$$\begin{aligned}\left\|\psi(\overline{\mathbf{P}}) - \psi(\mathbf{P})\right\|_F &\le \epsilon' + 2\mu^2 r\left(7\left\|\widetilde{\mathbf{N}}\right\|_2 + \frac{8n}{r}\left\|\widetilde{\mathbf{N}}\right\|_\infty\right)\\ &\overset{\xi_1}{\le} \epsilon' + 32\mu^2 n\left\|\widetilde{\mathbf{N}}\right\|_2 \overset{\xi_2}{\le} \epsilon' + 32\mu^2 n\tau\\ &\overset{\xi_3}{\le} n\sqrt{\frac{\epsilon}{1+\gamma}}\frac{\Delta}{2}\end{aligned}$$

where ξ_1 follows by using $r \le n$ and $\left\|\widetilde{\mathbf{N}}\right\|_\infty \le \left\|\widetilde{\mathbf{N}}\right\|_2$, ξ_2 follows by substituting for $\widetilde{\mathbf{N}}$ from Lemma 2 with $K \ge \frac{L^2n^2\log^2 n}{\tau^2}$, and ξ_3 is obtained using $\epsilon' = n\sqrt{\frac{\epsilon}{1+\gamma}}\frac{\Delta}{4}$, $\tau = \min\left(\sigma_{\min}(\psi(\mathbf{P}))/100, \sqrt{\frac{\epsilon}{1+\gamma}}\frac{\Delta}{128\mu^2}\right)$. Then using similar arguments as proof of Theorem 13 in (Rajkumar and Agarwal, 2016), we obtain our result. □

Lemma 2 (Concentration of sampling noise). *Under the conditions of Theorem 1, let each response pair be compared such that the number of comparisons per response pair is $K \ge \frac{L^2n^2\log(n)}{\tau^2}$; with probability at least $1 - 1/n^3$, $\left\|\widetilde{\mathbf{N}}\right\|_2 \le \tau$.*

Proof. Let L be the Lipschitz constant of ψ and set $K \ge \frac{L^2n^2\log(n)}{\tau^2}$. Using the inequality that $\left\|\widetilde{\mathbf{N}}\right\|_2 \le n\left\|\widetilde{\mathbf{N}}\right\|_\infty$,

$$\begin{aligned}\Pr(\|\widetilde{\mathbf{N}}\|_2 \ge \tau) &\le \Pr\left(\left\|\widetilde{\mathbf{N}}\right\|_\infty \ge \frac{\tau}{n}\right)\\ &= \Pr\left(\exists(i,j) : \left|\psi(\widehat{P}_{ij}) - \psi(P_{ij})\right| \ge \frac{\tau}{n}\right)\\ &\le \sum_{i,j}\Pr\left(\left|\psi(\widehat{P}_{ij}) - \psi(P_{ij})\right| \ge \frac{\tau}{n}\right)\\ &\le \sum_{i,j}\Pr\left(\left|\widehat{P}_{ij} - P_{ij}\right| \ge \frac{\tau}{nL}\right) \le \frac{1}{n^3}\end{aligned}$$

□

Next, for completeness, we recall the following lemma (proved in Theorem 8 and Lemma 14 of (Rajkumar and Agarwal, 2016)) which characterizes the incoherence constant μ of $\mathbf{P} \in (\mathcal{P}_n^{LR(\psi,2)} \cap \mathcal{P}_n^{ST})$ in Assumption 1.

Lemma 3 (Incoherence of BTL and LR models). *We have $\mathbf{P} \in (\mathcal{P}_n^{LR(\psi,2)} \cap \mathcal{P}_n^{ST})$ if and only if $\psi(\mathbf{P}) = \mathbf{u}\mathbf{v}^\top - \mathbf{v}\mathbf{u}^\top$ for $\mathbf{u} \in \mathbb{R}_+^n$ and $\mathbf{v} \in \mathbb{R}^n$ where $\mathbf{u}^\top\mathbf{v} = 0$. Moreover, $\psi(\mathbf{P})$ is μ-incoherent where $\mu = \sqrt{\frac{n}{2}}\left(\frac{u_{\max}^2}{u_{\min}^2} + \frac{v_{\max}^2}{v_{\max}^2}\right)^{1/2}$ where $u_{\min} = \min_i|u_i|$, $u_{\max} = \max_i|u_i|$, $v_{\min} = \min_i|v_i|$ and $v_{\max} = \min_i|v_i|$. We also have $\mathcal{P}_n^{BTL} \subset (\mathcal{P}_n^{LR(\psi,2)} \cap \mathcal{P}_n^{ST})$ since we may set $\mathbf{u} = \mathbf{1}$ where $\mathbf{1}$ is the all-ones vector and $\mathbf{v} = \mathbf{w}$ where $\mathbf{w}$ is the BTL parameter vector. In this case, we may rewrite $\mu = \sqrt{\frac{n}{2}}\left(1 + \frac{(w_{\max}-\overline{w})^2}{(w_{\min}-\overline{w})^2}\right)$ where $\overline{w} = \frac{1}{n}\sum_{i=1}^n w_i$.*

The following corollary makes precise our claim that up to $O(n^2)$ response pairs may be subject to adversarial corruption, but our RORATRON algorithm still recovers a good ranking.

Corollary 1 (Recovery result for BTL model). *Consider $\mathbf{P} \in \mathcal{P}_n^{BTL}$. Using Assumption 1, let the adversarial matrix be $\mathbf{S} \in [0,1]^{n\times n}$ satisfying $\|\mathbf{S}\|_0 \le n/1024\mu^2$ where μ is characterized as in Lemma 3. Then, with probability $1 - 1/n^3$, the output of Algorithm 1 with input $\widehat{\mathbf{P}}$ computed using $\aleph = \{(i,j,\{y_{ij}^k\})\}$ satisfies and $r = 2$, $\mathrm{dist}(\widehat{\sigma}, \mathbf{P}) \le \epsilon$.*

Algorithm 2 CURATRON: Complete Robust Preference Data for Rigorous Alignment

Input: Comparison dataset $\aleph = \{(i, j, \{y_{ij}^k\})\}$, true rank r.
Output: Ranking of n responses, $\widehat{\sigma} \in \mathcal{S}_n$.
1: Estimate entries of $\widehat{\mathbf{P}}$ for $i \le j$ as:

$$\widehat{P}_{ij} = \begin{cases} \frac{1}{K}\sum_{k=1}^{K} y_{ij}^k & \text{if } i < j \text{ and } (i,j) \in \Omega \\ 1/2 & \text{if } i = j \text{ and } (i,j) \in \Omega \\ 1/2 & \text{if } (i,j) \notin \Omega \end{cases}$$

2: Set $\widehat{P}_{ij} = 1 - \widehat{P}_{ji}$ for all $i > j$.
3: Set $\mathbf{R} \leftarrow \text{OptSpace}(\psi(\widehat{\mathbf{P}})_\Omega)$.
4: Use a robust PCA procedure: $\psi(\overline{\mathbf{P}}) \leftarrow \text{RPCA}(\mathbf{R})$.
5: Using a pairwise ranking procedure after taking the inverse transform: $\widehat{\sigma} \leftarrow \text{PR}(\overline{\mathbf{P}})$.
6: **return** $\widehat{\sigma}$.

5 Partially Observered Adversarial Setting

In this section, we consider the partially observed and adversarially corrupted comparison results setting. Both factors can be modeled in a unified manner by setting the corresponding missing entries of the preference matrix to zero (or a specific constant to account for numerical stability). We present our robust ranking algorithm for this setting in Algorithm 2 – this essentially involves using the 'OptSpace' matrix completion algorithm of (Keshavan et al., 2010) followed by using the robust PCA algorithm of (Netrapalli et al., 2014) as sub-routines. We now derive the recovery guarantees as follows.

Theorem 2 (Provably good estimation of ranking in BTL model in the presence of adversarial noise as well as missing data). *Consider a similar notation as in Theorem 1 but let* $\mathbf{P} \in \mathcal{P}_n^{BTL}$. *Let* $\Omega \subseteq [n] \times [n]$ *be a set of compared response pairs. Assume* Ω *is drawn uniformly from all subsets of* $[n] \times [n]$ *of size* $|\Omega|$ *such that* $|\Omega| \ge C'' n \log(n)$ *and let the sparse noise satisfy* $\|\mathbf{S}\|_\infty \le \Delta_w \frac{\log(n)}{C_\Delta n}$ *where* $\Delta_w := \min_{i,j} |w_i - w_j|$. *Let the number of comparisons per pair be* $K \ge cn^4/\Delta_w$. *Then with probability at least* $1 - 2/n^3$, *Algorithm 2 returns a ranking that satisfies* $\text{dist}(\widehat{\sigma}, \mathbf{P}) \le \epsilon$.

Remark 3 (Robust Estimation of BTL Model in the Partially Observed Case). *For the BTL model, Theorem 2 says* $O(n \log n)$ *pairs suffice to estimate the BTL model, which matches bounds from (Rajkumar and Agarwal, 2016). Further, even in this incomplete comparison data case, we are able to tolerate uniformly random additive sparse noise with its maximum absolute entry scaling as the order of the BTL 'score-gap' divided by the number of responses up to logarithmic factors, ie,* $\widetilde{O}(\Delta_w/n)$.

Proof. From Lemma 3, we have $\psi(\mathbf{P}) = \mathbf{1}\mathbf{w}^\top - \mathbf{w}\mathbf{1}^\top$ for the BTL model where ψ is the logit function. Clearly, in this case, $\psi(\mathbf{P})$ is a real skew-symmetric matrix of rank $r = 2$. Since it is skew-symmetric, its eigenvalues, which are the roots of its charateristic polynomial, are of the form $\pm\lambda i$ for some $\lambda \in \mathbb{R}$ and $i = \sqrt{-1}$, and hence, $\sigma_{\min}(\psi(\mathbf{P})) = \sigma_{\max}(\psi(\mathbf{P}))$, ie, the condition number of $\psi(\mathbf{P})$, $\kappa = 1$. Now, we recall the spectral-lower bound from Corollary 2 of (Horne, 1997),

$$\sigma_{\min}(\psi(\mathbf{P})) \ge \frac{\|\psi(\mathbf{P})\|_F}{\sqrt{r(r-1)}} \ge \sqrt{\frac{n(n-1)}{2}}\Delta_w \tag{1}$$

where $\Delta_w = \min_{i,j} |w_i - w_j|$.

Let $\Omega \subseteq [n] \times [n]$ be a subset of all the response pairs with comparison results among which some might be corrupted by sparse noise, ie, $\psi(\widehat{\mathbf{P}}_\Omega) = \psi(\mathbf{P}_\Omega) + \widetilde{\mathbf{S}}_\Omega + \widetilde{\mathbf{N}}_\Omega$. Let $\mathbf{T} := \widetilde{\mathbf{S}}_\Omega + \widetilde{\mathbf{N}}_\Omega$. From Theorem 1.2 of (Keshavan et al., 2010), we have $\frac{1}{n}\left\|\psi(\widehat{\mathbf{P}}) - \psi(\mathbf{P})\right\|_F = \frac{1}{n}\|\mathbf{T} + \mathbf{M}\|_F \le C\kappa^2 \frac{n\sqrt{r}}{|\Omega|}\|\mathbf{T}\|_2$ where $\mathbf{M}$ is the noise matrix after obtaining the completed matrix $\psi(\widehat{\mathbf{P}})$ from $\psi(\widehat{\mathbf{P}}_\Omega)$ using OptSpace. Using triangle inequality and noting that $|\Omega| \ge C'' n \log(n)$, the noise may be bounded as

$$\begin{aligned} \|_{\widetilde{\mathbf{N}}_\Omega + \mathbf{M}}\|_\infty \le \left\|\widetilde{\mathbf{N}}_\Omega + \mathbf{M}\right\|_F &\le \|\mathbf{T}\|_2 \frac{\sqrt{2}Cn^2}{|\Omega|} + \left\|\widetilde{\mathbf{S}}_\Omega\right\|_F \\ &\overset{\zeta_1}{\le} C' \frac{n}{\log(n)} \left\|\widetilde{\mathbf{S}}_\Omega\right\|_2 \end{aligned} \tag{2}$$

where C, C' and C'' are constants and ζ_1 is obtained by using the triangle inequality that $\|\mathbf{T}\|_2 \le \left\|\widetilde{\mathbf{S}}_\Omega\right\|_2 + \left\|\widetilde{\mathbf{N}}_\Omega\right\|_2$, followed by setting $K \ge cn^4/\Delta_w$ for constant c and finally using $\left\|\widetilde{\mathbf{S}}_\Omega\right\|_F \le \sqrt{n}\left\|\widetilde{\mathbf{S}}_\Omega\right\|_2$. Then, combining Equations 2 and 1, we have if

$$\begin{aligned} \tfrac{\log(n)}{C_\Delta n}\Delta_w \ge \left\|\widetilde{\mathbf{S}}_\Omega\right\|_2 &= \left\|\psi(\widehat{\mathbf{P}}) - \psi(\widetilde{\mathbf{P}})\right\|_2 \\ &\ge \left\|\psi(\widehat{\mathbf{P}}) - \psi(\widetilde{\mathbf{P}})\right\|_\infty \ge L\left\|\widehat{\mathbf{P}} - \widetilde{\mathbf{P}}\right\|_\infty \ge \|\mathbf{S}\|_\infty \end{aligned}$$

where C_Δ is a global constant and using Lemma 2, then we have the guarantee (along similar lines as that of Theorem 1 that Algorithm 2 returns an estimated permutation which satisfies $\text{dist}(\widehat{\sigma}, \mathbf{P}) \le \epsilon$. □

6 Experiments

We now perform simulations in order to understand the performance of our robust ranking approach in practice in both general and LLM preference dataset settings.

6.1 Performance of Robust Ranking in LLM Preference Dataset

In this illustrative experiment, from the MT-Bench dataset (Zheng et al., 2023), we collect the data of the first prompt "Compose an engaging travel blog post about a recent trip to Hawaii, highlighting cultural experiences and must-see attractions" and its six responses from `GPT-3.5`, `GPT-4` (OpenAI et al., 2023), `Claude-v1` (Anthropic, 2023), `Vicuna-13B` (Chiang et al., 2023), `Alpaca-13B` (Taori et al., 2023), and `LLaMA-13B` (Touvron et al., 2023a). Additionally, we generated nine responses to the same prompt using `Llama-2-70B-chat-hf` (Touvron et al., 2023b), `Falcon-180B-chat` (Almazrouei et al., 2023), `Openchat-3.5` (Wang et al., 2023), `Mixtral-8x7B-Instruct-v0.1` (Jiang et al., 2024), `Mistral-7B-Instruct-v0.2` (Jiang et al., 2023), `Gemini-pro` (Gemini et al., 2023), `Dolphin-2.2.1-mistral-7B` (Hartford, 2023), `Solar-10.7B-instruct-v1.0` (Kim et al., 2023), `Yi-34B-chat` (01.ai, 2023) from Hugging Face's HuggingChat (Hugging Face, 2023) and LMSYS's Chatbot Arena (Zheng et al., 2023). So we have $n = 15$ responses.

Next, we rank the responses using OpenAI's GPT-4 Turbo `GPT-4-1106-preview` (OpenAI et al., 2023). This ranking helps us create the BTL parameter vector $\mathbf{w}$. We then sort this vector descendingly for visually accessible when building the corresponding preference matrix $\mathbf{P} \in \mathbb{R}^{n \times n}$. With $\binom{n}{2}$ comparisons in $\mathbf{P}$, we randomly remove entries based on a specified deletion probability parameter, dp, to simulate unobserved comparisons. We then create an adversarial skew-symmetric sparse matrix, $\mathbf{S}$, using the given matrix $\mathbf{P}$ and an adversarial corruption probability parameter ap. When corruption is applied, it involves randomly selecting a value from $U(-5, 5)$ and then adding to the $\mathbf{P}$ to give $\mathbf{P}^c$, which then becomes the input of our algorithm. It's important to note that $\mathbf{P}$ is a skew-symmetric matrix, any corruption must be applied to both ij and ji values.

Our experiment results visualized in Figure 2 show that $dp = 10\%$ and $ap = 10\%$ can significantly affect the ranking of different models and the rank of the matrix when performing logit link transformation. The ranking can get altered quite badly when compared to the original matrix. Also, the logit link transformation of the corrupted matrix is high-rank, which indicates that there are noises in the matrix. By using `CURATRON` to impute the missing comparisons and filter out the noisy sparse matrix, we successfully reconstruct the original matrix, which is low-rank when in logit link transformed form. As a result, we obtain the correct ranking. We also obtain noisy comparisons that can be used to identify responders with malicious intent and prevent them from continuing to alter results.

We now examine how our algorithm performs across different levels of unobserved and adversarially corrupted comparisons. In the plots shown in Figure 3, we compare the performance of our approach by varying two parameters, dp and ap. We use normalized Frobenius error, correlation, and ranking distance as evaluation metrics. Our results are averaged over 5 runs. When there is no adversarial noise, we can recover the original $\mathbf{P}$ with no normalized Frobenius error and perfect correlation and ranking, even if 50% of the comparison data was missing. This suggests that we may not need to collect all comparisons from humans to obtain the entire data. We observe that, with $n = 15$, we only need to obtain about $50 - 55\%$ of the 105 comparisons and fill in the rest with our algorithm to achieve a strict 0% NFE, perfect correlation, and ranking. On the other hand, when missing data is absent, our algorithm performs well with NFE of approximately 6%, even when 35% of the comparison data is adversarially corrupted. When both adversarial noise and missing data are present, we can achieve a low NFE of around 4% when both 15% of the comparison data is missing and 15% of adversarially corrupted comparisons (30% in total) affect $\mathbf{P}$.

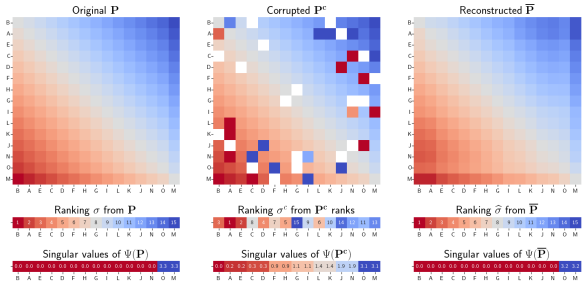

Figure 2: Left: Original matrix. Middle: corrupted matrix. Right: reconstructed matrix. The corrupted matrix has 10% adversarial corruptions and 10% of unobserved comparisons. We use our `CURATRON` algorithm to successfully recover the original matrix.

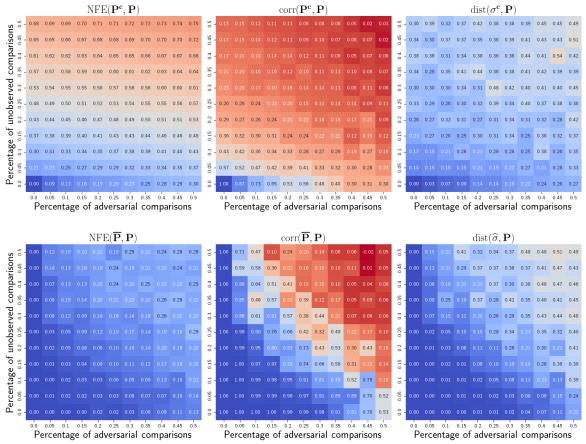

Figure 3: Average over 5 runs of reconstruction error, correlation, and distance between reconstructed ranking and original ranking for different percentages of unobserved and adversarial comparisons.

7 Conclusion

Our study examines how missing information and distorted feedback can impact LLMs, potentially compromising their performance in terms of alignment with human values. We have proposed a robust algorithm for provably correct and efficient ranking responses in the BTL, LR, and general binary choice models. This robust ranking data is then input in the PL step. Further, we also handled the partially observed setting, wherein only some response pairs are compared, by integrating matrix completion techniques into our robust learning algorithm. In all cases, we provided statistical and computational guarantees using novel techniques. Through our comprehensive analysis, we hope to contribute to the ongoing discussion on AI safety by helping to create and scale LLMs/AGI models that align with human values and expectations. Some future research directions include tightening the recovery results for partially observed settings under weaker conditions (possibly using noisy-case extensions of (Yi et al., 2016)), exploring other notions of adversarial noise, and understanding the minimax optimal rates for ranking estimators under various noise models. We also plan to study the parametric non-active pairwise ranking setting, studying lower bounds and practical algorithms in the active setting similar to (Heckel et al., 2016). Furthermore, it would be interesting to investigate whether we can extend this approach to solve the entity corruption problem in retrieval models, as shown in (Naresh et al., 2022). Another research direction could be defining an alignment framework that expands DPO to various objective functions based on Rank Centrality (Negahban et al., 2017). Finally, we aim to examine the relationship between robust PL and model capacity, as this can shed light on the trade-offs between model complexity and generalization performance.

References

01.ai. 2023. Yi-34b. https://www.01.ai. Accessed 03-03-2024.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models. *Preprint*, arXiv:2311.16867.

Anthropic. 2023. Introducing Claude. https://www.anthropic.com/news/introducing-claude. Accessed 03-03-2024.

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

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://lmsys.org/blog/2023-03-30-vicuna/. Accessed 03-03-2024.

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

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. GPTs are GPTs: An early look at the labor market impact potential of Large Language Models. *Preprint*, arXiv:2303.10130.

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

Eric Hartford. 2023. Dolphin-2.2.1-mistral-7b. https://huggingface.co/cognitivecomputations/dolphin-2.2.1-mistral-7b. Accessed 03-03-2024.

Reinhard Heckel, Nihar B Shah, Kannan Ramchandran, and Martin J Wainwright. 2016. Active ranking from pairwise comparisons and when parametric assumptions don't help. *arXiv preprint arXiv:1606.08842*.

Bill G Horne. 1997. Lower bounds for the spectral radius of a matrix. *Linear algebra and its applications*, 263:261–273.

Daniel Hsu, Sham M Kakade, and Tong Zhang. 2011. Robust matrix decomposition with sparse corruptions. *IEEE Transactions on Information Theory*, 57(11):7221–7234.

Hugging Face. 2023. Huggingchat. https://huggingface.co/chat. Accessed 03-03-2024.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. 2010. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11(Jul):2057–2078.

Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. 2023. Solar 10.7b: Scaling large language models with simple yet effective depth up-scaling. *Preprint*, arXiv:2312.15166.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Niranjan Uma Naresh, Ziyan Jiang, Ankit, Sungjin Lee, Jie Hao, Xing Fan, and Chenlei Guo. 2022. PENTATRON: Personalized context-aware transformer for retrieval-based conversational understanding. *Preprint*, arXiv:2210.12308.

Sahand Negahban, Sewoong Oh, and Devavrat Shah. 2017. Rank centrality: Ranking from pairwise comparisons. *Operations Research*, 65(1):266–287.

Praneeth Netrapalli, UN Niranjan, Sujay Sanghavi, Animashree Anandkumar, and Prateek Jain. 2014. Non-convex robust PCA. In *Advances in Neural Information Processing Systems*, pages 1107–1115.

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

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Arun Rajkumar and Shivani Agarwal. 2016. When can we rank well from comparisons of $o(n \log(n))$ non-actively chosen pairs? In *29th Annual Conference on Learning Theory*, pages 1376–1401.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca. Accessed 03-03-2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of LM alignment. *Preprint*, arXiv:2310.16944.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023. Openchat: Advancing open-source language models with mixed-quality data. *Preprint*, arXiv:2309.11235.

Xinyang Yi, Dohyung Park, Yudong Chen, and Constantine Caramanis. 2016. Fast algorithms for Robust PCA via Gradient Descent. *arXiv preprint arXiv:1605.07784*.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J. Liu. 2023. SLiC-HF: Sequence likelihood calibration with human feedback. *Preprint*, arXiv:2305.10425.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Preprint*, arXiv:2306.05685.