Prediction-Powered Ranking of Large Language Models

Ivi Chatzi
ichatzi@mpi-sws.org
Max Planck Institute for Software Systems
Germany

Eleni Straitouri
estraitouri@mpi-sws.org
Max Planck Institute for Software Systems
Germany

Suhas Thejaswi
thejaswi@mpi-sws.org
Max Planck Institute for Software Systems
Germany

Manuel Gomez Rodriguez
manuelgr@mpi-sws.org
Max Planck Institute for Software Systems
Germany

ABSTRACT

Large language models are often ranked according to their level of alignment with human preferences—a model is better than other models if its outputs are more frequently preferred by humans. One of the most popular ways to elicit human preferences utilizes pairwise comparisons between the outputs provided by different models to the same inputs. However, since gathering pairwise comparisons by humans is costly and time-consuming, it has become a very common practice to gather pairwise comparisons by a strong large language model—a model strongly aligned with human preferences. Surprisingly, practitioners cannot currently measure the uncertainty that any mismatch between human and model preferences may introduce in the constructed rankings. In this work, we develop a statistical framework to bridge this gap. Given a small set of pairwise comparisons by humans and a large set of pairwise comparisons by a model, our framework provides a rank-set—a set of possible ranking positions—for each of the models under comparison. Moreover, it guarantees that, with a probability greater than or equal to a user-specified value, the rank-sets cover the true ranking consistent with (the distribution of) human pairwise preferences. Our framework is computationally efficient, easy to use, and does not make any assumption about the distribution of human preferences nor about the degree of alignment between the pairwise comparisons by the humans and the strong large language model.

1 INTRODUCTION

During the last years, large language models (LLMs) have shown a remarkable ability to generate and understand general-purpose language [7]. As a result, there has been an increasing excitement in their potential to help humans solve a variety of open-ended, complex tasks across many application domains such as coding [32], health [18] and scientific discovery [37], to name a few. However, evaluating and comparing the performance of different LLMs has become very challenging [8]. The main reasons is that, in contrast to traditional machine learning models, LLMs can solve a large number of different tasks and, in many of these tasks, there is not a unique, structured solution. As a consequence, there has been a paradigm shift towards evaluating their performance according to their level of alignment with human preferences—a model is better than other models if its outputs are more frequently preferred by humans [19, 28, 34, 45, 47].

One of the most popular paradigms to rank a set of LLMs according to their level of alignment with human preferences utilizes pairwise comparisons [6, 8, 23, 24, 28, 39, 41, 52]. Under this paradigm, each pairwise comparison comprises the outputs of two different models picked uniformly at random to an input sampled from a given distribution of inputs. Moreover, the pairwise comparisons are used to rank the models according to the (empirical) probability that they are preferred over other models. While it is widely agreed that, given a sufficiently large set of pairwise comparisons, higher (lower) ranking under this paradigm corresponds to better (worse) human alignment, there have also been increasing concerns that this paradigm is too costly and time-consuming to be practical, especially given the pace at which models are updated and new models are developed.

To lower the cost and increase the efficiency of ranking from pairwise comparisons, it has become a common practice to ask a strong LLM—a model known to strongly align with human preferences—to perform pairwise comparisons [10, 11, 14, 20, 35, 42, 44, 46, 52]. The rationale is that, if a model strongly aligns with human preferences, then, the distributions of pairwise comparisons by the model and by the human should in principle match [10, 42, 43]. Worryingly, there are multiple lines of evidence showing that the rankings constructed using pairwise comparisons made by a strong LLM are sometimes different to those constructed using pairwise comparisons by humans [6, 12, 13, 24, 39, 52], questioning the rationale above. In this work, we introduce a statistical framework to measure the uncertainty in the rankings constructed using pairwise comparisons made by a model, which may be introduced by a mismatch between human and model preferences or by the fact that we use a finite number of pairwise comparisons.

Our contributions. Our framework measures uncertainty using rank-sets—sets of possible ranking positions that each model can take. If the rank-set of a model is large (small), it means that there is high (low) uncertainty in the ranking position of the model. To construct the rank-sets, our framework first leverages a small set of pairwise comparisons by humans and a large set of pairwise comparisons by a strong LLM to create a confidence ellipsoid. By using prediction-powered inference [2, 3, 53], this confidence ellipsoid is guaranteed to contain the vector of (true) probabilities that each model is preferred over others by humans with a user-specified coverage probability $1 - \alpha$. Then, it uses the distance between this

1st HEAL Workshop at CHI Conference on Human Factors in Computing Systems, May 12 2024, Honolulu, HI, USA

ellipsoid and the hyperplanes under which pairs of models have the same probability values of being preferred over others to efficiently construct the rank-sets. Importantly, we can show that, with probability greater than or equal to $1-\alpha$, the constructed rank-sets are guaranteed to cover the ranking consistent with the (true) probability that each model is preferred over others by humans.

We will provide an open-source implementation of our statistical framework as well as case studies in the accompanying GitHub repository.[1]

Further related work. Our work builds upon recent work on prediction-powered inference, ranking under uncertainty, and ranking of LLMs.

Prediction-powered inference [2, 3, 53] is a recently introduced statistical framework to obtain valid p-values and confidence intervals about a population-level quantity such as the mean outcome or a regression coefficient using a small labeled dataset and a large unlabeled dataset, whose labels are imputed using a black-box machine learning model. However, our work is the first to use prediction-powered inference (as a subroutine) to construct rank-sets with coverage guarantees. In this context, it is worth acknowledging that a very recent work by Saad-Falcon et al. [38] has used prediction-powered inference to construct a (single) ranking, rather than rank-sets. However, the ranking constructed by Saad-Falcon et al. does not enjoy coverage guarantees with respect to the true ranking consistent with (the distribution of) the human preferences.

The vast majority of the literature on ranking under uncertainty has focused on confidence intervals for individual ranking positions [16, 17, 22, 30, 49–51]. Only recently, a paucity of work has focused on joint measures of uncertainty for rankings [1, 21, 33, 36]. Similarly as in our work, this line of work also seeks to construct rank-sets with coverage guarantees. However, in contrast to our work, it estimates the quality metric (in our work, the probability that an LLM is preferred over others) and the confidence intervals separately for each of the items (in our work, LLMs) using independent samples. As a consequence, it needs to perform multiple comparison correction to create the rank-sets.

In addition to the related work on ranking of LLMs discussed previously, it is worth highlighting that, in recent years, there has been a flurry of work on ranking LLMs using benchmark datasets with manually hand-crafted inputs and ground-truth outputs [4, 9, 19, 26, 29, 31, 40, 48]. However, it has become increasingly clear that oftentimes rankings derived from benchmark datasets do not correlate well with rankings derived from human preferences—an improved ranking position in the former does not lead to an improved ranking position in the latter [11, 23, 24, 52].

2 LLM RANKING UNDER UNCERTAINTY

Let $\mathcal{M}$ be a set of k large language models (LLMs) and $P(Q)$ be a distribution of inputs on a discrete set of inputs $\mathcal{Q}$. Moreover, assume that, for each input $q \sim P(Q)$,[2] each model $m \in \mathcal{M}$ may provide an output $r \sim P_m(R \mid Q = q)$ from a discrete set of outputs $\mathcal{R}$. Further, given two outputs $r, r' \in \mathcal{R}$ from two different models, the (binary) variable $w \sim P(W \mid Q = q, R = r, R' = r')$ indicates whether a human prefers r over r' ($w = 1$) or viceversa ($w = 0$). In what follows, we use $m(r)$ and $m(r')$ to denote the models that provide outputs r and r' respectively, and without loss of generality, we assume that the output r is shown first. Then, our goal is to rank all models according to the (empirical) probability θ_m that their outputs are preferred over the outputs of any other model picked uniformly at random.

To this end, we start by writing the probability θ_m as an expectation over the distribution of inputs, outputs and pairwise preferences:

$$\theta_m = \frac{1}{k-1} \sum_{\tilde{m} \in \mathcal{M} \setminus \{m\}} \mathbb{E}_Q \left[\frac{1}{2} \mathbb{E}_{R \sim P_m, R' \sim P_{\tilde{m}}} \left[\mathbb{E}_W [W \mid Q, R, R'] \right] + \frac{1}{2} \mathbb{E}_{R \sim P_{\tilde{m}}, R' \sim P_m} \left[1 - \mathbb{E}_W [W \mid Q, R, R'] \right] \right], \tag{1}$$

where note that the order of the pairs of outputs is picked at random. Next, following previous work [1, 33], we formally characterize the ranking position of each model $m \in \mathcal{M}$ in the ranking induced by the probabilities θ_m using a rank-set $[l(m), u(m)]$, where

$$\begin{aligned} l(m) &= 1 + \sum_{\tilde{m} \in \mathcal{M} \setminus \{m\}} \mathbf{1}\{\theta_m < \theta_{\tilde{m}}\} \\ u(m) &= k - \sum_{\tilde{m} \in \mathcal{M} \setminus \{m\}} \mathbf{1}\{\theta_m > \theta_{\tilde{m}}\}, \end{aligned} \tag{2}$$

are the lower and upper ranking position respectively and smaller ranking position indicates better alignment with human preferences. Here, note that it often holds that $\theta_m \neq \theta_{\tilde{m}}$ for all $\tilde{m} \in \mathcal{M} \setminus \{m\}$ and then the rank-set reduces to a singleton, *i.e.*, $l(m) = u(m)$.

In general, we cannot directly construct the rank-sets as defined above because the probabilities θ_m are unknown. Consequently, the typical strategy reduces to first gathering pairwise comparisons by humans to compute unbiased estimates $\hat{\theta}_m$ of the above probabilities using sample averages and then construct estimates $[\hat{l}(m), \hat{u}(m)]]$ of the rank-sets using Eq. 2 with $\hat{\theta}_m$ rather than θ_m. Under this strategy, if the amount of pairwise comparisons we gather is sufficiently large, the estimates of the rank-sets will closely match the true rank-sets. However, since gathering pairwise comparisons from humans is costly and time-consuming, it has become a very common practice to gather pairwise comparisons $\hat{w}$ by a strong LLM, rather than pairwise comparisons w by humans [5, 14, 15, 20, 23–25, 27, 46, 52], and then utilize them to compute unbiased estimates of the probabilities $\breve{\theta}_m$ that the outputs provided by each model is preferred over others by the strong LLM, which can be written in terms of expectations as follows:

$$\breve{\theta}_m = \frac{1}{k-1} \sum_{\tilde{m} \in \mathcal{M} \setminus \{m\}} \mathbb{E}_Q \left[\frac{1}{2} \mathbb{E}_{R \sim P_m, R' \sim P_{\tilde{m}}} \left[\mathbb{E}_{\hat{W}} [\hat{W} \mid Q, R, R'] \right] + \frac{1}{2} \mathbb{E}_{R \sim P_{\tilde{m}}, R' \sim P_m} \left[1 - \mathbb{E}_{\hat{W}} [\hat{W} \mid Q, R, R'] \right] \right], \tag{3}$$

In general, one can only draw valid conclusions about θ using (an estimate of) $\breve{\theta}$ if the distribution of the pairwise comparisons by the strong LLM $P(\hat{W} \mid Q = q, R = r, R' = r')$ closely matches

[1] https://github.com/Networks-Learning/prediction-powered-ranking
[2] We denote random variables with capital letters and realizations of random variables with lower case letters.

Algorithm 1 It constructs C_α using prediction-powered inference

Input: k, $\mathcal{D}_N$, $\mathcal{D}_n$, α
Output: $\hat{\theta}$, C_α

$\hat{\boldsymbol{w}}_N, \boldsymbol{M}_N, \boldsymbol{M}'_N \leftarrow \text{SUMMARIZE}(\mathcal{D}_N, k)$
$\boldsymbol{w}_n, \hat{\boldsymbol{w}}_n, \boldsymbol{M}_n, \boldsymbol{M}'_n \leftarrow \text{SUMMARIZE}(\mathcal{D}_n, k)$
$\boldsymbol{a} \leftarrow \left(\mathbf{1}_k\left(\left(\boldsymbol{M}_N + \boldsymbol{M}'_N\right)\mathbf{1}_N\right)^\top \odot \mathbb{I}_k\right)^{-1}\left(\boldsymbol{M}_N \cdot \hat{\boldsymbol{w}}_N + \boldsymbol{M}'_N(\mathbf{1}_N - \hat{\boldsymbol{w}}_N)\right)$
$\boldsymbol{b} \leftarrow \left(\mathbf{1}_k\left(\left(\boldsymbol{M}_n + \boldsymbol{M}'_n\right)\mathbf{1}_n\right)^\top \odot \mathbb{I}_k\right)^{-1}\left(\boldsymbol{M}_n(\hat{\boldsymbol{w}}_n - \boldsymbol{w}_n) + \boldsymbol{M}'_n(\boldsymbol{w}_n - \hat{\boldsymbol{w}}_n)\right)$
$\hat{\boldsymbol{\theta}} \leftarrow \boldsymbol{a} - \boldsymbol{b}$
$\boldsymbol{A} \leftarrow \left(\mathbf{1}_k(\hat{\boldsymbol{w}}_N - \boldsymbol{M}_N^\top \boldsymbol{a})^\top\right) \odot \boldsymbol{M}_N + \left(\mathbf{1}_k(\mathbf{1}_N - \hat{\boldsymbol{w}}_N - \boldsymbol{M}'^\top_N \boldsymbol{a})^\top\right) \odot \boldsymbol{M}'_N$
$\boldsymbol{B} \leftarrow \left(\mathbf{1}_k(\hat{\boldsymbol{w}}_n - \boldsymbol{w}_n - \boldsymbol{M}_n^\top \boldsymbol{b})^\top\right) \odot \boldsymbol{M}_n + \left(\mathbf{1}_k(\boldsymbol{w}_n - \hat{\boldsymbol{w}}_n - \boldsymbol{M}'^\top_n \boldsymbol{b})^\top\right) \odot \boldsymbol{M}'_n$
$\widehat{\Sigma} \leftarrow \frac{1}{N^2}\boldsymbol{A}\boldsymbol{A}^\top + \frac{1}{n^2}\boldsymbol{B}\boldsymbol{B}^\top$
$C_\alpha \leftarrow \left\{\boldsymbol{x} \in \mathbb{R}^k \mid \left(\boldsymbol{x} - \hat{\boldsymbol{\theta}}\right)^\top \left(\frac{\widehat{\Sigma}^{-1}}{\chi^2_{k,1-\alpha}}\right)\left(\boldsymbol{x} - \hat{\boldsymbol{\theta}}\right) \leq 1\right\}$
return $\hat{\boldsymbol{\theta}}, \widehat{\Sigma}, C_\alpha$

the distribution of pairwise comparisons by the humans $P(W \mid Q = q, R = r, R' = r')$ for any $q \in \mathcal{Q}$ and $r, r' \in \mathcal{R}$. However, there are multiple lines of evidence showing that there is a mismatch between the distributions, questioning the validity of the conclusions drawn by a myriad of papers. In what follows, we introduce a statistical framework that, by complementing a (large) set of $N + n$ pairwise comparisons $\hat{w}$ by a strong large language model with a small set of n pairwise comparisons w by humans, is able to construct estimates $[\hat{l}(m), \hat{u}(m)]$ of the rank-sets with provable coverage guarantees. More formally, given a user-specified value $\alpha \in (0, 1)$, the estimates of the rank-sets satisfy that

$$\lim_n \mathbb{P}\left(\bigcap_{m \in \mathcal{M}} [l(m), u(m)] \subseteq [\hat{l}(m), \hat{u}(m)]\right) \geq 1 - \alpha. \quad (4)$$

To this end, we will first use prediction-powered inference [2, 3] to construct a confidence ellipsoid that, with probability $1 - \alpha$, is guaranteed to contain the (column) vector of (true) probabilities $\boldsymbol{\theta} = (\theta_m)_{m \in \mathcal{M}}$. Then, we will use the distance between this ellipsoid and the hyperplanes under which each pair of models $m, \tilde{m} \in \mathcal{M}$ have the same probability values of being preferred over others to efficiently construct the estimates $[\hat{l}(m), \hat{u}(m)]$ of the rank-sets.

3 CONSTRUCTING CONFIDENCE REGIONS WITH PREDICTION-POWERED INFERENCE

Let the set $\mathcal{D}_N = \{(q_i, r_i, r'_i, m(r_i), m(r'_i), \hat{w}_i)\}_{i=1}^N$ comprise pairwise comparisons by a strong LLM to N inputs and the set $\mathcal{D}_n = \{(q_i, r_i, r'_i, m(r_i), m(r'_i), w_i, \hat{w}_i)\}_{i=1}^n$ comprise pairwise comparisons by the same strong LLM and by humans to n inputs, with $n \ll N$. In what follows, for each pairwise comparison, we will refer to the models $m(r)$ and $m(r')$ that provided the first and second output using one-hot (column) vectors $\boldsymbol{m}$ and $\boldsymbol{m}'$, respectively. Moreover, to summarize the pairwise comparisons[3] in $\mathcal{D}_N$ and $\mathcal{D}_n$, we will stack the one-hot vectors $\boldsymbol{m}$ and $\boldsymbol{m}'$ into four matrices, $\boldsymbol{M}_N$ and $\boldsymbol{M}'_N$ for $\mathcal{D}_N$ and $\boldsymbol{M}_n$ and $\boldsymbol{M}'_n$ for $\mathcal{D}_n$, where each column corresponds to a one-hot vector, and the indicators w and $\hat{w}$ into three (column) vectors, $\hat{\boldsymbol{w}}_N$ for $\mathcal{D}_N$ and $\hat{\boldsymbol{w}}_n$ and $\boldsymbol{w}_n$ for $\mathcal{D}_n$.

[3]We assume that each model $m \in \mathcal{M}$ participates in at least one pairwise comparison in both $\mathcal{D}_N$ and $\mathcal{D}_n$.

Then, building upon the recent framework of prediction-powered inference [2], we compute the an unbiased estimate $\hat{\boldsymbol{\theta}}$ of the vector of (true) probabilities $\boldsymbol{\theta}$:

$$\hat{\boldsymbol{\theta}} = \underbrace{\left(\mathbf{1}_k\left(\left(\boldsymbol{M}_N + \boldsymbol{M}'_N\right)\mathbf{1}_N\right)^\top \odot \mathbb{I}_k\right)^{-1}\left(\boldsymbol{M}_N \cdot \hat{\boldsymbol{w}}_N + \boldsymbol{M}'_N(\mathbf{1}_N - \hat{\boldsymbol{w}}_N)\right)}_{\boldsymbol{a}}$$
$$- \underbrace{\left(\mathbf{1}_k\left(\left(\boldsymbol{M}_n + \boldsymbol{M}'_n\right)\mathbf{1}_n\right)^\top \odot \mathbb{I}_k\right)^{-1}\left(\boldsymbol{M}_n(\hat{\boldsymbol{w}}_n - \boldsymbol{w}_n) + \boldsymbol{M}'_n(\boldsymbol{w}_n - \hat{\boldsymbol{w}}_n)\right)}_{\boldsymbol{b}}, \quad (5)$$

where $\mathbf{1}_d$ denotes a d-dimensional column vector where each dimension has value 1 and $\mathbb{I}_k$ denotes a k-dimensional identity matrix. Here, note that the first term $\boldsymbol{a}$ utilizes the pairwise comparisons by the strong LLM from $\mathcal{D}_N$ to compute a unbiased estimate of the vector of probabilities $\check{\boldsymbol{\theta}}$ defined in Eq. 3 using sample averages, and the second term $\boldsymbol{b}$ utilizes the pairwise comparisons by the strong LLM and by humans from $\mathcal{D}_n$ to compute an unbiased estimate of the difference of probabilities $\boldsymbol{\theta} - \check{\boldsymbol{\theta}}$ defined in Eqs. 1 and 3, also using sample averages.

Further, as shown in Angelopoulos et al. [3], the difference of probabilities $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}$ converges in distribution to a k-dimensional normal $\mathcal{N}_k(0, \Sigma)$, where $\Sigma = \mathbb{E}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T]$, and thus the confidence region

$$C_\alpha = \left\{\boldsymbol{x} \in \mathbb{R}^k \mid \left(\boldsymbol{x} - \hat{\boldsymbol{\theta}}\right)^\top \left(\frac{\widehat{\Sigma}^{-1}}{\chi^2_{k,1-\alpha}}\right)\left(\boldsymbol{x} - \hat{\boldsymbol{\theta}}\right) \leq 1\right\}, \quad (6)$$

where $\widehat{\Sigma}$ is an empirical estimate of the covariance matrix Σ using pairwise comparisons from $\mathcal{D}_N$ and $\mathcal{D}_n$, *i.e.*,

$$\widehat{\Sigma} = \frac{1}{N^2}\boldsymbol{A}\boldsymbol{A}^\top + \frac{1}{n^2}\boldsymbol{B}\boldsymbol{B}^\top, \quad (7)$$

with

$$\boldsymbol{A} = \left(\mathbf{1}_k(\hat{\boldsymbol{w}}_N - \boldsymbol{M}_N^\top \boldsymbol{a})^\top\right) \odot \boldsymbol{M}_N + \left(\mathbf{1}_k(\mathbf{1}_N - \hat{\boldsymbol{w}}_N - \boldsymbol{M}'^\top_N \boldsymbol{a})^\top\right) \odot \boldsymbol{M}'_N,$$
$$\boldsymbol{B} = \left(\mathbf{1}_k(\hat{\boldsymbol{w}}_n - \boldsymbol{w}_n - \boldsymbol{M}_n^\top \boldsymbol{b})^\top\right) \odot \boldsymbol{M}_n + \left(\mathbf{1}_k(\boldsymbol{w}_n - \hat{\boldsymbol{w}}_n - \boldsymbol{M}'^\top_n \boldsymbol{b})^\top\right) \odot \boldsymbol{M}'_n,$$

and $\chi^2_{k,1-\alpha}$ is the $1-\alpha$ quantile of the χ^2 distribution with k degrees of freedom, satisfies that

$$\lim_n \mathbb{P}(\theta \in C_\alpha) = 1 - \alpha \quad (8)$$

Algorithm 1 summarizes the overall procedure to compute $\hat{\boldsymbol{\theta}}$, $\widehat{\Sigma}$ and C_α.

4 CONSTRUCTING RANK-SETS WITH COVERAGE GUARANTEES

For each pair of models $m, \tilde{m} \in \mathcal{M}$ such that $m \neq \tilde{m}$, we first define a hyperplane $H_{m,\tilde{m}} \subseteq \mathbb{R}^k$ as follows:

$$H_{m,\tilde{m}} = \{\boldsymbol{x} \in \mathbb{R}^k \mid x_m = x_{\tilde{m}}\}. \quad (9)$$

Then, for each of these hyperplanes $H_{m,\tilde{m}}$, we calculate the distance $d(C_\alpha, H_{m,\tilde{m}})$ between $H_{m,\tilde{m}}$ and the confidence region C_α defined

Algorithm 2 It constructs $[\hat{l}(m), \hat{u}(m)]$ for all $m \in \mathcal{M}$

Input: $\mathcal{M}$, $\mathcal{D}_N$, $\mathcal{D}_n$, α
Output: $\{[\hat{l}(m), \hat{u}(m)]\}_{m\in\mathcal{M}}$

```
k ← |M|
θ̂, Σ̂, C_α ← CONFIDENCE-ELLIPSOID(k, D_N, D_n, α)        ▷ Algorithm 1
for m ∈ M do
    l̂(m) ← 1
    û(m) ← k
    for m̃ ∈ M \ {m} do
        d ← |θ̂_m − θ̂_m̃|/√2 − sqrt( 1/2 (Σ̂_{m,m} + Σ̂_{m̃,m̃} − 2Σ̂_{m,m̃}) χ²_{k,1−α} )
        if d > 0 and θ̂_m < θ̂_m̃ then
            l̂(m) ← l̂(m) + 1
        else if d > 0 and θ̂_m > θ̂_m̃ then
            û(m) ← û(m) − 1
return {[l̂(m), û(m)]}_{m∈M}
```

by Eq. 6, *i.e.*,

$$d(C_\alpha, H_{m,\tilde{m}}) = \frac{|\hat{\theta}_m - \hat{\theta}_{\tilde{m}}| - \sqrt{(\widehat{\Sigma}_{m,m} + \widehat{\Sigma}_{\tilde{m},\tilde{m}} - 2\widehat{\Sigma}_{m,\tilde{m}})\chi^2_{k,1-\alpha}}}{\sqrt{2}}, \tag{10}$$

where $\widehat{\Sigma}$ is the empirical covariance matrix defined by Eq. 7.

Now, for each pair of models $m, m' \in \mathcal{M}$, we can readily conclude that, if the distance $d(C_\alpha, H_{m,\tilde{m}}) > 0$, then, the confidence region C_α lies in the half-space of $\mathbb{R}^k$ where $x_m > x_{\tilde{m}}$ if $\hat{\theta}_m > \hat{\theta}_{\tilde{m}}$ and it lies in the half space of $\mathbb{R}^k$ where $x_m < x_{\tilde{m}}$ if $\hat{\theta}_m < \hat{\theta}_{\tilde{m}}$. Building upon this observation, for each model $m \in \mathcal{M}$, we construct the following estimates $[\hat{l}(m), \hat{u}(m)]$ of the rank-sets $[l(m), u(m)]$:

$$\begin{aligned} \hat{l}(m) &= 1 + \sum_{\tilde{m}\in\mathcal{M}\setminus\{m\}} \mathbf{1}\{d(C_\alpha, H_{m,\tilde{m}}) > 0\} \cdot \mathbf{1}\{\hat{\theta}_m < \hat{\theta}_{\tilde{m}}\} \\ \hat{u}(m) &= k - \sum_{\tilde{m}\in\mathcal{M}\setminus\{m\}} \mathbf{1}\{d(C_\alpha, H_{m,\tilde{m}}) > 0\} \cdot \mathbf{1}\{\hat{\theta}_m > \hat{\theta}_{\tilde{m}}\}. \end{aligned} \tag{11}$$

Importantly, using a similar proof technique as in Lemma 1 in the recent paper by Neuhof and Benjamini [33], we can show that the estimates $[\hat{l}(m), \hat{u}(m)]$ as defined above enjoy provable coverage guarantees with respect to the rank-sets $[l(m), u(m)]$ induced by the probabilities $\boldsymbol{\theta}$ that the outputs of each model is preferred over any other model by humans:

Theorem 4.1. *The estimates $[\hat{l}(m), \hat{u}(m)]$ of the rank-set defined by Eq. 11 satisfy that*

$$\lim_n \mathbb{P}\left(\bigcap_{m\in\mathcal{M}} [l(m), u(m)] \subseteq [\hat{l}(m), \hat{u}(m)]\right) \geq 1 - \alpha. \tag{12}$$

Proof. Note that Eq. 12 holds if and only if:

$$\lim_n \mathbb{P}\left(\exists m \in \mathcal{M} : [l(m), u(m)] \not\subseteq [\hat{l}(m), \hat{u}(m)]\right) \leq \alpha \tag{13}$$

Therefore, to prove the theorem, it is sufficient to prove that Eq. 13 holds. Now, to prove that Eq. 13, we first show that the probability on the left hand side of the above equation is smaller than or equal to the probability $\mathbb{P}(\boldsymbol{\theta} \notin C_\alpha)$.

To this end, first note that, if for at least one model $m \in \mathcal{M}$, we have that $\hat{l}(m) > l(m)$ or $\hat{u}(m) < u(m)$, then it holds that

$$\bigcap_{m\in\mathcal{M}} [l(m), u(m)] \not\subseteq [\hat{l}(m), \hat{u}(m)].$$

Next, without loss of generality, assume that, for model m, we have that $\hat{l}(m) > l(m)$. In this case, from Eqs. 11 and 2 we get:

$$\sum_{\tilde{m}\in\mathcal{M}\setminus\{m\}} \mathbf{1}\{d(C_\alpha, H_{m,\tilde{m}}) > 0\} \cdot \mathbf{1}\{\hat{\theta}_m < \hat{\theta}_{\tilde{m}}\} > \sum_{\tilde{m}\in\mathcal{M}\setminus\{m\}} \mathbf{1}\{\theta_m < \theta_{\tilde{m}}\},$$

which means that there must be at least one model $\tilde{m} \in \mathcal{M}$ such that $x_m < x_{\tilde{m}}$ $\forall \boldsymbol{x} \in C_\alpha$ and $\theta_m > \theta_{\tilde{m}}$, which implies that $\boldsymbol{\theta} \notin C_\alpha$.

As a result, we can immediately conclude that,

$$\begin{aligned} \lim_n \mathbb{P}\left(\exists m \in \mathcal{M} : [l(m), u(m)] \not\subseteq [\hat{l}(m), \hat{u}(m)]\right) &\leq \lim_n \mathbb{P}(\boldsymbol{\theta} \notin C_\alpha) \\ &= \alpha. \end{aligned}$$

This concludes the proof. □

Algorithm 2 summarizes the overall procedure to construct the rank-sets $[\hat{l}(m), \hat{u}(m)]$ for all $m \in \mathcal{M}$.

5 CONCLUSIONS

In this work, we have introduced a statistical framework to construct a ranking of a collection of large language models consistent with their level of alignment with human preferences using a small set of pairwise comparisons by humans and a large set of pairwise comparisons by a strong large language model. Our framework quantifies uncertainty in the ranking by providing a rank-set—a set of possible ranking positions—for each of the models under comparison. Moreover, it guarantees that, with a probability greater than or equal to a user-specific value, the rank-sets cover the ranking consistent with the (true) probability that each model is preferred over others by humans. Our work opens up many interesting avenues for future work. For example, it would be important to validate our framework using real pairwise comparison data from humans and strong large language models. Moreover, it would be interesting to derive PAC-style, finite-sample coverage guarantees. Further, in our work, we have assumed that the pairwise comparisons by humans and by the strong LLM comprise the same distribution of inputs. However, in practice, the distribution of inputs may be different and thus it would be important to extend our framework to allow for distribution shifts. Finally, it would be worthwhile to explore other measures of uncertainty for rankings beyond rank-sets.

REFERENCES

[1] Diaa Al Mohamad, Jelle J Goeman, and Erik W van Zwet. 2022. Simultaneous confidence intervals for ranks with application to ranking institutions. *Biometrics* 78, 1 (2022), 238–247.

[2] Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. 2023. Prediction-powered inference. *Science* 382, 6671 (2023), 669–674.

[3] Anastasios N. Angelopoulos, John C Duchi, and Tijana Zrnic. 2023. PPI++: Efficient Prediction-Powered Inference. *arXiv preprint arXiv:2311.01453* (2023).

[4] Stephen Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-david, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Fries, Maged Al-shaibani, Shanya Sharma,

Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-jian Jiang, and Alexander Rush. 2022. PromptSource: An Integrated Development Environment and Repository for Natural Language Prompts. In *Proceedings of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Dublin, Ireland, 93–104.

[5] Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, Jiayin Zhang, Juanzi Li, and Lei Hou. 2024. Benchmarking Foundation Models with Language-Model-as-an-Examiner. In *Advances in Neural Information Processing Systems*.

[6] Meriem Boubdir, Edward Kim, Beyza Ermis, Sara Hooker, and Marzieh Fadaee. 2023. Elo Uncovered: Robustness and Best Practices in Language Model Evaluation. *arXiv preprint arXiv:2311.17295* (2023).

[7] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712* (2023).

[8] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology* (2024).

[9] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating Large Language Models Trained on Code. *arXiv preprint arXiv:2107.03374* (2021).

[10] Cheng-Han Chiang and Hung-yi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluations?. In *Proceedings of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 15607–15631.

[11] 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://vicuna.lmsys.org. Online; accessed 24 February 2024.

[12] Tim R Davidson, Veniamin Veselovsky, Martin Josifoski, Maxime Peyrard, Antoine Bosselut, Michal Kosinski, and Robert West. 2024. Evaluating Language Model Agency Through Negotiations. In *Proceedings of the International Conference on Learning Representations*.

[13] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. QLoRA: Efficient Finetuning of Quantized LLMs. In *Advances in Neural Information Processing Systems*.

[14] Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2024. AlpacaFarm: A Simulation Framework for Methods that Learn from Human Feedback. In *Advances in Neural Information Processing Systems*.

[15] Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like Summarization Evaluation with ChatGPT. *arXiv preprint arXiv:2304.02554* (2023).

[16] Harvey Goldstein and David J Spiegelhalter. 1996. League Tables and Their Limitations: Statistical Issues in Comparisons of Institutional Performance. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 159, 3 (1996), 385–409.

[17] Peter Hall and Hugh Miller. 2009. Using the bootstrap to quantify the authority of an empirical ranking. *The Annals of Statistics* 37 (2009), 3929–3959.

[18] Claudia E Haupt and Mason Marks. 2023. AI-Generated Medical Advice—GPT and Beyond. *Journal of American Medical Association* 329, 16 (2023), 1349–1350.

[19] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *Proceedings of the International Conference on Learning Representations*.

[20] Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. LLM-Blender: Ensembling Large Language Models with Pairwise Ranking and Generative Fusion. In *Proceedings of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 14165–14178.

[21] Martin Klein, Tommy Wright, and Jerzy Wieczorek. 2020. A Joint Confidence Region for an Overall Ranking of Populations. *Journal of the Royal Statistical Society Series C: Applied Statistics* 69, 3 (2020), 589–606.

[22] Oscar Lemmers, Jan AM Kremer, and George F Borm. 2007. Incorporating natural variation into IVF clinic league tables. *Human Reproduction* 22, 5 (2007), 1359–1362.

[23] Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2024. Generative Judge for Evaluating Alignment. In *Proceedings of the International Conference on Learning Representations*.

[24] Ruosen Li, Teerth Patel, and Xinya Du. 2023. PRD: Peer Rank and Discussion Improve Large Language Model Based Evaluations. *arXiv preprint arXiv:2307.02762* (2023).

[25] Zongjie Li, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, and Yang Liu. 2023. Split and Merge: Aligning Position Biases in Large Language Model Based Evaluators. *arXiv preprint arXiv:2310.01432* (2023).

[26] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research* (2023).

[27] Adian Liusie, Potsawee Manakul, and Mark JF Gales. 2023. LLM Comparative Assessment: Zero-shot NLG Evaluation through Pairwise Comparisons using Large Language Models. *arXiv preprint arXiv:2307.07889* (2023).

[28] LMSYS. 2023. Chatbot Arena: Benchmarking LLMs in the Wild with Elo Ratings. https://lmsys.org/. Online; accessed 20 February 2024.

[29] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The Flan Collection: Designing Data and Methods for Effective Instruction Tuning. In *Proceedings of the International Conference on Machine Learning*. PMLR, 22631–22648.

[30] E Clare Marshall, Colin Sanderson, David J Spiegelhalter, and Martin McKee. 1998. Reliability of league tables of in vitro fertilisation clinics: retrospective analysis of live birth rates. *British Medical Journal* 316, 7146 (1998), 1701–1705.

[31] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-Task Generalization via Natural Language Crowdsourcing Instructions. In *Proceedings of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 3470–3487.

[32] Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. 2024. Reading Between the Lines: Modeling User Behavior and Costs in AI-Assisted Programming. In *Proceedings of the Conference on Human Factors in Computing Systems*.

[33] Bitya Neuhof and Yuval Benjamini. 2023. Confident Feature Ranking. In *ICML workshop on Spurious Correlations, Invariance and Stability*.

[34] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, 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. In *Advances in Neural Information Processing Systems*.

[35] Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a General-Purpose Natural Language Processing Task Solver?. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[36] Justin Rising. 2021. Uncertainty in Ranking. *arXiv preprint arXiv:2107.03459* (2021).

[37] Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. 2023. Mathematical discoveries from program search with large language models. *Nature* 625, 7995 (2023), 468–475.

[38] Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2023. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. *arXiv preprint arXiv:2311.09476* (2023).

[39] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. Large language models encode clinical knowledge. *Nature* 620, 7972 (July 2023), 172–180.

[40] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short*

Papers). Association for Computational Linguistics, 4149–4158.

[41] 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. Online; accessed 20 February 2024.

[42] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2023. Large Language Models Can Accurately Predict Searcher Preferences. *arXiv preprint arXiv:2309.10621* (2023).

[43] Mudit Verma, Siddhant Bhambri, and Subbarao Kambhampati. 2023. Preference Proxies: Evaluating Large Language Models in Capturing Human Preferences in Human-AI Tasks. In *ICML Workshop The Many Facets of Preference-Based Learning*.

[44] Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large Language Models are not Fair Evaluators. *arXiv preprint arXiv:2305.17926* (2023).

[45] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In *Proceedings of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 13484–13508.

[46] Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2024. PandaLM: An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization. In *Proceedings of the International Conference on Learning Representations*.

[47] Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning Large Language Models with Human: A Survey. *arXiv preprint arXiv:2307.12966* (2023).

[48] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned Language Models are Zero-Shot Learners. In *Proceedings of the International Conference on Learning Representations*.

[49] Tommy Wright, Martin Klein, and Jerzy Wieczorek. 2014. Ranking Populations Based on Sample Survey Data. *Statistics* (2014), 12.

[50] Minge Xie, Kesar Singh, and Cun-Hui Zhang. 2009. Confidence intervals for population ranks in the presence of ties and near ties. *Journal of the American Statistical Association* 104, 486 (2009), 775–788.

[51] Shunpu Zhang, Jun Luo, Li Zhu, David G Stinchcomb, Dave Campbell, Ginger Carter, Scott Gilkeson, and Eric J Feuer. 2014. Confidence intervals for ranks of age-adjusted rates across states or counties. *Statistics in Medicine* 33, 11 (2014), 1853–1866.

[52] 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. In *Advances in Neural Information Processing Systems, data track*.

[53] Tijana Zrnic and Emmanuel J Candès. 2023. Cross-Prediction-Powered Inference. *arXiv preprint arXiv:2309.16598* (2023).