arXiv:2611.13485v1 [stat.ML] 10 Nov 2026

Differentially Private Statistical Inference for Rankings from Pairwise Comparisons

Xintao Xia* Jia Gu*†

Abstract

Reliable inference for rankings of large language models based on human pairwise comparisons must preserve the privacy of individual preference records. To meet this need, we develop methods for score and rank inference under the Bradley–Terry model with edge differential privacy. For an objective-perturbed estimator, a new leave-one-out analysis establishes finite-sample bounds on Euclidean and entrywise estimation errors. For sufficiently dense comparison graphs, the error bounds extend to smaller privacy budgets, including regimes in which its privacy term dominates its statistical term. A uniform asymptotic expansion yields asymptotically valid confidence intervals for individual scores and pairwise score differences. To improve finite-sample performance, we account for both regularization bias and privacy-induced randomness by combining a plug-in ridge-bias correction with Gaussian–Laplace calibration. We further establish a bootstrap procedure for simultaneous rank inference, providing asymptotically valid confidence sets for individual ranks and tests of top-K membership. When the comparison probability is public and known, all proposed inferential procedures require no additional privacy budget. We evaluate finite-sample performance through simulations and apply the proposed methods to rank large language models for creative writing.

Keywords: Bradley–Terry model; Differential privacy; Parametric bootstrap; Rank confidence intervals; Top-K inference

*Center for Data Science, Zhejiang University
†Corresponding author. Email: gujia@zju.edu.cn.

1 Introduction

Ranking from pairwise comparison data is a fundamental statistical problem (Bradley and Terry, 1952). For example, pairwise comparisons have been used to assess consumer preferences for beverages (Duineveld et al., 2000) and to evaluate travel packages (Hatzinger and Mazanec, 2007). With rapid advances in artificial intelligence (AI), ranking has become increasingly important for evaluating large language models (LLMs). As academic institutions and companies release a growing number of models, reliable rankings help identify top models, providing a basis for assessing research progress and guiding model selection for users. These rankings also have commercial implications, as companies with top-ranked AI models may attract greater investor interest (Aradau and Blanke, 2026). Conventional benchmarks provide standardized assessments of capabilities such as knowledge and problem solving (Hendrycks et al., 2021). However, benchmark scores may not fully capture how users experience model responses, especially in terms of perceived empathy and emotional support (Zheng et al., 2023). Human feedback therefore provides a basis for rankings that reflect user experience, as illustrated by Chatbot Arena, which ranks LLMs by aggregating human preferences between anonymous pairs of model responses (Chiang et al., 2024). A central statistical challenge is therefore to estimate LLM rankings from human pairwise comparisons and quantify the uncertainty in these rankings.

However, releasing rankings derived from human comparisons may reveal individual preferences, even when the underlying comparison records are not released. Data protection laws, including the General Data Protection Regulation in the European Union and the California Consumer Privacy Act in the United States, further emphasize the need to protect personal information during data analysis. These concerns motivate ranking procedures with privacy guarantees while ensuring statistically valid inference. Differential pri-

vacy (DP) provides a mathematical framework for data analysis under privacy constraints (Dwork et al., 2006). From a hypothesis-testing perspective, DP limits the power of any test for distinguishing neighboring datasets based on the released output (Wasserman and Zhou, 2010). Achieving DP generally requires introducing randomness into the analysis (Dwork and Roth, 2014). The challenge in ranking inference under privacy constraints is to account for both sampling variability and privacy-induced randomness.

1.1 Related works

The Bradley–Terry (BT) model provides a standard framework for analyzing pairwise comparison data by associating each item with a latent parameter (Bradley and Terry, 1952). These parameters determine pairwise comparison probabilities and induce a ranking. A substantial literature establishes estimation error rates and top-K recovery guarantees for likelihood and spectral methods (Negahban et al., 2017; Chen et al., 2019, 2022). Complementing these results, Han et al. (2020, 2023) establish consistency guarantees for maximum likelihood estimation under sparse networks. Recent developments extend the theory to general comparison graphs (Han et al., 2026) and incorporate covariates within the Plackett–Luce framework (Dong et al., 2026). Beyond estimation, statistical inference seeks to quantify uncertainty in the estimated rank (Han et al., 2024). Gao et al. (2023) derive asymptotic distributions for score estimators and construct confidence intervals for individual ranks, while Liu et al. (2023) develop inference for ranking properties, including pairwise ordering and top-K membership. Fan et al. (2025) consider top-choice multiway comparison and construct simultaneous confidence intervals for score differences and individual ranks. These methods address ranking uncertainty without privacy constraints.

Estimation and inference under DP have gained attention in statistics. For exam-

ple, output and objective perturbation provide general approaches to private empirical risk minimization (Chaudhuri et al., 2011). Asymptotic guarantees for private estimators and hypothesis tests have been established through bounded-influence methods and noisy optimization (Avella-Medina, 2021; Avella-Medina et al., 2023). More recently, this line of research has expanded to false discovery rate control and variable selection in high-dimensional regression (Cai et al., 2026b; Gu and Cai, 2026), as well as M-estimation by differentially private stochastic gradient descent (Xia et al., 2025). However, extending classical M-estimation methods to rank inference in our setting requires additional analysis to account for the identifying constraint on the latent scores, the growing dimension of the score vector, and dependence among score estimates.

More directly related to our setting, Xu et al. (2025) develop an adaptive debiasing procedure for locally privatized pairwise comparisons and establish estimation and ranking recovery guarantees. On the inferential side, Ouyang et al. (2025) establish consistency and asymptotic normality for BT parameter estimators based on degree sequences perturbed by Laplace noise. Cai et al. (2026a) study ranking under edge and individual differential privacy and propose a perturbed likelihood with guarantees for estimation error and top-K recovery. Our work differs from that of Cai et al. (2026a) in three respects. First, although both approaches use objective perturbation, our ridge penalty depends only on the privacy budget, whereas theirs also depends on the number of items and the comparison probability. Second, whereas their entrywise analysis tracks leave-one-out gradient descent iterates, ours directly compares the full estimator with estimators obtained by removing the comparisons involving each item. Combined with our tuning rule, this approach establishes entrywise error bounds for smaller privacy budgets when the comparison graph is sufficiently dense. Third, and most importantly, we go beyond point estimation by establishing

a uniform asymptotic expansion that enables confidence intervals for individual scores and simultaneous confidence sets for ranks. To the best of our knowledge, this provides the first unified framework for inference on scores and ranks under the BT model subject to edge differential privacy.

1.2 Main contributions

We develop a unified framework for inference on individual scores, individual ranks, and top-K membership under edge differential privacy. Our contributions are threefold.

First, we develop a new leave-one-out analysis for objective-perturbed ranking estimation and establish finite-sample bounds for both the ℓ_2 and entrywise estimation errors. For sufficiently dense comparison graphs, our entrywise guarantee extends to smaller privacy budgets, including regimes in which privacy-induced error dominates statistical error.

Second, we establish a uniform asymptotic expansion for the private estimator and derive asymptotically valid inference for individual scores and pairwise score differences. To improve finite-sample performance, we introduce a bias-aware procedure that combines a plug-in ridge-bias correction with Gaussian–Laplace calibration, jointly addressing regularization bias and privacy-induced randomness without additional privacy cost.

Third, we extend private score inference to simultaneous uncertainty quantification for individual ranks and top-K membership. Our bootstrap procedure yields asymptotically valid rank confidence sets and top-K membership tests over a potentially growing collection of items. When the comparison probability is public and known, these inferential procedures require no additional privacy budget beyond the initial private release.

Organization. The paper is organized as follows. Section 2 introduces the BT model and edge differential privacy. Section 3 presents the private estimator and procedures for score

and rank inference. Section 4 establishes privacy and estimation guarantees, a uniform asymptotic expansion, and the asymptotic validity of the inference procedures. Sections 5 and 6 present simulation studies and a real data analysis, respectively. Section 7 concludes with a discussion. The supplementary material (SM) provides background on differential privacy, proofs and technical lemmas, and additional numerical results.

Notation. We use $\mathbf{1}$ and $\mathbf{I}$ for the all-ones vector and identity matrix, with dimensions indicated by subscripts when needed. We use $\mathbf{1}\{\cdot\}$ for the indicator function. For a vector $\mathbf{u}$, $\|\mathbf{u}\|_2$ and $\|\mathbf{u}\|_\infty$ denote its Euclidean and maximum absolute entry norms, respectively. All limits are taken as $n \to \infty$. For positive sequences a_n and b_n, $a_n \gg b_n$ means $b_n/a_n \to 0$, and $X_n = o_p(a_n)$ means $X_n/a_n \to 0$ in probability.

2 Preliminaries

2.1 Ranking from pairwise comparisons

We consider a ranking problem with n items indexed by $[n] = \{1, \ldots, n\}$. Pairwise comparisons are observed at random. Specifically, each unordered pair $\{i, j\}$, $1 \le i < j \le n$, is included independently with probability $p \in (0, 1]$. The observed comparisons therefore form an Erdős–Rényi graph $G = ([n], \mathcal{E})$, where $\mathcal{E}$ denotes the set of observed pairs. Let A_{ij} denote the sampling indicator, where $A_{ij} = 1$ if $\{i, j\} \in \mathcal{E}$ and $A_{ij} = 0$ otherwise. Set $A_{ji} = A_{ij}$ and $A_{ii} = 0$, and write $\mathbf{A} = (A_{ij})$. The sampling probability p is public and known. For each observed pair $\{i, j\} \in \mathcal{E}$, we record a binary comparison outcome $Y_{ij} \in \{0, 1\}$, where $Y_{ij} = 1$ indicates that item i is preferred to item j, and $Y_{ij} = 0$ otherwise. We set $Y_{ji} = 1 - Y_{ij}$, so that $Y_{ij} + Y_{ji} = 1$. Conditional on the comparison graph, the observed outcomes are independent, with $Y_{ij} \sim \text{Bernoulli}(\pi_{ij})$ for $\{i, j\} \in \mathcal{E}$, where

$\pi_{ij} \in [0,1]$ is the probability that item i defeats item j. Terms multiplied by $A_{ij} = 0$ are defined to be zero. Thus, all data-dependent sums can be restricted to observed edges.

We model pairwise preferences using the BT model (Bradley and Terry, 1952), which assigns each item $i \in [n]$ a latent score $\theta_i^* \in \mathbb{R}$. The probability that i is preferred to j is

$$\pi_{ij} = F(\theta_i^* - \theta_j^*), \qquad F(t) = \frac{1}{1+\exp(-t)},$$

where $F(\cdot)$ is the standard logistic distribution function. The latent score vector $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_n^*)^\top$ determines the population ranking and provides the basis for ranking inference. Since the model depends only on pairwise score differences, we center the scores for identifiability and impose the following boundedness condition.

Assumption 1. *The true score vector $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_n^*)^\top$ satisfies $\sum_{i=1}^n \theta_i^* = 0$. Moreover, there exists a constant $\kappa > 0$, independent of n, such that $\|\boldsymbol{\theta}^*\|_\infty \le \kappa$.*

Bounded scores keep the pairwise comparison probabilities away from zero and one and ensure nonvanishing likelihood curvature, as commonly assumed in ranking estimation and inference (Chen et al., 2022; Gao et al., 2023; Fan et al., 2025).

2.2 Edge differential privacy

We study statistical inference for the latent ranking parameter and the induced ranking under (ε, δ)-edge differential privacy, which we define formally in this section.

Definition 1 ((ε,δ)-differential privacy)**.** *A randomized algorithm $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ satisfies (ε,δ)-differential privacy if, for every measurable set $B \subseteq \mathcal{R}$ and every pair of adjacent datasets $D, D' \in \mathcal{D}$ differing in at most one record,*

$$\Pr\{\mathcal{M}(D) \in B\} \le e^{\varepsilon} \Pr\{\mathcal{M}(D') \in B\} + \delta,$$

where the probability is taken over the randomness of $\mathcal{M}$.

The specified adjacency relation determines the privacy guarantee. While classical differential privacy for tabular data typically protects a single record, our pairwise-comparison setting treats an unordered pair, together with its comparison information, as the protected unit. This leads to the notion of *edge adjacency*. Let $\mathcal{P}_n = \{(i,j) : 1 \le i < j \le n\}$ denote the set of unordered item pairs. For each pair $1 \le i < j \le n$, define

$$Z_{ij} = Y_{ij} \text{ if } A_{ij} = 1, \qquad Z_{ij} = \varnothing \text{ if } A_{ij} = 0,$$

where $\varnothing$ denotes the absence of a comparison. We write the pairwise comparison dataset as $D = (Z_{ij})_{1 \le i < j \le n}$. We next define the adjacency relation for pairwise-comparison data.

Definition 2 (Edge adjacency)**.** *Two comparison datasets $D = (Z_{ij})_{1 \le i < j \le n}$ and $D' = (Z'_{ij})_{1 \le i < j \le n}$ are* edge-adjacent *if exactly one coordinate differs. Specifically, a unique pair (i^*, j^*) with $1 \le i^* < j^* \le n$ satisfies $Z_{i^*j^*} \ne Z'_{i^*j^*}$, while $Z_{ij} = Z'_{ij}$ for all other pairs.*

Thus, edge-adjacent datasets differ by the addition, removal, or modification of a comparison for a single unordered pair of items. Differential privacy under this relation protects both whether a comparison is observed and its outcome. This privacy notion is called *edge differential privacy* (Karwa et al., 2011; Sala et al., 2011).

3 Private estimation and uncertainty quantification

In this section, we construct a private estimator of the latent scores and decompose the effects of regularization and privacy perturbation. We then develop confidence intervals for individual scores, followed by simultaneous inference for ranks and top-K membership.

3.1 Private estimation

We first introduce a differentially private estimator of the latent score vector $\boldsymbol{\theta}^*$. The estimator is based on objective perturbation (Chaudhuri et al., 2011; Iyengar et al., 2019) and follows the construction of Cai et al. (2026a). For $\boldsymbol{\theta} \in \mathbb{R}^n$, define $F_{ij}(\boldsymbol{\theta}) = F(\theta_i - \theta_j)$ for $1 \le i < j \le n$. The negative log-likelihood based on the observed comparisons is

$$\mathcal{L}_n(\boldsymbol{\theta}) = \sum_{1 \le i < j \le n} A_{ij}\big[- Y_{ij} \log F_{ij}(\boldsymbol{\theta}) - (1 - Y_{ij}) \log\{1 - F_{ij}(\boldsymbol{\theta})\}\big].$$

Equivalently, since $Y_{ji} = 1 - Y_{ij}$, this can be written as

$$\mathcal{L}_n(\boldsymbol{\theta}) = \sum_{1 \le i < j \le n} A_{ij}\big[- Y_{ij} \log F_{ij}(\boldsymbol{\theta}) - Y_{ji} \log\{1 - F_{ij}(\boldsymbol{\theta})\}\big].$$

To privatize the estimator, we add a random linear perturbation to the regularized objective. Let $\mathbf{w} = (w_1, \ldots, w_n)^\top$, where $w_1, \ldots, w_n \overset{\text{ind}}{\sim} \text{Laplace}(0, \lambda)$, independently of the comparison data, with scale parameter $\lambda > 0$. For a regularization parameter $\gamma > 0$, define

$$\widetilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^n}{\arg\min} \left\{ \mathcal{L}_n(\boldsymbol{\theta}) + \frac{\gamma}{2} \|\boldsymbol{\theta}\|_2^2 + \mathbf{w}^\top \boldsymbol{\theta} \right\}. \tag{3.1}$$

The ridge penalty ensures strong convexity of the objective, while the random linear term masks the change in the score equation caused by modifying a single comparison. These two additional terms together guarantee the privacy protection for the estimated scores.

In non-private settings, existing likelihood-based inference procedures for ranking models (Gao et al., 2023; Fan et al., 2025) are typically based on the unregularized maximum likelihood estimator (MLE) under the centering constraint,

$$\widehat{\boldsymbol{\theta}}_{\text{MLE}} = \underset{\mathbf{1}_n^\top \boldsymbol{\theta} = 0}{\operatorname{argmin}} \, \mathcal{L}_n(\boldsymbol{\theta}).$$

The constraint respects the identification condition of the scores, since the likelihood depends only on their pairwise differences. The differentially private estimator $\widetilde{\boldsymbol{\theta}}$ is instead

obtained by minimizing a regularized likelihood with an additional random linear perturbation. To understand its asymptotic behavior, we compare the estimators under the same centering normalization and discuss the effects of these two modifications separately.

Effects of regularization. The ℓ_2 penalty ensures strong convexity of the penalized loss function, but it also introduces shrinkage relative to the MLE. To isolate this effect, let

$$\widehat{\boldsymbol{\theta}}_\gamma := \underset{\mathbf{1}_n^\top \boldsymbol{\theta}=0}{\arg\min} \left\{ \mathcal{L}_n(\boldsymbol{\theta}) + \frac{\gamma}{2}\|\boldsymbol{\theta}\|_2^2 \right\}, \qquad \gamma > 0.$$

This estimator coincides with the unconstrained ridge estimator studied for ranking estimation (Chen et al., 2019). To see this, shift invariance implies that $\mathbf{1}_n^\top \nabla \mathcal{L}_n(\boldsymbol{\theta}) = 0$ for every $\boldsymbol{\theta}$. Summing the first-order conditions for the unconstrained ridge estimator then yields $\mathbf{1}_n^\top \widehat{\boldsymbol{\theta}}_\gamma = 0$. Thus, the ridge penalty automatically enforces the centering constraint required for identifiability, making an explicit constraint unnecessary. The same shift invariance implies that the Lagrange multiplier for the constrained MLE is zero, so $\nabla \mathcal{L}_n(\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}) = \mathbf{0}$. By the mean-value theorem, a first-order expansion of $\nabla \mathcal{L}_n(\widehat{\boldsymbol{\theta}}_\gamma)$ around $\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ then gives $\nabla \mathcal{L}_n(\widehat{\boldsymbol{\theta}}_\gamma) = \mathbf{H}_n^\dagger(\widehat{\boldsymbol{\theta}}_\gamma - \widehat{\boldsymbol{\theta}}_{\mathrm{MLE}})$, where $\mathbf{H}_n^\dagger$ is the integrated Hessian matrix of the form

$$\mathbf{H}_n^\dagger := \int_0^1 \nabla^2 \mathcal{L}_n\big\{\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}} + t(\widehat{\boldsymbol{\theta}}_\gamma - \widehat{\boldsymbol{\theta}}_{\mathrm{MLE}})\big\}\, dt.$$

Combining this identity with the first-order condition for the penalized estimator yields

$$\widehat{\boldsymbol{\theta}}_\gamma = (\mathbf{H}_n^\dagger + \gamma \mathbf{I})^{-1} \mathbf{H}_n^\dagger \widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}. \tag{3.2}$$

To illustrate the effect on individual scores, let $h_i = (\mathbf{H}_n^\dagger)_{ii}$ denote the integrated likelihood curvature for item i. Replacing $\mathbf{H}_n^\dagger$ by its diagonal in (3.2) gives the heuristic approximation

$$\widehat{\theta}_{\gamma,i} \approx \frac{h_i}{h_i + \gamma} \widehat{\theta}_{\mathrm{MLE},i},$$

which suggests stronger shrinkage for items with less comparison information. The omitted off-diagonal terms couple changes in all fitted scores, so their combined effect must be

controlled to assess the inferential consequences of regularization. For the private estimator, the regularization term in (3.5) accounts for this coupling through the full Hessian.

Effects of perturbation. The linear perturbation can shift the mean of the estimated scores from zero, so the private estimator need not satisfy the centering constraint. To compare it with the penalized MLE, define the constrained perturbed penalized estimator

$$\widehat{\boldsymbol{\theta}} = \underset{\mathbf{1}_n^\top \boldsymbol{\theta}=0}{\operatorname{argmin}} \left\{ \mathcal{L}_n(\boldsymbol{\theta}) + \frac{\gamma}{2}\|\boldsymbol{\theta}\|_2^2 + \mathbf{w}^\top \boldsymbol{\theta} \right\}. \tag{3.3}$$

We decompose the privacy perturbation into its average and centered components, $\bar{w} = \sum_{i=1}^n w_i/n$ and $\mathbf{w}^\perp = \mathbf{w} - \bar{w}\mathbf{1}_n$. On the centered parameter space, $\mathbf{w}^\top\boldsymbol{\theta} = (\mathbf{w}^\perp)^\top\boldsymbol{\theta}$. The constrained first-order condition has the form $\nabla\mathcal{L}_n(\widehat{\boldsymbol{\theta}}) + \gamma\widehat{\boldsymbol{\theta}} + \mathbf{w} = a\mathbf{1}_n$, where a is the Lagrangian multiplier. Taking coordinate sums gives $a = \bar{w}$, and hence $\nabla\mathcal{L}_n(\widehat{\boldsymbol{\theta}}) + \gamma\widehat{\boldsymbol{\theta}} + \mathbf{w}^\perp = \mathbf{0}$. By strong convexity, $\widehat{\boldsymbol{\theta}}$ is therefore also the unconstrained minimizer of

$$\mathcal{L}_n(\boldsymbol{\theta}) + \frac{\gamma}{2}\|\boldsymbol{\theta}\|_2^2 + (\mathbf{w}^\perp)^\top\boldsymbol{\theta}.$$

The following lemma relates this centered estimator to the original private release.

Lemma 1 (Decomposition of the privacy perturbation)**.** *Let* $\widetilde{\boldsymbol{\theta}}$ *and* $\widehat{\boldsymbol{\theta}}$ *be defined in* (3.1) *and* (3.3)*, respectively. Then, we have*

$$\widetilde{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}} - \frac{\bar{w}}{\gamma}\mathbf{1}_n.$$

Consequently, $\widetilde{\theta}_i - \widetilde{\theta}_j = \widehat{\theta}_i - \widehat{\theta}_j$ *for all* $1 \le i \ne j \le n$.

Lemma 1 shows that the average component of the privacy noise is absorbed entirely as a uniform translation of the estimator. The centered estimator $\widehat{\boldsymbol{\theta}}$ can be obtained directly from the private release $\widetilde{\boldsymbol{\theta}}$ as $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}} - \mathrm{ave}(\widetilde{\boldsymbol{\theta}})\mathbf{1}_n$, where ave denotes the average of the coordinates. Centering is a post-processing operation and preserves the privacy guarantee.

It removes the common translation, while regularization and the centered perturbation remain. Inference for normalized individual scores is based on $\widehat{\boldsymbol{\theta}}$, whereas inference for pairwise score differences $\theta_i^* - \theta_j^*$ can use $\widetilde{\theta}_i - \widetilde{\theta}_j$ or, equivalently, $\widehat{\theta}_i - \widehat{\theta}_j$. These observations motivate the confidence intervals for individual scores and pairwise differences.

3.2 Private inference for scores and pairwise differences

After removing the common shift, the private estimator still differs from the non-private MLE $\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ because of regularization and the centered privacy perturbation. To construct private confidence intervals, we separate these effects from statistical randomness. Let

$$S_i = \sum_{j \neq i} A_{ij}\{Y_{ij} - F(\theta_i^* - \theta_j^*)\} \quad \text{and} \quad g_i = \sum_{j \neq i} A_{ij} F'(\theta_i^* - \theta_j^*), \tag{3.4}$$

where S_i is the sampling score and d_i is the penalized curvature for item i on the observed graph. Further define the penalized loss Hessian and the associated regularization term by

$$\mathbf{H}_\gamma(\boldsymbol{\theta}) = \nabla^2 \mathcal{L}_n(\boldsymbol{\theta}) + \gamma \mathbf{I}_n \quad \text{and} \quad \boldsymbol{\beta}_\gamma = \gamma \mathbf{H}_\gamma(\boldsymbol{\theta}^*)^{-1}\boldsymbol{\theta}^*.$$

By the proof of Theorem 2, the analysis underlying its expansion gives, for each fixed i,

$$\widehat{\theta}_i - \theta_i^* = \frac{S_i}{g_i} - \beta_{\gamma,i} - \frac{w_i^{\perp}}{g_i} + o_p\{(np)^{-1/2}\}. \tag{3.5}$$

The first term captures sampling variability and, under the stated conditions, is asymptotically equivalent to the leading fluctuation of the non-private MLE. The second and third terms represent regularization bias and centered privacy perturbation, respectively. Under the stated conditions, regularization bias and centered privacy perturbation are asymptotically negligible but can affect finite-sample coverage.

Wald confidence intervals. Under regularity conditions, the regularization bias and centered privacy perturbation are asymptotically negligible relative to statistical sampling

variation. The centered private estimator $\widehat{\boldsymbol{\theta}}$ therefore has the same first-order normal approximation as the constrained non-private MLE coordinate-wise. Such first-order equivalence motivates Wald inference for private M-estimators. Specifically, we estimate the standard deviation of the leading statistical term S_i/g_i using the plug-in estimator $\widetilde{\rho}_i := \widetilde{g}_i^{-1/2}$, where $\widetilde{g}_i = p\sum_{j\neq i} F'(\widetilde{\theta}_i - \widetilde{\theta}_j)$. For a prespecified pair (i, j), the respective leading statistical terms are asymptotically independent, giving the pairwise standard error estimator $\widetilde{\rho}_{ij} := (\widetilde{\rho}_i^2 + \widetilde{\rho}_j^2)^{1/2}$. For $\alpha \in (0, 1)$, let $z_{1-\alpha/2}$ denote the $(1-\alpha/2)$ standard normal quantile. The two-sided Wald intervals for a prespecified score θ_i^* and pairwise difference $\theta_i^* - \theta_j^*$ are

$$\mathrm{CI}_i(1-\alpha) = \left[\widehat{\theta}_i - z_{1-\alpha/2}\widetilde{\rho}_i, \widehat{\theta}_i + z_{1-\alpha/2}\widetilde{\rho}_i\right] \quad \text{and} \tag{3.6}$$

$$\mathrm{CI}_{ij}(1-\alpha) = \left[(\widetilde{\theta}_i - \widetilde{\theta}_j) - z_{1-\alpha/2}\widetilde{\rho}_{ij}, (\widetilde{\theta}_i - \widetilde{\theta}_j) + z_{1-\alpha/2}\widetilde{\rho}_{ij}\right], \tag{3.7}$$

respectively. These intervals take the standard Wald form, with their centers and standard errors obtained by post-processing the private estimator.

In finite samples, particularly when comparison information is limited, or the privacy budget ε is small, regularization bias and privacy perturbation may be non-negligible relative to statistical sampling variation. Ignoring these effects may lead to undercoverage of the Wald confidence intervals. We therefore develop a bias-corrected procedure that accounts for both effects in the subsequent calibrations.

Bias correction. The leading regularization bias is $-\beta_{\gamma,i}$, so we correct the center by adding a plug-in estimate of $\beta_{\gamma,i}$. To estimate $\boldsymbol{\beta}_\gamma$, we replace each graph indicator A_{ij} in $\mathbf{H}_\gamma(\boldsymbol{\theta})$ by its public sampling probability p and evaluate the resulting matrix at the centered private estimator $\widehat{\boldsymbol{\theta}}$. Denote this matrix by $\widetilde{\mathbf{H}}_\gamma$, whose entries are

$$(\widetilde{\mathbf{H}}_\gamma)_{ii} = \widetilde{g}_i + \gamma, \qquad (\widetilde{\mathbf{H}}_\gamma)_{ij} = -pF'(\widehat{\theta}_i - \widehat{\theta}_j), \quad i \neq j.$$

The diagonal entries use the previously defined $\widetilde{g}_i$ because centering preserves score differences: $\widehat{\theta}_i - \widehat{\theta}_j = \widetilde{\theta}_i - \widetilde{\theta}_j$. We then estimate $\boldsymbol{\beta}_\gamma$ by $\widetilde{\boldsymbol{\beta}}_\gamma = \gamma\widetilde{\mathbf{H}}_\gamma^{-1}\widehat{\boldsymbol{\theta}}$, and the bias-corrected

estimator is constructed as

$$\widetilde{\theta}_i^{(\mathrm{bc})} = \widehat{\theta}_i + \widetilde{\beta}_{\gamma,i}, \qquad i \in [n]. \tag{3.8}$$

Gaussian–Laplace calibration. Correcting the regularization bias leaves the random privacy perturbation in the expansion (3.5). To account for privacy-induced randomness in finite samples, we use Gaussian—Laplace calibration motivated by the expansion. Although the centered noise coordinates $\{w_i^{\perp}\}_{i=1}^n$ are dependent and are not exactly Laplace, the contribution of the noise average $\bar{w}$ is asymptotically negligible on the standard-error scale for a fixed score or pair. This permits calibration using independent Laplace draws. Specifically, with the private estimator held fixed, let $Z \sim N(0,1)$ and $W, W_1, W_2 \overset{\text{i.i.d.}}{\sim} \mathrm{Laplace}(0,\lambda)$ be fresh, mutually independent draws, independent of the fit. For each prespecified score or pair, define its Gaussian–Laplace calibration variable on the sampling-standard-error scale:

$$T_i^{\mathrm{mix}} = Z + \widetilde{\rho}_i W, \qquad T_{ij}^{\mathrm{mix}} = Z + \frac{\widetilde{\rho}_i^2 W_1 + \widetilde{\rho}_j^2 W_2}{\widetilde{\rho}_{ij}}.$$

The symmetry of the Laplace distribution justifies the plus signs. Let $q_i^{\mathrm{mix}}(u)$ and $q_{ij}^{\mathrm{mix}}(u)$ denote the conditional u-quantiles. Since both distributions are symmetric about zero, the bias-corrected two-sided confidence intervals for θ_i^* and $\theta_i^* - \theta_j^*$ are constructed as

$$\mathrm{CI}_i^{(\mathrm{bc})}(1-\alpha) = \left[\widetilde{\theta}_i^{(\mathrm{bc})} - q_i^{\mathrm{mix}}(1-\alpha/2)\widetilde{\rho}_i, \widetilde{\theta}_i^{(\mathrm{bc})} + q_i^{\mathrm{mix}}(1-\alpha/2)\widetilde{\rho}_i\right] \quad \text{and} \tag{3.9}$$

$$\mathrm{CI}_{ij}^{(\mathrm{bc})}(1-\alpha) = \left[(\widetilde{\theta}_i^{(\mathrm{bc})} - \widetilde{\theta}_j^{(\mathrm{bc})}) - q_{ij}^{\mathrm{mix}}(1-\alpha/2)\widetilde{\rho}_{ij}, (\widetilde{\theta}_i^{(\mathrm{bc})} - \widetilde{\theta}_j^{(\mathrm{bc})}) + q_{ij}^{\mathrm{mix}}(1-\alpha/2)\widetilde{\rho}_{ij}\right], \tag{3.10}$$

respectively. Theorem 3 establishes asymptotic coverage $1-\alpha$ for both procedures.

Remark 1 (No additional privacy cost)**.** *A notable feature of the proposed inference procedure is that neither the variance estimation nor the bias correction requires an additional privacy budget. This contrasts with private inference for general M-estimation, where estimating covariance matrices often requires separate privatization steps (Avella-Medina et al.,*

2023; Xia et al., 2025). In our setting, all quantities entering the confidence intervals are functions of the private estimator $\widetilde{\boldsymbol{\theta}}$ and public parameters. Bias correction, plug-in variance estimation, and interval construction are post-processing operations and therefore preserve the original $(\varepsilon, 0)$-edge DP guarantee.

3.3 Private inference for ranks

Section 3.2 develops confidence intervals for prespecified scores and pairwise score differences. Inference for an item's rank instead requires simultaneous control of its comparisons with the other items. We therefore use bootstrap calibration to obtain simultaneous bounds on the relevant score differences, which in turn yield two-sided confidence intervals for the rank of a prespecified item $i \in [n]$ and one-sided lower bounds for top-K inference. Let

$$r(i) = 1 + \sum_{k \neq i} \mathbf{1}\{\theta_k^* > \theta_i^*\}$$

denote the population rank of item i. Smaller rank values correspond to higher scores, and this definition allows items with equal scores to share a rank. Let $\varnothing \neq \mathcal{I} \subseteq [n]$ be a prespecified set of items whose ranks are of interest, whose size may grow with n: Choosing $\mathcal{I} = \{i\}$ targets an individual rank, while $\mathcal{I} = [n]$ gives simultaneous inference for all ranks.

Bootstrap construction. Consider the ordered comparison family and contrasts by

$$\mathcal{P}_{\mathcal{I}} := \{(k, i) : i \in \mathcal{I},\ k \neq i\}, \quad \Delta_{ki}^* := \theta_k^* - \theta_i^*, \quad \text{and} \quad \widetilde{\Delta}_{ki} := \widetilde{\theta}_k - \widetilde{\theta}_i,$$

respectively, where $(k, i) \in \mathcal{P}_{\mathcal{I}}$. We use the pairwise standard error estimator from Section 3.2, denoted here by $\widetilde{\sigma}_{ki} = \widetilde{\rho}_{ki} = \left(\widetilde{\rho}_k^2 + \widetilde{\rho}_i^2\right)^{1/2}$. To calibrate simultaneous bounds, we approximate the joint distribution of the contrast errors using the leading terms in the asymptotic expansion of the private estimator. Let $g_i := \sum_{j \neq i} A_{ij} F'(\theta_i^* - \theta_j^*)$, which is the ith diagonal entry of the Hessian $\nabla^2 \mathcal{L}_n(\boldsymbol{\theta}^*)$. With S_i defined as in (3.4), the uniform

expansion in Theorem 2 identifies S_i/g_i as the leading estimation error after removal of the common shift. Since this shift cancels in pairwise differences, the expansion motivates the approximation

$$\widetilde{\Delta}_{ki} - \Delta^*_{ki} \approx \frac{S_k}{g_k} - \frac{S_i}{g_i}, \qquad (k,i) \in \mathcal{P}_{\mathcal{I}}.$$

The simultaneous inference result of Fan et al. (2025) uses a Gaussian multiplier bootstrap and requires a growing number of observations per sampled comparison set. In our pairwise setting, we construct a parametric bootstrap of these leading terms that allows one outcome per sampled pair. Specifically, let B denote the number of bootstrap replications. Conditional on the private estimator $\widetilde{\boldsymbol{\theta}}$, the replications are independent. For each $b = 1, \ldots, B$ and $1 \le i < j \le n$, generate $A^{(b)}_{ij} \sim \text{Bernoulli}(p)$ independently, where p is public and known. Set $A^{(b)}_{ji} = A^{(b)}_{ij}$ and $A^{(b)}_{ii} = 0$. Conditional on the simulated graph, generate independent outcomes by

$$Y^{(b)}_{ij} \sim \text{Bernoulli}\{F(\widetilde{\theta}_i - \widetilde{\theta}_j)\}, \qquad Y^{(b)}_{ji} = 1 - Y^{(b)}_{ij}.$$

For each simulated sample $(\mathbf{A}^{(b)}, \mathbf{Y}^{(b)})$, compute the likelihood scores and the diagonal entries of the negative log-likelihood Hessian at the private estimator:

$$S^{(b)}_i = \sum_{j \neq i} A^{(b)}_{ij}\{Y^{(b)}_{ij} - F(\widetilde{\theta}_i - \widetilde{\theta}_j)\}, \quad g^{(b)}_i = \sum_{j \neq i} A^{(b)}_{ij} F'(\widetilde{\theta}_i - \widetilde{\theta}_j).$$

Set $S^{(b)}_i/g^{(b)}_i = 0$ when $g^{(b)}_i = 0$. Under the stated conditions, the probability that a replication has at least one isolated vertex is at most $n(1-p)^{n-1}$, so this convention has a negligible distributional effect. The ratio $S^{(b)}_i/g^{(b)}_i$ simulates the leading estimation error for item i under the model evaluated at $\widetilde{\boldsymbol{\theta}}$. Within each replication, all contrasts are computed from the same simulated graph and outcomes, reproducing their dependence under the fitted model. The original plug-in standard errors are held fixed across replications.

With p public and known, the bootstrap uses only the private estimator and independent

simulation randomness, so the resulting critical values and inferential outputs incur no additional privacy cost. If p is unknown, it can be estimated using a separate differentially private procedure; see, for example, Ullman and Sealfon (2019). Incorporating such an estimate requires accounting for its privacy cost and estimation error. The same bootstrap replications are used for the two-sided and one-sided calibrations below.

(i) Two-sided rank confidence intervals. For two-sided inference, define the maximum absolute standardized contrast error by

$$T_{\mathcal{I}} = \max_{(k,i)\in\mathcal{P}_{\mathcal{I}}} \left| \frac{\widetilde{\Delta}_{ki} - \Delta^*_{ki}}{\widetilde{\sigma}_{ki}} \right|.$$

Using the simulated replications above, compute the bootstrap statistic

$$T^{(b)}_{\mathcal{I}} = \max_{(k,i)\in\mathcal{P}_{\mathcal{I}}} \left| \frac{S^{(b)}_k/g^{(b)}_k - S^{(b)}_i/g^{(b)}_i}{\widetilde{\sigma}_{ki}} \right|.$$

The simulated critical value is the empirical $(1-\alpha)$-quantile,

$$\widehat{q}_{1-\alpha} = \inf\left\{ t : \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\{T^{(b)}_{\mathcal{I}} \le t\} \ge 1-\alpha \right\}.$$

Using the bootstrap critical value, construct the pairwise confidence intervals for each pair $(k,i) \in \mathcal{P}_{\mathcal{I}}$ with endpoints

$$\widetilde{C}_L(k,i) = \widetilde{\Delta}_{ki} - \widehat{q}_{1-\alpha}\widetilde{\sigma}_{ki}, \quad \widetilde{C}_U(k,i) = \widetilde{\Delta}_{ki} + \widehat{q}_{1-\alpha}\widetilde{\sigma}_{ki}.$$

If these intervals simultaneously cover their respective score differences, a positive lower endpoint certifies that item k ranks above item i, while a negative upper endpoint certifies that it ranks below. We therefore count the items certified above and below each item of interest: $\widetilde{n}_1(i) = \sum_{k\neq i} \mathbf{1}\{\widetilde{C}_L(k,i) > 0\}$ and $\widetilde{n}_2(i) = \sum_{k\neq i} \mathbf{1}\{\widetilde{C}_U(k,i) < 0\}$. On the simultaneous coverage event, these counts imply $\widetilde{n}_1(i) + 1 \le r(i) \le n - \widetilde{n}_2(i)$ for every $i \in \mathcal{I}$. The resulting rank confidence sets are

$$\widetilde{\mathcal{R}}(i) = \{\widetilde{n}_1(i)+1, \ldots, n-\widetilde{n}_2(i)\}, \qquad i \in \mathcal{I}. \tag{3.11}$$

(ii) One-sided rank bounds and top-K inference. To certify that item i lies outside the top-K group, it suffices to obtain a lower confidence bound on $r(i)$ that exceeds K. We therefore focus on one-sided inference and construct such a rank bound from simultaneous lower bounds on Δ^*_{ki}. The corresponding one-sided target statistic is

$$T_{\mathcal{I},+} = \max_{(k,i)\in\mathcal{P}_\mathcal{I}} \frac{\widetilde{\Delta}_{ki} - \Delta^*_{ki}}{\widetilde{\sigma}_{ki}}.$$

Using the same simulated replications as above, its bootstrap analog is

$$T^{(b)}_{\mathcal{I},+} = \max_{(k,i)\in\mathcal{P}_\mathcal{I}} \frac{S^{(b)}_k/g^{(b)}_k - S^{(b)}_i/g^{(b)}_i}{\widetilde{\sigma}_{ki}}.$$

The sign agrees with the leading error in $\widetilde{\Delta}_{ki} - \Delta^*_{ki}$. Let

$$\widehat{q}^{\,\mathrm{lo}}_{1-\alpha} = \inf\left\{ t : \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\{T^{(b)}_{\mathcal{I},+} \le t\} \ge 1-\alpha \right\}$$

be the corresponding one-sided empirical quantile. The simultaneous lower confidence bounds for the score differences are $\widetilde{C}^{\,\mathrm{lo}}_L(k,i) = \widetilde{\Delta}_{ki} - \widehat{q}^{\,\mathrm{lo}}_{1-\alpha}\widetilde{\sigma}_{ki}$ for $(k,i) \in \mathcal{P}_\mathcal{I}$. On the simultaneous coverage event, $\widetilde{C}^{\,\mathrm{lo}}_L(k,i) > 0$ certifies that item k ranks above item i. Thus, define $\widetilde{n}^{\,\mathrm{lo}}_+(i) := \sum_{k\neq i} \mathbf{1}\{\widetilde{C}^{\,\mathrm{lo}}_L(k,i) > 0\}$. Since at least $\widetilde{n}^{\,\mathrm{lo}}_+(i)$ items rank above item i on this event, the lower rank bound and its associated confidence set are

$$\widetilde{R}^{\,\mathrm{lo}}(i) = 1 + \widetilde{n}^{\,\mathrm{lo}}_+(i), \qquad \widetilde{\mathcal{R}}^{\,\mathrm{lo}}(i) = \{\widetilde{R}^{\,\mathrm{lo}}(i), \dots, n\}. \tag{3.12}$$

For $K \in \{1,\dots,n\}$, consider the hypothesis testing problem

$$H_0 : r(i) \le K \qquad \text{versus} \qquad H_1 : r(i) > K. \tag{3.13}$$

We reject H_0 whenever $\widetilde{R}^{\,\mathrm{lo}}(i) > K$. By Theorem 4, the two-sided pairwise intervals and the one-sided lower bounds each have asymptotic simultaneous coverage $1-\alpha$ over $\mathcal{P}_\mathcal{I}$. Consequently, the induced rank confidence sets satisfy

$$\liminf_{n\to\infty} \min\left\{ \mathbb{P}\{r(i) \in \widetilde{\mathcal{R}}(i) \text{ for all } i \in \mathcal{I}\}, \quad \mathbb{P}\{r(i) \in \widetilde{\mathcal{R}}^{\,\mathrm{lo}}(i) \text{ for all } i \in \mathcal{I}\} \right\} \ge 1-\alpha.$$

4 Theoretical results

This section establishes theoretical guarantees for the procedures introduced in Section 3. We first establish privacy guarantees and estimation-error bounds, then derive the asymptotic expansion underlying inference for individual scores and pairwise differences, and finally establish simultaneous rank coverage and its implication for top-K inference.

The following proposition provides a privacy guarantee for the proposed estimators and the associated inferential procedures under the edge adjacency relation in Definition 2. The inferential outputs use only the private release, public parameters, and independent simulation randomness. They therefore inherit its privacy guarantee by post-processing, without an additional privacy allocation.

Proposition 1 (Privacy guarantee for the estimator and inferential procedures)**.** *If $\lambda > 8/\varepsilon$ and $\gamma > 1/\varepsilon$, then the estimator $\widetilde{\boldsymbol{\theta}}$ defined in* (3.1) *satisfies $(\varepsilon, 0)$-edge differential privacy. Moreover, the score confidence intervals in* (3.6) *and* (3.9)*, the pairwise confidence intervals in* (3.7) *and* (3.10)*, and the rank confidence sets in* (3.11) *and* (3.12) *are also $(\varepsilon, 0)$-edge differentially private, as are the associated top-K testing decisions.*

We use the fixed tuning for the statistical guarantees and inference validity,

$$\lambda = c_\lambda/\varepsilon, \qquad \gamma = c_\gamma/\varepsilon, \qquad c_\lambda > 8, \quad c_\gamma > 1, \tag{4.1}$$

where the constants are fixed. The constants in the probability and error bounds may depend on $\kappa, c_\lambda, c_\gamma$, but not on n, p, ε. Privacy alone permits larger tuning parameters; the fixed tuning in (4.1) supplies the upper bounds needed for statistical accuracy. We take limits along $n \to \infty$ with $p = p_n$ and $\varepsilon = \varepsilon_n$ as specified in each result. Statistical probability statements account for the randomness of the comparison graph, outcomes, and privacy perturbation, together with bootstrap simulation randomness where applicable.

Theorem 1 (Finite-sample ℓ_2 and entrywise error bounds)**.** *Under Assumption 1 and* (4.1)*, assume* $p \geq c_0 \log(n)/n$ *and* $\varepsilon \geq c_0 \log(n)^{1.5}/(np)$ *for a sufficiently large constant* $c_0 > 0$*. Then there exists a constant* $C > 0$ *such that the following inequalities hold with probability at least* $1 - O(n^{-7})$*:*

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2 \leq C\left(\frac{1}{p} + \frac{\log(n)^2}{np^2\varepsilon^2}\right) \quad \text{and} \quad \|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty^2 \leq C\left(\frac{\log(n)}{np} + \frac{\log(n)^2}{n^2p^2\varepsilon^2}\right).$$

Theorem 1 gives the same sampling and privacy rates for entrywise estimation as Theorem 2.1 of Cai et al. (2026a), under different tuning choices and conditions on the comparison graph and privacy budget. At a fixed privacy budget $0 < \varepsilon < 1$, their sufficient lower bound on p is smaller than ours by a factor of $\sqrt{\log n}$. When $np \gg (\log n)^2$, our theorem permits smaller privacy budgets, including values of ε for which the privacy term dominates the statistical term in the entrywise bound. This regime still includes sparse graphs, for example $p = (\log n)^3/n$. Theorem 1 also controls the squared Euclidean error of the centered scores. Its statistical term is of order $1/p$, compared with $\log n/p$ obtained by converting the entrywise guarantee of Cai et al. (2026a), while the privacy term has the same order in both bounds.

Cai et al. (2026a) derive their entrywise bound by tracking leave-one-out gradient iterates, following Chen et al. (2019). We instead use such iterates only to establish preliminary boundedness and then adapt the approach of Chen et al. (2022) to the perturbed likelihood. For each item, we compare the full fit with the reduced fit obtained by removing that item and its comparisons. Because the reduced fit is independent of the omitted comparisons, we can isolate the individual sampling fluctuation in its score equation. This decomposition yields the entrywise bound. Under the stronger conditions of Theorem 2, we use a further leave-two-out argument to derive the uniform asymptotic expansion.

Theorem 2 (Asymptotic expansion for private estimator)**.** *Let* $\widetilde{\boldsymbol{\theta}}$ *be the estimator defined*

in (3.1)*, let* $\bar{w} = n^{-1}\sum_{i=1}^{n} w_i$ *be the average privacy noise, and write*

$$S_i = \sum_{j \neq i} A_{ij}\{Y_{ij} - F(\theta_i^* - \theta_j^*)\}, \qquad g_i = \sum_{j \neq i} A_{ij} F'(\theta_i^* - \theta_j^*).$$

Under Assumption 1 and (4.1)*, assume* $np \gg (\log n)^{3/2}$ *and* $np\varepsilon^2 \gg (\log n)^3$*. Then, with probability tending to one, uniformly over* $i \in [n]$*,*

$$\widetilde{\theta}_i - \theta_i^* = (1 + r_{1,i})\frac{S_i}{g_i} + r_{2,i} - \frac{\bar{w}}{\gamma},$$

where the remainder vectors satisfy $\|\mathbf{r}_1\|_\infty = o_p(1)$ *and* $\|\mathbf{r}_2\|_\infty = o_p\{(np)^{-1/2}\}$*.*

Centering removes the common shift $-\bar{w}/\gamma$, which also cancels in pairwise differences. For the centered estimator, the leading term S_i/g_i and the graph condition $np \gg (\log n)^{3/2}$ agree with those of Gao et al. (2023). To control the expansion remainder uniformly, we adapt their leave-two-out argument to the perturbed likelihood. Within each leave-one-out fit, removing a second item makes the remaining fitted scores independent of that item's incident comparisons. In contrast, the reduced score equations preserve the ridge penalty and centered privacy perturbation. Under $np\varepsilon^2 \gg (\log n)^3$, the resulting ridge bias and centered perturbation are asymptotically negligible. The fixed-coordinate representation (3.5) motivates our bias correction and calibration, and the following proposition gives the finite-dimensional normal approximation.

Proposition 2 (Finite-dimensional Central Limit Theorem)**.** *Suppose the conditions of Theorem 2 hold. Then, for any prespecified subset* $\mathcal{I} \subseteq [n]$ *of fixed cardinality,*

$$\rho_i(\boldsymbol{\theta}^*)^{-1}\{\widetilde{\theta}_i - \mathrm{ave}(\widetilde{\boldsymbol{\theta}}) - \theta_i^*\} \xrightarrow{d} N(0,1) \text{ and } \rho_{ij}(\boldsymbol{\theta}^*)^{-1}\{(\widetilde{\theta}_i - \widetilde{\theta}_j) - (\theta_i^* - \theta_j^*)\} \xrightarrow{d} N(0,1),$$

uniformly over $i, j \in \mathcal{I}$ *with* $i \neq j$ *in the pairwise statement, where*

$$\rho_i(\boldsymbol{\theta}) = \Big\{\sum_{k \neq i} pF'(\theta_i - \theta_k)\Big\}^{-1/2} \text{ and } \rho_{ij}(\boldsymbol{\theta}) = \sqrt{\rho_i(\boldsymbol{\theta})^2 + \rho_j(\boldsymbol{\theta})^2}.$$

Proposition 2 uses oracle standard errors. The next theorem establishes coverage for the feasible Wald and bias-corrected Gaussian–Laplace intervals constructed in Section 3.2.

Theorem 3 (Plug-in Wald and bias-corrected mixture inference)**.** *Suppose that the conditions of Theorem 2 hold. For any fixed* $\alpha \in (0,1)$ *and prespecified* $\mathcal{I} \subseteq [n]$ *with fixed cardinality, the following statements hold.*

(i) Wald intervals. *The confidence intervals defined in* (3.6) *and* (3.7) *satisfy*

$$\max_{i\in\mathcal{I}} \left|\mathbb{P}\{\theta_i^* \in \mathrm{CI}_i(1-\alpha)\} - (1-\alpha)\right| \longrightarrow 0,$$

$$\max_{\substack{i,j\in\mathcal{I}\\ i\neq j}} \left|\mathbb{P}\{\theta_i^* - \theta_j^* \in \mathrm{CI}_{ij}(1-\alpha)\} - (1-\alpha)\right| \longrightarrow 0.$$

(ii) Bias-corrected mixture intervals. *The confidence intervals defined in* (3.9)–(3.10) *satisfy*

$$\max_{i\in\mathcal{I}} \left|\mathbb{P}\{\theta_i^* \in \mathrm{CI}_i^{(\mathrm{bc})}(1-\alpha)\} - (1-\alpha)\right| \to 0,$$

$$\max_{\substack{i,j\in\mathcal{I}\\ i\neq j}} \left|\mathbb{P}\{\theta_i^* - \theta_j^* \in \mathrm{CI}_{ij}^{(\mathrm{bc})}(1-\alpha)\} - (1-\alpha)\right| \to 0.$$

Theorem 3 establishes that both the classical plug-in Wald procedure and the bias-corrected Gaussian–Laplace construction attain asymptotic marginal coverage $1-\alpha$ for prespecified scores and pairwise differences. Our coverage guarantees assume negligible asymptotic privacy and regularization effects, and we assess finite-sample performance in Section 5. Beyond inference for individual pairwise comparisons, inference for even a single item's rank depends on how its score compares with the scores of all $n-1$ opponents. Thus, our bootstrap procedure controls the relevant family of score differences simultaneously.

Theorem 4 (Simultaneous rank inference)**.** *Under Assumption 1 and* (4.1)*, assume that* $np \gg (\log n)^5$ *and* $np\varepsilon^2 \gg (\log n)^3$*, and that the number of bootstrap replications satisfies* $B \to \infty$*. Fix* $\alpha \in (0,1)$ *and a nonempty prespecified set* $\mathcal{I} \subseteq [n]$*, whose cardinality may grow with* n*. Then the following statements hold.*

(i) Two-sided inference. *The simultaneous intervals with endpoints* $\widetilde{C}_L(k,i)$, $\widetilde{C}_U(k,i)$ *satisfy*

$$\mathbb{P}\left\{\Delta^*_{ki} \in [\widetilde{C}_L(k,i), \widetilde{C}_U(k,i)] \text{ for all } (k,i) \in \mathcal{P}_{\mathcal{I}}\right\} \to 1-\alpha.$$

Consequently, the rank confidence sets $\widetilde{\mathcal{R}}(i)$ *satisfy*

$$\liminf_{n\to\infty} \mathbb{P}\left\{r(i) \in \widetilde{\mathcal{R}}(i) \text{ for all } i \in \mathcal{I}\right\} \geq 1-\alpha.$$

(ii) One-sided inference. *The simultaneous lower bounds* $\widetilde{C}^{\mathrm{lo}}_L(k,i)$ *satisfy*

$$\mathbb{P}\left\{\Delta^*_{ki} \geq \widetilde{C}^{\mathrm{lo}}_L(k,i) \text{ for all } (k,i) \in \mathcal{P}_{\mathcal{I}}\right\} \to 1-\alpha.$$

Consequently, the one-sided rank confidence sets $\widetilde{\mathcal{R}}^{\mathrm{lo}}(i)$ *satisfy*

$$\liminf_{n\to\infty} \mathbb{P}\left\{r(i) \in \widetilde{\mathcal{R}}^{\mathrm{lo}}(i) \text{ for all } i \in \mathcal{I}\right\} \geq 1-\alpha.$$

The contrast bounds have asymptotic simultaneous coverage $1-\alpha$, while the induced rank coverage can be conservative because simultaneous contrast coverage is sufficient for rank coverage. We achieve this via a Gaussian approximation combined with a bootstrap for the maximum of pairwise score contrasts. Theorem 4 uses the same privacy condition $np\varepsilon^2 \gg (\log n)^3$ as Theorem 2 and strengthens the graph condition to $np \gg (\log n)^5$, which together ensure that both the Gaussian approximation error and privacy-induced perturbation are negligible relative to the anti-concentration scale of the max statistic (Chernozhukov et al., 2022). The empirical bootstrap approximation requires $B \to \infty$ without additional privacy composition cost across the bootstrapped samples.

5 Simulation results

In this section, we evaluate the finite-sample performance of the confidence intervals for scores and pairwise score differences, rank confidence sets, and top-K tests developed in

Section 3. We generated comparison graphs from $G(n,p)$ and, conditional on each graph, generated one independent Bernoulli comparison on each observed edge under the BT model. We set $K = 10$ and constructed the true scores as

$$a_i = \begin{cases} \dfrac{1}{2} + s\dfrac{K-i}{K-1}, & 1 \le i \le K, \\ -\dfrac{1}{2} - s\dfrac{i-K-1}{n-K-1}, & K < i \le n, \end{cases} \qquad \theta_i^* = a_i - \frac{1}{n}\sum_{j=1}^{n} a_j. \tag{5.1}$$

This construction gives $r(i) = i$, a boundary gap $\theta_K^* - \theta_{K+1}^* = 1$, and a score span of s within each group. Moreover, $\|\boldsymbol{\theta}^*\|_\infty \le 1 + 3s/2$, so the choices of s specified below satisfy Assumption 1 with $\kappa = 4$. The true score vector was held fixed across replications within each configuration. Throughout, we set $\alpha = 0.05$ and varied $\varepsilon \in \{\infty, 5, 2.5, 1\}$. For finite ε, we took $c_\gamma = 4$ and $c_\lambda = 8.1$ in (4.1). The non-private benchmark, denoted by $\varepsilon = \infty$, used $\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ with $\gamma = \lambda = 0$. We evaluated each configuration over $R = 500$ independent Monte Carlo replications. Appendix D provides additional results on estimation bias and distribution diagnostics of both uncorrected and bias-corrected estimators, and examines the sensitivity of rank inference and top-K placement to the edge sampling probability p.

Individual and pairwise inference. We first examined confidence intervals for the boundary score θ_K^* and the boundary contrast $\theta_K^* - \theta_{K+1}^*$, with $n \in \{500, 700, 1000\}$, $p = 0.4$, and $s = 0.5$. We compared the Wald intervals in (3.6)–(3.7) with the bias-corrected (BC) intervals in (3.9)–(3.10). For the BC intervals, we approximated the Gaussian–Laplace quantiles using $M = 50{,}000$ Gaussian and Laplace reference draws, generated once independently of the experimental data and reused across replications. We computed the plug-in quantities used to transform these draws separately for each fit. Table 1 reports empirical coverage, with mean interval length in parentheses. Coverage was calculated as the proportion of replications in which the interval contained its target, and length was averaged over the R replications. The procedures coincide in the non-private benchmark.

Table 1: Empirical coverage probabilities and mean lengths of the nominal 95% Wald and bias-corrected (BC) intervals for the individual score θ^*_K and score contrast $\theta^*_K - \theta^*_{K+1}$, with $K = 10$, $p = 0.4$ and $s = 0.5$. Each entry reports empirical coverage, with mean interval length in parentheses. Results are based on $R = 500$ replications.

n	ε	Individual score θ^*_K		Score contrast $\theta^*_K - \theta^*_{K+1}$	
		Wald	BC	Wald	BC
500	∞	0.968 (0.669)	0.968 (0.669)	0.960 (0.876)	0.960 (0.876)
	5	0.930 (0.665)	0.948 (0.716)	0.940 (0.873)	0.952 (0.929)
	2.5	0.914 (0.662)	0.958 (0.850)	0.926 (0.871)	0.978 (1.092)
	1	0.720 (0.663)	0.960 (1.551)	0.708 (0.877)	0.946 (1.937)
700	∞	0.944 (0.565)	0.944 (0.565)	0.946 (0.739)	0.946 (0.739)
	5	0.938 (0.560)	0.952 (0.589)	0.944 (0.736)	0.956 (0.768)
	2.5	0.892 (0.560)	0.942 (0.675)	0.916 (0.736)	0.954 (0.873)
	1	0.740 (0.560)	0.946 (1.128)	0.758 (0.738)	0.952 (1.415)
1000	∞	0.948 (0.472)	0.948 (0.472)	0.944 (0.617)	0.944 (0.617)
	5	0.938 (0.471)	0.944 (0.487)	0.954 (0.616)	0.960 (0.634)
	2.5	0.916 (0.470)	0.944 (0.540)	0.912 (0.616)	0.940 (0.695)
	1	0.736 (0.469)	0.926 (0.822)	0.764 (0.616)	0.938 (1.036)

At each value of n, the Wald interval lengths changed little as ε decreased, while coverage deteriorated substantially, falling to 0.708–0.764 at $\varepsilon = 1$. This pattern was consistent with the Wald calibration accounting only for the leading sampling variation, although regularization bias and privacy perturbation could remain significant in finite samples. The BC procedure combined regularization-bias correction with convolution calibration and produced wider intervals with coverage substantially closer to the nominal level. At fixed p, increasing n provided more comparisons per item and shortened the intervals for both procedures. For the BC procedure, the proportional increase in mean length over the non-private benchmark also declined at each fixed privacy level, indicating a smaller relative cost of privacy. We assessed this cost against the non-private benchmark because the shorter private Wald intervals suffered from undercoverage.

Rank inference. We next examined the bootstrap rank confidence sets $\widetilde{\mathcal{R}}(i)$ in (3.11), with $n \in \{600, 1000, 1400\}$, $p = 0.6$, and $s = 2$. For each item $i \in \{2, 5, 8\}$, we set $\mathcal{I} = \{i\}$

and calibrated the comparisons with all other items using $B = 7000$ bootstrap draws conditional on the fitted estimator. Coverage therefore refers to each rank, rather than simultaneous coverage of the three displayed ranks. Table 2 reports empirical coverage and mean rank endpoints for each item. The bracketed endpoint averages describe both the locations and sizes of the rank confidence sets.

Table 2: Empirical coverage probabilities and mean lower and upper endpoints of the nominal 95% rank confidence sets $\widetilde{\mathcal{R}}(i)$ for $i \in \{2, 5, 8\}$, with $K = 10$, $p = 0.6$ and $s = 2$. Each target was calibrated separately using $\mathcal{I} = \{i\}$. For each item, the two columns report empirical coverage and the mean lower and upper endpoints, respectively; the endpoint pair is shown in brackets. Results are based on $R = 500$ replications and $B = 7000$ bootstrap draws per replication.

n	ε	$i = 2$		$i = 5$		$i = 8$	
		Coverage	Mean endpoints	Coverage	Mean endpoints	Coverage	Mean endpoints
600	∞	1.000	[1.00, 6.36]	1.000	[1.11, 8.56]	1.000	[3.24, 9.96]
	5	1.000	[1.00, 6.74]	0.998	[1.07, 8.75]	1.000	[2.88, 9.95]
	2.5	0.996	[1.00, 7.04]	0.994	[1.13, 8.86]	1.000	[2.49, 9.95]
	1	0.962	[1.14, 7.53]	0.946	[1.44, 9.14]	0.970	[2.56, 12.74]
1000	∞	0.998	[1.00, 5.46]	1.000	[1.40, 7.88]	1.000	[4.59, 9.84]
	5	1.000	[1.00, 5.80]	1.000	[1.29, 7.97]	1.000	[4.38, 9.87]
	2.5	0.998	[1.02, 6.06]	0.998	[1.30, 7.99]	1.000	[4.12, 9.85]
	1	0.962	[1.11, 6.37]	0.982	[1.51, 8.28]	0.986	[3.67, 9.85]
1400	∞	1.000	[1.00, 4.95]	1.000	[1.66, 7.33]	1.000	[5.30, 9.73]
	5	0.998	[1.00, 5.09]	1.000	[1.62, 7.49]	1.000	[5.16, 9.72]
	2.5	1.000	[1.01, 5.38]	1.000	[1.53, 7.57]	1.000	[4.98, 9.75]
	1	0.966	[1.10, 5.70]	0.994	[1.66, 7.82]	0.994	[4.44, 9.64]

Rank coverage was close to or above the nominal level throughout Table 2, approaching one even in the non-private benchmark. This pattern is consistent with the distinction between score-difference coverage and rank coverage in Theorem 4, as also documented in non-private rank inference (Fan et al., 2025). In particular, converting simultaneous confidence intervals for score differences into rank confidence sets can introduce additional conservativeness. The rank confidence sets narrowed as n increased. Stronger privacy generally enlarged the sets, although at $n = 1400$ the mean cardinality under $\varepsilon = 1$

exceeded the non-private benchmark by less than one rank for each of the three displayed items. Even without privacy, the sets still spanned several rank positions on average, indicating that uncertainty about the items' exact ranks remained.

Top-K inference. Finally, we evaluated the proposed tests for the individual top-K problem in (3.13), with $n = 400$, $p = 0.6$, and $s = 0.5$. We used the one-sided rank bound $\widetilde{R}^{\mathrm{lo}}(i)$ in (3.12) and rejected when $\widetilde{R}^{\mathrm{lo}}(i) > K$. Each item was calibrated separately using $\mathcal{I} = \{i\}$ and $B = 7000$ bootstrap draws. Figure 1 shows the empirical rejection probabilities near the top-K boundary and over the full range of ranks, calculated as the proportion of rejections across the R replications. Rejection probability measures type I error for the items with $r(i) \leq K$ and power for those with $r(i) > K$.

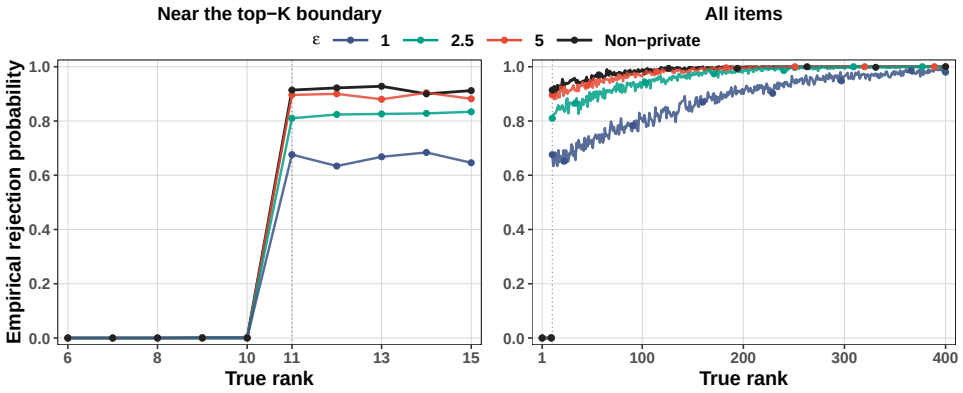


Figure 1: Empirical rejection probabilities of the individual top-K placement tests over $R = 500$ replications, with $n = 400$, $p = 0.6$, $K = 10$, $s = 0.5$, and $B = 7000$ bootstrap draws per replication. The left panel shows ranks 6–15 and the right panel shows all items. Colored curves give results for finite ε; the black curve gives the non-private benchmark. Vertical dotted lines mark the first rank outside the top-K group. Each item was calibrated separately.

Empirical rejection probabilities for the ten null items were below the nominal level at every privacy budget, consistent with the asymptotic type I error control implied by Theorem 4. Under moderate privacy, power was close to the non-private benchmark over much of the rank range. More pronounced losses occurred under stronger privacy, particularly for items near the top-K boundary, whereas power approached one farther from the boundary

even at the strongest privacy level considered. In this configuration, the principal cost of stronger privacy was therefore a reduced ability to identify and exclude items near the boundary, while erroneous exclusions of true top-K items remained rare.

6 Real data analysis

Creative-writing platforms face the practical problem of screening a growing pool of language models before committing resources to detailed editorial evaluation. For short-form drama, such as that distributed through platforms including *Hongguo* and *ReelShort*, language models could assist by drafting scenes for subsequent editorial development. We consider a creative-writing task that asks for a dramatic scene of 500–700 words based on *Hamlet* under the premise that Claudius neither killed King Hamlet nor arranged his death. The familiar characters and specified premise provide a common narrative foundation for comparing models' ability to develop coherent dramatic conflict, characterization, and dialogue. Pairwise comparisons of the generated scenes provide evidence for selecting models suited to this task. An initial screening stage may exclude individual models whose performance falls outside a target top-K group, while limiting the risk of discarding a genuinely top-K candidate. When screening reports are shared with model providers or publishing partners, the underlying editorial or reader assessments may need to remain confidential to protect individual comparison records. This motivates top-K inference that accounts for both sampling uncertainty and privacy perturbation under edge differential privacy.

Using the October 15, 2026 snapshot of the Arena text leaderboard (Chiang et al., 2024), we analyzed $n = 99$ models with accessible application programming interfaces (APIs) from the 402 listed models. The original pool contained 100 models, but one lost every comparison and was excluded; the privacy guarantee applies to each fit on the $n = 99$

models. We constructed a complete set of pairwise comparisons among these models. For each pair (i, j), both models generated fresh responses independently of those used for other pairs. The responses were then evaluated by GPT-6.1 Sol using the task prompt and task-specific criteria, with model identities replaced by anonymous labels. We recorded $Y_{ij} = 1$ if model i was preferred over model j, and $Y_{ij} = 0$ otherwise. Here we used an automated judge for simplicity, and the proposed procedure also accommodates human pairwise judgments and does not require the prompt or evaluation criteria to be public.

We then illustrate the proposed inference procedures using the generated comparison data under a working BT model. With the observed judgments fixed, we sampled $R = 500$ graphs from $G(n, p)$ at each $p \in \{0.6, 0.8\}$ and fitted the model on each graph at $\varepsilon \in \{\infty, 5, 2.5, 1\}$, using the tuning constants in Section 5. Each finite ε specifies the privacy budget for one fit, with fresh perturbations across repetitions. For $K = 10$ and $\alpha = 0.05$, we constructed the one-sided lower rank bound $\widetilde{R}^{\mathrm{lo}}(i)$ in (3.12) and excluded model i when $\widetilde{R}^{\mathrm{lo}}(i) > K$. Each model was calibrated separately using $\mathcal{I} = \{i\}$ and $B = 7000$ bootstrap draws. These repetitions assess the sensitivity of individual rank inference to comparison subsampling and privacy perturbation, conditional on the observed judgments. Table 3 presents four illustrative models spanning different positions in the non-private ordering estimated from all observed comparisons. The non-private summaries include only repetitions yielding finite MLEs, whereas the private summaries use all 500 repetitions.

At each fixed p, GPT-5.5 and Claude Sonnet 4.6 were occasionally excluded under stronger privacy, although their median lower rank bounds remained at 1. For Kimi-K2, at $p = 0.8$, the exclusion frequency increased from 2.2% at $\varepsilon = 5$ to 15.4% at $\varepsilon = 1$, while the median bound remained at 4. Thus, the increase in exclusions reflected changes in the upper tail rather than a shift in the typical bound beyond the top-10 threshold. For DeepSeek-

Table 3: One-sided lower rank bounds $\widetilde{R}^{\mathrm{lo}}(i)$ and top-10 exclusion rates for four selected models at $\alpha = 0.05$. For each model and setting, we report the median and interquartile range $[Q_{25}, Q_{75}]$ of $\widetilde{R}^{\mathrm{lo}}(i)$ across repetitions. Exclusion (%) is the percentage of repetitions in which $\widetilde{R}^{\mathrm{lo}}(i) > 10$. Results for each finite ε use all 500 fits, whereas the non-private benchmark ($\varepsilon = \infty$) includes only the 233 and 382 fits yielding finite MLEs at $p = 0.6$ and 0.8, respectively.

Model	ε	$p = 0.6$		$p = 0.8$	
		Median $[Q_{25}, Q_{75}]$	Exclusion (%)	Median $[Q_{25}, Q_{75}]$	Exclusion (%)
GPT-5.5	∞	1 [1, 1]	0.0	1 [1, 1]	0.0
	5	1 [1, 1]	0.0	1 [1, 1]	0.0
	2.5	1 [1, 1]	0.0	1 [1, 1]	0.0
	1	1 [1, 1]	4.6	1 [1, 1]	3.6
Claude Sonnet 4.6	∞	1 [1, 1]	0.0	1 [1, 2]	0.0
	5	1 [1, 1]	0.0	1 [1, 1]	0.0
	2.5	1 [1, 1]	0.8	1 [1, 1]	0.4
	1	1 [1, 3]	4.6	1 [1, 3]	3.8
Kimi-K2	∞	6 [5, 7]	2.1	7 [7, 8]	2.4
	5	2 [1, 4]	0.8	4 [3, 6]	2.2
	2.5	2 [1, 3]	3.4	3 [2, 5]	3.8
	1	3 [1, 6]	14.0	4 [2, 7]	15.4
DeepSeek-V3.1	∞	21 [19, 23]	100.0	23 [22, 24]	100.0
	5	14 [11, 17]	80.6	19 [16, 20]	98.4
	2.5	11 [8, 16]	54.8	16 [13, 20]	86.8
	1	9 [4, 17]	44.0	13 [7, 19]	60.4

V3.1, whose bounds mostly exceeded 10 at $\varepsilon = 5$, stronger privacy reduced the median and widened the interquartile range across the threshold, leading to more non-exclusions. These patterns are consistent with the combined effects of stronger regularization and larger privacy perturbations in the private score estimators. Increasing p partly offset this loss of exclusion evidence for DeepSeek-V3.1: at $\varepsilon = 1$, its exclusion frequency rose from 44.0% to 60.4% as p increased from 0.6 to 0.8. This suggests that additional comparisons strengthen the information about score differences relative to the regularization and privacy perturbations at a fixed privacy budget. Appendix E.2 presents a complementary analysis of a story-beginning task from MT-Bench (Zheng et al., 2023) that examines the same four illustrative models. The comparison illustrates how the evidence for individual top-K

exclusion can vary across tasks, highlighting the importance of choosing evaluation tasks and criteria suited to the intended application.

7 Discussion

Motivated by the use of pairwise comparisons to evaluate LLMs, this paper develops a framework for ranking inference that protects individual preference records. Under the BT model, we establish estimation guarantees and asymptotically valid procedures for score and rank inference. Simulation studies demonstrate the benefits of the proposed bias correction and calibration for finite-sample coverage. The real-data application illustrates how the proposed procedures can support selecting LLMs for creative writing.

Several directions are worth further investigation. First, future work could extend current inference to stronger privacy regimes in which regularization bias can dominate statistical sampling variation. Establishing coverage guarantees for the proposed bias-aware confidence intervals in these regimes would require sharper distributional approximations and tighter control of bias-correction error. Second, heterogeneous comparison designs and covariate-dependent comparison models would accommodate more flexible evaluation settings. Incorporating prompt characteristics could capture variation in relative model performance across tasks, while online extensions would require accounting for adaptive comparison selection and privacy expenditure across successive releases. Third, extending privacy protection from individual comparison records to all judgments contributed by a participant would require user-level differential privacy. This extension would involve controlling the influence of multiple contributions, including repeated evaluations of the same pair, and accounting for dependence among judgments in the inferential analysis. We leave these to future work.

References

Aradau, C. and Blanke, T. (2026). Making GenAI valuable: Benchmarks, singularities, and the enrichment economy. *Big Data & Society*.

Avella-Medina, M. (2021). Privacy-preserving parametric inference: A case for robust statistics. *Journal of the American Statistical Association*, 116(534):969–983.

Avella-Medina, M., Bradshaw, C., and Loh, P.-L. (2023). Differentially private inference via noisy optimization. *The Annals of Statistics*, 51(5):2067–2092.

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3-4):324–345.

Cai, T. T., Chakraborty, A., and Wang, Y. (2026a). Optimal differentially private ranking from pairwise comparisons. *Journal of the American Statistical Association*, pages 1–15. Published online.

Cai, Z., Li, S., Xia, X., and Zhang, L. (2026b). Differentially private estimation and inference in high-dimensional regression with FDR control. *Journal of Machine Learning Research*, 27(81):1–54.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. (2011). Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(29):1069–1109.

Chen, P., Gao, C., and Zhang, A. Y. (2022). Partial recovery for top-K ranking: Optimality of MLE and suboptimality of the spectral method. *The Annals of Statistics*, 50(3):1618–1652.

Chen, Y., Fan, J., Ma, C., and Wang, K. (2019). Spectral method and regularized MLE are both optimal for top-K ranking. *The Annals of Statistics*, 47(4):2204–2235.

Chernozhukov, V., Chetverikov, D., Kato, K., and Koike, Y. (2022). Improved central limit theorem and bootstrap approximations in high dimensions. *The Annals of Statistics*, 50(5):2562–2586.

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhu, B., Zhang, H., Jordan, M., Gonzalez, J. E., and Stoica, I. (2024). Chatbot Arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 8359–8388. PMLR.

Dong, P., Han, R., Jiang, B., and Xu, Y. (2026). Statistical ranking with dynamic covariates. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 88(1):221–238.

Duineveld, C. A. A., Arents, P., and King, B. M. (2000). Log-linear modelling of paired comparison data from consumer tests. *Food Quality and Preference*, 11(1–2):63–70.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer.

Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407.

Fan, J., Lou, Z., Wang, W., and Yu, M. (2025). Ranking inferences based on the top choice of multiway comparisons. *Journal of the American Statistical Association*, 120(549):237–250.

Gao, C., Shen, Y., and Zhang, A. Y. (2023). Uncertainty quantification in the bradley–terry–luce model. *Information and Inference: A Journal of the IMA*, 12(2):1073–1140.

Gu, J. and Cai, T. T. (2026). Learning sparse support under differential privacy: Adaptive algorithms and minimax limits. *arXiv preprint arXiv:2610.05867*.

Han, R., Luo, Y., and Xu, Y. (2026). Bradley–Terry model under general comparison graphs. arXiv preprint arXiv:2610.06237.

Han, R., Tang, W., and Xu, Y. (2024). Statistical inference for pairwise comparison models. arXiv preprint arXiv:2401.08463.

Han, R., Xu, Y., and Chen, K. (2023). A general pairwise comparison model for extremely sparse networks. *Journal of the American Statistical Association*, 118(544):2422–2432.

Han, R., Ye, R., Tan, C., and Chen, K. (2020). Asymptotic theory of sparse Bradley–Terry model. *The Annals of Applied Probability*, 30(5):2491–2515.

Hatzinger, R. and Mazanec, J. A. (2007). Measuring the part worth of the mode of transport in a trip package: An extended Bradley–Terry model for paired-comparison conjoint data. *Journal of Business Research*, 60(12):1290–1302.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2021). Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Iyengar, R., Near, J. P., Song, D., Thakkar, O., Thakurta, A., and Wang, L. (2019). Towards practical differentially private convex optimization. In *2019 IEEE symposium on security and privacy (SP)*, pages 299–316. IEEE.

Karwa, V., Raskhodnikova, S., Smith, A., and Yaroslavtsev, G. (2011). Private analysis of graph structure. *Proceedings of the VLDB Endowment*, 4(11):1146–1157.

Liu, Y., Fang, E. X., and Lu, J. (2023). Lagrangian inference for ranking problems. *Operations Research*, 71(1):202–223.

Negahban, S., Oh, S., and Shah, D. (2017). Rank Centrality: Ranking from pairwise comparisons. *Operations Research*, 65(1):266–287.

Ouyang, Y., Jing, L., Qiuping, W., and Zhimeng, X. (2025). Asymptotics in the Bradley–Terry model for networks with a differentially private degree sequence. *Communications in Statistics—Theory and Methods*, 54(2):437–456.

Sala, A., Zhao, X., Wilson, C., Zheng, H., and Zhao, B. Y. (2011). Sharing graphs using differentially private graph models. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pages 81–98.

Ullman, J. and Sealfon, A. (2019). Efficiently estimating erdos-renyi graphs with node differential privacy. *Advances in Neural Information Processing Systems*, 32.

Wasserman, L. and Zhou, S. (2010). A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389.

Xia, X., Zhang, L., and Cai, Z. (2025). Statistical inference for differentially private stochastic gradient descent. *arXiv preprint arXiv:2507.20560*.

Xu, S., Sun, W. W., and Cheng, G. (2025). Rate-optimal rank aggregation with private pairwise rankings. *Journal of the American Statistical Association*, 120(550):737–750.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36.