

# 000 001 FEWER BATTLES, MORE GAIN: AN INFORMATION- 002 EFFICIENT FRAMEWORK FOR ARENA-BASED LLM 003 EVALUATION 004 005

006 **Anonymous authors**

007 Paper under double-blind review  
008  
009  
010  
011

## ABSTRACT

013 Arena-based evaluation has become a key method for assessing large language  
014 models (LLMs) through head-to-head model comparisons, closely reflecting hu-  
015 man preferences. However, current arena rating systems (e.g., ELO rating system)  
016 often suffer from inefficiencies due to exhaustive or random model pair annota-  
017 tions, leading to redundant evaluations, longer evaluation times, and lower over-  
018 all efficiency. To address these challenges, we propose a novel adaptive model-  
019 pair selection algorithm. By leveraging the asymptotic normality of LLM abil-  
020 ity estimation under sparse conditions, our approach strategically selects high-  
021 value model pairs, focusing on confrontations with the lowest variance. Specif-  
022 ically, we introduce Fisher information as a metric to guide model pair selec-  
023 tion, optimizing the evaluation process through A-optimality and D-optimality.  
024 A-optimality minimizes estimation variance, ensuring balanced reliability across  
025 models, while D-optimality reduces uncertainty by maximizing the determinant  
026 of the Fisher Information Matrix. Extensive experiments on both simulated  
027 and real-world datasets demonstrate that our method outperforms existing ap-  
028 proaches in terms of information efficiency and result reliability. Notably, our  
029 method offers a flexible, general toolkit that can be easily integrated into existing  
030 arena-based platforms, greatly improving scalability and efficiency for large-scale  
031 LLM evaluations. Our code is publicly available to promote reproducibility at  
032 <https://anonymous.4open.science/r/Adaptive-Arena-870D>.  
033

## 034 1 INTRODUCTION

035  
036 The rapid advancements in large language models (LLMs) have led to their widespread use in vari-  
037 ous fields, including content generation (Wang et al., 2024a), customer service, and intelligent tutor-  
038 ing (Kasneci et al., 2023). As LLM architectures (Liu et al., 2024a), fine-tuned variants (Hu et al.,  
039 2023), and domain-specific models (Patil & Gudivada, 2024) continue to evolve, thus the demand  
040 for efficient and scalable evaluation frameworks is growing (Mizrahi et al., 2024; Zhou et al., 2023;  
041 Li et al., 2024b). Among existing evaluation methods, arena-based evaluation is notable (Chiang  
042 et al., 2024): unlike traditional benchmark-based approaches (Wang et al., 2024b; Mirzadeh et al.,  
043 2024) that rely on predefined prompts and fixed answers, arena-based frameworks assess models  
044 through head-to-head confrontations. This approach captures real-world performance dynamics and  
045 produces rankings that closely align with human judgments (Zheng et al., 2023; Luo et al., 2024),  
046 making it a valuable tool for assessing LLMs' practical utility.

047 However, current arena-based evaluation methods face significant issues. First, is the **efficiency** (Luo  
048 et al., 2024; Spangher et al., 2025). As shown in Figure 1, these frameworks often rely on exhaustive  
049 or random pairwise annotations to generate reliable rankings, which becomes increasingly resource-  
050 intensive. Second, is the **scalability**. As the number of newly proposed models grows, platforms  
051 like Chatbot Arena (Chiang et al., 2024), hosting over 190 models, require tens of thousands of  
052 confrontations to produce reliable rankings (Patil & Gudivada, 2024). This not only consumes  
053 considerable resources but also results in substantial evaluation cycles, making it difficult to keep up  
with the rapid development of LLMs (Min et al., 2025).

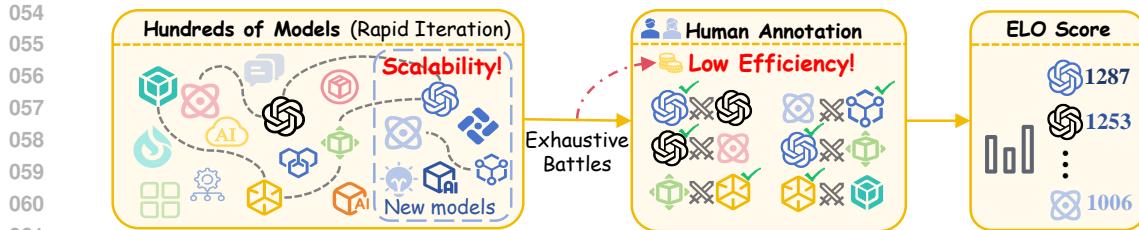


Figure 1: The diagram of Arena-based Evaluation. Models engage in pairwise battles, annotators label the outcomes, and the ELO ratings are then estimated based on these annotations.

To address these challenges, we propose a novel adaptive arena-based evaluation method. Leveraging the asymptotic normality of ability estimation under sparse conditions, we can strategically select high-value model pairs for evaluation and improve efficiency by minimizing the variance of the asymptotic distribution. Specifically, we introduce two key methods: A-optimality, which minimizes the variance in model estimates, and D-optimality, which reduces uncertainty by maximizing the determinant of the Fisher Information Matrix (Rissanen, 1996; Ly et al., 2017). Through these techniques, we achieve *more gain* in each annotation, leading to *fewer battles*.

Our contributions are as follows:

- We first introduce the concept of efficiency in arena-based LLM evaluation, using statistical uncertainty to minimize redundant evaluations, significantly improving both speed and resource usage in ranking.
- We apply A-optimality and D-optimality to model-pair selection, backed by rigorous theoretical analysis, to reduce annotation costs while preserving ranking reliability.
- Through extensive experiments, we show that our method outperforms existing approaches in efficiency and reliability, and can seamlessly integrate with current arena-based platforms. This makes our method not just a theoretical innovation, but a practical and scalable solution for large-scale LLM evaluations.

## 2 RELATED WORK

**Arena-based Evaluation** Arena-based evaluation is a key approach in LLM evaluation, distinguished from traditional benchmark methods by its focus on direct head-to-head model competitions (Wang et al., 2025). This design provides a more realistic measure of models' relative abilities (Zellers et al., 2019; Hendrycks et al., 2020; Cobbe et al., 2021). Unlike traditional approaches that use predefined datasets (Liang et al., 2022; Liu et al., 2024b), arena settings emphasize dynamic model confrontations that better reflect real-world usage scenarios. Ranking systems are central to Arena-based evaluation (Busa-Fekete et al., 2014; Szörényi et al., 2015; Chernoff, 1992), with most methods relying on the ELO Rating System to estimate model proficiency (Coulom, 2007; Pelánek, 2016). However, ELO's sensitivity to sample order can lead to instability in performance estimates (Boubdir et al., 2024). To address this, recent studies have proposed adaptations, such as statistical refinements (Ameli et al., 2024; Yin et al., 2024; Liu et al., 2025) and bias-mitigation methods like UDA (Zhang et al., 2025). Notably, all existing Arena-based methods aim to improve ranking accuracy but overlook the critical issue of annotation efficiency. Whether using standard ELO, adapted variants, or frameworks like UDA, these approaches often involve unnecessary confrontations, leading to low annotation efficiency and high evaluation overhead.

**Efficiency of LLM Evaluation** Efficiency has become a key focus in LLM evaluation due to the high costs of annotating large datasets and conducting extensive tests (Marion et al., 2023). Existing methods (Xie et al., 2023; Chung et al., 2023; Saranathan et al., 2024; Li et al., 2024a) share the goal of reducing the number of evaluation examples while maintaining accuracy, but they are designed primarily for traditional static question-answering (QA) tasks, where models are assessed against predefined prompts and answers. For example, Polo et al. (2024a) introduced TinyBenchmarks, which evaluates LLMs with fewer examples than traditional benchmarks by prioritizing information-dense samples, thus reducing annotation and computational costs while maintaining

108 reliable performance estimation. Similarly, Kipnis et al. (2024) proposed MetaBench, which selects  
 109 diverse, compact prompts based on statistical analysis, avoiding redundant evaluations while cap-  
 110 turing core model abilities. Based on this, Polo et al. (2024b) developed an efficient multi-prompt  
 111 evaluation method, which optimizes prompt selection across rounds to reduce redundant data collec-  
 112 tion and computation without sacrificing accuracy. However, these efficiency-driven methods are not  
 113 applicable to dynamic Arena-based evaluation for two key reasons: First, Arena scenarios involve  
 114 pairwise head-to-head model confrontations rather than static QA tasks; second, rankings in Arena  
 115 settings rely on iterative human judgments (Kruglanski & Ajzen, 1983), breaking the assumption of  
 116 fixed criteria inherent in traditional QA methods. Therefore, the dynamic and interactive nature of  
 117 Arena evaluation makes current efficiency methods unsuitable.

118 Given these gaps—existing efficiency methods being incompatible with dynamic Arena scenarios,  
 119 and Arena-based methods neglecting annotation efficiency—our study aims to address these issues.  
 120 We propose optimizing Arena-based evaluation through strategic sample selection techniques to  
 121 both accelerate system convergence and improve overall evaluation efficiency.

122

### 123 3 ADAPTIVE ARENA METHOD

124

#### 125 3.1 MODELING ELO RATINGS

126

127 Arena-based evaluation has emerged as a key approach for LLM evaluation, where models com-  
 128 pete head-to-head on benchmarks or datasets, with outcomes annotated by evaluators. Let  $\mathcal{S} =$   
 129  $\{(i, j, w_{ij}) | i < j \in [N]\}$  represent the comparative records, where  $N$  is the total number of mod-  
 130 els being evaluated. Each element  $(i, j, w_{ij}) \in \mathcal{S}$  contains important information: it indicates that  
 131 model  $i$  and model  $j$  have engaged in a competition, and  $w_{ij}$  is the annotation result. Specifically,  
 132 if  $w_{ij} = 1$ , model  $i$  wins; if  $w_{ij} = 0$ , model  $j$  wins; and if  $w_{ij} = 0.5$ , it indicates a tie. The goal of  
 133 arena-based evaluation is to estimate the ranking scores  $(r_1, \dots, r_N)$ . Since the ranking is relative,  
 134 we assume that the ability of the last model is 0, *i.e.*,  $r_N = 0$ , and in the following, we focus on the  
 135 ability estimation of the first  $N - 1$  models, denoted as  $\mathbf{r} = (r_1, \dots, r_{N-1})$ .

136

137 To model the win probabilities, we use the Bradley-Terry model (Rao & Kupper, 1967), which  
 138 forms the statistical foundation for modern ELO-style rating systems. It posits that the probability  
 139 of model  $i$  winning against model  $j$ , denoted  $P_{ij}$ , is a logistic function of their ability difference:

140

$$P_{ij} = \frac{1}{1 + e^{-C(r_i - r_j)}}, \quad (1)$$

141

142 where  $C$  is a scaling constant.

143

144 With the obtained winning probability, many methods can be adopted to get the ranking scores.  
 145 Among these, Maximum Likelihood Estimation (MLE)-based **ELO** finds the ratings that best ex-  
 146 plain all observed match results simultaneously. It is a more statistically robust batch approach to  
 147 estimate the abilities  $\mathbf{r}$  by MLE. It is achieved by minimizing the negative log-likelihood function:

148

$$\hat{\mathbf{r}} = \arg \min_{\mathbf{r}} \mathcal{L}_{\mathcal{S}}(\mathbf{r}), \quad \text{where} \quad \mathcal{L}_{\mathcal{S}}(\mathbf{r}) = - \sum_{(i, j, w_{ij}) \in \mathcal{S}} [w_{ij} \ln P_{ij} + w_{ji} \ln P_{ji}]. \quad (2)$$

149

150

151 Adopting the MLE framework is crucial as it allows us to rigorously analyze the statistical proper-  
 152 ties of the estimated ratings, which is essential for developing our adaptive selection strategy. The  
 153 detailed properties are as follows.

154

155 **Statistical Properties of ELO.** Consider modeling the arena as an Erdős-Rényi (E-R) Graph  
 156  $G(N, q_N)$  (Kumar et al., 2000; Guimera et al., 2004), where  $N$  is variable and the sample size  $|\mathcal{S}|$   
 157 increases with  $N$ . Here,  $q_N$  represents the probability that a confrontation occurs between models  $i$   
 158 and  $j$ . Under this framework, the estimation results of ELO exhibit both uniqueness and asymptotic  
 159 normality (Simons & Yao, 1999; Han et al., 2020):

160

161 **Lemma 1.** Let  $G(N, q_N)$  be the Erdős-Rényi Graph, and  $\hat{\mathbf{r}}_n, \mathbf{r}_n^*$  be the estimated and true abilities  
 162 of the first  $n$  models, respectively. We have:

163

164 (1) (**Uniqueness**) If  $q_N = \omega\left(\frac{\log N}{N}\right)$ , then ELO almost surely has a unique solution  $\hat{\mathbf{r}}_{N-1}$ .

162 (2) **(Asymptotic Normality)** If  $q_N \cdot \frac{N^{1/10}}{(\log N)^{1/5}} \rightarrow \infty$  as  $N \rightarrow \infty$ , then for any  $n < N$ ,  
 163  $\sqrt{|\mathcal{S}|}(\hat{\mathbf{r}}_n - \mathbf{r}_n^*) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_{\mathcal{S}}(\mathbf{r}_n^*)^{-1})$ , where  $\mathcal{I}_{\mathcal{S}}(\mathbf{r}^*)$  is the **Fisher information matrix**.  
 164

165 In the ELO method, Fisher information (Frieden, 2004) is calculated by solving the Hessian matrix:  
 166

$$\mathcal{I}_{\mathcal{S}}(\mathbf{r}) = -\mathbb{E}[\mathcal{H}(\mathcal{L}_{\mathcal{S}}(\mathbf{r}))] = C^2 \sum_{(i,j, w_{ij}) \in \mathcal{S}} P_{ij} P_{ji} (\mathbf{e}_i - \mathbf{e}_j)(\mathbf{e}_i - \mathbf{e}_j)^\top, \quad (3)$$

170 where  $\mathbf{e}_i$  is the  $i$ -th standard basis vector, and  $\mathbf{e}_N = \mathbf{0}$ . To explain the Fisher information matrix  
 171 intuitively, consider a toy example with three models,  $X$ ,  $Y$ , and  $Z$ , and two battle records  $\mathcal{S} =$   
 172  $\{(X > Y), (X = Z)\}$ . Assuming all models have an initial ability of 0 and that model  $Z$  is fixed  
 173 during the estimation process, the calculation of the information quantity is as follows:

$$C^2 \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \xrightarrow{X > Y} C^2 \begin{bmatrix} 0.25 & -0.25 \\ -0.25 & 0.25 \end{bmatrix} \xrightarrow{X = Z} C^2 \begin{bmatrix} 0.5 & -0.25 \\ -0.25 & 0.25 \end{bmatrix}, \quad (4)$$

177 and we also provide a case study of the information matrix based on real dataset in Appendix E.  
 178

179 Considering Lemma 1 with Fisher information matrix like Equation 4, it can be confirmed that  
 180 ELO has a unique estimated solution and this estimation is asymptotically unbiased and normally  
 181 distributed, even in sparse scenarios. Therefore, from a statistical perspective, **the efficiency problem**  
 182 **in arena evaluation essentially revolves around how to reduce variance more quickly** during the  
 183 evaluation process. Specifically, the key lies in **reducing the inverse of the Fisher information**  
 184 **matrix  $\mathcal{I}_{\mathcal{S}}(\mathbf{r}^*)$**  under ELO context.  
 185

### 3.2 SELECTION STRATEGY FOR ARENA-BASED EVALUATION

187 As mentioned earlier, our direct goal is to reduce the inverse of the information matrix. However,  
 188 the true ability values of candidate models,  $\mathbf{r}^*$ , are unknown, making it impossible to directly se-  
 189 lect the model pairs. Therefore, we dynamically select each subsequent model pair based on the  
 190 accumulated records  $\mathcal{S}_{t-1}$ . Specifically, by approximating the true abilities  $\mathbf{r}^*$  with the estimated  
 191 abilities  $\hat{\mathbf{r}}^{t-1}$  derived from  $\mathcal{S}_{t-1}$ , we can compute an approximate information matrix. Using the  
 192 Fisher information of the current abilities, we propose the following A-optimality method (Chan,  
 193 1982; Wanyonyi et al., 2021) for Arena-based Evaluation:

194 **The A-optimality method.** This method aims to select the model pair that minimizes the sum of  
 195 the sampling variances of the ability estimators, which is equivalent to minimizing the trace of the  
 196 inverse of the information matrix:  
 197

$$(i_t, j_t) = \arg \min_{1 \leq i < j \leq N} \text{tr} \left[ (\mathcal{I}_{\mathcal{S}_{t-1}}(\hat{\mathbf{r}}^{t-1}) + \mathcal{I}_{\{(i,j)\}}(\hat{\mathbf{r}}^{t-1}))^{-1} \right]. \quad (5)$$

200 For this method to be feasible, the Fisher information matrix must be invertible. We have the fol-  
 201 lowing theorem:

202 **Theorem 1.** Consider the Erdős-Rényi Graph  $G(N, q_N)$  for ELO. If  $q_N = \omega\left(\frac{\log N}{N}\right)$ , then the  
 203 information matrix  $\mathcal{I}_{\mathcal{S}}(\mathbf{r})$  is almost surely positive definite.  
 204

205 From Theorem 1, it follows that as long as there are sufficient battle samples, the information matrix  
 206 is guaranteed to be positive definite (Johnson, 1970). This not only ensures that the information  
 207 matrix is invertible but also guarantees that the determinant is positive.  
 208

209 However, A-optimality method still requires calculating the inverse of a matrix, which increases  
 210 computational complexity. To address this, we consider maximizing the determinant of information  
 211 matrix and propose the D-optimality method (John & Draper, 1975) for Arena-based Evaluation:  
 212

213 **The D-optimality method.** The D-optimality method aims to select the model pair that maximizes  
 214 the determinant of the information matrix in the current state, expressed as:  
 215

$$(i_t, j_t) = \arg \max_{1 \leq i < j \leq N} |\mathcal{I}_{\mathcal{S}_{t-1}}(\hat{\mathbf{r}}^{t-1}) + \mathcal{I}_{\{(i,j)\}}(\hat{\mathbf{r}}^{t-1})|. \quad (6)$$

216

217  
218  
219  
Table 1: This table summarizes the characteristics of each strategy. Using  $\hat{r}$  indicates whether the  
strategy utilizes currently estimated abilities; Global Info indicates whether global information is  
used, and  $\min q_N$  represents the minimum edge probability that ensures effective selection.

220

221

222

223

224

225

226

227

228

229

230

231

232

233

Method	Random	Nearest	Ada-Pair	A-optimality	D-optimality
Using $\hat{r}$	✗	✓	✓	✓	✓
Global Info	✗	✗	✗	✓	✓
$\min q_N$	0	$w\left(\frac{\log N}{N}\right)$	1	$w\left(\frac{\log N}{N}\right)$	$w\left(\frac{\log N}{N}\right)$
Complexity	$O(1)$	$O(N)$	$O(N^2)$	$O(N^{\beta+3})$	$O(N^{\beta+2})$

## 3.3 THEORETICAL ANALYSIS

In this subsection, we analyze our proposed method from three perspectives: algorithmic prerequisites, numerical computation, and time complexity.

**Algorithmic Prerequisites.** Different selection strategies depend on distinct prerequisites. Table 1 summarizes the conditions for each method.

For example, the Random selection method requires no additional variables, so it is often used to initialize other strategies. In contrast, methods that consider model abilities  $\hat{r}$  must meet at least the conditions outlined in Lemma 1(1), which represents the theoretical lower bound of  $q_N$  for these methods. More stringent conditions apply to methods like Ada-Pair, which considers the confidence interval of each model pair  $(i, j)$ . To ensure the existence of confidence intervals, Ada-Pair requires that each pair has competed at least once, implying that the density of the E-R graph must satisfy  $q_N = 1$ . This requirement becomes restrictive as the number of models  $N$  increases. This highlights a key advantage of our proposed method: it only requires the minimal condition,  $\omega\left(\frac{\log N}{N}\right)$ , to ensure the positive definiteness of the information matrix (Theorem 1) without requiring additional samples for initialization. This feature demonstrates the broad applicability of our approach.

**Numerical Computation.** To avoid numerical instability and ensure computational validity, we need to consider the relationship between Fisher information and the number of samples. We present the following theorem to address this:

**Theorem 2.** *Let  $\mathbf{r}$  be the ability parameter vector, and let  $\mathcal{I}_{\mathcal{S}_t}(\mathbf{r})$  denote the Fisher information matrix based on the subset  $\mathcal{S}_t$  with size  $t$ . Define:*

- $\mathcal{A}(\mathcal{S}_t) = \frac{1}{t\mathcal{I}_{\mathcal{S}_t}(\mathbf{r})^{-1}}$  (average information efficiency),
- $\mathcal{D}(\mathcal{S}_t) = |\mathcal{I}_{\mathcal{S}_t}(\mathbf{r})|^{\frac{1}{N-1}}$  (normalized geometric information density).

Assume that each element in  $\mathcal{S}_t$  contributes bounded positive information and that  $\mathcal{I}_{\mathcal{S}_t}(\mathbf{r})$  is positive definite for sufficiently large  $t$ . Then as  $t \rightarrow \infty$ :

1.  $\mathcal{A}(\mathcal{S}_t) = O(t)$  (there exist constants  $C_1$  and  $t_1$  such that  $\mathcal{A}(\mathcal{S}_t) \leq C_1 \cdot t$  for all  $t \geq t_1$ ),
2.  $\mathcal{D}(\mathcal{S}_t) = O(t)$  (there exist constants  $C_2$  and  $t_2$  such that  $\mathcal{D}(\mathcal{S}_t) \leq C_2 \cdot t$  for all  $t \geq t_2$ ).

Theorem 2 provides the upper bound for the variation of Fisher information with respect to the number of competitions. Specifically, the inverse of the A-optimality measure grows linearly with the increase in the number of samples, while the  $(N - 1)$ -th root of the D-optimality measure grows linearly. Therefore, when practically implementing the D-optimality method, it is advisable to transform the calculation into the form defined in Theorem 2 to prevent numerical explosion.

**Time Complexity.** Algorithm 1 presents the pseudo-code for model selection in the Arena. The main computational cost in the model pair selection phase comes from calculating the determinant or the inverse of the matrix. Both operations have a complexity of  $O(N^\beta)$  (Kaltofen & Villard, 2005), where  $N$  is the matrix size or the number of models in the arena. Therefore, for the D-optimality method, which traverses all model pairs and computes the determinant of the information matrix at each step, the complexity is  $O(N^{\beta+2})$ . For the A-optimality method, whether calculating the trace of the inverse matrix or computing  $N$  submatrices, it requires  $N$  additional calculations per iteration, resulting in a complexity of  $O(N \times N^2 \times N^\beta) = O(N^{\beta+3})$ . The complexity for calculating the

270

**Algorithm 1:** The Adaptive Arena Framework

271

**Initialize:** Step  $t \leftarrow 1$ , Model’s ability  $\hat{r}^0$ , Battle records  $\mathcal{S}_0 \leftarrow \emptyset$ .

272

```

1 while True do
2   Select model pair:  $(i_t, j_t) \leftarrow \arg \max_{1 \leq i < j \leq N} |\mathcal{I}_{\mathcal{S}_{t-1}}(\hat{r}^{t-1}) + \mathcal{I}_{\{(i, j)\}}(\hat{r}^{t-1})|$ .
3   Get Annotation for model pair:  $w_{i_t j_t} \leftarrow \text{Anno}(i_t, j_t)$ .
4   Add record in  $\mathcal{S}$ :  $\mathcal{S}_t \leftarrow \mathcal{S}_{t-1} \cup \{(i_t, j_t, w_{i_t j_t})\}$ .
5   Update model’s ability  $\hat{r}$ :  $\hat{r}^t \leftarrow \arg \max_{\mathbf{r}} \mathcal{L}_{\mathcal{S}_t}(\mathbf{r})$ ,  $t \leftarrow t + 1$ .

```

273

**Output:** The models’ estimated abilities  $\hat{r}^t$ .

274

275

276

277

278

279

280

information matrix involves summing over  $t$  iterations, making the overall time complexity for the A-optimality method  $O(|S|^2 + |S| \cdot N^{\beta+3})$ , and for the D-optimality method,  $O(|S|^2 + |S| \cdot N^{\beta+2})$ .

Although our method has a higher time complexity than previous approaches, its primary goal is to reduce human-annotated samples by leveraging computational time, ultimately minimizing the overall annotation effort. Therefore, we only require our method to meet the operational needs of the Arena system—providing model pairs for evaluation within a limited time for each user request. To achieve this, we have implemented the following improvements: (1) The information matrix is updated directly by adding the current model pair’s information to the matrix from the previous step. (2) To address system concurrency, the top- $K$  model pairs with the highest information gain are selected for each request.

We discuss the time cost and the efficiency of these improvements in section 4.2. Although our method does not directly reduce the complexity related to the annotation volume, its value becomes evident within the iterative, ongoing Arena framework. For example, if our method improves efficiency by 50%, it means that the annotation results from our approach can achieve the same outcomes as twice the original annotation volume, given the Arena’s continuous operation. This offers a significant long-term advantage, enabling comparable or even superior results with fewer resources over extended periods.

298

299

## 4 EXPERIMENTS

300

In this section, we evaluate the performance of the proposed A-optimality and D-optimality methods through experiments on both simulated and real-world datasets, analyzing the abilities of each method and discussing the time cost for proposed method.

304

305

### 4.1 EXPERIMENTAL SETUP

306

**Evaluation Metric.** Since the ELO score is relative, we use the pairwise index (Hastie & Tibshirani, 1997) to measure the consistency between the estimated abilities and the true abilities:

307

308

309

310

$$\text{Pairwise} = \frac{2}{N(N-1)} \sum_{1 \leq i < j \leq N} \mathbb{I}[(r_i^* - r_j^*)(\hat{r}_i - \hat{r}_j) > 0], \quad (7)$$

311

where  $\mathbb{I}[\cdot]$  is the indicator function,  $r^*$  is true abilities of the models, and  $\hat{r}$  is the estimated abilities. Since the true abilities are not available in real-world scenarios, we use *the results estimated from the full dataset by ELO* (Equation 2) as the ground truth (Zhuang et al., 2023; Liu et al., 2024c). Lemma 1 ensures that  $r^*$  derived from this process is unique and unbiased, making this ground truth feasible. In other words, our task can be viewed as selecting a subset of matchups that achieves the same evaluation quality as using the full set, thereby significantly improving efficiency.

312

313

314

315

316

317

**Dataset.** We conducted experiments on two real-world datasets: **Chatbot** (Chiang et al., 2024; Zheng et al., 2023) and **PPE** (Frick et al., 2024). The Chatbot dataset<sup>1</sup> was collected from 13,000 distinct IP addresses on the Chatbot Arena platform, while the PPE dataset<sup>2</sup> contains human preference evaluations for Preference Proxy Evaluations. Detailed statistics and experimental settings are provided in Appendix C.1.

322

323

<sup>1</sup>[https://huggingface.co/datasets/lmsys/chatbot\\_arena\\_conversations](https://huggingface.co/datasets/lmsys/chatbot_arena_conversations)

<sup>2</sup><https://huggingface.co/datasets/lmarena-ai/PPE-Human-Preference-V1>

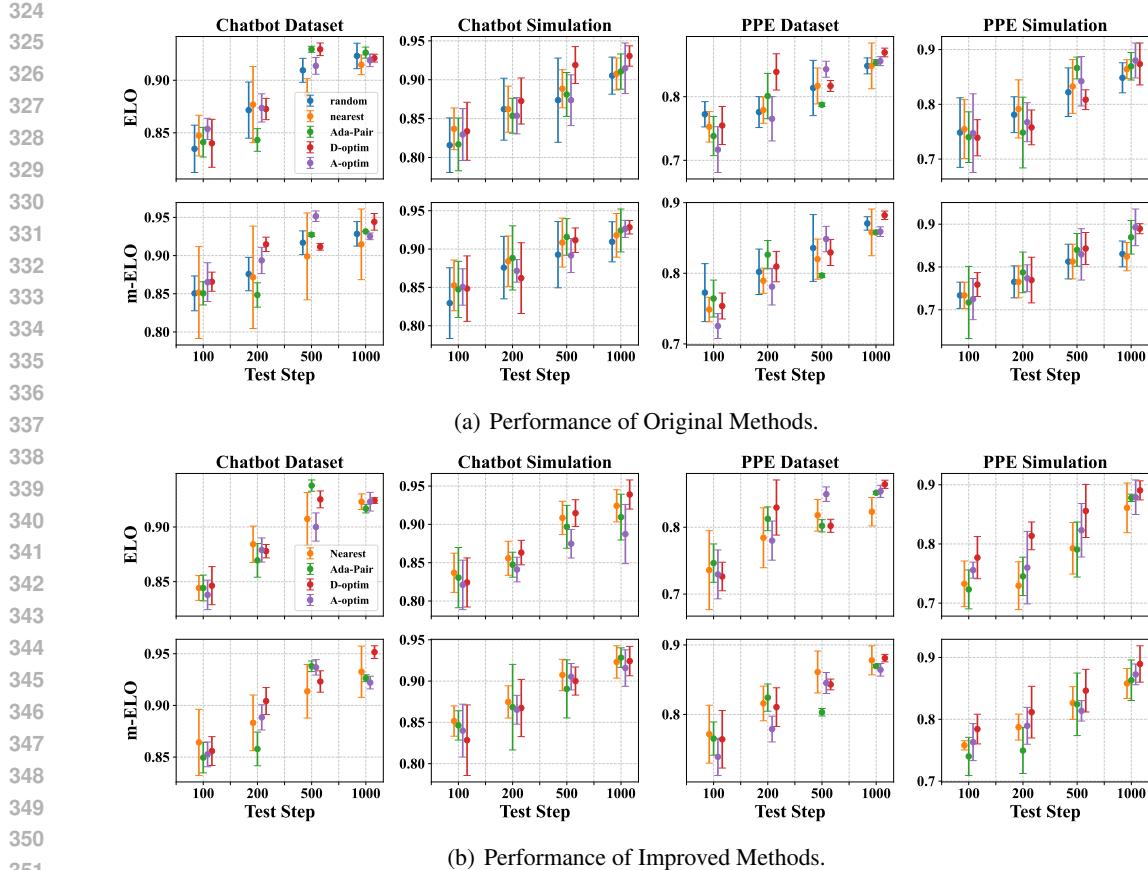


Figure 2: This Error bar illustrates pairwise metrics for each dataset and ELO method. "Dataset" indicates randomly sampled competition results, while "Simulation" denotes results simulated based on real model abilities. Error bars represent standard deviation.

**Compared Approaches.** This paper compares two model ability evaluation methods: **ELO** (Albers & Vries, 2001) and **m-ELO** (Liu et al., 2025). ELO refers to the traditional ELO, it updates each model’s rating iteratively by adding the difference between the annotation  $w_{ij}$  and the prediction  $P_{ij}$ , while m-ELO uses a MLE process like Equation 2. For model pair selection, we compare several common strategies used in both traditional competitions and modern LLM arenas. These strategies include: **Random**, which randomly selects a model pair for annotation at each step and serves as a baseline; **Nearest**, which selects the two models with the closest abilities at each step, a typical strategy in competitive games; and **Ada-Pair** (Chiang et al., 2024), an official selection strategy from Chatbot Arena, which takes into account the confidence interval of each model pair and selects the one that maximizes the reduction of the interval. Additionally, for each method, we consider both the original version and the time-improved version discussed in Section 3.3, referred to as **Original** and **Improved**, respectively.

## 4.2 EXPERIMENT RESULTS AND ANALYSIS

To evaluate the effectiveness of our proposed selection method, we conducted experiments on real-world datasets. The following conclusions were drawn:

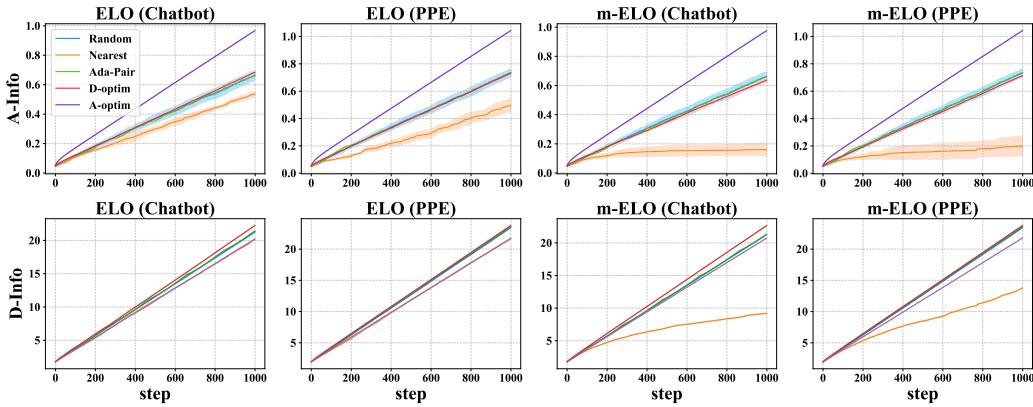
**Performance on Pairwise Metrics.** Figure 2(a) shows the performance of different selection strategies at each step, and Table 2 presents the average pairwise for each strategy. Our proposed method, D-optim, outperforms all others across all scenarios and is the only method that significantly surpasses the Random baseline. Additionally, as shown in Figure 2(a), D-optim effectively reduces variance in most cases, indicating that it lowers evaluation uncertainty and enhances efficiency.

378

379  
380  
Table 2: Performance of Pairwise Metrics for Different Strategies. Results are averaged over five  
experiments and four steps. Bold values indicate significant improvements over the baseline.

381

382	Method	Dataset		Is Simu		ELO Method	Total										
		383	384	385	386	387											
388	389	390	391	392	393	394	395	396	397	398	399	400	401	402	403	404	
Original	Random	0.8797	0.8021	0.8501	0.8316	0.8379	0.8439	0.8409									
	Nearest	0.8825	0.7995	0.8422	0.8398	0.8412	0.8407	0.8410									
	Ada-Pair	0.8835	0.8040	0.8452	0.8423	0.8379	0.8496	0.8437									
	A-optim	0.8880	0.8033	0.8494	0.8418	0.8406	0.8506	0.8456									
	D-optim	<b>0.8941</b>	<b>0.8122</b>	<b>0.8596</b>	<b>0.8466</b>	<b>0.8486</b>	<b>0.8577</b>	<b>0.8531</b>									
Improved	Nearest	0.8897	0.8020	0.8525	0.8392	0.8351	0.8566	0.8459									
	Ada-Pair	0.8849	0.7993	0.8509	0.8332	0.8376	0.8465	0.8421									
	A-optim	0.8807	0.8061	0.8488	0.8380	0.8372	0.8495	0.8434									
	D-optim	<b>0.8918</b>	<b>0.8243</b>	<b>0.8581</b>	<b>0.8581</b>	<b>0.8546</b>	<b>0.8616</b>	<b>0.8581</b>									

405  
406  
Figure 3: This line chart shows A-Info and D-Info for each dataset and ELO method.407  
408  
409  
410  
411  
412  
413  
The Nearest method, on the other hand, performs suboptimally in the LLM arena. This is due to  
the nature of the arena: rather than actual battles, models answer a single question, making ties  
more likely than in traditional competitions. This diminishes the effectiveness of pairing models  
with similar abilities. We also observe that while m-ELO incurs additional time costs, its impact on  
selection strategies that rely on current model abilities is notable. In practical applications, engineers  
need to balance computational efficiency and accuracy, such as using m-ELO for high-precision  
tasks and ELO in resource-constrained settings.415  
416  
417  
418  
**Performance of Time-Improved Method on Pairwise Metrics.** To assess the impact of our  
improvements on proposed methods, we compared the Nearest, Ada-Pair, A-optimality, and D-  
optimality methods, where the Top-10 samples were selected in each sampling round (i.e.,  $K = 10$ ).  
The experimental results are presented in Figure 2(b) and Table 2.419  
420  
421  
422  
423  
We can observe that our D-optimality method not only maintains its original performance but also  
shows improvements, where a total 0.5% gain is observed compared to the time-improved version.  
This suggests that model pairs with higher D-Info are indeed prioritized for selection, supporting the  
effectiveness of our D-Info design. This result confirms the effectiveness of our improved D-optim.  
In the appendix C.3, we also present its performance on larger-scale datasets.425  
426  
427  
428  
429  
430  
431  
**Performance on Fisher Information Gain.** Figure 3 shows the variation of  $\mathcal{A}(\mathcal{S}_t) = \frac{1}{\mathbb{E}[\mathcal{I}_{\mathcal{S}_t}(\mathbf{r})^{-1}]}$   
(A-Info) and  $\mathcal{D}(\mathcal{S}_t) = |\mathcal{I}_{\mathcal{S}_t}(\mathbf{r})|^{\frac{1}{N-1}}$  (D-Info) as the number of samples increases. As shown in  
Theorem 2, both A-Info and D-Info exhibit linear growth. The experimental results confirm that  
under most strategies, their growth is indeed linear. Furthermore, our method achieves the fastest  
improvement rate (steepest slope) in both A-Info and D-Info. In contrast, Ada-Pair shows similar  
information gain to the Random method. This suggests that our method, despite using estimated  
abilities, **enhances information gain** from samples, improving evaluation efficiency.

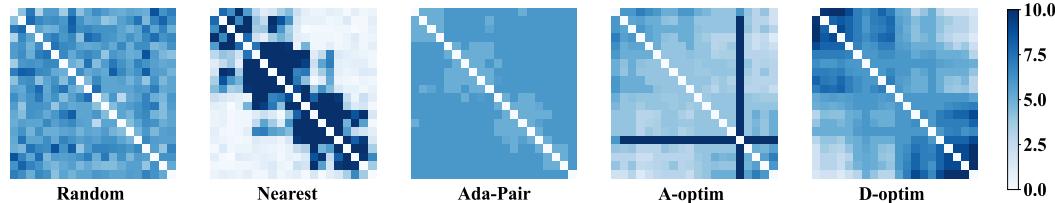


Figure 4: Heatmap of model pair sampling frequency for different selection methods. Each cell represents the number of battles between Model  $i$  (ranked by ability) and Model  $j$  (ranked by ability).

**Visualization of Sampling Distribution.** To visualize the sampling distribution of model pairs selected by each strategy, we sorted the models by true abilities and used heatmaps to represent the sampling frequency of each pair. Figure 4 shows that D-optim selects model pairs with similar true abilities more often (darker colors near the diagonal), while ensuring diverse coverage. In contrast, Ada-Pair samples more uniformly across all pairs, and Nearest focuses too much on similar models.

Interestingly, the A-optimality method results in many samplings involving a specific model, since this model’s ability is fixed in the ELO system. This creates a unique role for the model pair selection: The A-optimality method, while ensuring sparsity (i.e., most pairs have battled), prefer to use this model as an “anchor” to estimate others’ abilities. As shown in Figure 3, this method actually leads to high A-Info, but may introduce bias and impact the Arena’s sustainability. An improvement could be to fix each model sequentially, calculate A-Info for each, and average the results, but this would increase complexity. In contrast, D-optimality is minimally impacted by this constraint by ELO, making D-Info a more suitable metric for Arena-based evaluation.

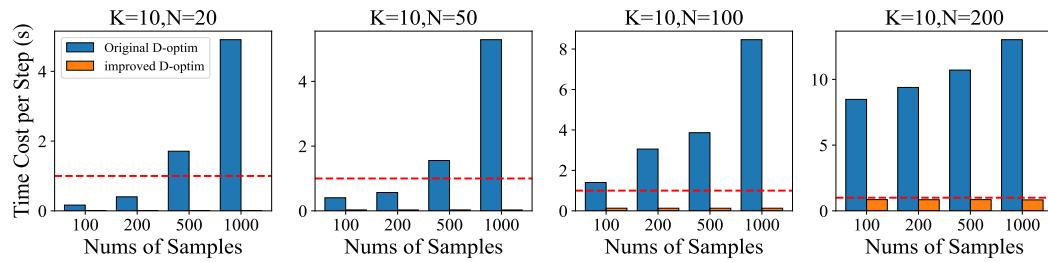


Figure 5: Time costs for Original D-optimality and Improved D-optimality Methods as the number of samples and models increases.

**Performance on Time Cost.** Figure 5 reports the time required for each sample selection as the number of samples grows, with the red line marking 1 second. Although the D-optimality method does not always achieve the absolute minimum time, its cost remains far lower than that of LLM API calls or human annotations in realistic settings. Given the relatively low computational cost of our method, in practice, we can leverage computation to reduce the reliance on human annotation wherever possible. These results suggest that our method is efficient in practice and can be seamlessly incorporated into existing arena-based evaluation platforms.

## 5 CONCLUSION

In this work, we addressed the efficiency bottleneck in arena-based LLM evaluation. By leveraging the asymptotic normality of ELO-based ability estimation and introducing Fisher information as a guiding principle, we proposed adaptive model-pair selection strategies grounded in A-optimality and D-optimality. These methods systematically reduce redundant battles while preserving ranking reliability, thereby accelerating evaluation and lowering annotation costs. Extensive experiments on both simulated and real-world datasets confirm that our framework consistently outperforms existing approaches in efficiency and robustness. Beyond improving scalability for current arenas, our approach can be regarded as a general and flexible toolkit that can be readily integrated into future evaluation platforms. We believe this work opens a promising direction toward information-efficient, scalable, and reliable assessment of large language models. In future work, an important direction is to incorporate instance-level factors into the m-ELO framework, paving the way for fully adaptive evaluation across test instances, models, and even annotators.

486 **6 ETHICS STATEMENT**

487

488 This research does not involve any human subjects, animals, or sensitive data. The proposed methods  
 489 and experiments focus solely on the evaluation of LLMs using publicly available datasets. We  
 490 ensure that all data used in our experiments are publicly accessible, adhering to ethical standards  
 491 and relevant data usage policies. Furthermore, our work does not raise any concerns related to bias,  
 492 fairness, or social impact, as it is focused on improving the efficiency of model evaluations in a  
 493 controlled, academic setting. In summary, our research does not present any ethical issues, and we  
 494 are committed to conducting future work in line with ethical best practices and standards.

495

496 **7 REPRODUCIBILITY STATEMENT**

497

498 To ensure the reproducibility of our results, we provide the complete implementation of our meth-  
 499 ods, along with all necessary code and data, on an anonymous GitHub repository. The repository  
 500 includes scripts to replicate the experiments presented in the paper, including data preprocessing,  
 501 model training, and evaluation procedures. All hyperparameters, random seeds, and environment  
 502 configurations are specified to ensure consistent results across different platforms. The repository  
 503 can be accessed at the following anonymous link: <https://anonymous.4open.science/r/Adaptive-Arena-870D>.

505

506 **REFERENCES**

507

508 Paul CH Albers and Han de Vries. Elo-rating as a tool in the sequential estimation of dominance  
 509 strengths. *Animal behaviour*, pp. 489–495, 2001.

510 Siavash Ameli, Siyuan Zhuang, Ion Stoica, and Michael W Mahoney. A statistical framework for  
 511 ranking llm-based chatbots. *arXiv preprint arXiv:2412.18407*, 2024.

512 Meriem Boubdir, Edward Kim, Beyza Ermis, Sara Hooker, and Marzieh Fadaee. Elo uncovered:  
 513 Robustness and best practices in language model evaluation. *Advances in Neural Information  
 514 Processing Systems*, 37:106135–106161, 2024.

516 Róbert Busa-Fekete, Eyke Hüllermeier, and Balázs Szörényi. Preference-based rank elicitation using  
 517 statistical models: The case of mallows. In *International conference on machine learning*, pp.  
 518 1071–1079. PMLR, 2014.

519 NN Chan. A-optimality for regression designs. *Journal of Mathematical Analysis and Applications*,  
 520 87(1):45–50, 1982.

522 Sourav Chatterjee and SR Srinivasa Varadhan. The large deviation principle for the erdős-rényi  
 523 random graph. *European Journal of Combinatorics*, 32(7):1000–1017, 2011.

525 Herman Chernoff. *Sequential design of experiments*. Springer, 1992.

526 Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li,  
 527 Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Sto-  
 528 ica. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first  
 529 International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*.  
 530 OpenReview.net, 2024. URL <https://openreview.net/forum?id=3MW8GKNyzI>.

532 Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and  
 533 Orhan Firat. Unimax: Fairer and more effective language sampling for large-scale multilingual  
 534 pretraining. *arXiv preprint arXiv:2304.09151*, 2023.

535 Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser,  
 536 Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to  
 537 solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

539 Rémi Coulom. Computing “elo ratings” of move patterns in the game of go. *ICGA journal*, 30(4):  
 198–208, 2007.

540 Joel N Franklin. *Matrix theory*. Courier Corporation, 2000.  
 541

542 Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios N. Angelopoulos, Jiantao Jiao,  
 543 Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. How to evaluate reward models for rlhf, 2024.  
 544 URL <https://arxiv.org/abs/2410.14872>.

545 B Roy Frieden. *Science from Fisher information*, volume 974. Citeseer, 2004.  
 546

547 Roger Guimera, Marta Sales-Pardo, and Luís A Nunes Amaral. Modularity from fluctuations in  
 548 random graphs and complex networks. *Physical Review E—Statistical, Nonlinear, and Soft Matter  
 549 Physics*, 70(2):025101, 2004.

550 Ruijian Han, Rougang Ye, Chunxi Tan, and Kani Chen. Asymptotic theory of sparse bradley–terry  
 551 model. 2020.  
 552

553 Trevor Hastie and Robert Tibshirani. Classification by pairwise coupling. *Advances in neural  
 554 information processing systems*, 10, 1997.

555 Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and  
 556 Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint  
 557 arXiv:2009.03300*, 2020.

558 Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya  
 559 Poria, and Roy Ka-Wei Lee. Llm-adapters: An adapter family for parameter-efficient fine-tuning  
 560 of large language models. *arXiv preprint arXiv:2304.01933*, 2023.

561 RC St John and Norman R Draper. D-optimality for regression designs: a review. *Technometrics*,  
 562 17(1):15–23, 1975.

563 Charles Royal Johnson. Positive definite matrices. *The American Mathematical Monthly*, 77(3):  
 564 259–264, 1970.

565 Erich Kaltofen and Gilles Villard. On the complexity of computing determinants. *Computational  
 566 complexity*, 13(3):91–130, 2005.

567 Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank  
 568 Fischer, Urs Gasser, Georg Groh, Stephan Günemann, Eyke Hüllermeier, et al. Chatgpt for  
 569 good? on opportunities and challenges of large language models for education. *Learning and  
 570 individual differences*, 103:102274, 2023.

571 Alex Kipnis, Konstantinos Voudouris, Luca M Schulze Buschoff, and Eric Schulz. metabench—  
 572 a sparse benchmark of reasoning and knowledge in large language models. *arXiv preprint  
 573 arXiv:2407.12844*, 2024.

574 Arie W Kruglanski and Icek Ajzen. Bias and error in human judgment. *European Journal of Social  
 575 Psychology*, 13(1):1–44, 1983.

576 Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, D Sivakumar, Andrew Tomkins, and Eli  
 577 Upfal. Stochastic models for the web graph. In *Proceedings 41st Annual Symposium on Foundations  
 578 of Computer Science*, pp. 57–65. IEEE, 2000.

579 Conglong Li, Zhewei Yao, Xiaoxia Wu, Minjia Zhang, Connor Holmes, Cheng Li, and Yuxiong  
 580 He. Deepspeed data efficiency: Improving deep learning model quality and training efficiency  
 581 via efficient data sampling and routing. In *Proceedings of the AAAI Conference on Artificial  
 582 Intelligence*, volume 38, pp. 18490–18498, 2024a.

583 Jiatong Li, Renjun Hu, Kunzhe Huang, Yan Zhuang, Qi Liu, Mengxiao Zhu, Xing Shi, and Wei  
 584 Lin. Perteval: Unveiling real knowledge capacity of llms with knowledge-invariant perturbations.  
 585 *arXiv preprint arXiv:2405.19740*, 2024b.

586 Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian  
 587 Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language  
 588 models. *arXiv preprint arXiv:2211.09110*, 2022.

594 Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao,  
 595 Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint*  
 596 *arXiv:2412.19437*, 2024a.

597 Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan,  
 598 Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around  
 599 player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.

600 601 Zirui Liu, Zhuang Yan, Qi Liu, Jiatong Li, Yuren Zhang, Zhenya Huang, Jinze Wu, and Shijin Wang.  
 602 Computerized adaptive testing via collaborative ranking. In *Neural Information Processing Sys-*  
 603 *tems*, 2024c. URL <https://api.semanticscholar.org/CorpusID:276259892>.

604 605 Zirui Liu, Jiatong Li, Yan Zhuang, Qi Liu, Shuanghong Shen, Jie Ouyang, Mingyue Cheng, and  
 606 Shijin Wang. am-elo: A stable framework for arena-based llm evaluation, 2025. URL <https://arxiv.org/abs/2505.03475>.

607 608 Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Qingwei Lin, Jianguang Lou, Shifeng Chen, Yan-  
 609 song Tang, and Weizhu Chen. Arena learning: Build data flywheel for llms post-training via  
 610 simulated chatbot arena. *arXiv preprint arXiv:2407.10627*, 2024.

611 612 Alexander Ly, Maarten Marsman, Josine Verhagen, Raoul PPP Grasman, and Eric-Jan Wagenmak-  
 613 ers. A tutorial on fisher information. *Journal of Mathematical Psychology*, 80:40–55, 2017.

614 615 Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker.  
 616 When less is more: Investigating data pruning for pretraining llms at scale. *arXiv preprint*  
*arXiv:2309.04564*, 2023.

617 618 Rui Min, Tianyu Pang, Chao Du, Qian Liu, Minhao Cheng, and Min Lin. Improving your model  
 619 ranking on chatbot arena by vote rigging. In *ICLR 2025 Workshop on Foundation Models in the*  
*Wild*, 2025.

620 621 Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad  
 622 Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large  
 623 language models. *arXiv preprint arXiv:2410.05229*, 2024.

624 625 Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State  
 626 of what art? a call for multi-prompt llm evaluation. *Transactions of the Association for Compu-*  
*tational Linguistics*, 12:933–949, 2024.

627 628 Onorato Timothy O’Meara. *Introduction to quadratic forms*, volume 117. Springer, 2013.

629 630 Rajvardhan Patil and Venkat Gudivada. A review of current trends, techniques, and challenges in  
 631 large language models (llms). *Applied Sciences*, 14(5):2074, 2024.

632 633 Radek Pelánek. Applications of the elo rating system in adaptive educational systems. *Computers*  
*& Education*, 98:169–179, 2016.

634 635 Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail  
 636 Yurochkin. tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint*  
*arXiv:2402.14992*, 2024a.

637 638 Felipe Maia Polo, Ronald Xu, Lucas Weber, Mírian Silva, Onkar Bhardwaj, Leshem Choshen,  
 639 Allysson Flávio Melo de Oliveira, Yuekai Sun, and Mikhail Yurochkin. Efficient multi-prompt  
 640 evaluation of llms. *arXiv preprint arXiv:2405.17202*, 2024b.

641 642 PV Rao and Lawrence L Kupper. Ties in paired-comparison experiments: A generalization of the  
 643 bradley-terry model. *Journal of the American Statistical Association*, 62(317):194–204, 1967.

644 645 Jorma J Rissanen. Fisher information and stochastic complexity. *IEEE transactions on information*  
 646 *theory*, 42(1):40–47, 1996.

647 648 Gayathri Saranathan, Mohammad Parwez Alam, James Lim, Suparna Bhattacharya, Soon Yee  
 649 Wong, Martin Foltin, and Cong Xu. Dele: Data efficient llm evaluation. In *ICLR 2024 Work-*  
*shop on Navigating and Addressing Data Problems for Foundation Models*, 2024.

648 Gordon Simons and Yi-Ching Yao. Asymptotics when the number of parameters tends to infinity in  
 649 the bradley-terry model for paired comparisons. *The Annals of Statistics*, 27(3):1041–1060, 1999.  
 650

651 Lucas Spangher, Tianle Li, William F Arnold, Nick Masiewicki, Xerxes Dotiwalla, Rama Kumar  
 652 Pasumarthi, Peter Grabowski, Eugene Ie, and Daniel Gruhl. Chatbot arena estimate: towards a  
 653 generalized performance benchmark for llm capabilities. In *Proceedings of the 2025 Conference  
 654 of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human  
 655 Language Technologies (Volume 3: Industry Track)*, pp. 1016–1025, 2025.

656 Balázs Szörényi, Róbert Busa-Fekete, Adil Paul, and Eyke Hüllermeier. Online rank elicitation for  
 657 plackett-luce: A dueling bandits approach. *Advances in neural information processing systems*,  
 658 28, 2015.

659 Kelly Tang, Wei-Lin Chiang, and Anastasios N. Angelopoulos. Arena explorer: A topic modeling  
 660 pipeline for llm evals & analytics, 2025.

661

662 Kangyu Wang, Hongliang He, Lin Liu, Ruiqi Liang, Zhenzhong Lan, and Jianguo Li. Inclusion  
 663 arena: An open platform for evaluating large foundation models with real-world apps. *arXiv  
 664 preprint arXiv:2508.11452*, 2025.

665 Xianquan Wang, Likang Wu, Shukang Yin, Zhi Li, Yanjiang Chen, Hufeng Hufeng, Yu Su, and  
 666 Qi Liu. I-am-g: Interest augmented multimodal generator for item personalization. In *Pro-  
 667 ceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp.  
 668 21303–21317, 2024a.

669

670 Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming  
 671 Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-  
 672 task language understanding benchmark. *Advances in Neural Information Processing Systems*,  
 673 37:95266–95290, 2024b.

674 Samson W Wanyonyi, Ayubu A Okango, and Julius K Koech. Exploration of d-, a-, i-and g-  
 675 optimality criteria in mixture modeling. 2021.

676 Hermann Weyl. Inequalities between the two kinds of eigenvalues of a linear transformation. *Pro-  
 677 ceedings of the national academy of sciences*, 35(7):408–411, 1949.

678

679 Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy S Liang. Data selection for language  
 680 models via importance resampling. *Advances in Neural Information Processing Systems*, 36:  
 681 34201–34227, 2023.

682

683 Yanbin Yin, Zhen Wang, Kun Zhou, Xiangdong Zhang, Shibo Hao, Yi Gu, Jieyuan Liu, Somanshu  
 684 Singla, Tianyang Liu, Eric P. Xing, Zhengzhong Liu, Haojian Jin, and Zhiting Hu. Decentralized  
 685 arena via collective llm intelligence: Building automated, robust, and transparent llm evaluation  
 686 for numerous dimensions, 10 2024. URL <https://de-arena.maitrix.org/>.

687

688 Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a ma-  
 chine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

689

690 Yang Zhang, Cunxiang Wang, Lindong Wu, Wenbo Yu, Yidong Wang, Guangsheng Bao, and Jie  
 691 Tang. Uda: Unsupervised debiasing alignment for pair-wise llm-as-a-judge. 2025. URL <https://api.semanticscholar.org/CorpusID:280642317>.

692

693 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang,  
 694 Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica.  
 695 Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

696

697 Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin,  
 698 Ji-Rong Wen, and Jiawei Han. Don’t make your llm an evaluation benchmark cheater. *arXiv  
 699 preprint arXiv:2311.01964*, 2023.

700

701 Yan Zhuang, Qi Liu, GuanHao Zhao, Zhenya Huang, Weizhe Huang, Zachary Pardos, Enhong Chen,  
 Jinze Wu, and Xin Li. A bounded ability estimation for computerized adaptive testing. *Advances  
 in Neural Information Processing Systems*, 36, 2023.

702 A PROOFS OF THEOREM 1  
703704 **Theorem 1.** Consider the Erdős-Renyi Graph  $G(N, q_N)$ , if  $q_N = \omega(\frac{\log N}{N})$ , then the infomation  
705 matrix  $\mathcal{I}_S(\mathbf{r})$  is almost surely positive definite.  
706707 *Proof.* Given that  $q_N = \omega(\frac{\log N}{N})$  implies that the graph  $G(N, q_N)$  is almost surely con-  
708 nected (Guimera et al., 2004; Chatterjee & Varadhan, 2011), we next prove the following: if there ex-  
709 ist a path between any two models  $i$  and  $j$ , then the information matrix is positive definite. Assume  
710  $r_N = 0$  and consider the remaining variables  $\mathbf{r} = (r_1, \dots, r_{N-1})$ . Consider the log-likelihood  
711 function for records  $\mathcal{S}$ :  
712

713 
$$\mathcal{I}_S(\mathbf{r}) = C^2 \sum_{(i,j, w_{ij}) \in \mathcal{S}} P_{ij} P_{ji} (\mathbf{e}_i - \mathbf{e}_j)(\mathbf{e}_i - \mathbf{e}_j)^\top \quad (8)$$
  
714

715 where  $\mathbf{e}_i$  is the  $i$ -st standard basis vector and let  $\mathbf{e}_N = \mathbf{0}$ . Let the number of matches between model  
716  $i$  and model  $j$  be  $\delta_{ij}$  and define  $a_{ij} = \delta_{ij} C^2 P_{ij} P_{ji}$ , then the quadratic form (O’Meara, 2013) for  
717  $\mathcal{I}_S(\mathbf{r})$  can be expressed as:  
718

719 
$$\mathbf{x} \mathcal{I}_S(\mathbf{r}) \mathbf{x}^\top = \sum_{1 \leq i < j \leq N-1} a_{ij} (x_i - x_j)^2 + \sum_{i=1}^{N-1} a_{iN} x_i^2 + \sum_{j=1}^{N-1} a_{jN} x_j^2.$$
  
720  
721

722 Note that  $a_{ij} \geq 0$ , therefore:  
723

724 
$$\mathbf{x} \mathcal{I}_S(\mathbf{r}) \mathbf{x}^\top \geq 0.$$

725 The following proves that the condition for the equality of this inequality can only be achieved at  $\mathbf{0}$ .  
726727 Suppose there exists a non-zero vector  $\hat{\mathbf{x}} = (\hat{x}_1, \dots, \hat{x}_{N-1})$  such that  $\hat{\mathbf{x}} \mathcal{I}_S(\mathbf{r}) \hat{\mathbf{x}}^\top = 0$ . Without  
728 loss of generality, assume that  $\hat{x}_1 \neq 0$ . Then, since the graph formed by the pairwise battles among  
729 the models is connected, it means that there must exist a path  $(c_0 = 1, c_1, c_2, \dots, c_{k-1}, c_k = N)$   
730 that connects 1 and  $N$ . Then  $a_{c_i c_{i+1}} > 0$ , from which we can deduce that  $\hat{x}_1 = \hat{x}_{c_1} = \dots = \hat{x}_{c_{k-1}}$ .  
731 Since  $c_{k-1}$  and  $N$  are connected, it indicated that  $a_{c_{k-1} N} > 0$ . Therefore,  $\hat{x}_{c_{k-1}} = 0$ , and thus  
732  $\hat{x}_1 = 0$ , which is a contradiction! So  $\mathbf{x} \mathcal{I}_S(\mathbf{r}) \mathbf{x}^\top = 0$  if and only if  $\mathbf{x}$  is the zero vector. Therefore,  
733  $\mathcal{I}_S(\mathbf{r})$  is a positive definite matrix (Johnson, 1970).  $\square$   
734735 B PROOFS OF THEOREM 2  
736737 **Lemma 2.** Let  $0 < \lambda_1^t \leq \lambda_2^t \leq \dots \leq \lambda_N^t$  denote the eigenvalues of the information matrix  $\mathcal{I}_{S_t}(\mathbf{r})$ ,  
738 then  $\lambda_N^t = O(t)$ .  
739740 *Proof.* Without loss of generality, we can assume that model  $i$  and  $j$  compete against each other at  
741 time  $t+1$ . It is not hard to see that the information increment  $C^2 P_{ij} P_{ji} (\mathbf{e}_i - \mathbf{e}_j)(\mathbf{e}_i - \mathbf{e}_j)^\top$  is a  
742 rank-1 matrix and its only positive eigenvalue  $a$  satisfies that  $0 < a \leq 2C^2 P_{ij} P_{ji} \leq C^2$ . Therefore,  
743 by Weyl’s Inequality for matrix perturbations (Weyl, 1949; Franklin, 2000), the eigenvalues of the  
744 information matrix at time  $t+1$  satisfy  $\lambda_N^t \leq \lambda_{N-1}^{t+1} \leq \lambda_N^t + a \leq \lambda_N^t + C^2$ . By recursively expanding  
745 this inequality, we obtain that  $\lambda_N^t \leq \lambda_{N-1}^{t-1} + C^2 \leq \dots \leq N \cdot C^2$ , i.e.  $\lambda_N^t = O(t)$ .  $\square$   
746747 **Theorem 2.** Let  $\mathcal{A}(\mathcal{S}_t) = \frac{1}{\text{tr}[\mathcal{I}_{S_t}(\mathbf{r})^{-1}]}$  and  $\mathcal{D}(\mathcal{S}_t) = |\mathcal{I}_{S_t}(\mathbf{r})|^{\frac{1}{N-1}}$ , then it can be proved that:  
748 (1)  $\mathcal{A}(\mathcal{S}_t) = O(t)$  (2)  $\mathcal{D}(\mathcal{S}_t) = O(t)$ .  
749750 *Proof.* Since  $\text{tr}(\mathcal{I}_{S_t}(\mathbf{r})^{-1}) = \frac{1}{\lambda_1^t} + \frac{1}{\lambda_2^t} + \dots + \frac{1}{\lambda_N^t} \geq \frac{N}{\lambda_N^t}$  and  $|\mathcal{I}_{S_t}(\mathbf{r})| = \lambda_1^t \lambda_2^t \dots \lambda_N^t \leq (\lambda_N^t)^N$ ,  
751 we can derive the following bounds for the optimality criteria:  
752753 For the A-optimality criterion  $\mathcal{A}(\mathcal{S}_t) = \frac{1}{\text{tr}[\mathcal{I}_{S_t}(\mathbf{r})^{-1}]} \leq \frac{\lambda_N^t}{N}$ .  
754755 For the D-optimality criterion  $\mathcal{D}(\mathcal{S}_t) = |\mathcal{I}_{S_t}(\mathbf{r})|^{\frac{1}{N-1}} \leq \lambda_N^t$ .  
756 By Lemma 2, we know  $\lambda_N^t = O(t)$ , substituting this into the above inequalities yields  $\mathcal{A}(\mathcal{S}_t) = O(t)$   
757 and  $\mathcal{D}(\mathcal{S}_t) = O(t)$ .  $\square$

756 

## C DETAILS OF EXPERIMENTS

757 

### C.1 STATISTICS OF THE DATASET

760 In addition to the two datasets (**Chatbot** and **PPE**) mentioned in the main text, we have also included  
 761 a larger dataset **Arena-100k** (Chiang et al., 2024; Tang et al., 2025) in the appendix. The Arena-  
 762 100k<sup>1</sup> dataset comprises conversation data collected from June 2024 to August 2024, as well as  
 763 English human preference evaluations, which were used to develop Arena Explorer. Each record  
 764 in datasets includes a question ID, the names of the two competing models, their full conversation  
 765 transcripts and the annotator’s vote. The statistics of the datasets can be seen in Table 3.

766 **Table 3: Statistics of the dataset**

767 Dataset	768 Chatbot	769 PPE	770 Arena-100k
768 #Models	769 20	770 20	771 55
769 #Response logs	770 33000	771 16038	772 106134
770 #Response logs per model	771 1650.0	772 801.9	773 1929.7
771 #Response logs per model pair	772 173.68	773 89.10	774 71.47

772 

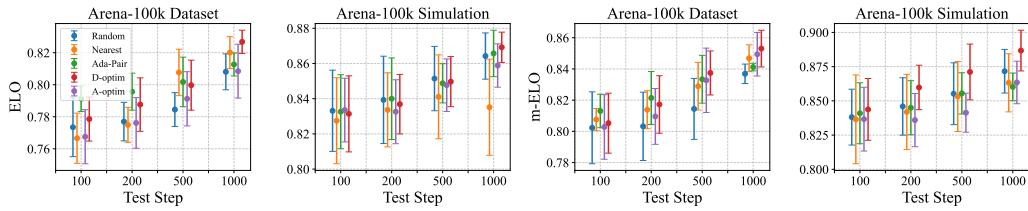
### C.2 DETAILS OF EXPERIMENTAL SETTING.

773 In this experiment, 100 sample pairs will be randomly selected during the model initialization phase  
 774 to initialize ELO, and the performance of different selection strategies will be recorded at steps 100,  
 775 200, 500, and 1000. After a selection strategy chooses a model pair, we consider two methods to  
 776 simulate the competition result: (1) Conducting without-replacement random sampling on the real  
 777 competition results in the dataset; (2) Simulating based on the real model abilities. If the abilities of  
 778 the two models are  $r_i^*, r_j^*$  and  $p = \sigma(C(r_i^* - r_j^*))$ , then each battle will return 1 with a probability  
 779 of  $p^2$ , 0 with a probability of  $(1 - p)^2$  and 0.5 with a probability of  $2p(1 - p)$ . As for ELO rating  
 780 system, we set the hyperparameter  $C$  of ELO to  $\frac{\log 10}{400}$  and epochs to 50 for each battle in m-ELO.  
 781 All experiments described in this paper can be executed on one 8G NVIDIA GeForce RTX 4060.  
 782

783 

### C.3 PERFORMANCE ON LARGE DATASET

784 To further validate the effectiveness of the D-optimality method, we also conducted experiments on  
 785 a larger dataset, namely Arena-100k. The key difference from the previous experiments lies in the  
 786 following: as the number of models increased to 55, the number of initial samples we used was also  
 787 increased to 1000. All other experimental settings remained unchanged. The experimental results  
 788 are presented in Figure 6 and Table 4.



789 **Figure 6:** This line chart illustrates the pairwise metrics for Arena-100k dataset and each ELO  
 790 method. Here, “Dataset” indicates that the competition results are randomly sampled from the  
 791 dataset, while “Simulation” denotes that the results are simulated based on the real model abilities.  
 792 The error bars represent the standard deviation.

803 It can be observed that with an increase in the number of models, the errors introduced by the  
 804 traditional ELO method into the information matrix become amplified, leading to inaccurate information  
 805 estimation. This explains why the D-optim method exhibits superior performance specifically in  
 806 terms of m-ELO. However, at step 1000, the D-optim method achieves the best performance across  
 807 all experimental settings. Therefore, overall speaking, as long as the current abilities of the models  
 808 can be effectively estimated, the D-optimality method remains fully effective even in scenarios with  
 809 a relatively large number of models.

<sup>1</sup><https://huggingface.co/datasets/1marena-ai/arena-human-preference-100k>

810  
 811 Table 4: The Performance of pairwise for Different Selection Algorithms. Each experimental re-  
 812 sult represents the average of five experiments and four selection steps, the bold font represents a  
 813 significant improvement in statistics compared to the baseline.

Method	Is Simu		ELO Method		Total
	Real	Simu	ELO	m-ELO	
Random	0.8000	0.8499	0.8164	0.8335	0.8249
Nearest	0.8083	0.8415	0.8133	0.8365	0.8249
Ada-Pair	<b>0.8138</b>	0.8486	<b>0.8236</b>	0.8388	0.8312
A-optimal	0.8047	0.8438	0.8145	0.8340	0.8243
D-optimal	0.8132	<b>0.8561</b>	0.8225	<b>0.8468</b>	<b>0.8346</b>

## D USAGE OF LLM

824  
 825 According to the rules about LLM use set by the ICLR 2026 conference, when we worked on this  
 826 paper, we only used LLMs to polish the text. This includes making the flow of language smoother  
 827 and improving how well sentences connect logically. We did not use LLMs for key steps like  
 828 coming up with main points, making research content, or handling data. Also, we made sure that the  
 829 polishing process does not change the main ideas or weaken the academic strictness of the original  
 830 text, and we fully follow the standards for proper LLM use set by ICLR 2026.

## E CASE STUDY OF FISHER INFOMATION MATRIX

834 To facilitate a better understanding of our method, we directly visualize the changes in the information  
 835 matrix throughout the experimental process. Specifically, we adopt the D-optimality method:  
 836 in each iteration, a model pair is selected, and the model ability estimates are updated in real time.  
 837 Below, we present the information matrix of 5 iterations under the condition of 100 initial samples:

$$\begin{pmatrix} 3.7 & -0.4 & -0.2 & -0.2 & 0.0 & 0.0 & 0.0 & -0.5 & -0.2 & -0.2 & -0.5 & 0.0 & -0.5 & -0.2 & -0.2 & 0.0 & 0.0 & -0.2 \\ -0.4 & 2.3 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & -0.5 & 0.0 & -0.4 & 0.0 & -0.2 & 0.0 & 0.0 & 0.0 & 0.0 & -0.2 \\ -0.2 & 0.0 & 2.0 & -0.2 & 0.0 & 0.0 & -0.5 & 0.0 & 0.0 & -0.2 & 0.0 & 0.0 & 0.0 & -0.2 & -0.2 & 0.0 & 0.0 & -0.2 \\ -0.2 & 0.0 & -0.2 & 2.9 & 0.0 & -0.2 & 0.0 & -0.2 & -0.2 & -0.2 & -0.4 & 0.0 & 0.0 & -0.5 & -0.2 & 0.0 & 0.0 & -0.2 \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & -0.2 & 0.0 & 0.0 & -0.2 & -0.2 & -0.2 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & -0.2 & 0.0 & 1.6 & -0.2 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & -0.2 & -0.4 & 0.0 & 0.0 & -0.2 \\ 0.0 & 0.0 & -0.5 & 0.0 & 0.0 & -0.2 & 2.7 & -0.5 & 0.0 & 0.0 & -0.4 & -0.2 & -0.4 & 0.0 & -0.2 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & -0.2 & 0.0 & 0.0 & -0.5 & 2.7 & -0.2 & 0.0 & 0.0 & 0.0 & 0.0 & -0.4 & -0.4 & 0.0 & 0.0 & -0.2 \\ -0.5 & 0.0 & 0.0 & -0.2 & 0.0 & 0.0 & 0.0 & -0.2 & 1.6 & 0.0 & 0.0 & 0.0 & -0.4 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ -0.2 & -0.5 & -0.2 & -0.2 & 0.0 & 0.0 & 0.0 & -0.2 & 0.0 & 2.7 & -0.2 & -0.6 & 0.0 & 0.0 & 0.0 & 0.0 & -0.2 & 0.0 \\ -0.2 & 0.0 & 0.0 & -0.4 & 0.0 & 0.0 & -0.4 & 0.0 & 0.0 & -0.2 & 2.8 & -0.4 & 0.0 & 0.0 & -0.2 & -0.7 & 0.0 & 0.0 \\ -0.5 & -0.4 & 0.0 & 0.0 & -0.2 & 0.0 & -0.2 & 0.0 & 0.0 & -0.6 & -0.4 & 3.1 & -0.2 & 0.0 & 0.0 & -0.2 & -0.2 & 0.0 \\ 0.0 & -0.2 & 0.0 & 0.0 & 0.0 & -0.4 & 0.0 & -0.4 & 0.0 & 0.0 & -0.2 & 1.9 & -0.2 & -0.5 & 0.0 & 0.0 & 0.0 & 0.0 \\ -0.5 & 0.0 & -0.2 & -0.5 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & -0.2 & 0.0 & -0.2 & 1.6 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ -0.2 & -0.2 & -0.2 & -0.2 & -0.2 & 0.0 & 0.0 & 0.0 & -0.7 & 0.0 & 0.0 & -0.5 & 0.0 & 3.1 & 0.0 & -0.2 & 0.0 & 0.0 \\ -0.2 & 0.0 & 0.0 & 0.0 & -0.2 & -0.4 & 0.0 & -0.4 & 0.0 & 0.0 & -0.2 & 0.0 & 0.0 & 0.0 & 2.2 & -0.2 & -0.2 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & -0.2 & 0.0 & 0.0 & -0.4 & 0.0 & 0.0 & 0.0 & -0.2 & 0.0 & 0.0 & -0.2 & 1.8 & 0.0 & -0.2 \\ 0.0 & 0.0 & 0.0 & -0.2 & 0.0 & 0.0 & 0.0 & -0.2 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & -0.2 & 0.0 & 0.9 & -0.2 & 0.0 \\ -0.2 & -0.2 & -0.2 & 0.0 & 0.0 & -0.2 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & -0.2 & -0.2 & 0.0 & 2.1 & 0.0 \end{pmatrix}$$

849  $\downarrow$  Step 1:(17,4,1)

$$\begin{pmatrix} 3.7 & -0.4 & -0.2 & -0.2 & 0.0 & 0.0 & 0.0 & -0.5 & -0.2 & -0.2 & -0.5 & 0.0 & -0.5 & -0.2 & -0.2 & 0.0 & 0.0 & -0.2 \\ -0.4 & 2.3 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & -0.5 & 0.0 & -0.4 & 0.0 & -0.2 & 0.0 & 0.0 & 0.0 & 0.0 & -0.2 \\ -0.2 & 0.0 & 2.0 & -0.2 & 0.0 & 0.0 & -0.5 & 0.0 & 0.0 & -0.2 & 0.0 & 0.0 & 0.0 & -0.2 & -0.2 & 0.0 & 0.0 & -0.2 \\ -0.2 & 0.0 & -0.2 & 2.9 & 0.0 & -0.2 & 0.0 & -0.2 & -0.2 & -0.2 & -0.4 & 0.0 & 0.0 & -0.5 & -0.2 & 0.0 & 0.0 & -0.2 \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & -0.2 & 0.0 & 0.0 & -0.2 & -0.2 & -0.2 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & -0.2 & 0.0 & 1.6 & -0.2 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & -0.2 & -0.4 & 0.0 & 0.0 & -0.2 \\ 0.0 & 0.0 & -0.5 & 0.0 & 0.0 & -0.2 & 2.7 & -0.5 & 0.0 & 0.0 & -0.4 & -0.2 & -0.4 & 0.0 & -0.2 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & -0.2 & 0.0 & 0.0 & -0.5 & 2.7 & -0.2 & 0.0 & 0.0 & 0.0 & 0.0 & -0.4 & -0.4 & 0.0 & 0.0 & -0.2 \\ -0.5 & 0.0 & 0.0 & -0.2 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & -0.2 & 0.0 & -0.2 & 1.6 & 0.0 & 0.0 & 0.0 & 0.0 \\ -0.2 & -0.2 & -0.2 & -0.2 & -0.2 & 0.0 & 0.0 & 0.0 & -0.7 & 0.0 & 0.0 & -0.5 & 0.0 & 3.1 & 0.0 & -0.2 & 0.0 & 0.0 \\ -0.2 & 0.0 & 0.0 & 0.0 & -0.2 & -0.4 & 0.0 & -0.4 & 0.0 & 0.0 & -0.2 & 0.0 & 0.0 & 0.0 & 2.2 & -0.2 & -0.2 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & -0.2 & 0.0 & 0.0 & -0.4 & 0.0 & 0.0 & 0.0 & -0.2 & 0.0 & 0.0 & -0.2 & 1.8 & 0.0 & -0.2 \\ 0.0 & 0.0 & 0.0 & -0.2 & 0.0 & 0.0 & 0.0 & -0.2 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & -0.2 & 0.0 & 0.9 & -0.2 & 0.0 \\ -0.2 & -0.2 & -0.2 & 0.0 & 0.0 & -0.2 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & -0.2 & -0.2 & 0.0 & 2.1 & 0.0 \end{pmatrix}$$

850  $\downarrow$  Step 2:(17,13,0)

