# BEYOND ARROW: FROM IMPOSSIBILITY TO POSSIBILITY IN MULTI-CRITERIA BENCHMARKING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Modern benchmarks such as HELM MMLU account for multiple metrics like accuracy, robustness and efficiency. When trying to turn these metrics into a single ranking, natural aggregation procedures can produce cyclic pairwise comparisons or rankings that are unstable to changes in the model set. We formalize this aggregation as a social choice problem where each metric induces a preference ranking over models on each dataset, and a benchmark operator aggregates these votes across metrics. While prior work has focused on Arrow's impossibility result, we identify sufficient conditions under which the impossibility disappears and meaningful multi-criteria benchmarking becomes possible. In particular, we deal with three restrictions on the combinations of rankings and show that on single-peaked, group-separable and distance-restricted preferences, the benchmark operator allows for the construction of well-behaved rankings of the involved models. Empirically, we investigate several modern benchmark suites like HELM MMLU and verify which structural conditions are fulfilled on which benchmark problems.

## 1 INTRODUCTION

Benchmarks are fundamental to progress in modern machine learning, for they provide standardized frameworks for model evaluation and comparison (Hardt & Recht, 2022). Within the past decades, evaluation methods in the field have expanded from single-task benchmarks such as leaderboards based on the ImageNet dataset to multi-task suites including GLUE (Wang et al., 2018), SuperGLUE (Wang et al., 2019), MMLU (Hendrycks et al., 2021) and BIG-Bench (Srivastava et al., 2023). More recently, holistic evaluation frameworks like HELM (Liang et al., 2023) have been introduced that combine hundreds of tasks to assess language models across a broad range of scenarios and along different metrics that go beyond accuracy. Yet, as these benchmark suites have grown more complex, a central challenge has emerged: how should we aggregate performance across multiple evaluation criteria, each measured over several datasets, into one meaningful ranking of models?

The prevailing approach in benchmarking is based on the "unspoken utilitarian principles" (Rofin et al., 2023), ranking systems by the arithmetic mean of their scores across task-specific metrics (Wang et al., 2018; 2019; Hendrycks et al., 2021). Nonetheless, this aggregation procedure has been increasingly viewed as inadequate, for instance, due to its sole focus on the value of model predictions and disregard for the cost (e.g., model size, training time) of those predictions (Ethayarajh & Jurafsky, 2020). Moreover, mean aggregation can be easily dominated by performance on a few outlier tasks (Agarwal et al., 2021). In general, averaging methods can mislead when samples vary in difficulty, resulting in inflated model performance and unreliable rankings (Mishra & Arunkumar, 2021). The problem is amplified when benchmarks include ordinal criteria such as interpretability, since mean aggregation presupposes a meaningful scale of differences that such metrics do not provide. Further challenges such as task selection bias (Dehghani et al., 2021), saturation of benchmarks over time (Kiela et al., 2021) as well as limitations of dynamic benchmarks (Shirali et al., 2023) have been identified. However, it has also been argued that a focus on failures merely incentivizes the development of models that sidestep those mistakes by shifting errors elsewhere, creating only the appearance of progress (Bowman & Dahl, 2021). Instead of focusing on limitations, our work identifies conditions under which meaningful multi-criteria benchmarking becomes possible. Rather than asking "what goes wrong?", we ask "under what structural conditions can benchmarking succeed?".



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

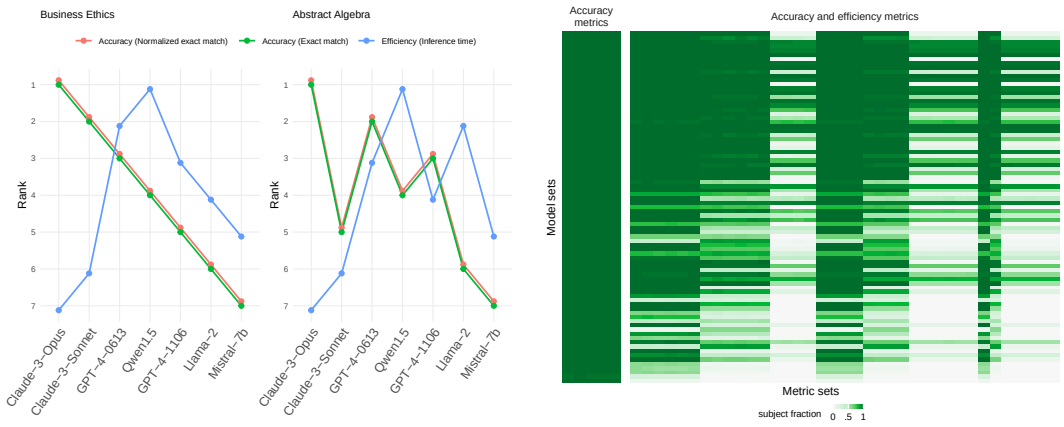



Figure 1: Single-peaked structure in HELM MMLU. Left: for each MMLU subject dataset, each of the three HELM metrics induces a ranking of models; we study when these rankings are consistent with a single shared ordering of models (x-axis) so that each metric has one "sweet spot" (*peak*) along that ordering. *Business Ethics* satisfies this structure, *Abstract Algebra* does not. Right: each tile is one combination of model set and metric set; color shows the fraction of 57 MMLU subjects on which metric rankings are single-peaked. In 1426 out of 3486 evaluation settings, single-peakedness holds for all subjects.

Social choice theory – the "study of collective decision procedures and mechanisms" (List, 2022) – offers a precise vocabulary to assess aggregation processes. While Arrow's theorem 1950 establishes that no aggregation function can satisfy all desirable properties simultaneously over the *universal domain* (i.e., the function works for every possible pattern of preferences), the social choice literature has long recognized that *domain restrictions* – natural constraints on the inputs of the aggregation rule – can restore the possibility of meaningful aggregation (Black, 1948; Inada, 1964; Sen, 1966; Dietrich & List, 2010; Puppe & Slinko, 2024). In the context of benchmarking, this raises a fundamental question that has so far received surprisingly little attention: **Do multi-criteria benchmarks naturally exhibit structure in the domain that restores possibility?** In contrast to prior work focusing on impossibility (Zhang & Hardt, 2024), we deal with sufficient conditions under which well-behaved multi-criteria benchmarking becomes possible.

**Contributions.** We make the following three contributions.

1. We formalize multi-criteria benchmarking as a preference aggregation problem, where metrics (e.g., accuracy, fairness, efficiency) act as voters and models act as alternatives.

2. We prove that if metric rankings satisfy *single-peakedness*, *group separability* or *distance-restrictedness*, then pairwise majority yields well-behaved leaderboard rankings. Intuitively, single-peakedness means that there exists a one-dimensional ordering of models such that each metric has a single "sweet spot", and moving away from it along the ordering makes models consistently less preferred by that metric (see Figure 1). When it holds empirically, single-peakedness gives an ordinal counterpart to recent evidence that benchmark performance often has a low-dimensional, sometimes nearly one-dimensional, structure (Ruan et al., 2024; Hardt, 2025).

3. We test these structural assumptions on HELM MMLU across all 83 model sets and all 42 metric subsets of sizes 3-6, under two deterministic tie-breaking rules (6972 evaluation settings) and ten repetitions of random tie-breaking. We compare with matched uniformly random profiles and conduct further experiments on PMLB and OpenML.

## 2 Related Work

A growing body of literature has turned to social choice theory to formalize the aggregation problem in benchmarks. Early contributions include Eugster et al. (2012) who frame benchmarking as consensus over dataset-level preference relations as well as Mersmann et al. (2015) who study

benchmarking of optimization algorithms as consensus-ranking problem over many test functions. Colombo et al. (2022) propose an aggregation method based on Kemeny consensus, while Himmi et al. (2024) develop a ranking approach using the Borda count for managing missing scores in benchmarks. Rofin et al. (2023) introduce VOTE'N'RANK, a framework applying eight social choice procedures onto multi-task benchmarks. Zhang & Hardt (2024) use Arrow's theorem to analyze the aggregation problem in multi-task benchmarks, highlighting a trade-off between diversity and stability. Lanctot et al. (2025) employ principles from social choice and game theory for building a framework to evaluate general agents. Complementary work Demšar (2006); Benavoli et al. (2017); Jansen et al. (2023; 2024); Longjohn et al. (2025) proposes statistically sound methods for aggregating performance across multiple tasks.

## 3 Pathologies in Multi-Criteria Benchmarking

We deal with multi-criteria benchmarking that aims to evaluate models along different criteria/metrics, each measured over multiple datasets. The central challenge is how to obtain a well-behaved, overall ranking of models from criteria across qualitatively different dimensions. But what does "well-behaved" mean? We argue that two intuitive properties are *transitivity* and *stability*.

**Transitivity.** One plausible constraint is that the overall ranking of models should not contradict itself. For three models $A, B, C$, it may happen that metrics collectively prefer $A$ to $B$, $B$ to $C$, and yet $C$ to $A$. Even if each metric induces a sound ordering of models, there does not necessarily exist a transitive aggregated ranking over all metrics. We illustrate this tension on MMLU under HELM, which evaluates 23 language models across 57 subjects and records multiple metrics per subject (Hendrycks et al., 2021; Liang et al., 2023). Each metric generates a ranking of models for each subject. A natural aggregation method that treats all metrics equally is pairwise majority. We consider each metric as one *vote* and declare that $A$ beats $B$ if more metrics prefer $A$ to $B$ than prefer $B$ to $A$. Table 1 (left) illustrates that these pairwise comparisons may result in cyclic majority rankings such as GPT-4 $\succ$ Qwen1.5 $\succ$ GPT-3.5 $\succ$ GPT-4 (where $\succ$ is short for "is preferred to"). The majority ranking is not well-behaved: even this basic aggregation method fails at delivering a coherent ranking of models across three metrics. Since the pairwise majority relation is cyclic on a subset of models, there is no transitive ranking over the full model set that agrees with all majority comparisons at once. This phenomenon is not rare. Searching over triples of accuracy and efficiency metrics in HELM MMLU, we find such cycles in 51 of 57 subjects (Appendix F).

Table 1: Two pathologies in multi-criteria benchmarking. Left: on *Formal Logic*, columns list the three models from best to worst under that metric, with values in parentheses; for *Accuracy* higher means better, for *Time* and *Length* lower means better. Pairwise majority yields GPT-4 $\succ$ Qwen1.5 $\succ$ GPT-3.5 $\succ$ GPT-4. Right: on *High School World History*, rankings are computed across *Accuracy*, *Time* and *Length* by average rank before and after Llama-2 is added. The relative order of Claude-3-Opus and GPT-3.5 flips. Exact model identifiers can be found in Appendix F.

| **Accuracy** | **Time (s)** | **Length (bytes)** |
|---|---|---|
| GPT-4 (0.65) | Qwen1.5 (0.32) | GPT-3.5 (1.00) |
| Qwen1.5 (0.49) | GPT-3.5 (0.41) | GPT-4 (1.17) |
| GPT-3.5 (0.40) | GPT-4 (0.49) | Qwen1.5 (2.00) |

| **Before** | **After** |
|---|---|
| 1. GPT-4-0613 | 1. GPT-4-0613 |
| 2. GPT-4-1106 | 2. GPT-4-1106 |
| 3. Claude-3-Opus | 3. Qwen1.5 |
| 4. Qwen1.5 | 4. GPT-3.5 |
| 5. GPT-3.5 | 5. Claude-3-Opus |
| ⋮ | ⋮ |
| 15. Google-Text-Bison | 15. Google-Text-Bison |
| | 16. Llama-2 |

**Stability.** We can avoid cycles in benchmark rankings by using another aggregation method, for instance, any rule based on combining ranks of models across different datasets/tasks or metrics. One example is Borda score (de Borda, 1781), which is being increasingly used in benchmarks such as MTEB (Chung et al., 2025) and yields identical rankings as mean win rate, a widely used benchmarking rule (Liang et al., 2023; Lee et al., 2023; Zhang & Hardt, 2024; Hardt, 2025). Then, however, we face another problem: benchmarks may become unstable to irrelevant changes in the model set. Stability means that adding or removing a model $C$ should not flip the ranking between

two existing models $A$ and $B$. Table 1 (right) demonstrates the violation of stability on one dataset in HELM MMLU. We consider 15 models with the highest accuracy on this subject among those with complete measurements for the three metrics, and then add one additional model. We compute the overall rankings of models before and after the change across metrics by averaging the metrics' ranks of models (Borda score over ranks). Note that the new model, Llama-2, is worse on *Accuracy* compared to all existing 15 models, yet relatively efficient with an *Time* faster than Claude-3-Opus but slower than GPT-3.5. Crucially, the addition of Llama-2 changes the rankings of models near the top of the leaderboard. We find such flips in 44 out of 57 MMLU subjects (see Appendix F).

By no means do these failures imply the impossibility of well-behaved multi-criteria benchmarking. Rather, these examples tell us what must be true of the rankings induced by each metric if transitive and stable leaderboard rankings are to be possible. For the aforementioned pathologies arise only in the *unstructured* combinations of rankings. In this paper, we show meaningful aggregation can be achieved and analyze which structures are sufficient for transitive and stable leaderboards.

# 4 Benchmarking as a Social Choice Problem

## 4.1 Notation

**Multi-criteria benchmarking.** Let $\mathcal{D}$ be some fixed universe of instances/datasets/tasks and let $\mathcal{A}$ be some fixed and finite set of algorithms/models with $k := |\mathcal{A}| \in \mathbb{N}$. Let $(\phi_i)_{i\in[n]}$ denote a family of metrics/criteria $\phi_i : \mathcal{A} \times \mathcal{D} \to [0, 1]$, for some $n \in \mathbb{N}$, where $[n] := \{1, \ldots, n\}$. Let $\Phi = (\phi_1, \ldots, \phi_n) : \mathcal{A} \times \mathcal{D} \to [0, 1]^n$ be a multivariate metric.[1] The goal of multi-criteria benchmarking is to search for a model from $\mathcal{A}$ that best balances the metrics $\phi_1, \ldots, \phi_n$ across the various datasets $D \in \mathcal{D}$.

**Preference aggregation problem.** Let pref($\mathcal{A}$) be the set of all complete and transitive *preference relations* on $\mathcal{A}$. Each fixed metric $i \in [n]$ and dataset $D \in \mathcal{D}$ induce a preference relation $R_{D,i} \in$ pref($\mathcal{A}$), defined by setting $A \succeq_{R_{D,i}} A' :\Leftrightarrow \phi_i(A, D) \geq \phi_i(A', D)$.[2] Hence, $R_{D,i}$ corresponds to the ranking of models in $\mathcal{A}$ induced by the scores of the $i$th performance metric, evaluated on instance $D$. We denote by $P_{D,i}$ the strict part of $R_{D,i}$ defined by $A \succ_{P_{D,i}} A' :\Leftrightarrow \phi_i(A, D) > \phi_i(A', D)$. Since metric scores can tie, $R_{D,i}$ may include indifference between models; we use $P_{D,i}$ whenever strict comparisons are necessary. We obtain $P_{D,i}$ from $R_{D,i}$ by breaking ties according to a fixed deterministic rule. Furthermore, let prof($\mathcal{A}, \mathcal{D}, \Phi$) denote the set of *preference profiles* induced by $\mathcal{D}$, i.e., the set prof($\mathcal{A}, \mathcal{D}, \Phi$) = $\{R_D := (R_{D,1}, \ldots, R_{D,n}) | D \in \mathcal{D}\}$. Each profile $R_D$ arises by collecting the preference rankings induced by metrics $\phi_1, \ldots, \phi_n$ evaluated on instance $D$ in an $n$-tuple. Let $B$ : prof($\mathcal{A}, \mathcal{D}, \Phi$) $\to$ pref($\mathcal{A}$) be a *benchmark operator* – a map returning a preference order among the algorithms of interest for every profile of such rankings that is inserted to it. Importantly, the domain of the operator $B$ depends on each of the objects $\mathcal{A}, \mathcal{D}, \Phi$.

In HELM MMLU, $\mathcal{D}$ corresponds to the set of 57 MMLU subject datasets, while each $D \in \mathcal{D}$ is a single subject (e.g. *Formal Logic*). The set of evaluated language models is denoted by $\mathcal{A}$, while $\Phi$ is the collection of all reported HELM metrics, with each $\phi_i$ corresponding to the metric $i$. The relation $R_{\textit{Formal Logic, Time}}$ denotes the ranking of models according to metric *Time* on subject *Formal Logic*, while the profile $R_{\textit{Formal Logic}}$ includes the rankings induced by all metrics on this subject. The benchmark's overall ranking of models on that subject is given by $B(R_{\textit{Formal Logic}})$.

There are many kinds of benchmark operators, including pairwise majority and mean win rate mentioned in Section 3. Pairwise majority aggregation is the most reasonable local aggregation rule, though its downside – the possibility of cyclic majority rankings – is well-known. We show that there are many realistic benchmarking scenarios where pairwise majority not only produces a transitive overall ranking of models but also satisfies further desirable properties that we introduce in the next section. This makes it, under the structural assumptions we study, more attractive for

[1] Note that our framework is more general than a mere multi-criteria benchmarking problem: any problem that produces evaluations over a finite set of alternatives can be cast in terms of $\mathcal{D}$, $\mathcal{A}$ and $\Phi$, e.g., one may treat datasets or tasks as voters (Zhang & Hardt, 2024) or consider test functions as voters in optimizer benchmarking (Rodemann & Blocher, 2024).

[2] In general, we assume that larger values of metric $\phi$ mean better performance on that metric. Otherwise, e.g, for metrics *Time* and *Length*, we flip the sign of the values.

benchmarking than rules like mean win rate.[3] For any models $A, A' \in \mathcal{A}$ and any tasks $D \in \mathcal{D}$, $M_D(A, A') := |\{i \in [n] : A \succeq_{R_{D,i}} A'\}|$ counts how many metrics rank $A$ at least as high as $A'$ and $A \succeq_M^D A' \Leftrightarrow M_D(A, A') \geq M_D(A', A)$.

### 4.2 Social Choice: From Impossibility to Possibility

**Arrow's theorem.** Social choice is known for its impossibility results, most prominently Arrow's theorem 1950. Yet, the problem behind impossibilities in voting was recognized much earlier, by Condorcet 1785 who formalized a situation where three voters rank the alternatives $x$, $y$, and $z$ and use majority voting on each pair of alternatives. Even though each voter's preference is rational, majority prefers $x$ over $y$, $y$ over $z$, and $z$ over $x$, yielding a *Condorcet cycle* and an irrational collective ranking. The cycle in Table 1 corresponds to the Condorcet paradox. Arrow 1950 generalized it by showing that, given a set of individual preferences over three or more alternatives, there exists no aggregation procedure that satisfies a few reasonable assumptions (called *axioms*) concerning the autonomy of voters and the rationality of their preferences Morreau (2025).

**Theorem 4.1.** *Arrow (1950) Suppose $k \geq 3$ and $n \geq 2$. Let $F$ be an operator that maps each profile from $(\text{pref}(\mathcal{A}))^n$ (**Universal Domain**) to a preference relation in $\text{pref}(\mathcal{A})$ (**Social Ordering**). For profiles $R = (R_1, \ldots, R_n)$ and $R' = (R'_1, \ldots, R'_n)$, let $P_i, P'_i$ be the strict parts of $R_i$ and $P, P'$ the strict parts of $F(R), F(R)'$. Then $F$ cannot satisfy the following conditions simultaneously:*

***Non-Dictatorship:** There is no $i \in [n]$ such that $\forall R$ and $\forall A, A' \in \mathcal{A}$, if $A \succ_{P_i} A'$, then $A \succ_P A'$.*

***Weak Pareto:** $\forall R$ and $\forall A, A' \in \mathcal{A}$, if $\forall i \in [n]\ \ A \succ_{P_i} A'$, then $A \succ_P A'$.*

***Independence of Irrelevant Alternatives (IIA):** $\forall R, R'$ and $\forall A, A' \in \mathcal{A}$, if $\forall i \in [n]\ \ A \succeq_{P_i} A' \Leftrightarrow A \succeq_{P'_i} A'$, then $A \succeq_P A' \Leftrightarrow A \succeq_{P'} A'$.*[4]

*Social Ordering* requires the benchmark to output a coherent ranking, i.e. a complete and transitive binary relation on $\mathcal{A}$. *Universal Domain* demands that the aggregation rule admits any logically possible combination of metric rankings over $\mathcal{A}$. This implies that we are not allowed to make any assumptions about metrics' behavior or relationships between metrics in advance. In benchmarking, however, *Universal Domain* is not a reasonable assumption. For metrics do exhibit systematic behavior, and trade-offs across families of metrics are well-known (e.g., Kaplan et al. (2020); Ang et al. (2022)). *Weak Pareto* asks that if every metric ranks model $A$ over model $A'$, then the collective ranking across metrics must prefer $A$ to $A'$. *Non-Dictatorship* rules out any single metric that always determines the overall preferential ranking across all metrics. Finally, the theorem's most demanding condition of *IIA* states that the social comparison of models $A$ and $A'$ depends only on how each metric compares the two models and not on how any metric ranks other models. Thus, the intransitivity in Table 1 is a violation of *Social Ordering*, while the instability to changes in model set implies a violation of the benchmark analogue of *IIA*.

**Restricted preference domains.** A longstanding literature (Black, 1948; Sen, 1966; Inada, 1969; Dietrich & List, 2010; Elkind et al., 2025) investigates *possibilities*, under which majority cycles can be avoided. The idea is to relax or violate one of Arrow's axioms in a plausible manner under which the remaining conditions can become jointly satisfiable. One "escape route" from impossibility is to violate *Universal Domain* and restrict attention to profiles with certain structure rather than allowing all logically possible profiles of complete and transitive preferences. If a voter's preferences fall into a restricted domain, the possibility of meaningful aggregation can be restored. The well-behaved restricted domains are called *Condorcet domains* and examples include single-peaked, group-separable and value-restricted preferences (Puppe & Slinko, 2024). Sen's theorem 1966 provides a sufficient condition under which the pairwise majority relation yields a transitive ranking. While Sen's condition is very general, it is comparatively technical (Elsholtz & List, 2005) and not

[3] In benchmarking, rules based on combining ranks of models across tasks (e.g. mean win rate) are common. However, these procedures are unstable to changes in model sets (e.g., see Table 1) and violate Arrow's *IIA* (Benavoli et al., 2016).

[4] Note that the operator $F$ generally differs from the operator $B$ as defined earlier, since in most cases the benchmark suite $B$ does not induce the full domain of preference profiles $(\text{pref}(\mathcal{A}))^n$. This is why Arrow's impossibility theorem cannot be explicitly applied to the benchmark operator $B$.

interpretable in the benchmarking context. Thus, we focus on restricted domains with clearer meaning and investigate domains of single-peaked, group-separable and distance-restricted preferences.

### 4.3 Aggregation Across Datasets

So how can we go from a benchmark operator to a ranking of the models over all datasets? Prior to the aggregation across datasets, we impose the domain restrictions on the profile $R_D$ of rankings induced by metrics within each dataset. If $R_D$ is single-peaked, group-separable or distance-restricted to degree 1, then the benchmark operator yields a coherent ranking. We then consider the family of orders generated by $B$, given by $\text{ord}(B, \mathcal{D}) := \big(B(R_D)\big)_{D \in \mathcal{D}}$, which contains one aggregate ranking $B(R_D)$ for each dataset. The domain restrictions in the previous step ensure that each $B(R_D)$ is a complete and transitive ranking. If we now want to choose *one* ranking from this list, we should choose one that best represents the family. Under a statistical perspective, every dataset $D$ and its ranking $B(R_D)$ can be treated as a data point of some population (i.e., of $\mathcal{D}$) of interest. A natural approach is based on *depth functions* (Zuo & Serfling, 2000) and its generalizations to non-standard data spaces, such as relations. There are different depth functions available, for example the ufg depth (Blocher et al., 2024; Blocher & Schollmeyer, 2024) or a generalization of Tukey depth 1975, firstly used in Jansen et al. (2018) under the name *commonality sharing rule* and analyzed in Blocher & Schollmeyer (2025), for other depth based approaches, cf., e.g., Goibert et al. (2022); Nehring & Puppe (2023). In this paper, we use the generalized Tukey depth where the objects of interest are rankings. The idea is to choose from the population $\text{ord}(B, \mathcal{D})$ of rankings the so-called *commonality sharing ranking(s)* – the ranking(s) that share(s) with every subpopulation of minimum-size $k$ the commonalities of the subpopulation, where $k$ is chosen as small as possible. Here, commonalities are all pairs $(A, A^{'})$ which are ranked identically by all members of the subpopulation. This step requires the benchmark operator to produce exclusively acyclic relations: otherwise, the commonality sharing rule is not well defined.[5]. Thus, the role of the domain restrictions is to ensure that the objects being aggregated across datasets are coherent rankings

## 5 Restricted Preference Domains in Benchmarking

### 5.1 Single-Peaked Preferences

Given fixed objects $\mathcal{A}$, $\mathcal{D}$ and $\Phi$, we now examine the structure in the set of preference profiles $\text{prof}(\mathcal{A}, \mathcal{D}, \Phi)$. We consider three types of domain restrictions: single-peakedness, group separability and distance-restrictedness. Conditional on suites with this specific structure, we show that pairwise majority satisfies *Non-Dictatorship*, *IIA*, *Weak Pareto* and *Social Ordering*. Hence, this section turns the two properties in Section 3 into testable sufficient conditions.

The first and most famous restricted preference domain was established by Black (1948) and independently discovered by Arrow (1951). Single-peakedness means that there exists a way to arrange models along some spectrum so that each metric's ranking of models has a single *peak* on this spectrum. This can be viewed as the "sweet spot" of each metric where the most preferred models for this metric lie, while performance decreases when moving further away from that region on the spectrum. In benchmark practice, this one-dimensional axis can be observed in several contexts, reflecting different trade-offs of model design. For instance, along an axis ordered by model size, interpretability and efficiency may prefer smaller models, while accuracy may peak at an intermediate or larger model size.

Our definition of single-peaked preference profiles follows Ballester & Haeringer (2011, Sec. 3). Assume $S \subseteq \mathcal{A}$ (with $|S| \geq 3$). Let $\mathcal{L}$ be the set of all strict linear orders over $S$ and set $L \in \mathcal{L}$. For any two models $A, A^{'} \in \mathcal{A}$, we write $A \succ_L A^{'}$ if $A$ precedes $A^{'}$ in the ordering $L$.

**Definition 5.1.** For any set $S \subseteq \mathcal{A}$ let $w(S, P_{D,i})$ denote the **least preferred model** in $S$ according to the strict preference relation $P_{D,i}$ and let $b(S, P_{D,i})$ be the **most preferred model** in $S$ according to $P_{D,i}$. Let $L$ be an **admissible orientation of** $S$ **with respect to** $P_{D,i}$ if for any three models $A, A^{'}, A^{''} \in S$ with $A^{'} = b(S, P_{D,i})$ being the most preferred model in $S$ according to metric $i$

[5]The generality of the commonality sharing rule (and other methods like the ufg depth) would in principle allow one to extend the aggregation ideas to acyclic relations. However, the aggregated result would then generally be only a cyclic relation, which is not satisfying for a leaderboard ranking.

such that $A^{'} \succ_L A \succ_L A^{''}$ or $A^{''} \succ_L A \succ_L A^{'}$, we have $\forall A, A^{''} \in S$, $A \succ_{P_{D,i}} A^{''}$. Denote the set of all admissible orientations of $S \subseteq \mathcal{A}$ with respect to $P_{D,i}$ by $\mathcal{L}_S(P_{D,i})$ and set $\mathcal{L}_S(R_D) = \bigcap_{i\in[n]} \mathcal{L}_S(P_{D,i})$. A profile $R_D$ is called **single-peaked** if there exists an ordering $L$ over the set of algorithms $\mathcal{A}$ with $L \in \mathcal{L}_{\mathcal{A}}(R_D)$ such that $\mathcal{L}_{\mathcal{A}}(R_D) \neq \varnothing$.

**Theorem 5.2.** *Fix a dataset $D \in \mathcal{D}$ and a finite set of algorithms $\mathcal{A}$. Let $R_D$ be a single-peaked profile induced by $D \in \mathcal{D}$. Assume $n$ is odd, with $n \geq 3$. Then the weak pairwise majority relation $\succeq_M^D$ satisfies Non-Dictatorship, Weak Pareto, IIA and yields a complete and transitive order on $\mathcal{A}$.*[6]

### 5.2 Group-Separable Preferences

Inada (1964; 1969) has introduced the class of group-separable domains which are sufficient for a transitive social ranking and fulfillment of all Arrow's axioms except for *Universal Domain*, provided that the number of voters is odd. In a group-separable profile, for any subset of models, there exists at least one group of models that every metric regards as entirely above (or entirely below) the rest. This is very plausible in cases where the metrics measure the same latent construct, for instance, when using several accuracy metrics. Even if the metrics are not identical, they typically agree that some models are better than the others. But the condition can also hold in multi-objective comparisons, such as combining accuracy and efficiency metrics when the benchmark contains models that are consistently dominated on the chosen metric set, e.g. models that tend to be less accurate and less efficient than several others. Whenever such dominated models are present, the metrics agree on a *separation* that places them entirely below the rest (and similarly, if a subset contains models that are both accurate and efficient, they can form a block that all metrics place above the rest). We define group-separability following Ballester & Haeringer (2011, Sec. 4).

**Definition 5.3.** Assume $S \subseteq \mathcal{A}$ (with $|S| \geq 3$). Let a non-empty set $E \subset S$ be a **separation with respect to** $P_{D,i}$ if either $\forall A \in E, \forall A^{'} \in S \backslash E : A \succ_{P_{D,i}} A^{'}$ or $\forall A \in E, \forall A^{'} \in S \backslash E : A^{'} \succ_{P_{D,i}} A$. Let $\mathcal{S}_{P_{D,i}}$ be the set of all separations of $S$ with respect to $P_{D,i}$ and $\mathcal{S}_{R_D} = \bigcap_{i\in[n]} \mathcal{S}_{P_{D,i}}$. A profile $R_D$ is called **group-separable** if for any $S \subseteq \mathcal{A}$, $\mathcal{S}_{R_D} \neq \varnothing$.

**Theorem 5.4** (Inada, 1964; 1969)**.** *Fix a dataset $D \in \mathcal{D}$ and a finite set of algorithms $\mathcal{A}$. Assume $n$ is odd, with $n \geq 3$. Let $R_D$ be a group-separable profile induced by $D \in \mathcal{D}$. Then the strict pairwise majority relation $\succ_M^D$ satisfies Non-Dictatorship, Weak Pareto, IIA and yields a complete and transitive order on $\mathcal{A}$.*

### 5.3 Distance-Restricted Preferences

So far we have investigated two different origins of structure in benchmarks: structure from an axis over models resulting in each metric's "sweet spot" and structure from clustering of models with respect to their performance on different metrics. Now we deal with structure emerging from a situation of *near-consensus* among metrics. This is particularly valuable when benchmarking aims at comparing models in terms of a single evaluation criterion such as robust accuracy, which is too complex to be expressed in terms of a single metric.

**Definition 5.5.** A profile $R_D$ is called **distance-restricted to degree** $p \in \mathbb{N}$, if for all $i, j \in [n]$ we have that $d_S(P_{D,i}, P_{D,j}) \leq p$, where $d_S(P_{D,i}, P_{D,j})$ is defined as $|\{(A, A^{'}) \mid A \succ_{P_{D,i}} A^{'} \text{ and } A^{'} \succ_{P_{D,j}} A\}|$, where $d_S(P_{D,i}, P_{D,j})$ is the **swap distance** counting the number of disagreeing pairs between $P_{D,i}$ and $P_{D,j}$.

The distance $d_S$ coincides with the Kendall $\tau$ distance between strict rankings (Kendall, 1948). In social choice literature, it is also known as the Kemeny distance between two preference orders, measuring the number of pairwise comparisons on which two orders disagree (Kemeny, 1959; Kemeny & Snell, 1962; Baigent, 1987; Bossert & Storcken, 1992; Can & Storcken, 2018).

A $p$-distance-restricted profile means that any two metrics disagree on at most $p$ pairwise comparisons of models. This is plausible when metrics assess the same latent construct. For instance, *exact_match*, *prefix_exact_match* and *quasi_exact_match* from HELM all measure whether the correct answer matches the model's prediction, differing only in how much extra text or small formatting variation they allow. Hence, we expect a near-consensus: the rankings differ only on a few pairwise disagreements. In the case of $p = 1$, Condorcet cycles cannot appear and we get the following result.

[6]The proof of this and the following theorems can be found in Appendix A-C.

**Theorem 5.6.** *Fix a dataset $D \in \mathcal{D}$ and a finite set of algorithms $\mathcal{A}$. Suppose $n \geq 3$. Assume the induced profile $R_D$ is distance-restricted to degree 1. Then, the weak pairwise majority relation $\succeq_M^D$ is a transitive and complete relation on $\mathcal{A}$. It satisfies Non-Dictatorship, Weak Pareto and IIA.*

# 6 EXPERIMENTS

## 6.1 EXPERIMENTAL SETUP

We use the MMLU benchmark under HELM (v1.0.0) which includes 23 models evaluated on 57 subjects across multiple evaluation metrics. For each run directory, we identify the MMLU subject as well as the evaluated model and extract the subject-specific mean metric values reported by HELM. When multiple runs exist for the same model-subject pair under different evaluation settings, we keep the most frequent setting and average across repeats.

We consider 83 model sets: 23 sets are selected to represent common benchmark comparisons, such as comparing strong systems, models from the same provider and mixtures of open and closed models; the remaining 60 sets are random, sampled without replacement from the 22 models with full MMLU coverage, with 30 sets of size 3, 20 of size 4 and 10 of size 5 (see Appendix D for the full list of models; the model *google/gemini-pro* was excluded as HELM reports only 5 of 57 MMLU subjects for it). We use all 42 metric subsets of sizes 3–6 drawn from the set of six metrics, which includes four accuracy metrics *exact_match*, *prefix_exact_match*, *quasi_exact_match*, *quasi_prefix_exact_match* and two efficiency metrics *inference_runtime* and *num_bytes*. Metrics are voters, while models are alternatives. For each subject and each metric we sort the fixed models by their scores to obtain a preference ranking. We consider both the alphabetical and the reverse alphabetical tie-breaking rule for resolving equal scores. We test the three domain restrictions across all combinations of model sets, metric subsets and tie-breaking rules, resulting in $83 \times 42 \times 2 = 6972$ settings. We also repeat the analysis ten times with random tie-breaking and compare with matched uniformly random profiles.

For each configuration $(\mathcal{A}, \Phi, \text{tie})$, we compute the fraction of subjects with complete data on which the induced preference profile is single-peaked, group-separable and distance-restricted to degree 1. We also compute the mean Kendall $\tau$ between the metrics' rankings on a subject, averaged across the 57 subjects. We test for single-peakedness by enumerating all possible permutations of the models that correspond to potential axes and checking whether each metric's ranking is single-peaked with respect to that axis. Using the strict metrics' rankings obtained after tie-breaking, we test for group separability. Within any set $S$ of models, we repeatedly search for a nontrivial subset $E \subset S$ that all metrics rank entirely above or entirely below $S \setminus E$. If such a separation exists, we split $S$ into $E$ and $S \setminus E$ and continue on each part until only sets of fewer than three models remain. To test for distance-restrictedness, we compute the swap distance $d_S$ between every pair of metrics' rankings for each task. A subject is then declared distance-restricted to degree 1 if the maximum swap distance is at most 1. Further experiments on suites such as PMLB and OpenML are reported in Appendix E. For some domains which satisfy our structural assumptions, we also compute the overall ranking across all datasets by using the commonality sharing aggregation rule described in Section 4.3.

For each setting, the number of subjects (out of those with complete data for that configuration) on which a property holds is a binomial count and we report Wilson 95% confidence intervals (Wilson, 1927; Brown et al., 2001) on the corresponding probability – the chance that the property holds on a single MMLU subject under that configuration. The Wilson interval reflects uncertainty at the level of each configuration: treating the 57 MMLU subjects as Bernoulli trials drawn from a hypothetical population of benchmarks like MMLU, it bounds the probability at which the property holds in that configuration. To measure whether the existence of structure in the preference domains is associated with agreement between metrics' rankings, we fit binomial mixed-effects regressions with a logit link for metric subsets of sizes 3 and 4. For each configuration, the response is the number of MMLU subjects on which the property holds out of the number of subjects with complete data. The predictors are the mean Kendall $\tau$, a dummy variable for whether the metric subset includes an efficiency metric, the model set size and the metric subset size. We add random intercepts for the model set and the metric subset, as the same models and metrics appear in multiple configurations, so they cannot be treated as independent observations.

### 6.2 RESULTS

We first report the results under alphabetical tie-breaking, thus referring to the $83 \times 42 = 3486$ combinations of model and metric sets. All configurations have complete data for all $57$ MMLU subjects. Each structural restriction occurs more often than under the matched uniformly random baseline at the level of profiles on individual subjects. Single-peakedness holds on all $57$ subjects in $1426$ configurations ($40.9\%$), group separability in $1208$ configurations ($34.7\%$) and distance-restrictedness to degree $1$ in $461$ configurations ($13.2\%$). For many choices of models and metrics, these structural assumptions hold for every MMLU subject. For any such configuration in which the property holds on all 57 subjects, the Wilson 95% confidence interval is $[0.937, 1.000]$, which means that observing the property on all MMLU subjects is compatible with an underlying probability as low as $0.937$. The strongest pattern appears for metric sets containing exclusively accuracy metrics. Among $415$ such combinations, single-peakedness holds throughout the suite in $411$ cases ($99.0\%$), group separability in $365$ cases ($88.0\%$) and distance-restrictedness in $329$ cases ($79.3\%$). This is consistent with interpretation that the accuracy metrics measure the same latent construct and therefore induce highly aligned rankings of models. Among the $3071$ configurations containing efficiency metrics such as inference time or output length, single-peakedness still holds on all subjects in $1015$ cases ($33.1\%$) and group separability in $843$ cases ($27.5\%$). Distance-restrictedness holds in only $132$ cases ($4.3\%$). For $461$ out of $3486$ evaluation settings ($13.2\%$), all three domain restrictions are satisfied on all subjects; while for $1620$ ($46.5\%$) configurations, at least one of the three restrictions holds for all subjects.

These empirical patterns indicate when the sufficient conditions from Section 5 are likely to appear in practice. If evaluation criteria measure the same underlying capability, one can often obtain the structure needed for well-behaved pairwise majority aggregation. Mixing criteria across qualitatively different dimensions makes distance-restrictedness rare and shifts attention to single-peakedness and group separability. The choice of both models and metrics is therefore an inherent part of the benchmark design. The mixed-effects regressions for metric subsets of sizes $3$ and $4$ support this interpretation. Higher mean Kendall $\tau$ between metric rankings is positively associated with all three domain restrictions, while the inclusion of efficiency metrics and larger model sets are negatively associated with them. For single-peakedness and distance-restrictedness, larger metric sets are also negatively associated. All of these associations are statistically significant in the mixed-effects models and the signs of the coefficients remain unchanged in the quasi-binomial regression conducted for sensitivity analysis to allow for overdispersion. All aggregated rankings, the results for the reverse alphabetical tie-breaker and a detailed account of our regression results can be found in Appendix D.

## 7 DISCUSSION

What conditions must multi-criteria benchmarks satisfy to produce well-behaved rankings? Our results suggest that leaderboards require a structural diagnostic before they can be treated as meaningful model rankings. When benchmark designers want one ranking across competing objectives, the key question is whether the metric rankings have enough structure to make a single ordering defensible. Rank-based rules such as Borda score and mean win rate evaluate a model through its position relative to the whole comparison set; consequently, the ordering of two fixed models can flip when a third model enters. Pairwise majority avoids this, and its transitivity can be justified by structure: without a restricted domain, those pairwise wins can form cycles. The design recommendation is therefore a diagnostic protocol: convert each metric on each task into a ranking over the candidate models and test whether the resulting profile satisfies one of the discussed domain restrictions. If this diagnostic passes across all instances, pairwise majority gives coherent per-dataset rankings that can be combined across datasets by commonality sharing. If it fails, this does not by itself rule out a coherent ranking; the benchmark should investigate whether aggregation can still be justified before deciding to report separate dimensions or revise the model or metric set. Several limitations qualify these conclusions. First, our results are conditional: the restricted domains are sufficient conditions for well-behaved pairwise majority aggregation, but they are not necessary. Second, our empirical analysis is limited mainly to HELM MMLU, while the PMLB and OpenML results suggest that such structure is not equally prevalent across domains. Third, near-ties and deterministic tie-breaking can affect the induced rankings. Future work should study approximate versions of these domain restrictions that remain robust under score perturbations or pairwise ranking reversals.

## REPRODUCIBILITY STATEMENT

Proofs are provided in Appendices A –C. Section 6.1 and Appendix D describe preprocessing, model and metric sets and tie-breaking. Appendices E and F describe the additional experiments. Code is available at `https://anonymous.4open.science/r/Beyond-Arrow-From-Impossibility-to-Possibility-in-Multi-Criteria-Benchmarking-118F`.

## AI USE STATEMENT

We used generative AI to assist with editing, brainstorming, information searches and analysis of existing literature. We take responsibility for the final content of this work.

## REFERENCES


Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C. Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. *Advances in Neural Information Processing Systems*, 34, 2021.

Phyllis Ang, Bhuwan Dhingra, and Lisa Wu Wills. Characterizing the efficiency vs. accuracy trade-off for long-context NLP models. *arXiv preprint*, abs/2204.07288, 2022. URL `https://arxiv.org/abs/2204.07288`.

Kenneth J. Arrow. A difficulty in the concept of social welfare. *Journal of Political Economy*, 58 (4):328–346, 1950.

Kenneth J. Arrow. *Social Choice and Individual Values*. John Wiley and Sons, 1951.

Nick Baigent. Preference proximity and anonymous social choice. *The Quarterly Journal of Economics*, 102(1):161–169, 1987.

Miguel A. Ballester and Guillaume Haeringer. A characterization of the single-peaked domain. *Social Choice and Welfare*, 36:305–322, 2011.

Alessio Benavoli, Giorgio Corani, and Francesca Mangili. Should we really use post-hoc tests based on mean-ranks? *Journal of Machine Learning Research*, 17(5):1–10, 2016.

Alessio Benavoli, Giorgio Corani, Janez Demšar, and Marco Zaffalon. Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis. *Journal of Machine Learning Research*, 18(77):1–36, 2017.

Duncan Black. On the rationale of group decision-making. *Journal of Political Economy*, 56(1): 23–34, 1948.

Hannah Blocher and Georg Schollmeyer. Union-free generic depth for non-standard data. *arXiv preprint*, 2024. URL `https://arxiv.org/abs/2412.14745`.

Hannah Blocher and Georg Schollmeyer. Data depth functions for non-standard data by use of formal concept analysis. *Journal of Multivariate Analysis*, 205:105372, 2025.

Hannah Blocher, Georg Schollmeyer, and Christoph Jansen. *ddandrda: Data Depth and Relational Data Analysis*, 2023. R package version 0.0.0.9000.

Hannah Blocher, Georg Schollmeyer, Malte Nalenz, and Christoph Jansen. Comparing machine learning algorithms by union-free generic depth. *International Journal of Approximate Reasoning*, 169:109166, 2024.

Walter Bossert and Ton Storcken. Strategy-proofness of social welfare functions: The use of the Kemeny distance between preference orderings. *Social Choice and Welfare*, 9:345–360, 1992.

Samuel Bowman and George Dahl. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4843–4855, 01 2021.

Lawrence D. Brown, T. Tony Cai, and Anirban DasGupta. Interval estimation for a binomial proportion. *Statistical Science*, 16(2):101–133, 2001.

Burak Can and Ton Storcken. A re-characterization of the Kemeny distance. *Journal of Mathematical Economics*, 79(C):112–116, 2018.

Isaac Chung, Imene Kerboua, Márton Kardos, Roman Solomatin, and Kenneth C. Enevoldsen. Maintaining MTEB: Towards long term usability and reproducibility of embedding benchmarks. *arXiv preprint*, abs/2506.21182, 2025. URL https://arxiv.org/abs/2506.21182.

Pierre Colombo, Nathan Noiry, Ekhine Irurozki, and Stephan Clemencon. What are the best systems? New perspectives on NLP benchmarking. *Advances in Neural Information Processing Systems*, 2022.

J. C. de Borda. *Memoire sur les elections au scrutin*. Historie de l'Academie Royale des Sciences, 1781.

Nicolas de Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Imprimerie Royale, 1785.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery. *arXiv preprint*, 2107.07002, 2021. URL https://arxiv.org/abs/2107.07002.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(1):1–30, 2006.

Franz Dietrich and Christian List. Majority voting on restricted domains. *Journal of Economic Theory*, 145(2):512–543, 2010.

Edith Elkind, Martin Lackner, and Dominik Peters. Preference restrictions in computational social choice: A survey. *arXiv preprint*, 2025. URL https://arxiv.org/abs/2205.09092.

Christian Elsholtz and Christian List. A simple proof of Sen's possibility theorem on majority decisions. *Elemente der Mathematik*, 60(1):1–7, 2005.

Kawin Ethayarajh and Dan Jurafsky. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4846–4853, 01 2020.

M. Eugster, T. Hothorn, and F. Leisch. Domain-based benchmark experiments: Exploratory and inferential analysis. *Austrian Journal of Statistics*, 41:5–26, 2012.

Morgane Goibert, Stephan Clemencon, Ekhine Irurozki, and Pavlo Mozharovskyi. Statistical depth functions for ranking distributions: Definitions, statistical learning and applications. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 10376–10406. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/goibert22a.html.

Moritz Hardt. The emerging science of machine learning benchmarks. Online at https://mlbenchmarks.org, 2025. Manuscript.

Moritz Hardt and Benjamin Recht. *Patterns, predictions, and actions: Foundations of machine learning*. Princeton University Press, 2022.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Xiaodong Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *International Conference on Learning Representations*, 2021.

Anas Himmi, Ekhine Irurozki, Nathan Noiry, Stephan Clémençon, and Pierre Colombo. Towards more robust NLP system evaluation: Handling missing scores in benchmarks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 11759–11785, 01 2024.

Kenichi Inada. A note on the simple majority decision rule. *Econometrica*, 32:525–531, 1964.

Kenichi Inada. The simple majority decision rule. *Econometrica*, 37:490–506, 1969.

Christoph Jansen, Georg Schollmeyer, and Thomas Augustin. A probabilistic evaluation framework for preference aggregation reflecting group homogeneity. *Mathematical Social Sciences*, 96:49–62, 2018.

Christoph Jansen, Malte Nalenz, Georg Schollmeyer, and Thomas Augustin. Statistical comparisons of classifiers by generalized stochastic dominance. *Journal of Machine Learning Research*, 24 (231):1–37, 2023.

Christoph Jansen, Georg Schollmeyer, Julian Rodemann, Hannah Blocher, and Thomas Augustin. Statistical multicriteria benchmarking via the GSD-front. *Advances in Neural Information Processing Systems*, 37:98143–98179, 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint*, abs/2001.08361, 2020. URL https://arxiv.org/abs/2001.08361.

J.G. Kemeny and J.L. Snell. *Mathematical Models in the Social Sciences*. Ginn, 1962.

John G. Kemeny. Mathematics without numbers. *Daedalus*, 88(4):577–591, 1959.

Maurice George Kendall. *Rank Correlation Methods*. Charles Griffin, 1948.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4110–4124, 2021.

Marc Lanctot, Kate Larson, Yoram Bachrach, Luke Marris, Zun Li, Avishkar Bhoopchand, Thomas W. Anthony, Brian Tanner, and Anna Koop. Evaluating agents using social choice theory. *arXiv preprint*, 2312.03121, 2025. URL https://arxiv.org/abs/2312.03121.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Benita Teufel, Marco Bellagente, Minguk Kang, Taesung Park, Jure Leskovec, Jun-Yan Zhu, Fei-Fei Li, Jiajun Wu, Stefano Ermon, and Percy Liang. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 2023.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, and et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

Christian List. Social Choice Theory. In Edward N. Zalta and Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2022 edition, 2022.

Rachel Longjohn, Giri Gopalan, and Emily Casleton. Statistical uncertainty quantification for aggregate performance metrics in machine learning benchmarks. *arXiv preprint*, 2501.04234, 2025. URL https://arxiv.org/abs/2501.04234.

O. Mersmann, M. Preuss, H. Trautmann, B. Bischl, and C. Weihs. Analyzing the BBOB results by means of benchmarking concepts. *Evolutionary Computation*, 23:161–185, 2015.

Swaroop Mishra and Anjana Arunkumar. How robust are model rankings : A leaderboard customization approach for equitable evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 13561–13569, 05 2021.

Michael Morreau. Arrow's Theorem. In Edward N. Zalta and Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2025 edition, 2025.

Klaus Nehring and Clemens Puppe. Multi-dimensional social choice under frugal information: The tukey median as condorcet winner ex ante by. Technical report, KIT Working Paper Series in Economics, 2023.

Clemens Puppe and Arkadii Slinko. Maximal condorcet domains. a further progress report. *Games and Economic Behavior*, 145:426–450, 2024.

Julian Rodemann and Hannah Blocher. Partial rankings of optimizers. *arXiv preprint*, abs/2402.16565, 2024. URL https://arxiv.org/abs/2402.16565.

Mark Rofin, Vladislav Mikhailov, Mikhail Florinsky, Andrey Kravchenko, Tatiana Shavrina, Elena Tutubalina, Daniel Karabekyan, and Ekaterina Artemova. Vote'n'rank: Revision of benchmarking with social choice theory. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2023.

Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto. Observational scaling laws and the predictability of language model performance. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Amartya K. Sen. A possibility theorem on majority decisions. *Econometrica*, 34(2):491–499, 1966.

Ali Shirali, Rediet Abebe, and Moritz Hardt. A theory of dynamic benchmarks. *arXiv preprint*, 2210.03165, 2023. URL https://arxiv.org/abs/2210.03165.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, and et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research (TMLR)*, 2023.

John W Tukey. Mathematics and the picturing of data. In *Proceedings of the international congress of mathematicians*, volume 2, pp. 523–531. Vancouver, 1975.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, 01 2018.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems*, 32, 2019.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Guanhua Zhang and Moritz Hardt. Inherent trade-offs between diversity and stability in multi-task benchmarks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 58984–59002, 2024.

Yijun Zuo and Robert Serfling. General notions of statistical depth function. *The Annals of Statistics*, 28(2):461–482, 2000.

## A SINGLE-PEAKED PREFERENCES

Fix a dataset $D$, a set of models $\mathcal{A}$, and a multivariate performance measure $\Phi = (\phi_1, \ldots, \phi_n)$. After applying a fixed deterministic tie-breaking rule, each metric $i$ induces a strict linear order $P_i$ on $\mathcal{A}$.

**Lemma A.1.** *Black (1948) Let $R_D$ be a single-peaked profile. Assume $n$ is odd (with $n \geq 3$). Then the pairwise majority relation $\succeq_M^D$ is transitive.*

*Proof.* Since $R_D$ is single-peaked, there exists an ordering $L$ over $\mathcal{A}$ with $L \in \mathcal{L}_{\mathcal{A}}(R_D)$. By Black (1948) the pairwise majority is transitive for profiles that are single-peaked with respect to a common axis. Hence, $\succeq_M^D$ is transitive. □

**Theorem A.2.** *Let $R_D$ be a single-peaked profile induced by dataset $D \in \mathcal{D}$. Assume $n$ is odd (with $n \geq 3$). Then the pairwise majority relation $\succeq_M^D$ satisfies Non-Dictatorship, Weak Pareto, IIA and yields a complete and transitive relation on $\mathcal{A}$.*

*Proof.* The pairwise majority relation $\succeq_M^D$ is complete by definition. By Lemma A.1 it is transitive. Furthermore, the relation $\succeq_M^D$ satisfies *IIA* by construction. *Weak Pareto* holds because if every metric prefers model $A$ to model $A^{'}$, then the pairwise majority count for $A$ over $A^{'}$ is $n$, and $A \succ_M^D A^{'}$. For *Non-Dictatorship* consider the following. Fix a metric $i$. Since $R_D$ is single-peaked, there exists an ordering $L$ over $\mathcal{A}$ with $L \in \mathcal{L}_{\mathcal{A}}(R_D)$. Assume that three models $A, A^{'}, A^{''}$ are ordered on $L$ such that $A \succ_L A^{'} \succ_L A^{''}$. Further assume that a metric $i$ has its peak at $A$ so that $A \succ_{P_{D,i}} A^{'} \succ_{P_{D,i}} A^{''}$ and every other metric $j \neq i$ has its peak at $A^{''}$ so that $A^{''} \succ_{P_{D,j}} A^{'} \succ_{P_{D,j}} A$. Then, pairwise majority over all metrics does not rank $A$ over $A^{''}$. It follows that metric $i$ is not a dictator. Since $i$ was arbitrarily chosen, no dictator exists. □

## B GROUP-SEPARABLE PREFERENCES

Fix a dataset $D$, a set of models $\mathcal{A}$, and a multivariate performance measure $\Phi = (\phi_1, \ldots, \phi_n)$. After applying a fixed deterministic tie-breaking rule, each metric $i$ induces a strict linear order $P_i$ on $\mathcal{A}$.

**Lemma B.1.** *Inada (1964; 1969) Let $R_D$ be a group-separable profile. Assume $n$ is odd (with $n \geq 3$). Then the strict pairwise majority relation $\succ_M^D$ is transitive.*

*Proof.* By Ballester & Haeringer (2011), the profile $R_D$ is medium-restricted, since this condition is implied by group separability. By Inada (1964; 1969), the pairwise majority relation is transitive for a medium-restricted profile, given an odd number of metrics under consideration. Hence, $\succ_M^D$ is transitive. □

**Theorem B.2.** *Let $R_D$ be a group-separable profile induced by dataset $D \in \mathcal{D}$. Assume $n$ is odd (with $n \geq 3$). Then the pairwise majority relation $\succ_M^D$ satisfies Non-Dictatorship, Weak Pareto, IIA and yields a complete and transitive order on $\mathcal{A}$.*

*Proof.* Fix $A \neq A^{'}$. Since each metric induces a strict linear order $P_i$, for every $i \in [n]$, exactly one of $A \succ_{P_i} A^{'}$ and $A^{'} \succ_{P_i} A$ holds. Hence, the two majority counts sum up to $n$. Since $n$ is odd, either $A \succ_M^D A^{'}$ or $A^{'} \succ_M^D A$ holds. It follows that the pairwise majority relation $\succ_M^D$ is complete. By Lemma B.1 it is transitive. The relation $\succ_M^D$ fulfills the condition *IIA* by construction. The condition *Weak Pareto* is satisfied because if every metric prefers model $A$ to model $A^{'}$, then the pairwise majority count for $A$ over $A^{'}$ is exactly $n$, and, thus, $A \succ_M^D A^{'}$. For *Non-Dictatorship* consider the following. Fix a metric $i$. Consider three models $A, A^{'}, A^{''} \in \mathcal{A}$. Assume that metric $i$ ranks $A \succ_{P_{D,i}} A^{'} \succ_{P_{D,i}} A^{''}$ and every other metric $j \neq i$ ranks $A^{'} \succ_{P_{D,j}} A \succ_{P_{D,j}} A^{''}$. Since $A^{''}$ is ranked consistently at the bottom by every metric, $\{A^{''}\}$ is a separation of $\{A, A^{'}, A^{''}\}$ with respect to $P_{D,l}$ for all $l \in [n]$. Consider all other models $Z \in \mathcal{A} \backslash \{A, A^{'}, A^{''}\}$. Fix $Z = \{Z_1, \ldots Z_q\}$ and assume $Z_1 \succ_{P_{D,l}} \cdots \succ_{P_{D,l}} Z_q$ for all $l \in [n]$, with all $Z$ ranked above $A, A^{'}, A^{''}$ for every $l$.

Thus, the resulting profile is group-separable by definition. The pairwise majority over all metrics ranks $A^{'}$ over $A$. Therefore, metric $i$ is not a dictator. Since $i$ was arbitrarily chosen, no dictator exists. □

## C Distance-Restricted Preferences

Fix a dataset $D$, a set of models $\mathcal{A}$, and a multivariate performance measure $\Phi = (\phi_1, \ldots, \phi_n)$. Assume $n \geq 3$. After applying a fixed deterministic tie-breaking rule, each metric $i$ induces a strict linear order $P_i$ over the finite set of models $\mathcal{A}$. Let $P_i, P_j$ be such strict linear orders on $\mathcal{A}$ with $P_i := P_{D,i}$ for all $i$. Let $d_S(P_i, P_j)$ be the number of pairwise comparisons between models in $\mathcal{A}$ on which $P_i$ and $P_j$ disagree. Assume the induced profile $R_D$ is 1-distance-restricted so that $d_S(P_i, P_j) \leq 1$ for all $i, j$.

**Lemma C.1.** *If $d_S(P_i, P_j) \leq 1$, then either $P_i = P_j$ or $P_j$ is obtained from $P_i$ by swapping exactly two adjacent models in $P_i$.*

*Proof.* Assume $d_S(P_i, P_j) \leq 1$ and $P_i \neq P_j$. Then there exists a pair of models $A, A^{'} \in \mathcal{A}$ such that $P_i$ ranks $A$ above $A^{'}$ and $P_j$ ranks $A^{'}$ above $A$. If models $A$ and $A^{'}$ are not adjacent in the order $P_i$, then there must exist at least one model $A^{''}$ that lies between $A$ and $A^{'}$ in $P_i$. For $P_j$ to rank $A^{'}$ above $A$, $A^{'}$ must move past $A^{''}$ or $A$ must move past $A^{''}$. This, however, changes at least one additional pairwise comparison relative to $P_i$. We get two pairs on which $P_i$ and $P_j$ disagree which contradicts $d_S(P_i, P_j) \leq 1$. Therefore, $A$ and $A^{'}$ are adjacent in $P_i$. Since there is only one disagreement, $P_j$ is obtained from $P_i$ by swapping exactly that pair of models in $P_i$. □

**Lemma C.2.** *Fix an order $P := P_1$ such that $d_S(P_i, P_1) \leq 1$ for all $i$. There exists a pair $A, A^{'} \in \mathcal{A}$ such that for all $i$ either $P_i = P_1$ or $P_i$ is obtained from $P_1$ by swapping the pair $A, A^{'}$.*

*Proof.* By $d_S(P_i, P_1) \leq 1$ and Lemma C.1, we must have either $P_i = P_1$ or $P_i$ is obtained from $P_1$ by swapping exactly one adjacent pair of models in $P_1$. Assume for contradiction that there exist metrics $i \neq j$ such that $P_i$ is obtained from $P_1$ by swapping one adjacent pair $A, A^{'}$, and that $P_j$ is obtained from $P_1$ by swapping a different pair of models $Q, Q^{'}$. Then, $P_i$ and $P_j$ disagree on the comparison between $A$ and $A^{'}$ and, additionally, on the comparison between $Q$ and $Q^{'}$. Thus, there are two pairs of models on which $P_i$ and $P_j$ disagree, which contradicts the assumption $d_S(P_i, P_j) \leq 1$ for all $i, j$. Therefore, all orders other than $P_1$ must swap the same adjacent pair of models $A, A^{'}$. □

**Lemma C.3.** *The pairwise majority relation $\succeq_M^D$ is transitive.*

*Proof.* By Lemma C.2, for all $i$, $P_i$ coincides with $P_1$ on every pairwise comparison except possibly the single pair of models $A, A^{'}$. It follows that for any pair of models $Z, Z^{'} \in \mathcal{A}$ different from $A, A^{'}$, every metric agrees on the ranking of $Z$ compared to $Z^{'}$. Thus, majority agrees on the ranking of $Z$ compared to $Z^{'}$ as well and ranks both models exactly the same way as in the order $P_1$. Therefore, the only pair on which the majority outcome might differ from $P_1$ is $A, A^{'}$. The majority can either rank $A$ above $A^{'}$ or $A^{'}$ above $A$ or tie them. In all cases, for every third model $Z$, the comparisons between $Z$ and $A$ and between $Z$ and $A^{'}$ are exactly as in $P_1$. Since at most one pairwise comparison can differ from $P_1$, a cycle involving three distinct models cannot exist. Hence, the pairwise majority relation $\succeq_M^D$ is transitive. □

**Theorem C.4.** *Fix a dataset $D \in \mathcal{D}$ and a finite set of algorithms $\mathcal{A}$. Suppose $n \geq 3$. Assume the induced profile $R_D$ is distance-restricted to degree 1. Then, the pairwise majority relation $\succeq_M^D$ is a transitive and complete relation on $\mathcal{A}$. It satisfies Non-Dictatorship, Weak Pareto and IIA.*

*Proof.* The relation $\succeq_M^D$ is complete by definition. By Lemma C.3, $\succeq_M^D$ is transitive. The relation $\succeq_M^D$ fulfills *IIA* by construction. *Weak Pareto* holds because if all metrics strictly prefer model $A$ to model $A^{'}$, then the pairwise majority count for $A$ over $A^{'}$ is exactly $n$ and $A \succ_M^D A^{'}$. For *Non-Dictatorship* consider the following. Fix a metric $i$. Consider the fixed order $P_1$. Let $A, A^{'}$ be the

pair of models from Lemma C.2 such that the only possible deviation from $P_1$ is swapping $A$ and $A^{'}$. Consider a preference profile where metric $i$ ranks $A$ over $A^{'}$ and every other metric $j \neq i$ ranks $A^{'}$ over $A$. Since the orders differ on at most one swapped pair, the profile is distance-restricted to degree 1. The pairwise majority over all metrics does not rank $A$ over $A^{'}$: metric $i$ ranks $A$ over $A^{'}$ but the collective preference over all metrics is $A^{'}$ over $A$. Hence, metric $i$ is not a dictator. Since $i$ was arbitrarily chosen, there is no dictator. □

## D Experiments on Restricted Preference Domains: HELM MMLU

**Model sets.** The curated model sets $\mathcal{A}_1, \ldots, \mathcal{A}_{23}$ used in Section 6 are as follows. Several of these sets are intentionally nested. This lets us complement the random model sets with controlled perturbations of the alternative set, so that the experiments can probe how the observed domain structure changes by adding a small number of models.

Model set $\mathcal{A}_1 = \{$GPT-4-0613, GPT-4-1106-Preview, Claude-3-Opus-20240229, Claude-3-Sonnet-20240229, Qwen1.5-72b, Llama-2-70b, Claude-3-Haiku-20240307, Yi-34b, Yi-6b$\}$.

Model set $\mathcal{A}_2 = \{$GPT-4-0613, GPT-4-1106-Preview, Claude-3-Opus-20240229, Claude-3-Sonnet-20240229, Qwen1.5-72b, Llama-2-70b, Mistral-7b-v0.1$\}$.

Model set $\mathcal{A}_3 = \{$GPT-4-0613, Qwen1.5-72b, Llama-2-70b, Mistral-7b-v0.1$\}$.

Model set $\mathcal{A}_4 = \{$GPT-4-0613, GPT-4-1106-Preview, Claude-3-Opus-20240229, Claude-3-Sonnet-20240229, Qwen1.5-72b$\}$.

Model set $\mathcal{A}_5 = \{$Claude-3-Opus-20240229, GPT-4-0613, GPT-4-1106-Preview, Google-Text-Unicorn@001$\}$.

Model set $\mathcal{A}_6 = \{$Claude-3-Opus-20240229, GPT-4-0613, GPT-4-1106-Preview, Google-Text-Unicorn@001, Qwen1.5-72b$\}$.

Model set $\mathcal{A}_7 = \{$Claude-3-Opus-20240229, GPT-4-0613, GPT-4-1106-Preview, Google-Text-Unicorn@001, Qwen1.5-72b, Yi-34b$\}$.

Model set $\mathcal{A}_8 = \{$Claude-3-Opus-20240229, GPT-4-0613, GPT-4-1106-Preview, Google-Text-Unicorn@001, Qwen1.5-72b, Yi-34b, Claude-3-Sonnet-20240229$\}$.

Model set $\mathcal{A}_9 = \{$Claude-3-Opus-20240229, Mistral-7b-v0.1, GPT-4-0613, Llama-2-13b$\}$.

Model set $\mathcal{A}_{10} = \{$Claude-3-Opus-20240229, Mistral-7b-v0.1, GPT-4-0613, Llama-2-13b, GPT-4-1106-Preview$\}$.

Model set $\mathcal{A}_{11} = \{$Claude-3-Sonnet-20240229, Mistral-7b-v0.1, GPT-4-0613, Llama-2-70b, GPT-4-1106-Preview$\}$.

Model set $\mathcal{A}_{12} = \{$GPT-4-0613, GPT-4-1106-Preview, Claude-3-Opus-20240229, Qwen1.5-72b$\}$.

Model set $\mathcal{A}_{13} = \{$GPT-4-0613, Qwen1.5-72b, Yi-6b$\}$.

Model set $\mathcal{A}_{14} = \{$Qwen1.5-72b, Llama-2-70b, Mistral-7b-v0.1, Yi-34b, Yi-6b$\}$.

Model set $\mathcal{A}_{15} = \{$GPT-4-0613, GPT-4-1106-Preview, Claude-3-Opus-20240229, Claude-3-Sonnet-20240229, Claude-3-Haiku-20240307$\}$.

Model set $\mathcal{A}_{16} = \{$GPT-4-0613, GPT-4-1106-Preview, Claude-3-Opus-20240229, Claude-3-Sonnet-20240229, Qwen1.5-72b, Llama-2-70b$\}$.

Model set $\mathcal{A}_{17} = \{$Claude-3-Opus-20240229, Llama-2-70b, Mistral-7b-v0.1, Yi-34b, Yi-6b$\}$.

Model set $\mathcal{A}_{18} = \{$Claude-3-Haiku-20240307, Mistral-7b-v0.1, Llama-2-13b, Yi-6b$\}$.

Model set $\mathcal{A}_{19} = \{$Claude-3-Haiku-20240307, Claude-3-Sonnet-20240229, Claude-3-Opus-20240229$\}$.

Model set $\mathcal{A}_{20} = \{$Qwen1.5-14b, Qwen1.5-72b, Qwen1.5-7b$\}$.

Model set $\mathcal{A}_{21} = \{$GPT-4-0613, GPT-4-1106-Preview, GPT-3.5-Turbo-0613$\}$.

Table 2: Single-peakedness in HELM MMLU and matched uniformly random profiles. Entries are percentages of profiles on individual subjects. HELM MMLU uses alphabetical tie-breaking.

| Experimental setting | HELM MMLU | Uniformly random |
|---|---|---|
| 3 models, 3 metrics | 92.566 | 77.778 |
| 4 models, 3 metrics | 78.263 | 36.111 |
| 5 models, 3 metrics | 58.821 | 11.028 |
| 6 models, 3 metrics | 39.386 | 2.369 |
| 7 models, 3 metrics | 33.289 | 0.378 |
| 9 models, 3 metrics | 22.632 | 0.0038 |
| All model sets, 3 metrics | 77.387 | 46.133 |
| All model sets, 4 metrics | 66.271 | 27.840 |
| All model sets, 5 metrics | 53.629 | 17.607 |
| All model sets, 6 metrics | 39.400 | 11.425 |

Model set $\mathcal{A}_{22} = \{$Llama-2-13b, Llama-2-70b, Llama-2-7b$\}$.

Model set $\mathcal{A}_{23} = \{$Mistral-7b-v0.1, Llama-2-13b, Yi-34b, Llama-2-70b, Qwen1.5-72b$\}$.

The remaining 60 model sets are random sets generated in the code after setting `set.seed(13)`. They are denoted by their code identifiers `R_3_01`,...,`R_3_30`, `R_4_01`,...,`R_4_20` and `R_5_01`,...,`R_5_10`, corresponding to 30 random sets of size 3, 20 random sets of size 4 and 10 random sets of size 5. Each random set is sampled without replacement from the 22 models with complete MMLU coverage after excluding *google/gemini-pro*.

**Metric sets.** The main HELM MMLU analysis uses six metrics: four accuracy metrics $\{$*exact_match*, *prefix_exact_match*, *quasi_exact_match*, *quasi_prefix_exact_match*$\}$ and two efficiency metrics $\{$*inference_runtime*, *num_bytes*$\}$. Since lower values are preferable for the efficiency metrics, their signs are flipped before rankings are constructed. The metric sets are all subsets of these six metrics of size 3, 4, 5 and 6. Thus, we have $\binom{6}{3} = 20$ metric triples, $\binom{6}{4} = 15$ metric quadruples, $\binom{6}{5} = 6$ metric subsets of size 5 and one subset containing all six metrics, for 42 metric sets in total. Of these, $\binom{4}{3} + \binom{4}{4} = 5$ contain only accuracy metrics and the remaining 37 include at least one efficiency metric. Crossing these metric sets with the 83 model sets gives 3486 combinations and 198,702 profiles on individual subjects for each tie-breaking rule.

**Matched uniformly random profiles.** For each combination of model set size and metric set size, we consider profiles in which each metric independently induces a uniformly random strict ranking over the models. We compute the probabilities exactly where possible and use 1,000,000 uniformly random profiles for each remaining combination of sizes. The aggregate baseline uses the same distribution of model set and metric set sizes as the HELM MMLU analysis. The comparisons in Table 2 concern profiles on individual subjects and do not assume independence across MMLU subjects.

With three models and three metrics, a uniformly random profile is single-peaked with probability $168/216 = 77.778\%$. The corresponding HELM MMLU fraction is $36{,}934/39{,}900 = 92.566\%$. This comparison differs from the criterion that a restriction holds on all 57 subjects: among the 700 configurations with three models and three metrics, single-peakedness holds on all subjects in $530/700 = 75.714\%$.

Under alphabetical tie-breaking, across metric sets of sizes 3–6, the HELM MMLU fractions of profiles satisfying single-peakedness, group separability and distance-restrictedness to degree 1 are $69.119\%$, $76.920\%$ and $28.084\%$, respectively. The corresponding matched uniformly random fractions are $34.698\%$, $36.641\%$ and $5.379\%$.

Using the criterion that single-peakedness holds on all 57 subjects, the fractions of configurations are $52.651\%$, $35.904\%$, $20.281\%$ and $4.819\%$ for 3, 4, 5 and 6 metrics, respectively. Across all 42 metric subsets, single-peakedness holds on all 57 subjects in $1426/3486 = 40.906\%$ of configurations.

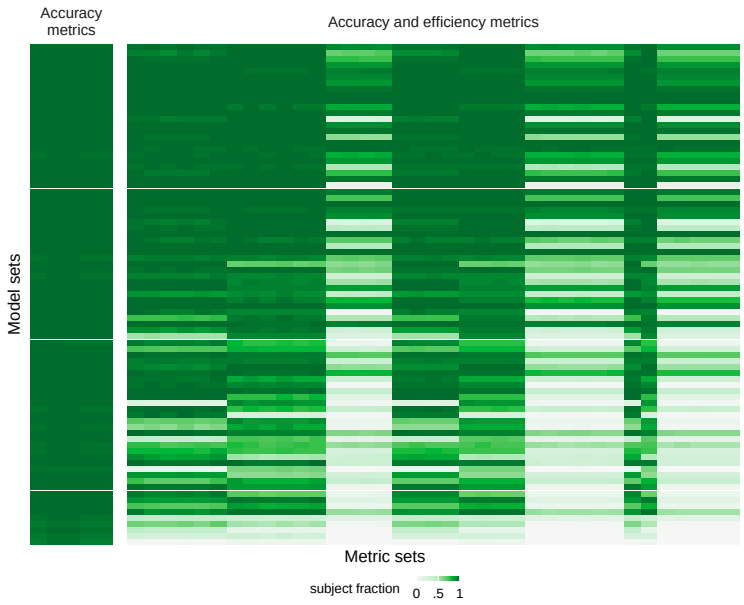



Figure 2: Group-separable structure in HELM MMLU. Each tile is one configuration of model and metric set under the alphabetical tie-breaker; the color gives the fraction of MMLU subjects whose induced preference profile is group-separable. Rows and columns use the same ordering as the single-peakedness heatmap in Figure 1.

The prevalence decreases as the number of models or metrics increases, and some of the largest combinations have no observed single-peaked profiles. Every subset of size 5 or 6 contains an efficiency metric, so the number of metrics and the inclusion of efficiency metrics cannot be treated as independent in these aggregate results.

**Results for group separability and distance-restrictedness.** Figure 2 is a heatmap over combinations of model set and metric set for group separability. Figure 3 illustrates the results for distance-restrictedness to degree 1. Group separability holds on all 57 MMLU subjects in 1208 out of 3486 configurations, while distance-restrictedness to degree 1 holds on all 57 subjects in 461 configurations. The low prevalence of distance-restrictedness in evaluation settings with accuracy and efficiency metrics is expected. This property requires metric rankings to be nearly identical. A highly accurate model can be slower or produce longer outputs, while a faster or more concise model can be efficient but less accurate. Such trade-offs change many pairwise comparisons at once, so the resulting rankings do not satisfy distance-restrictedness to degree 1. Thus, the low prevalence of distance-restrictedness in such configurations is consistent with trade-offs between these evaluation dimensions. Furthermore, note that group separability itself is a structural property of a preference profile and can be tested for metric sets of any size. However, the guarantees for properties of pairwise majority in Theorem B.2 assume an odd number of metrics and group separability regressions restrict attention to the 1660 configurations with three metrics (see Tables 3 and 4).

**Results under different tie-breaking rules.** Under reverse alphabetical tie-breaking, single-peakedness holds on all 57 subjects in 1342 configurations, group separability in 1205 and distance-restrictedness to degree 1 in 468, compared to 1426, 1208 and 461 under alphabetical tie-breaking. The respective numbers of configurations satisfying the property under both rules are 1208, 967 and 392. Thus, the aggregate prevalence is similar across the deterministic rules, although some individual configurations change status.

For random tie-breaking, we draw independent random keys for each model, metric and subject within each configuration and use them only to resolve equal scores. We repeat the analysis ten times across all 83 model sets and all 42 metric subsets. The mean fractions of profiles satisfying single-peakedness, group separability and distance-restrictedness to degree 1 are 70.011%, 74.925% and

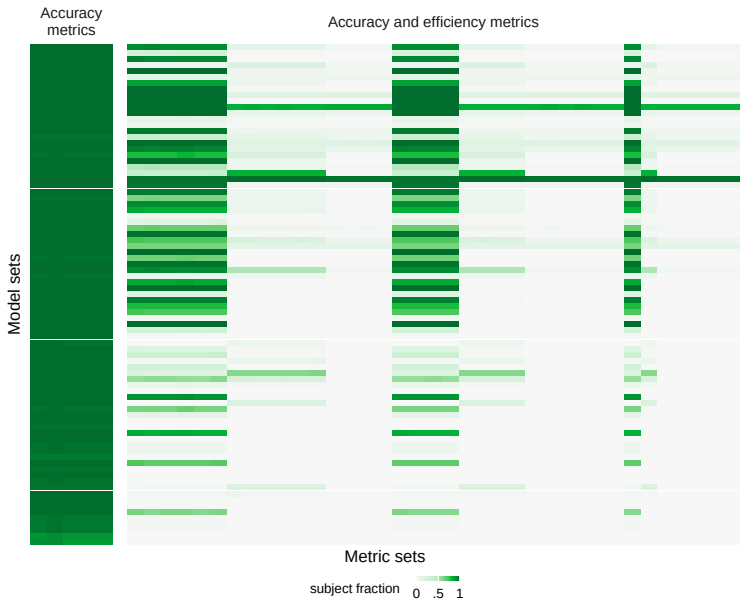



Figure 3: Distance-restrictedness to degree 1 in HELM MMLU. Each tile is one configuration of model and metric set under the alphabetical tie-breaker; the color gives the fraction of MMLU subjects whose induced preference profile is distance-restricted to degree 1. Rows and columns use the same ordering as the single-peakedness heatmap in Figure 1.

$27.952\%$, respectively. The corresponding mean fractions of configurations satisfying the properties on all 57 subjects are $30.341\%$, $24.656\%$ and $10.433\%$. The all-subject criterion is therefore more sensitive to tie-breaking than the aggregate fraction of profiles satisfying each restriction.

**Regressions.** As discussed in Section 6, we fit binomial mixed-effects logistic regressions to study which features of a configuration are associated with the fraction of MMLU subjects satisfying each domain restriction (see Table 3). These regressions use alphabetical tie-breaking and metric subsets of sizes 3 and 4, giving 2905 configurations. Within this analysis, the group-separability regressions use only the 1660 configurations with three metrics. The predictors are mean Kendall $\tau$ between the metric rankings, a dummy for whether the metric set includes an efficiency metric, the number of models in the model set and the number of metrics in the metric subset. For group separability, the number of metrics is therefore constant and is omitted as a predictor. As a sensitivity analysis, we also fit quasi-binomial logistic regressions. For each combination of model set and metric subset under alphabetical tie-breaking, the response is a count: how many subjects satisfy the domain restriction out of the MMLU subjects with complete data. Hence, a binomial model is a natural regression choice. However, the MMLU subjects cannot necessarily be treated as independent and identical Bernoulli trials. Some subjects may be more similar to each other than others and some configurations of model and metric sets may have heterogeneity that is not fully captured by the predictors. We therefore fit quasi-binomial regressions to allow for overdispersion. We find that the coefficient signs are unchanged (see Table 4), implying that our conclusions do not depend on the binomial model's variance assumption.

**Aggregation across datasets.** For aggregation across datasets, we report the commonality sharing aggregate only for a small subset of configurations for which the majority rankings are coherent on all 57 MMLU subjects. Figure 4 is therefore an illustration of the final aggregation step once the relevant structural condition has been verified. The commonality sharing ranking can be understood as a representative ranking in the whole benchmark suite. In statistical terms, it is a generalization of the concept of a median to the case of datasets, where each data point is a ranking. The deepest relation is computed by calculating the Tukey depth of all rankings using the R package ddandrda (Blocher et al., 2023).

Table 3: Generalized linear mixed models with binomial response and logit link fitted to the HELM MMLU model and metric set combinations. The regressions use alphabetical tie-breaking and metric subsets of sizes 3 and 4; for group separability, only subsets of size 3 are included. The table reports how metric agreement measured by Kendall $\tau$, the inclusion of efficiency metrics, model set size and metric set size are associated with each domain restriction. Coefficients are on the log-odds scale, with standard errors in parentheses. All reported fixed-effect coefficients are significant at $p < 0.001$.

| | Single-peaked | Group-separable | Distance-restricted |
|---|---|---|---|
| Mean Kendall $\tau$ | 6.649 (0.112) | 0.795 (0.112) | 9.556 (0.104) |
| Efficiency metric | −5.390 (0.671) | −5.413 (0.676) | −4.110 (0.183) |
| Number of models | −2.353 (0.182) | −1.751 (0.162) | −2.013 (0.110) |
| Number of metrics | −1.507 (0.303) | – | −1.451 (0.087) |
| Intercept | 20.176 (1.411) | 14.962 (0.911) | 10.956 (0.571) |
| Configurations | 2905 | 1660 | 2905 |
| Model set groups | 83 | 83 | 83 |
| Metric set groups | 35 | 20 | 35 |
| SD, model set intercept | 1.843 | 1.629 | 1.092 |
| SD, metric set intercept | 0.845 | 1.153 | 0.233 |
| AIC | 15139.2 | 8431.9 | 9336.9 |

Table 4: Quasi-binomial generalized linear models with logit link fitted to the HELM MMLU model and metric set combinations. The regressions use alphabetical tie-breaking and metric subsets of sizes 3 and 4; for group separability, only subsets of size 3 are included. Coefficients are on the log-odds scale, with standard errors in parentheses. All reported coefficients are significant at $p < 0.001$.

| | Single-peaked | Group-separable | Distance-restricted |
|---|---|---|---|
| Mean Kendall $\tau$ | 5.870 (0.144) | 2.304 (0.139) | 9.352 (0.487) |
| Efficiency metric | −8.010 (1.899) | −4.479 (0.620) | −7.832 (1.215) |
| Number of models | −1.833 (0.042) | −1.194 (0.043) | −2.250 (0.162) |
| Number of metrics | −1.402 (0.064) | – | −1.279 (0.193) |
| Intercept | 19.726 (1.936) | 10.421 (0.677) | 15.098 (1.745) |
| Configurations | 2905 | 1660 | 2905 |
| Dispersion | 14.055 | 14.523 | 74.670 |

**Licenses.** We only use existing benchmark outputs and software. HELM is credited through Liang et al. (2023) and used through its public framework and results; the HELM code is released under the Apache License 2.0. MMLU is credited through Hendrycks et al. (2021) and its public repository is released under the MIT License. For the additional PMLB and OpenML checks, we use the final results from Jansen et al. (2024). The R software is licensed under GPL-2 or GPL-3; the main R packages used in our scripts are `jsonlite`, `dplyr`, `purrr` and `tibble` under MIT, `digest`, `lme4`, `future.apply` and `relations` under GPL-compatible licenses, `future` and `ddandrda` under LGPL-compatible licenses and `Rgraphviz` under EPL.

## E Experiments on Restricted Preference Domains: PMLB and OpenML

For our verifications of restricted preference domains on PMLB and OpenML benchmarks we use the final results from Jansen et al. (2024) and then proceed exactly as in the tests on HELM MMLU.

**PMLB.** The benchmark suite includes six classifiers (cre, J48, glmnet, knn, ranger, svmRadial) evaluated on three metrics (predictive accuracy, robustness to feature perturbation and robustness to

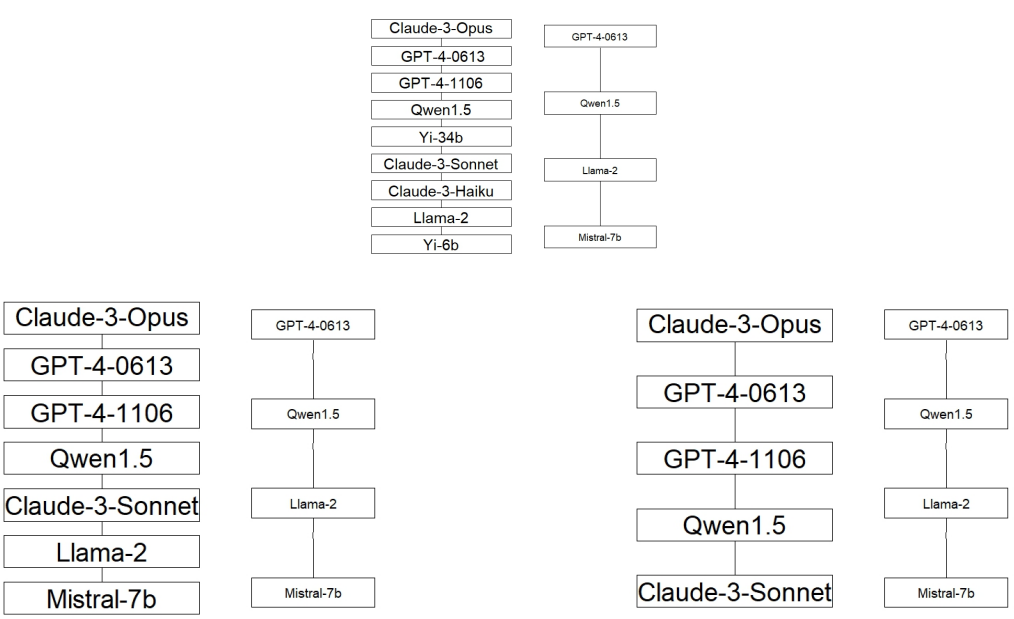



Figure 4: Aggregated rankings across all 57 datasets of HELM MMLU according to the commonality sharing rule (cf., Section 4.3). Top: single-peaked preferences for $\mathcal{A}_1$ with the accuracy metrics {*exact_match*, *prefix_exact_match*, *quasi_exact_match*} and for $\mathcal{A}_3$ with the mixed metrics {*exact_match*, *quasi_exact_match*, *inference_runtime*}. Bottom left: group separability for $\mathcal{A}_2$ with the same accuracy metrics and for $\mathcal{A}_3$ with the same mixed metric triple. Bottom right: distance-restrictedness to degree 1 for $\mathcal{A}_4$ and $\mathcal{A}_3$, both with the accuracy metric triple.

label perturbation) over 63 datasets. We find that seven out of 63 datasets are single-peaked. For group separability, we observe that 11 out of 63 datasets are group-separable. The results remain very similar after switching to the reverse alphabetical tie-breaking. Since the metrics measure related aspects of one latent concept (robust accuracy), it is plausible that we observe some single-peaked and group-separable datasets; however, the metrics are still different enough so that there is non-trivial disagreement between rankings which leads to multiple violations of our structural assumptions.

**OpenML.** The suite contains seven classifiers (classif.glmnet, classif.kknn, classif.multinom, classif.ranger, classif.rpart, classif.svm, classif.xgboost) evaluated on 80 datasets and three metrics (predictive accuracy, computation time on training data and computation time on test data). We find that none of 80 datasets is single-peaked, and four out of 80 datasets are group-separable. The results are unchanged under the reverse alphabetical tie-breaking. The combination of accuracy with training and test runtime induces a strong trade-off, meaning that different metrics disagree on many pairwise comparisons, which implies less structured preference profiles.

## F Experiments: Transitivity and Stability

**Condorcet Cycles.** We preprocess the HELM MMLU data exactly as in Section 6.1. We align metrics so that larger values mean better performance by flipping the sign for "lower-is-better" metrics such as *Time* and *Length*. Then, for each subject, we search over all triples of metrics in the fixed set {*exact_match*, *inference_runtime*, *num_bytes*, *logprob*, *perplexity*, *num_output_tokens*}. Within a metric triple, we restrict to models with complete observations and define pairwise majority comparisons. For each pair of models, each metric "votes" only when the two aligned scores differ by more than a small tolerance; ties and near-tie situations are treated as abstentions. Then, we search for cycles i.e., triples of models $A, A', A''$ such that $A$ is preferred to $A'$, $A'$ is preferred to $A''$, and $A''$ is preferred to $A$. A subject is counted as cyclic if at least one metric triple yields such

a cycle. For each subject, we keep the most robust witnessed cycle and assign it a buffer. For each of the three pairwise majority wins in the cycle, we take the smallest score gap among the metrics supporting that win so that the cycle's buffer is the minimum of these three values. Larger buffers therefore correspond to cycles that persist under small perturbations of the metric values. Under this procedure, we find at least one Condorcet cycle in 51 of 57 MMLU subjects. This suggests that cyclic preferences are widespread once accuracy and efficiency metrics are combined.

The full model names for the cycle presented in Table 1 are GPT-3.5-Turbo-0613, GPT-4-1106-Preview and Qwen1.5-14b. Note that in the experiments in Section 6.1 (also covered in Appendix D) the model name Qwen1.5 is used as the abbreviation for the model Qwen1.5-72b.

**Instability to changes in the model set.** For each MMLU subject, we fix the metric set {*exact_match*, *inference_runtime*, *num_bytes*} (flipping the sign for "lower-is-better" metrics as above). For each subject, we restrict to models with complete observations and focus only on the top 15 models by the metric *exact_match*. For each subject, we compute the overall ranking by averaging the metrics' ranks, then add one additional model and recompute the ranking. We record whether any pairwise order among the original 15 models flips under this change. We find such flips in 44 out of 57 MMLU subjects.

The full model names for all models mentioned in Table 1 are as follows: GPT-4-0613, GPT-4-1106-Preview, Claude-3-Opus-20240229, Qwen1.5-72b, GPT-3.5-Turbo-0613, Google-Text-Bison@001 and Llama-2-13b. Note that in the experiments in Section 6.1 (also covered in Appendix D) the model name Llama-2 is used as the abbreviation for the model Llama-2-70b.