🏆 COPA: Comparing the Incomparable to Explore the Pareto Front

Anonymous Authors[1]

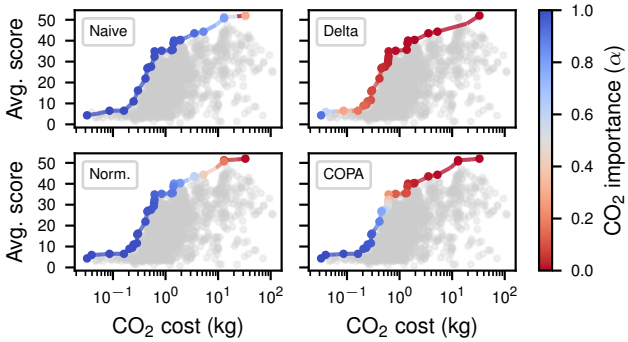
Abstract

In machine learning (ML), it is common to account for multiple objectives when, e.g., selecting a model to deploy. However, it is often unclear how one should *compare*, *aggregate* and, ultimately, *trade-off* these objectives, as they might be measured in different units or scales. For example, when deploying large language models (LLMs), we might not *only* care about their performance, but also their CO_2 consumption. In this work, we investigate *how* objectives can be sensibly compared and aggregated to navigate their Pareto front. To do so, we propose to make incomparable objectives comparable via their CDFs, approximated by their relative rankings. This allows us to aggregate them while matching user-specific preferences, allowing practitioners to meaningfully navigate and search for models in the Pareto front. We demonstrate the potential impact of our methodology in diverse areas such as LLM selection, domain generalization, and AutoML benchmarking, where classical ways to aggregate and normalize objectives fail.

1 Introduction

When evaluating machine learning (ML) models, one often has to account for many *objectives* at once. For example, in model selection for classification, we typically look for a compromise among objectives such as accuracy, sensitivity, or specificity (Japkowicz & Shah, 2011) and, as ML becomes widely adopted, other objectives beyond performance are considered as well. For example, the deployment of large language models (LLMs) at scale has opened new challenges regarding their robustness (Yuan et al., 2023), fairness (Huang et al., 2023), and CO_2 emissions (Coignion et al., 2024; Luccioni et al., 2023).

Consider the *Open LLM Leaderboard* (Fourrier et al., 2024) as an example, where LLMs are compared in terms of their performance across 6 benchmarks and inference CO_2 cost. How could we select the "best" LLM among the 2148 submitted models, if there are 7 objectives to consider? Among all LLMs, 487 present non-trivial trade-offs, i.e., for every pair of them, one is better in an objective but worse in another, see Fig. 1. Now, if a practitioner were to select an LLM that equally balances performance and CO_2 cost: How should they proceed? Should they manually inspect all of them? This simple scenario highlights two important limitations in multi-objective ML evaluation:



LLM base model	Method	Perf. score avg.	Perf. score top-%	CO_2 cost kg	CO_2 cost top-%	Rank $\alpha = 1/2$
`Qwen2.5-72B`	Naive	52.02	0.00	33.01	98.65	2090
`GPT-2`	Delta	5.98	90.87	0.04	0.05	1762
`Qwen2-72B`	Norm.	50.71	0.14	12.98	95.20	1944
`Qwen2.5-7B`	COPA	29.34	17.33	0.60	18.21	1

Figure 1: **The proposed COPA meaningfully navigates the performance-emissions trade-off of the Open LLM Leaderboard** (Fourrier et al., 2024), uniformly mapping the importance of CO_2 cost, α, to the Pareto front. In contrast, existing approaches are biased toward one of the objectives. This is reflected in the retrieved LLMs where, for $\alpha = 1/2$, COPA finds a top-18 % model for both objectives, while all other approaches select either a high-performing but CO_2-demanding model, or vice versa.

L1. Objectives with different semantics and domains, such as average performance score and CO_2 cost in Fig. 1, are not directly *comparable*, and thus cannot be properly aggregated nor traded-off. In physics, this would be akin to comparing metres and grams.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

L2. When we deal with many objectives (7 in our example), it is challenging for humans to translate their preferences into a concrete decision, as the number of plausible trade-offs quickly becomes overwhelming.

These limitations reinforce the idea that we need automatic tools to navigate the Pareto front (i.e., the set of optimal trade-offs) in high dimensions, tuning their parameters according to the user preferences. However, as we show in Fig. 1, directly aggregating the objectives (Naive), or normalizing them first using existing approaches (Norm. and Delta, see §2), fails to make objectives comparable and to uniformly explore the Pareto Front. In other words, they map most of the values of the objective importance weights α to a small region of the front. To overcome these issues, prior works had to devise heuristic approaches tailored to their specific problem instance (Nazabal et al., 2020; Caruana & Niculescu-Mizil, 2004). To this day, *there is a lack of a general and systematic approach to compare, aggregate and, ultimately, trade-off ML objectives.*

Contributions. In this work, we first motivate and establish the incomparability problem in multi-objective selection, highlighting why previous approaches fail (§2). Next, we introduce **COPA** 🏆,[1] a novel approach to multi-objective ML evaluation that *allows practitioners to navigate the Pareto front in a meaningful way*, so that they can compare and select models that reflect their preferences (§3). COPA accomplishes this goal with two components: **i)** a normalization function that *universally* makes all objectives comparable via the probability integral transform, which we approximate using relative rankings; and **ii)** a criterion function with two easily interpretable parameters controlling both the aggregation and the importance of each objective.

Finally, we discuss related works (§4), and demonstrate the potential impact of COPA (§5) in diverse and timely application domains such as domain generalization, multi-task learning, fair ML, AutoML benchmarking, and LLM selection as illustrated in Fig. 1. This figure demonstrates that COPA enables meaningful exploring the Pareto front via the importance of the CO_2 cost, controlled by α and uniformly distributed along the front. For instance, a practitioner equally interested in the performance and CO_2 emissions of the LLM, could use COPA with $\alpha = 1/2$ to pick the model in the middle of the Pareto front (last row in Fig. 1), which is ranked top-18 % for both objectives.

2 Problem Statement

We are given a population of models $\mathcal{H}$, typically obtained by changing hyperparameters, where each model $h \in \mathcal{H}$ is associated a vector of K metrics assessing its performance w.r.t. different evaluation objectives. In addition, we assume each objective to be a continuous random variable for which we have sampled observations in $\mathcal{H}$.

Without loss of generality, we assume that each individual objective has to be *minimized*, and we can thus frame the problem as a multi-objective optimization (MOO) problem of the following form:

$$\min_{h \in \mathcal{H}} \ \mathbf{y}(h) \coloneqq [\mathrm{y}_1(h), \mathrm{y}_2(h), \ldots, \mathrm{y}_K(h)]\,, \tag{1}$$

where $\mathbf{y}(h)$ is the objective vector of model h, and $\mathrm{y}_k(h)$ its performance on the k-th objective. For ease of notation, we omit the argument from here onwards, and write $\mathbf{y}$ and y_k directly when it is clear from the context.

How can we minimize a vector? A fundamental problem of Eq. 1 is that *minimizing the vector* $\mathbf{y}$ *is not well-defined*, as there is no canonical total order in high dimensions. Therefore, two models could yield objective vectors where one might not always be better than the other for all objectives. In the MOO literature, the set of all these optimal trade-off solutions is known as the *Pareto front* and, more formally, an objective vector $\mathbf{y}^*$ is in the Pareto front (and called *Pareto-optimal*) if there exists no other feasible vector $\mathbf{y}$ such that $\mathrm{y}_k \leq \mathrm{y}_k^*$ for all $k \in \{1, 2, \ldots, K\}$, and $\mathrm{y}_k < \mathrm{y}_k^*$ for at least one of the objectives.

While the Pareto front is theoretically appealing, in practice, the **decision maker** (DM) usually needs to navigate the Pareto front and, eventually, select one single model. That is, the DM needs to specify a total order in Eq. 1, which implies either: **i)** taking a total order directly in $\mathbb{R}^K$, e.g., the lexicographic order where $\mathbf{y} < \mathbf{y}^*$ iff $\mathrm{y}_k < \mathrm{y}_k^*$ and $\mathrm{y}_i = \mathrm{y}_i^*$ $\forall i < k$; or **ii)** defining a **criterion function** $\mathrm{C}\colon \mathbb{R}^K \to \mathbb{R}$ to rewrite Eq. 1 as a scalar-valued problem:

$$\min_{h \in \mathcal{H}} \quad \mathrm{C}(\mathbf{y}(h))\,. \tag{2}$$

One remarkable example of the latter are *global-criterion methods* (Zeleny, 1973), which map DM preferences to the problem geometry by interpreting Eq. 2 as selecting the model closest to the *ideal* one, i.e.,

$$\min_{h \in \mathcal{H}} \quad \|\mathbf{y}(h) - \mathbf{y}^{\text{ideal}}\|_*\,, \tag{3}$$

where $\|\cdot\|_*$ is typically a p-norm, and $\mathbf{y}^{\text{ideal}}$ is the ideal solution, $\mathbf{y}^{\text{ideal}} \coloneqq [\min_h \mathrm{y}_1, \min_h \mathrm{y}_2, \ldots, \min_h \mathrm{y}_K]$. However, naively solving Eq. 3 (and, more generally, Eq. 2) is well-known to be sensitive to the scaling of the objectives (Branke et al., 2008) (see **L1** in §1), and thus prevents us from properly accounting for DM preferences (**L2**).

In this work, we argue that the criterion function C should fulfil the following desiderata:

D1. It should reflect the DM preferences, translating their model expectations into an optimization problem.

D2. It should provide a simple way to tune its parameters to meaningfully explore the Pareto front.

[1] In Spanish, copa means trophy.

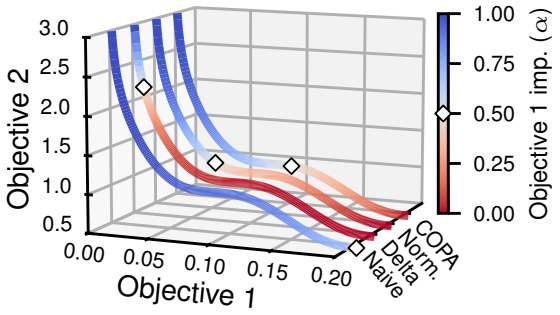


Figure 2: As we explore a synthetic Pareto front with different normalization functions to solve Eq. 5, only COPA meaningfully navigates it as we change α, and its min-max solution agrees with our prior expectations of a robust one.

***When* are objectives incomparable?** Similar to dimensional analysis in physics (Barenblatt, 1987)—which argues that we cannot combine incommensurable quantities, e.g., kilograms and meters—we argue that a second fundamental issue that we face in Eq. 2 is **semantic incomparability**, i.e., whether it is sensible to compare (and thus aggregate) the values of two different objectives.

For example, if objectives differ in their semantics they are hardly comparable in general, e.g.: despite both accuracy and ROC AUC lying in the unit interval, it does not make immediate sense to compare their values. There are, however, other aspects that are more subtle. To illustrate these, Fig. 2 presents a synthetic Pareto front from §5.1 where both objectives quantify prediction error in significantly different domains, namely, within the intervals $[0, 0.2]$ and $[0.5, 3.0]$. We navigate the Pareto front solving a weighted Tchebycheff problem (Bowman, 1976) of the form

$$\min_{h \in \mathcal{H}} \; \max\{\, \alpha|\mathrm{y}_1|,\, (1-\alpha)|\mathrm{y}_2|\,\}\;, \tag{4}$$

which is a particular case of Eq. 2 where C is a ∞-norm weighted by $0 \leq \alpha \leq 1$. Intuitively, Eq. 4 looks for robust solutions that account for the importance of solving the first objective over the second, seemingly satisfying our desiderata **D1**-**2**. However, the naive implementation using the original objectives in Eq. 4 clearly highlights how we are biasing model selection in favour of Objective 2, as it can be seen in Fig. 2.

***How* can we make objectives comparable?** As shown above, even if we use a well-designed criterion function, semantic incomparability can hinder our goal of meaningfully exploring the Pareto front. Historically, this has been addressed in the MOO literature by applying **component-wise transformations** to the objectives to normalize them (Miettinen, 1999), turning Eq. 2 into

$$\min_{h \in \mathcal{H}} \quad \mathrm{C}(\boldsymbol{\phi}(\mathbf{y})) \coloneqq \mathrm{C}\left([\phi_1(\mathrm{y}_1), \ldots, \phi_K(\mathrm{y}_K)]\right)\;. \tag{5}$$

Two classic examples of these transformations are

$$\Delta_k(\mathrm{y}_k) \coloneqq \frac{\mathrm{y}_k - \mathrm{y}_k^{\text{ideal}}}{\mathrm{y}_k^{\text{ideal}}}\;, \text{ and} \tag{6}$$

$$\mathrm{norm}_k(\mathrm{y}_k) \coloneqq \frac{\mathrm{y}_k - \mathrm{y}_k^{\text{ideal}}}{\mathrm{y}_k^{\text{nadir}} - \mathrm{y}_k^{\text{ideal}}}\;, \tag{7}$$

where $\mathrm{y}_k^{\text{nadir}} \coloneqq [\max_h \mathrm{y}_1, \max_h \mathrm{y}_2, \ldots, \max_h \mathrm{y}_K]$ is the worst plausible solution. Intuitively, Δ_k represents the difference relative to the ideal solution, and norm_k reweighs the objective to lie in the unit interval. Prior works have extensively used Δ_k, often replacing $\mathrm{y}_k^{\text{ideal}}$ with a reference vector, as computing it can be challenging (Miettinen, 1999; Maninis et al., 2019; Liu et al., 2023).

Back to our synthetic scenario, we now want to solve

$$\min_{h \in \mathcal{H}} \; \max\{\, \alpha|\phi_1(\mathrm{y}_1)|,\, (1-\alpha)|\phi_2(\mathrm{y}_2)|\,\}\;. \tag{8}$$

By testing different ϕ_k (see Eqs. 6 and 7), we can understand why classic approaches fail to make objectives comparable. More specifically, note that: **i)** using Δ_k biases the problem toward the first objective instead, since $\min_h \mathrm{y}_1 \approx 0$; and **ii)** using norm_k alleviates these problems, as the denominator is now bigger than the numerator, yet the differences between distributions (that of y_2 being heavy-tailed) still bias the optimization towards the first objective. Instead, we seek to explore the Pareto front with a more meaningful use of α, spreading it uniformly along the curve.

The *main goal* of the functions $\phi_k \colon \mathbb{R} \to \mathbb{R}$ is therefore to make the objectives semantically comparable, so that we can seamlessly aggregate them with the criterion function. To this end, we argue that the functions ϕ_k should be:

D3. Objective-agnostic, so that we can normalize any objective irrespectively of its specific nature.

D4. Order-preserving (e.g., strictly increasing), so that it preserves the Pareto-optimality of the models.

In summary, to meaningfully explore the Pareto front, it is important to design a a normalization function $\boldsymbol{\phi}$ that makes objectives semantically comparable (**D3**-**4**), and a criterion function C that translates well DM preferences into an optimization problem (**D1**-**2**).

These desiderata will blend in COPA, discussed in the next section. In the synthetic experiment above, COPA maps the value $\alpha = 1/2$, which induces a regular min-max problem in Eq. 8, to the flat region of the curve in Fig. 2, matching the intuition of what a robust solution should represent.

3 Methodology

Next, we introduce the proposed normalization and criterion functions fulfilling the desiderata **D1**-**4** described in §2. We refer to the problem resulting of solving Eq. 5 with the

proposed functions as ***cumulative-based optimization of the Pareto front*** or, in short, **COPA**.

3.1 Designing a Universal Normalization Function

We argued in §2 that the function ϕ should fulfil desiderata **D3** and **D4**, i.e., it should make any objectives semantically comparable, while preserving their Pareto-optimality. Taking advantage of our probabilistic perspective, we propose to design ϕ such that the resulting variables are all equally distributed and, w.l.o.g., uniformly distributed in the unit interval. That is, we propose to use $\mathbf{u} := [\mathrm{u}_1, \mathrm{u}_2, \dots, \mathrm{u}_K]$ instead of $\mathbf{y}$, where

$$\mathrm{u}_k := F_k(\mathrm{y}_k) \sim \mathcal{U}(0,1) \quad \forall k \in \{1, 2, \dots, K\}, \tag{9}$$

and $\phi_k = F_k$ is the marginal cumulative distribution function (CDF) of the k-th objective. Indeed, this transformation is known in statistics as the probability integral transform (Casella & Berger, 2021, Example 5.6.3), and u_k is guaranteed to follow a standard uniform if y_k is continuous.

Remarkably, Eq. 9 makes all criterion functions *marginal-distribution-free* in the sense of Kendall & Sundrum (1953), i.e., strips away all individual properties of the marginal distributions (e.g., the domain) of the individual objectives (**D3**). We note here that normalizing random variables this way is one of the fundamental building stones of copulae in statistics (Sklar, 1959; Geenens, 2024), ensuring that copula functions learn only the relationship between variables.

How can we interpret the values of u? One important advantage of using $\mathbf{u}$ in place of $\mathbf{y}$ in Eq. 5 is that it provides a common framework to think about all objectives, since all their values all are now framed as *elements within a population*. In practice, this means that the DM has a common language to express their expectations on the model. For example, a value $\mathrm{u} = 1/2$ corresponds for all objectives to the *the median value*, which divides $\mathcal{H}$ into two *halves* comprising the best and worst performing models.

However, there is one caveat we need yet to address: we have no access to the marginal CDF of each objective, but to samples of the joint distribution in $\mathcal{H}$. Next, we show how to robustly approximate u_k using relative rankings.

3.2 Rankings as Finite-Sample Approximations

As mentioned above, while we have no access to the CDFs themselves, we have samples from the joint distribution over the objectives, i.e., over, $p([\mathrm{y}_1, \mathrm{y}_2, \dots, \mathrm{y}_K])$. Namely, we can consider each model $h \in \mathcal{H}$ as a sample from the joint distribution and, by looking at each objective individually, as a sample from the marginal distributions.

Let us now focus on the k-th objective, y_k, and drop the subindex in the following paragraphs to ease notation. Say that we have $|\mathcal{H}| = N$ i.i.d. realizations of the objective, i.e., $\{\mathrm{y}_1, \mathrm{y}_2, \dots, \mathrm{y}_N\} \overset{\text{i.i.d.}}{\sim} P_k$. Then, we can approximate Eq. 9 for the i-th sample, $\mathrm{u}_i = F(\mathrm{y}_i)$, by computing its order statistic, i.e., the random variable representing its relative ranking within the population,[2] $R(i) := \sum_{j=1}^{N} [\,\mathrm{y}_j < \mathrm{y}_i\,]$, where Iverson brackets denote the indicator function, such that $\mathrm{y}_{R(1)} \leq \mathrm{y}_{R(2)} \leq \dots \leq \mathrm{y}_{R(N)}$. Specifically, since the *empirical CDF* is the fraction of samples smaller than the input, it is direct to show that

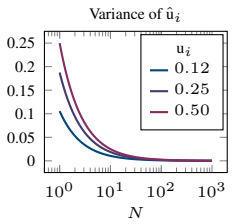
$$\hat{\mathrm{u}}_i = \hat{F}(i) := \frac{1}{N} \sum_{j=1}^{N} [\,\mathrm{y}_j < \mathrm{y}_i\,] = \frac{1}{N} R(i) \tag{10}$$

enjoys the following properties (Casella & Berger, 2021):

Proposition 3.1. *$\hat{\mathrm{u}}_i = \hat{F}(\mathrm{y}_i)$ is an unbiased estimator of the CDF at y_i, $\mathrm{u}_i = F(\mathrm{y}_i)$, with variance $\mathrm{u}_i(1-\mathrm{u}_i)/N$. The variance of $\hat{\mathrm{u}}_i$ decreases linearly with N, and has a maximum value of $0.25/N$ at the median.*

Proof. First, note that $[\,\mathrm{y}_j < \mathrm{y}_i\,] \sim \mathrm{Bern}(\mathrm{u}_i)$. Then, we have $R(i) \sim \mathrm{Bin}(N, \mathrm{u}_i)$ with mean $N\mathrm{u}_i$ and variance $N\mathrm{u}_i(1-\mathrm{u}_i)$. Hence, $\hat{\mathbf{u}}_i$ has mean $\frac{1}{N}\mathbb{E}[R(i)] = \mathrm{u}_i$, and variance $\frac{1}{N^2}\mathbb{V}[R(i)] = \mathrm{u}_i(1-\mathrm{u}_i)/N$ which, by taking derivatives w.r.t. u_i, $\partial_{\mathrm{u}_i}\mathbb{V}[\hat{\mathrm{u}}_i] = 1 - 2\mathrm{u}_i = 0 \Rightarrow \mathrm{u}_i = 1/2$, which is a maximum since $\partial^2_{\mathrm{u}_i}\mathbb{V}[1/2] < 0$. □

In other words, we can use the relative rankings of each objective to build an unbiased[3] estimator of the CDF, $\hat{\mathrm{u}}_i$, whose variance rapidly decreases as we increase the size of $\mathcal{H}$, i.e., $\mathbb{V}[\hat{\mathrm{u}}_i] \xrightarrow{N\to\infty} 0$. Indeed, the inset figure shows the variance of $\hat{\mathbf{u}}_i$ as a function of the sample size for three different values of u_i. Note that the relative ranking is strictly increasing, i.e., if $\mathrm{y}_i < \mathrm{y}_j$, then $\hat{F}(\mathrm{y}_i) < \hat{F}(\mathrm{y}_j)$ for any $\mathcal{H}$ containing both samples (**D4**). While this is an approximation of the true CDF, which would retain instead all the information about the joint distribution, it works egregiously well in our experiments (§5). Furthermore, by storing the original values, we can always invert the transform and project our rankings to the original space.



3.3 Incorporating Preferences into the Optimization

Now that we can effectively approximate our normalization function, we introduce a criterion function to translate DM preferences into an optimization problem (**D1**).

To do so, we start by looking back at global criterion methods, since plugging in our transformation $\mathbf{u} = \phi(\mathbf{y})$ simplifies the problem in Eq. 3 to $\min_h \|\mathbf{u}\|_*$ as the ideal point becomes the origin, i.e., $\mathbf{u}^{\text{ideal}} = \mathbf{0}$. Then, by using the approximation described in §3.2, the problem simply becomes

$$\min_{i \in \{1,2,\dots,N\}} \|\hat{\mathbf{u}}_i\|_* . \tag{11}$$

[2] When there is a tie, both elements get the minimum ranking.

[3] In fact, it is known to be a consistent estimator (Tucker, 1959).

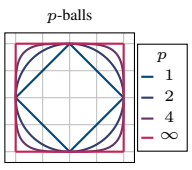
That is, we have reduced our problem to finding the model whose ranking vector is closest to the origin. Using this *marginal-free global-criterion method*, mapping the DM preferences now boils down to selecting an appropriate norm for the problem in Eq. 11. To this end, we propose to use a criterion function C a norm with parameters $p \geq 1$ and $\boldsymbol{\omega} \in \mathbb{R}_+^K$ defined as

$$\|\mathbf{u}\|_{p,\boldsymbol{\omega}} \coloneqq \left(\sum_{k=1}^{K} |\omega_k \cdot \mathrm{u}_k|^p \right)^{1/p}, \tag{12}$$

where $\sum_k \omega_k = 1$. This norm can be interpreted as a regular p-norm on a space with coordinates scaled by $\boldsymbol{\omega}$. However, note that this differs from the usual weighted p-norm, as the weights are *inside* the absolute value. We justify this choice given that the values of u_k lie in the unit interval, and the power would often make them vanish too quickly.

How can we interpret the parameters? Fortunately, the parameters of the proposed criterion function, p and $\boldsymbol{\omega}$, provide an easy and interpretable way for the DM to navigate the Pareto front **(D2)**. Regarding the interpretation of $\boldsymbol{\omega}$, as we apply them in Eq. 12 *before* taking the power, we can provide a clear interpretation of $\boldsymbol{\omega}$ in terms of ratio trade-offs. For example, if we had two objectives with $\boldsymbol{\omega} = [0.75, 0.25]$, then we see by equating the weighted objectives that minimizing the first objective to a value of u_1 is worth the same as minimizing the second objective to a value of $\mathrm{u}_2 = \omega_1/\omega_2 \mathrm{u}_1 = 3\mathrm{u}_1$, i.e., u_1 is three times more important than u_2. If we combine this with the interpretation of $\mathbf{u}$ given in §3.1 we could say, e.g., that we value being in the top-25 % of the models for the first objective the same as being in the top-75 % for the second objective.

For the interpretation of the p-norm, we can use the same intuition as in ML (Goodfellow et al., 2016): the models selected in Eq. 11 will be the first ones intersecting an ever-expanding p-ball centred at the origin, whose shape depends on p as depicted in the inset figure. Higher values of p lead to denser objective vectors, while smaller values lead instead to sparser ones. Moreover, specific values of p have clearer interpretations, e.g.: $p = 1$ is the average rank; $p = 2$ is the Euclidean distance we use in our daily life; and $p = \infty$ turns Eq. 11 into a min-max problem, typically used to formulate robust optimization problems (Verdu & Poor, 1984).



Does Eq. 12 enjoy theoretical guarantees? Finally, given the similarity with commonly-used norms, it is natural to wonder whether we can leverage existing results from the MOO literature and adapt them to the proposed norm. This is indeed the case, and we can easily guarantee, e.g., that the solutions found using Eq. 12 with $1 \leq p < \infty$ are always Pareto-optimal (Miettinen, 1999, Thm. 3.4.1). However, it might not reach all optima. Similarly, note that Eq. 12 with $p = \infty$ reduces Eq. 11 to a weighted Tchebycheff problem which reaches any Pareto-optimal solution (Miettinen, 1999, Thrm. 3.4.5), but also weakly optimal ones.

In practice, using a weighted Tchebycheff problem ($p = \infty$) can be a good practice when we have few objectives and a large budget on the weights $\boldsymbol{\omega}$ to explore. Instead, when interested in finding a particular model (i.e., solving Eq. 5 once), we suggest setting p based on the level of robustness desired (lower values of p lead to higher tolerance to bad performance on individual objectives), and $\boldsymbol{\omega}$ based on the importance of solving each objective given by the DM.

4 Related Works

Our work is nicely connected with other scientific domains, e.g., the notion of semantically incomparability is akin to that of incommensurability in dimensional analysis (Barenblatt, 1987). Similarly, using relative rankings to make better comparisons has been previously explored in microeconomics (Piggins, 2019), MOO (Kukkonen & Lampinen, 2007; Ibrahim et al., 2024), and in statistics, designing methods that avoid the normality assumption, e.g., the Friedman test (Friedman, 1937), Wilcoxon signed-rank test (Wilcoxon, 1945), or Kendall's τ coefficient (Kendall, 1938). Finally, as mentioned in §2, copulas exploit the probability integral transform to become marginal-distribution-free (Geenens, 2024), and the proposed criterion functions share similarities with weighted L_p-problems in MOO (Miettinen, 1999).

In ML, the closest work to ours is Park et al. (2024), which exploits a joint CDF, approximated through learned copulae, to recover a partial order for multi-objective Bayesian optimization. Differently from them, we employ marginal CDFs and provide a principled way to translate the DM preferences. ROC curves (Flach, 2010) provide a further connection, since their axes can be understood as the CDFs of the target classes (Hand, 2009). Many works proposed ad-hoc approaches to normalize and aggregate objectives using, e.g., normalized RMSEs (Nazabal et al., 2020)—we refer to §8.3 of Japkowicz & Shah (2011) for other references. Notoriously, some works in multitask learning (Navon et al., 2022; Liu et al., 2023) and domain generalization (Ramé et al., 2022) use rank averages to aggregate objectives, yet the standard is to use the average of Δ_k-normalized objectives (see Eq. 6). COPA can benefit these two areas, along any others accounting for several objectives such as fair ML (Martínez et al., 2020), federated learning (Kairouz et al., 2021), probabilistic ML (Javaloy et al., 2022), and multimodal learning (Baltrušaitis et al., 2018).

5 COPA in Action

In this section, we motivate the use of COPA by showing a range of practical scenarios which would benefit from

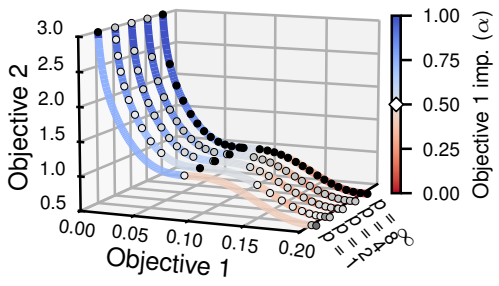


Figure 3: Using COPA with $p = 1$, the extrema are selected for most values of α. As we increase p, the distribution of solutions spreads out to reach denser solutions with different values of α. Here, circles represent selected solutions, and their darkness the amount of times they were selected.

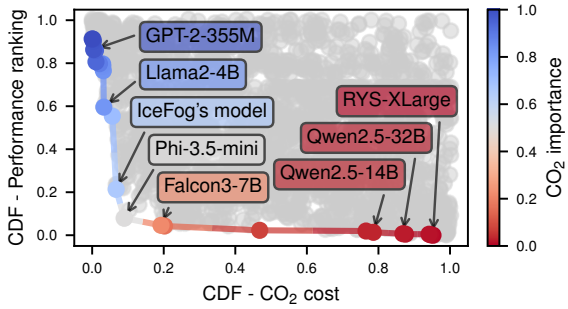


Figure 4: **With COPA, we can meaningfully navigate the Pareto-front of the Open LLM Leaderboard** (Fourrier et al., 2024). We use COPA with $p = \infty$ on all 7 objectives, and highlight some models selected as we change α.

adopting the proposed methodology. We defer additional details and results to App. A.

5.1 Synthetic Evaluation

To qualitative assess COPA, we first consider a synthetic experiment in which we parametrically simulate a Pareto front of the following form:

$$y_2 = 0.25\cos(39y_1^{0.85}) - \log(y_1) - 0.46\,, \qquad (13)$$

where $y_1 \sim \mathcal{U}(0.02, 0.2)$. The above formula results in a non-convex Pareto front with a flat area around $y_1 = 0.1$, and two objectives with significantly different distributions.

Does the parameter p match our intuitions? We corroborate the insights from §3.3 by showing in Fig. 3 the distribution of solutions found taking different values of p. First, note that since Eq. 13 is *strictly* increasing except in $[0.083, 0.091]$, we have that $u_1 \approx 1 - u_2$. As a result, we see that $p = 1$ finds only solutions on the extreme and middle points, most solutions being concentrated on the former. When we increase p, the distribution of solutions better spread along the curve and, as the p-balls become more squared, we gain finer control on the solution found by tuning α. It is however important to stress that the finer control of $p = \infty$ can be problematic at times: as we increase K, finding a proper $\boldsymbol{\omega}$ could prove challenging.

5.2 Case 1: Model Selection

First, we explore how the norm proposed in §3.3 can help us explore the Pareto front more meaningfully, i.e., how sensibly it maps the DM preferences to Eq. 5.

1. The performance-emissions trade-off. Despite LLMs recently showing outstanding performance (Naveed et al., 2023), their CO_2 emissions can be concerning and needs to be taken into account (Coignion et al., 2024). Next, we show how practitioners can leverage COPA to better navigate this crucial trade-off in the LLM space.

We gather the results of 2148 LLMs submitted to the Open LLM Leaderboard (Fourrier et al., 2024), and take as objectives their CO_2 cost and performance on 6 different datasets: IFEval (Zhou et al., 2023), BBH (Suzgun et al., 2023), MATH (Hendrycks et al., 2021), GPQA (Rein et al., 2023), MuSR (Sprague et al., 2023), and MMLU-Pro (Wang et al., 2024). Then, we use COPA with $p = \infty$ to select an LLM, changing $\boldsymbol{\omega}$ as we vary the importance given to their CO_2 emissions, denoted by α, as $\boldsymbol{\omega} := [\alpha, \frac{1-\alpha}{6}, \ldots, \frac{1-\alpha}{6}]$.

We highlight the selected LLMs in Fig. 4, which groups all benchmarks into one dimension as their ∞-norm for visualization purposes. We observe that the proposed norm enables the meaningful exploration of the Pareto front, with the values of α being uniformly spread-out across the front. Furthermore, not only can we sensibly explore the LLM space, but COPA enables interpreting these models in terms of the original objectives *and* the population they live in. For example, we can say that `GPT-2` is Pareto-optimal as it consumes the least, but it only achieves a 6 % average performance score, or that `Phi-3.5-mini` is a top-10 % model in both aspects, consuming 0.53 kg of CO_2 vs. the 13 kg consumed by the best-performing model.

2. The fairness-accuracy trade-off. Moving to a more classic example, we consider how a DM could use COPA to choose a trade-off between accuracy and fairness in a classification problem, two objectives which are defined in completely different ways (Zafar et al., 2017).

We reproduce the CelebA (Liu et al., 2015) experiment from Maheshwari & Perrot (2022) using FairGrad—an algorithm whose hyperparameter ϵ upper-bounds the unfairness of the classifier—and create a population of models by sweeping through values of ϵ and five random initializations.

Fig. 5 (left) shows the Pareto-front in the accuracy-fairness space, as we navigate it by changing α, clearly showing the difference between both objectives. Note that solving Eq. 3 directly would lead to the solution with maximum accuracy, as in §5.1. Instead, using COPA we can uniformly

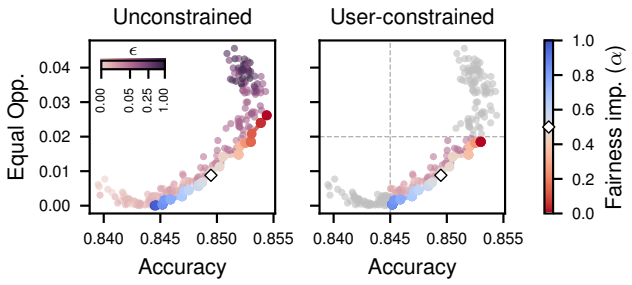


Figure 5: **COPA can be used to meaningfully explore the Pareto front between accuracy and fairness (equal opportunity)** in the CelebA experiment from Maheshwari & Perrot (2022) in unconstrained (left) as well as user-constrained scenarios (right).

navigate the Pareto front where, e.g., the robust min-max solution ($\alpha = 1/2$) lies precisely in the middle of the front. As a result, COPA offers a more reliable interpretation of its parameters than the upper-bound given by ϵ, which is clear by observing that, e.g., a value of $\epsilon = 1$ or 0.25 yields relatively similar solutions in Fig. 5.

In addition, we consider a more realistic scenario where DMs bargain on acceptable values for the objectives, e.g., a regulatory body could demand equal opportunity to never exceed 0.02 (MacCarthy, 2017). Despite constraining the Pareto front to consider only valid solutions,[4] COPA stills provides a sensible way to navigate the space of valid models, proving that we can easily combine rules on the original and CDF-transformed objective spaces.

5.3 Case 2: Comparative Model Analysis

Previously, we have explored how DMs can meaningfully explore the Pareto front. Now, we focus on a related but different question: *How much could semantic incomparability alter our analyses and conclusions?*

1. Incomparable objectives. First, we consider a multi-task learning (MTL) setting, where the heterogeneity of the tasks to solve makes it prone to face incomparable objectives. In fact, it is common to aggregate objectives with the average relative performance, as discussed in §4.

To clearly showcase the issue, we look at the multi-SVHN experiment from Javaloy & Valera (2022), which uses a modified version of SVHN (Netzer et al., 2011) with a digit on each side of the image, and where we solve three classification tasks: **i)** left digit; **ii)** right digit; and **iii)** parity of their product; and two regression tasks: **iv)** sum of digits; and **v)** number of active pixels in the image.

Fig. 6 shows the ranking of the 14 MTL methods considered by Javaloy & Valera (2022), if we were to use different criterion functions, namely: COPA with different values of

[4] We still use invalid models to approximate the CDF.

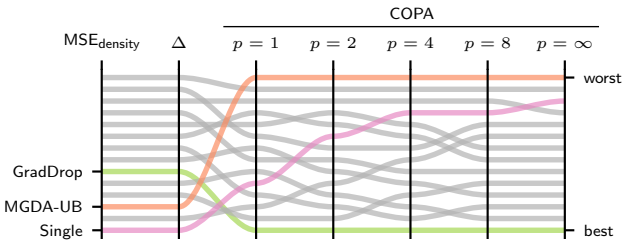


Figure 6: Ranking of MTL methods using different criterion functions to evaluate them. Those methods whose rankings change drastically with Δ are highlighted in colour.

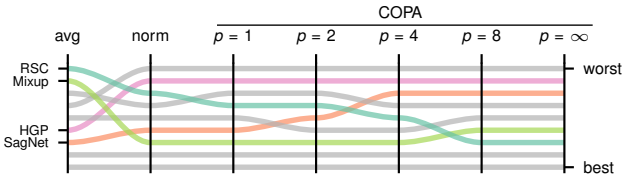


Figure 7: Ranking of domain generalization methods as we change the criterion function. Remarkably, the average accuracy is inconsistent with every COPA instance.

p and equal weights, the average relative performance, Δ, and the regression error over the density task. The first two columns of the plot make extremely clear how much the density task dominates the average relative performance, perfectly matching its ranking. Again, this is a result of the reference method having nearly zero regression error on this task, greatly magnifying its relative performance, Δ_k.

As expected, the outlined issue has a tremendous impact on the conclusions drawn, e.g.: **i)** the *worst* method for all COPA instances, `MGDA-UB` (Sener & Koltun, 2018), becomes the *best* one w.r.t. Δ; or **ii)** the best one for every COPA, `GradDrop` (Chen et al., 2020), becomes the 6th best. Fig. 6 also shows that the reference method (`Single`) is among the least robust models ($p = \infty$), and slowly improves as we prefer sparser solutions ($p = 1$).

It is worth-noting that the authors were aware of the issue and left the density task out when aggregating objectives, reporting both Δ and density MSE as a pair.

2. Seemingly comparable objectives. Sometimes, semantic incomparability can arise in unexpected scenarios. We take domain generalization as an example and, in particular, the DomainBed (Gulrajani & Lopez-Paz, 2021) experiment from Hemati et al. (2023). Here, the authors compare different methods by training them on some domains, and testing them on 4 unseen ones, reporting the average domain accuracy as commonly done in the literature.

Fig. 7 shows the ranking of the considered methods as we use different criterion functions, with the average accuracy in the first column. For two of the highlighted methods,

Table 1: Different effective ranges explain the differences in rankings of the domain generalization experiment. The table shows the effective range of each domain accuracy, and the performance of `Mixup` and `HGP` for the raw and normalized (Eq. 7) domain accuracies, respectively.

		VLC	PACS	OfficeHome	DomainNet	Avg
Min. acc.		76.30	78.80	60.20	23.40	-
Max. acc.		79.30	84.80	68.50	41.40	-
Acc.	Mixup	77.70	83.20	67.00	38.50	66.60
	HGP	76.70	82.20	67.50	41.10	66.88
Norm.	Mixup	46.67	73.33	81.93	83.89	71.45
	HGP	13.33	56.67	87.95	98.33	64.07

`RSC` (Huang et al., 2020) and `SagNet` (Nam et al., 2020), we observe their performance deteriorate and improve, respectively, as we consider less robust criteria, in accordance with the average accuracy. However, we see a different story with `HGP` (Hemati et al., 2023) and `Mixup` (Wang et al., 2020), whose rankings are consistent for all COPA instances, but drastically change when we average accuracies. This leads to significantly different analyses concluding, e.g., that `Mixup` is *worse* than `SagNet` and `HGP`, *in disagreement with every other criterion function*.

In fact, accuracies present *significantly different ranges* across domains, as we show in Tab 1, and differences in domains with the less variance are less important in the average. If we normalize the results using norm_k (Eq. 6), we see that `Mixup` significantly outperforms `HGP` in these domains, swapping their rankings. This can also be observed in Fig. 7, where `norm` aligns much better with COPA.

5.4 Case 3: Benchmarking

Finally, we motivate the use of COPA and CDF-normalized objectives in general benchmarking where, in contrast with the previous use cases, objectives are not necessarily aggregated into a scalar value, but plotted together.

We take the AutoML Benchmark (AMLB) (Gijsbers et al., 2024) as an example as it "follows best practices and avoids common mistakes when comparing frameworks." We reproduce all figures from the original publication, comparing 15 AutoML methods evaluated on 104 different objectives. Since objectives are incomparable, the authors scale them using norm_k (Eq. 7) with a random forest as reference model, providing different plots and analyses from this dvslrf objectives. It is worth-noting that the authors also encourage the use of CD diagrams and Friedman tests, two methods that are based on relative rankings.

A natural step is therefore to replace scaled objectives with CDF-normalized ones. As an example, we show in Fig. 8 the same AMLB boxplot using scaled and CCDF performance, i.e., $1 - F_k(\mathrm{y}_k)$. We find that using CCDFs comes with several benefits, e.g.: **i)** there are no outliers to report, unlike

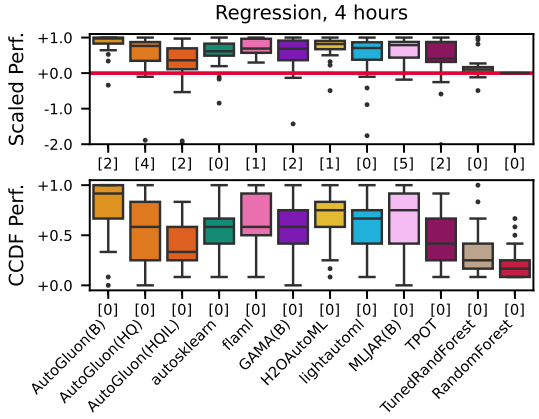


Figure 8: Comparison of different AutoML methods on AMLB (Gijsbers et al., 2024) using scaled performance, `norm`, with a default random forest as reference method (red line) to normalize objectives (top); and using CCDF-transformed performance instead (bottom). Brackets indicate the number of off-view outliers.

in the original plot; **ii)** all values lie in $[0, 1]$; **iii)** there is no need for an arbitrary reference model; and **iv)** we can provide clear population-based interpretations, e.g., "on average, `AutoGluon(B)` (Erickson et al., 2020) yields over top-10 % performance on the considered objectives."

As we report in App. A.5, these benefits extend to all AMBL plots, demonstrating that the proposed CDF transformation is a sensible way of normalizing objectives in general.

6 Concluding remarks

In this work, we have shown the importance of meaningfully navigating the Pareto front in multi-objective ML evaluation, allowing DMs to perform better-informed decisions regarding the trade-off they commit to. We have highlighted how crucial is to properly normalize all objectives—making them semantically comparable—and to have an interpretable criterion function that sensibly reflects DM preferences into an optimization problem—making managing multi-objective trade-offs feasible, especially in high dimensions. Finally, we have implemented these insights in COPA, and extensively shown the impact that it can have in areas as fundamental as model selection and benchmarking.

Our work opens many intriguing venues for future research. For example, we would be excited to see COPA adapted to active scenarios with humans-in-the-loop, criterion functions that parametrize other preference types, a systematization of model selection enabled by COPA, or the adoption of COPA in applications such as the Open LLM Leadearboard (Fourrier et al., 2024) with which we motivated this work.

Impact Statement

This paper presents work whose goal is to improve the way we evaluate machine learning models in multi-objectives scenarios. There are many potential societal consequences of our work inherent from the field itself, none of which we feel must be specifically highlighted here.

Bibliography

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41 (2):423–443, 2018. (Cited in page 5.)

Barenblatt, G. I. *Dimensional analysis*. CRC Press, 1987. (Cited in pages 3 and 5.)

Bowman, V. J. On the Relationship of the Tchebycheff Norm and the Efficient Frontier of Multiple-Criteria objectives. In Thiriez, H. and Zionts, S. (eds.), *Multiple Criteria Decision Making*, pp. 76–86, Berlin, Heidelberg, 1976. Springer Berlin Heidelberg. ISBN 978-3-642-87563-2. (Cited in page 3.)

Branke, J., Deb, K., Miettinen, K., and Roman, S. Multiobjective Optimization, Interactive and Evolutionary Approaches [outcome of Dagstuhl seminars]. 2008. (Cited in page 2.)

Caruana, R. and Niculescu-Mizil, A. Data mining in metric space: an empirical analysis of supervised learning performance criteria. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 69–78, 2004. (Cited in page 2.)

Casella, G. and Berger, R. L. *Statistical inference*. Cengage Learning, 2021. (Cited in page 4.)

Chen, Z., Ngiam, J., Huang, Y., Luong, T., Kretzschmar, H., Chai, Y., and Anguelov, D. Just Pick a Sign: Optimizing Deep Multitask Models with Gradient Sign dropout. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/16002f7a455a94aa4e91cc34ebdb9f2d-Abstract.html. (Cited in page 7.)

Coignion, T., Quinton, C., and Rouvoy, R. Green My LLM: Studying the key factors affecting the energy consumption of code assistants. *ArXiv preprint*, abs/2411.11892, 2024. URL https://arxiv.org/abs/2411.11892. (Cited in pages 1 and 6.)

Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., and Smola, A. Autogluon-tabular: Robust and accurate automl for structured data. *ArXiv preprint*, abs/2003.06505, 2020. URL https://arxiv.org/abs/2003.06505. (Cited in page 8.)

Flach, P. A. *ROC Analysis*, pp. 869–875. Springer US, Boston, MA, 2010. ISBN 978-0-387-30164-8. doi: 10.1007/978-0-387-30164-8_733. URL https://doi.org/10.1007/978-0-387-30164-8_733. (Cited in page 5.)

Fourrier, C., Habib, N., Lozovskaya, A., Szafer, K., and Wolf, T. Open LLM Leaderboard v2, 2024. URL https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard. (Cited in pages 1, 6, 8, 13, and 14.)

Friedman, M. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701, 1937. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2279372. (Cited in page 5.)

Geenens, G. (Re-)Reading Sklar (1959)–A Personal View on Sklar's theorem. *Mathematics*, 12(3):380, 2024. (Cited in pages 4 and 5.)

Gijsbers, P., Bueno, M. L. P., Coors, S., LeDell, E., Poirier, S., Thomas, J., Bischl, B., and Vanschoren, J. Amlb: an AutoML benchmark. *Journal of Machine Learning Research*, 25(101):1–65, 2024. URL http://jmlr.org/papers/v25/22-0493.html. (Cited in pages 8, 16, 17, and 18.)

Goodfellow, I. J., Bengio, Y., and Courville, A. C. *Deep Learning*. Adaptive computation and machine learning. MIT Press, 2016. ISBN 978-0-262-03561-3. URL http://www.deeplearningbook.org/. (Cited in page 5.)

Gulrajani, I. and Lopez-Paz, D. In Search of Lost Domain generalization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=lQdXeXDoWtI. (Cited in page 7.)

Hand, D. J. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine learning*, 77(1):103–123, 2009. (Cited in page 5.)

Hemati, S., Zhang, G., Estiri, A. H., and Chen, X. Understanding Hessian Alignment for Domain generalization. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 18958–18968. IEEE, 2023. doi: 10.1109/ICCV51070.

2023.01742. URL https://doi.org/10.1109/ICCV51070.2023.01742. (Cited in pages 7, 8, 15, and 16.)

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring Mathematical Problem Solving With the MATH dataset, 2021. URL https://arxiv.org/abs/2103.03874. (Cited in pages 6 and 14.)

Huang, D., Bu, Q., Zhang, J., Xie, X., Chen, J., and Cui, H. Bias assessment and mitigation in llm-based code generation. *ArXiv preprint*, abs/2309.14345, 2023. URL https://arxiv.org/abs/2309.14345. (Cited in page 1.)

Huang, Z., Wang, H., Xing, E. P., and Huang, D. Self-challenging improves cross-domain generalization. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part II 16*, pp. 124–140. Springer, 2020. (Cited in page 8.)

Ibrahim, A., Bidgoli, A. A., Rahnamayan, S., and Deb, K. A Novel Pareto-optimal Ranking Method for Comparing Multi-objective Optimization algorithms. *ArXiv preprint*, abs/2411.17999, 2024. URL https://arxiv.org/abs/2411.17999. (Cited in page 5.)

Japkowicz, N. and Shah, M. *Evaluating learning algorithms: a classification perspective*. Cambridge University Press, 2011. (Cited in pages 1 and 5.)

Javaloy, A. and Valera, I. Rotograd: Gradient Homogenization in Multitask learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=T8wHz4rnuGL. (Cited in pages 7 and 15.)

Javaloy, A., Meghdadi, M., and Valera, I. Mitigating Modality Collapse in Multimodal VAEs via Impartial optimization. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., and Sabato, S. (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9938–9964. PMLR, 2022. URL https://proceedings.mlr.press/v162/javaloy22a.html. (Cited in page 5.)

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konecný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Qi, H., Ramage, D., Raskar, R., Raykova, M., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. Advances and Open Problems in Federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021. ISSN 1935-8237. doi: 10.1561/2200000083. URL http://dx.doi.org/10.1561/2200000083. (Cited in page 5.)

Kendall, M. G. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938. ISSN 0006-3444. doi: 10.1093/biomet/30.1-2.81. URL https://doi.org/10.1093/biomet/30.1-2.81. (Cited in page 5.)

Kendall, M. G. and Sundrum, R. M. Distribution-Free Methods and Order properties. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, 21(3):124–134, 1953. ISSN 03731138. URL http://www.jstor.org/stable/1401424. (Cited in page 4.)

Krueger, D., Caballero, E., Jacobsen, J., Zhang, A., Binas, J., Zhang, D., Priol, R. L., and Courville, A. C. Out-of-Distribution Generalization via Risk Extrapolation (REx). In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5815–5826. PMLR, 2021. URL http://proceedings.mlr.press/v139/krueger21a.html. (Cited in page 15.)

Kukkonen, S. and Lampinen, J. Ranking-dominance and many-objective optimization. In *2007 IEEE Congress on Evolutionary Computation*, pp. 3983–3990. IEEE, 2007. (Cited in page 5.)

Liu, B., Feng, Y., Stone, P., and Liu, Q. FAMO: Fast Adaptive Multitask optimization. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/b2fe1ee8d936ac08dd26f2ff58986c8f-Abstract-Conference.html. (Cited in pages 3 and 5.)

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep Learning Face Attributes in the wild. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 3730–3738. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.425. URL

https://doi.org/10.1109/ICCV.2015.425. (Cited in pages 6 and 14.)

Luccioni, A. S., Viguier, S., and Ligozat, A.-L. Estimating the carbon footprint of bloom, a 176b parameter language model. *Journal of Machine Learning Research*, 24(253): 1–15, 2023. (Cited in page 1.)

MacCarthy, M. Standards of fairness for disparate impact assessment of big data algorithms. *Cumb. L. Rev.*, 48:67, 2017. (Cited in page 7.)

Maheshwari, G. and Perrot, M. Fairgrad: Fairness Aware Gradient descent. *ArXiv preprint*, abs/2206.10923, 2022. URL https://arxiv.org/abs/2206.10923. (Cited in pages 6, 7, and 14.)

Maninis, K., Radosavovic, I., and Kokkinos, I. Attentive Single-Tasking of Multiple tasks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 1851–1860. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00195. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Maninis_Attentive_Single-Tasking_of_Multiple_Tasks_CVPR_2019_paper.html. (Cited in page 3.)

Martínez, N., Bertrán, M., and Sapiro, G. Minimax Pareto Fairness: A Multi Objective perspective. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6755–6764. PMLR, 2020. URL http://proceedings.mlr.press/v119/martinez20a.html. (Cited in page 5.)

Miettinen, K. *Nonlinear multiobjective optimization*, volume 12. Springer Science & Business Media, 1999. (Cited in pages 3 and 5.)

Nam, J. H., Cha, H., Ahn, S., Lee, J., and Shin, J. Learning from Failure: De-biasing Classifier from Biased classifier. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/eddc3427c5d77843c2253f1e799fe933-Abstract.html. (Cited in page 8.)

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., and Mian, A. A comprehensive overview of large language models. *ArXiv preprint*, abs/2307.06435, 2023. URL https://arxiv.org/abs/2307.06435. (Cited in page 6.)

Navon, A., Shamsian, A., Achituve, I., Maron, H., Kawaguchi, K., Chechik, G., and Fetaya, E. Multi-Task Learning as a Bargaining game. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., and Sabato, S. (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16428–16446. PMLR, 2022. URL https://proceedings.mlr.press/v162/navon22a.html. (Cited in page 5.)

Nazabal, A., Olmos, P. M., Ghahramani, Z., and Valera, I. Handling incomplete heterogeneous data using VAEs. *Pattern Recognition*, 107:107501, 2020. (Cited in pages 2 and 5.)

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011. (Cited in page 7.)

Park, J. W., Tagasovska, N., Maser, M., Ra, S., and Cho, K. BOtied: Multi-objective Bayesian optimization with tied multivariate ranks. In *Forty-first International Conference on Machine Learning*, 2024. (Cited in page 5.)

Piggins, A. Collective Choice and Social Welfare–Expanded edition, 2019. (Cited in page 5.)

Ramé, A., Dancette, C., and Cord, M. Fishr: Invariant Gradient Variances for Out-of-Distribution generalization. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., and Sabato, S. (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18347–18377. PMLR, 2022. URL https://proceedings.mlr.press/v162/rame22a.html. (Cited in page 5.)

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. Gpqa: A Graduate-Level Google-Proof Q&A benchmark, 2023. URL https://arxiv.org/abs/2311.12022. (Cited in pages 6 and 14.)

Sener, O. and Koltun, V. Multi-Task Learning as Multi-Objective optimization. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 525–536, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/432aca3a1e345e339f35a30c8f65edce-Abstract.html. (Cited in page 7.)

Sklar, M. Fonctions de répartition à n dimensions et leurs marges. In *Annales de l'ISUP*, volume 8, pp. 229–231, 1959. (Cited in page 4.)

Sprague, Z., Ye, X., Bostrom, K., Chaudhuri, S., and Durrett, G. Musr: Testing the Limits of Chain-of-thought with Multistep Soft reasoning, 2023. URL https://arxiv.org/abs/2310.16049. (Cited in pages 6 and 14.)

Sun, B. and Saenko, K. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pp. 443–450. Springer, 2016. (Cited in page 15.)

Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q., Chi, E., Zhou, D., and Wei, J. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve them. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.824. URL https://aclanthology.org/2023.findings-acl.824. (Cited in pages 6 and 14.)

Tucker, H. G. A generalization of the Glivenko-Cantelli theorem. *The Annals of Mathematical Statistics*, 30(3): 828–830, 1959. (Cited in page 4.)

Verdu, S. and Poor, H. On minimax robustness: A general approach and applications. *IEEE transactions on Information Theory*, 30(2):328–340, 1984. (Cited in page 5.)

Wang, Y., Li, H., and Kot, A. C. Heterogeneous Domain Generalization Via Domain mixup. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pp. 3622–3626. IEEE, 2020. doi: 10.1109/ICASSP40776.2020.9053273. URL https://doi.org/10.1109/ICASSP40776.2020.9053273. (Cited in page 8.)

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., Li, T., Ku, M., Wang, K., Zhuang, A., Fan, R., Yue, X., and Chen, W. Mmlu-Pro: A More Robust and Challenging Multi-Task Language Understanding benchmark, 2024. URL https://arxiv.org/abs/2406.01574. (Cited in pages 6 and 14.)

Wilcoxon, F. Individual Comparisons by Ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945. ISSN 00994987. URL http://www.jstor.org/stable/3001968. (Cited in page 5.)

Yuan, L., Chen, Y., Cui, G., Gao, H., Zou, F., Cheng, X., Ji, H., Liu, Z., and Sun, M. Revisiting Out-of-distribution Robustness in NLP: Benchmarks, Analysis, and llms evaluations. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/b6b5f50a2001ad1cbccca96e693c4ab4-Abstract-Datasets_and_Benchmarks.html. (Cited in page 1.)

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. Fairness Constraints: Mechanisms for Fair classification. In Singh, A. and Zhu, X. J. (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pp. 962–970. PMLR, 2017. URL http://proceedings.mlr.press/v54/zafar17a.html. (Cited in page 6.)

Zeleny, M. Compromise programming. *Multiple criteria decision making*, 1973. (Cited in page 2.)

Zhou, J., Lu, T., Mishra, S., Brahma, S., Basu, S., Luan, Y., Zhou, D., and Hou, L. Instruction-Following Evaluation for Large Language models, 2023. URL https://arxiv.org/abs/2311.07911. (Cited in pages 6 and 14.)

Appendix

Table of Contents

A Experimental details and additional results

In this section, we provide all details to reproduce the experiments presented in the manuscript, as well as additional results which were omitted from the main paper due to space constraints.

A.1 Synthetic evaluation

As we describe in the main text, for the synthetic experiment we consider the following parametric curve:

$$y_2 = 0.25\cos(39y_1^{0.85}) - \log(y_1) - 0.46\,, \qquad (14)$$

where $y_1 \sim \mathcal{U}(0.02, 0.2)$. As a result, we end up with a non-convex Pareto front with a flat area around $y_1 = 0.1$, and two objectives with significantly different distributions. Moreover, the distribution of both objectives are significantly different. Specifically, the first objective is uniformly distributed, while the second one is precisely the plotted curve (if we flipped it to have the second objective as the x-axis), therefore being heavy tailed with most density lying in the $[0, 0.2]$ interval. The uneven and long-stretch of the domain of the second objective thus explains why, despite applying norm_k, we still get a biased optimization problem in Fig. 2, as discussed in the main text.

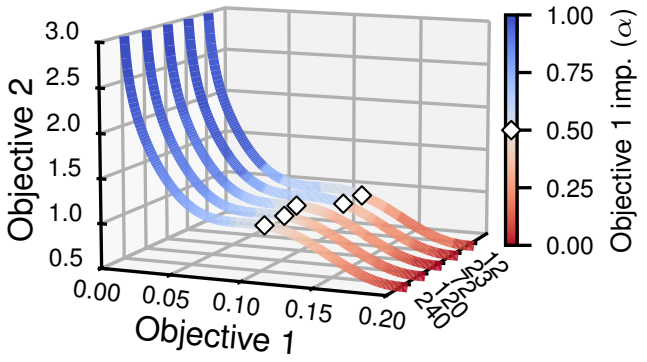


Figure 9: Synthetic Pareto front showing the Pareto front using COPA with $p = \infty$ as we change the number of sampled points. While it can be observed a deterioration on the estimated Pareto front (see quantized colours as we reduce N), COPA offers a robust estimator even with 12 datapoints.

A.1.1 Additional results

How robust are we to sample size? Despite having a closed-form expression for the variance of our estimator u_k in §3.2, we empirically show in Fig. 9 the estimated Pareto front using COPA with $p = \infty$ as a function of the first-objective importance, α, as we change the total number of points sampled to estimate it, N. We can observe that, despite considerably reducing the number of samples from 240 to 12 datapoints, the estimate given by COPA remains perfectly consistent.

A.2 Navigating the LLM Pareto front

Dataset details. In order to conduct our experiments, we retrieved the publicly available results from the Open LLM Leaderboard (Fourrier et al., 2024) using Huggingface's dataset Python package and, for reproducibility purposes, saved a local copy with the state as of the 9th-th of January 2025. From the 2929 total LLMs, we discard those which were not publicly available on Huggingface's hub. This leave us with a total of 2148 models, which we use to conduct the experiments described in this work.

Experimental details. As explained in the main text, we consider all reported values as objectives. Namely, we take as objectives the CO_2 emissions and all 6 benchmark performance scores computed on the following datasets: IFEval (Zhou et al., 2023), BBH (Suzgun et al., 2023), MATH (Hendrycks et al., 2021), GPQA (Rein et al., 2023), MuSR (Sprague et al., 2023), and MMLU-Pro (Wang et al., 2024). Then, we use COPA with $p = \infty$ to produce both Figs. 1 and 4, setting the values of $\boldsymbol{\omega}$ according to the importance given to CO_2 emissions, α, as $\boldsymbol{\omega} := [\alpha, \frac{1-\alpha}{6}, \dots, \frac{1-\alpha}{6}]$. To create these figures, we take 1000 values of α evenly-spaced in the unit interval and, since different values of α can provide us with the same model, use their range-average (Fig. 1) or maximum (Fig. 4) as the value to colour the selected LLMs in the figures. There are two more details worth-discussing. First, in Fig. 1 we use COPA over two objectives (the average score and CO_2 emissions) just so that the models selected by all criterion functions lied exactly in the plotted Pareto front, since Pareto-optimal models selected with all $K = 7$ objectives may not be Pareto-optimal when considering this bidimensional representation. Second, we use as y-axis for Fig. 4 the CDF of the p-norm computed using the CDF-transformed performance criteria (i.e., of the vector used with COPA, excluding the CO_2 dimension), since this represents much more closely the CDF-space that COPA navigates.

A.2.1 Additional results

Complementing Fig. 4, we present here the quantitative results of those LLMs selected with COPA. In the table we report the reported benchmark scores, a summary of their benchmark performance and CO_2, the CDF values found by COPA (same as in Fig. 4), and the value of α used to select these models. As it can be observed, COPA allows us to meaningfully navigate the performance-cost trade-off in the LLM space. Answering the initial question we posed in §1, if we were a practitioner trying to select a balanced LLM in terms of its performance and cost without further prior expectations, *we* would proceed in this case by using COPA with $p = \infty$ and $\alpha = 0.5$, which would yield us a model, unsloth/Phi-3-mini-4k-instruct, in the top-9 % of LLMs in terms of benchmark performance, and top-8 % in terms of CO_2 emissions.

Table 2: Quantitative results of the LLMs highlighted in Fig. 4 from the Open LLM Leadearboard (Fourrier et al., 2024) using COPA with $p = \infty$, as we change the importance of CO_2 consumption. Rather than using the average, the CDF value for the performance computes the weighted ∞-norm of the CDF-transformed benchmark results (i.e., the value used with COPA but separating CO_2 from the rest of objectives).

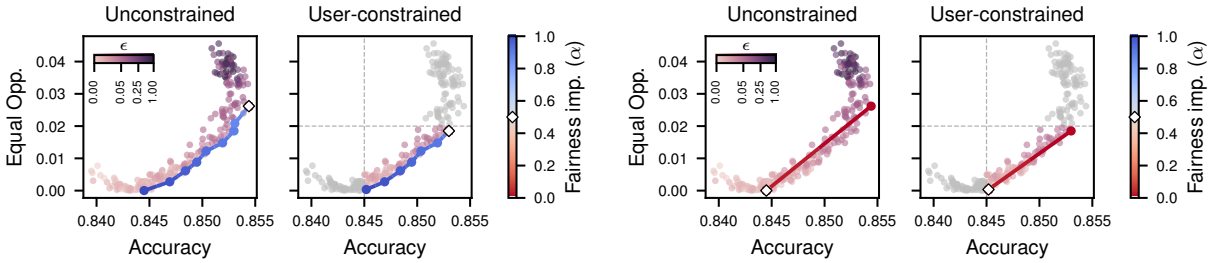
	Benchmarks scores						Summary		CDF values		
Full model name	IFEval (%)	BBH (%)	MATH (%)	GPQA (%)	MUSR (%)	MMLU-PRO (%)	Average (%)	CO_2 cost (kg)	Perf. ($p = \infty$)	CO_2 cost	α
dfurman/CalmeRys-78B-Orpo-v0.1	81.63	61.92	40.71	20.02	36.37	66.80	51.24	13.00	0.95	0.00	0.01
maldv/Qwentile2.5-32B-Instruct	73.93	57.21	38.07	17.90	19.96	54.21	43.55	3.53	0.87	0.01	0.02
sometimesanotion/Qwen2.5-14B-Vimarckoso-v3	72.57	48.58	34.44	17.34	19.39	48.26	40.10	1.93	0.79	0.01	0.03
hotmailuser/FalconSlerp3-7B	60.96	36.83	27.42	9.17	15.90	34.75	30.84	0.61	0.19	0.05	0.21
unsloth/Phi-3-mini-4k-instruct	54.40	36.73	15.41	9.73	13.12	33.68	27.18	0.47	0.09	0.08	0.50
icefog72/Ice0.37-18.11-RP	49.72	31.04	6.42	8.28	12.21	23.81	21.91	0.41	0.07	0.21	0.66
h2oai/h2o-danube3.1-4b-chat	50.21	10.94	2.11	4.70	10.20	19.10	16.21	0.30	0.03	0.60	0.82
postbot/gpt2-medium-emailgen	14.92	3.67	0.00	1.34	6.89	1.63	4.74	0.08	0.00	0.86	0.97

A.3 Navigating the fairness-accuracy trade-off

Experimental details. We reproduce the CelebA (Liu et al., 2015) experiment from (Maheshwari & Perrot, 2022) using their proposed FairGrad algorithm, which code is publicly available at github.com/saist1993/fairgrad, and run this experiment with 10 random initializations and 24 different values of ϵ (the hyperparameter of FairGrad that represents the desired fairness upper-bound). Namely, we consider the following values for ϵ:

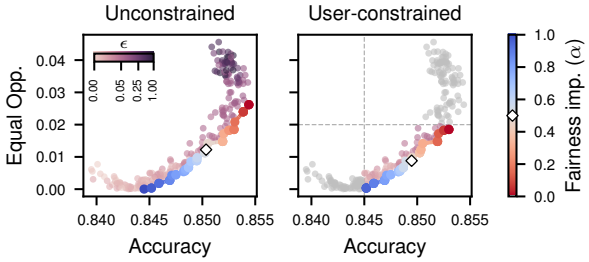
$$\{0.001, 0.002, 0.003, 0.004, 0.005, 0.006, 0.007, 0.008, 0.009,\\ 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09,\\ 0., 0.1, 0.2, 0.3, 0.5, 1.\}$$

This leave us with a total of 240 models. To produce Fig. 5, we use COPA with $p = \infty$ and 50 values of α evenly-spaced in the unit interval. For the constrained case, we simply drop those points that do not match the requirements for accuracy (being larger than 0.845) and fairness (having an equal opportunity value smaller than 0.02) before selecting any models with COPA. Of course, to compute the rankings of the accuracy, we take into account that it needs to be maximized and used the opposite order relation. Similarly, when we applied other normalization functions (see below), we employ the error rate (rather than the accuracy), so that it has to be minimized.

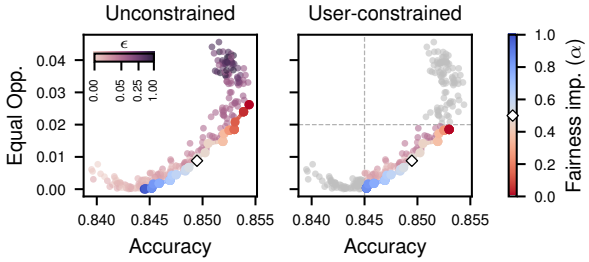


(a) Naively using the objectives.

(b) Using $\phi_k = \Delta_k$.



(c) Using $\phi_k = \text{norm}_k$.



(d) Using COPA.

Figure 10: We reproduce the fair ML experiment from §5.2 using different normalization functions. We can observe that only COPA meaningfully navigates the Pareto front, with all other approaches being biased towards one of the extreme solutions. Indeed, Δ_k only reaches the two extreme solution despite sampling 50 evenly-spaced values for α.

A.3.1 Additional results

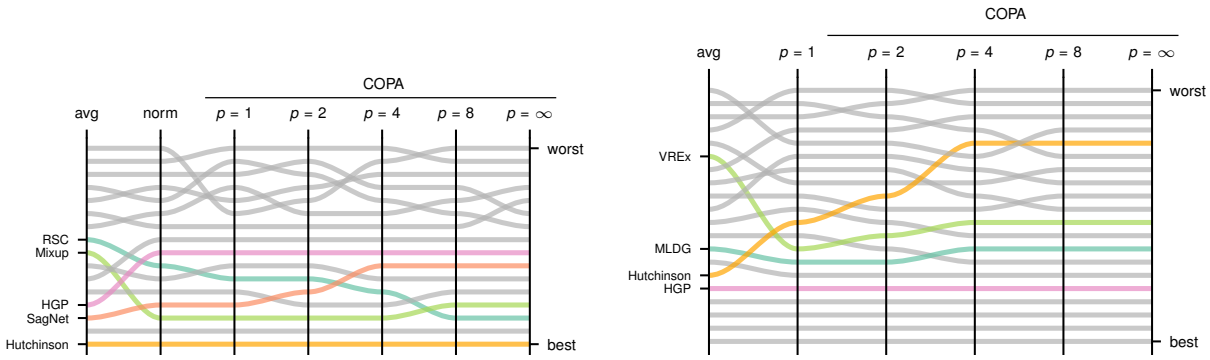
We show in Fig. 10 the same plot as in Fig. 5, but using all the considered normalization functions. Similarly to what we observed in the introductory example in Fig. 1, all other methods are biased towards minimizing one of the objectives.

A.4 Comparative model analysis experiments

Experimental details. For the figures shown in §5.3, we retrieved the results reported by the two selected works. In particular, we took the values reported in the second half of Table 5 from the work of Javaloy & Valera (2022) for the MTL experiment, and values reported in Table 4 of Hemati et al. (2023) for the domain generalization experiment of the main text. From these values, we simply re-rank them using the different criterion functions discussed in the main paper, and highlight those which we consider are interesting for the discussion we carry out in the main manuscript. We use equal weights for all versions of COPA. One important detail is that, for the domain generalization case, we kept only the top methods, as the rest do not add anything more to the discussion and make the plot more difficult to read.

A.4.1 Additional results

As mentioned just above, we discarded some methods in the domain generalization figure of the main text (i.e., Fig. 7). For completeness, we show in Fig. 11a the full figure with all methods included, and highlighting `Hutchinson`, the second method proposed by the authors, along `HGP`. Also, we show in Fig. 11b the same figure but using as data the one reported in Table 9 from Hemati et al. (2023) (instead of Table 4). This table was reported in the supplementary material, and the difference between both tables is the method used to select hyperparameters, with all methods but those proposed by this particular work (i.e., `HGP` and `Hutchinson`) improving their performance. More crucially, we show once again the huge discrepancies in ranking between using the average accuracy and any of the COPA versions. This time, we also report `Hutchinson`, which is the best method for all criterion functions in Fig. 11a, and the fourth to worst method in Fig. 11b. We can again observe how much our final conclusions can change in Fig. 11b, where the fourth to worst method in terms of average domain accuracy, `VREx` (Krueger et al., 2021), is better than `Hutchinson` in all instances of COPA. To finalize, we consider important to report that, in both figures, the first gray line (i.e., the second-best and best methods, respectively) correspond to the domain generalization method named `CORAL` (Sun & Saenko, 2016).



(a) Table 4 from Hemati et al. (2023).

(b) Table 9 from Hemati et al. (2023).

Figure 11: Ranking of the domain generalization methods considered by Hemati et al. (2023) as we use different criterion functions to rank them. We can appreciate a significant change of rankings, and the average accuracy in particular being highly inconsistent with all versions of COPA. We highlight those methods used for the discussion in the text.

A.5 AutoML Benchmarking (AMLB) experiment

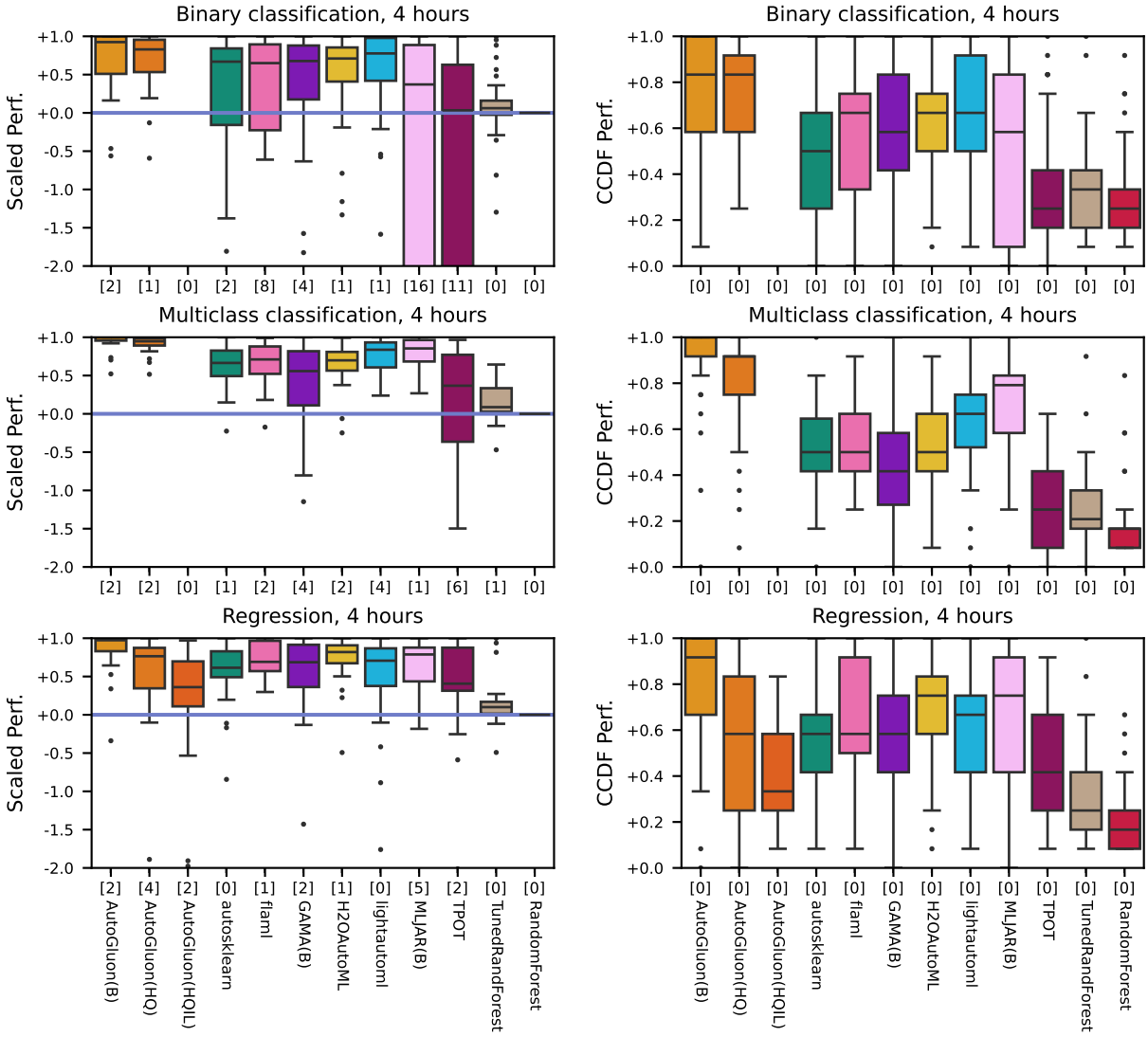
Experimental details. To demonstrate the out-of-the-box utility of COPA and its two components, we reproduce some of the plots from the AutoML Benchmark from Gijsbers et al. (2024). To achieve this, we simply modify the Jupyter notebook publicly available at github.com/PGijsbers/amlb-results, and add a few lines of code to compute COPA as proposed in this work.

A.5.1 Additional results

To complement Fig. 8 from the main text, we provide here side-by-side comparisons of more figures reported by Gijsbers et al. (2024), further reinforcing the argument of broadly adopting CDF-transformed objectives for general cases.

In particular, we show in Fig. 12 the same three figures as Figure 3 from the original work, where the same advantages when using the proposed CDF transformation, as those discussed in the main text (see §5.4), can be observed here. Furthermore, we show in Fig. 13 Figure 4 from the original work, where all 104 objectives are used, further showcasing the benefits of the proposed transformation.

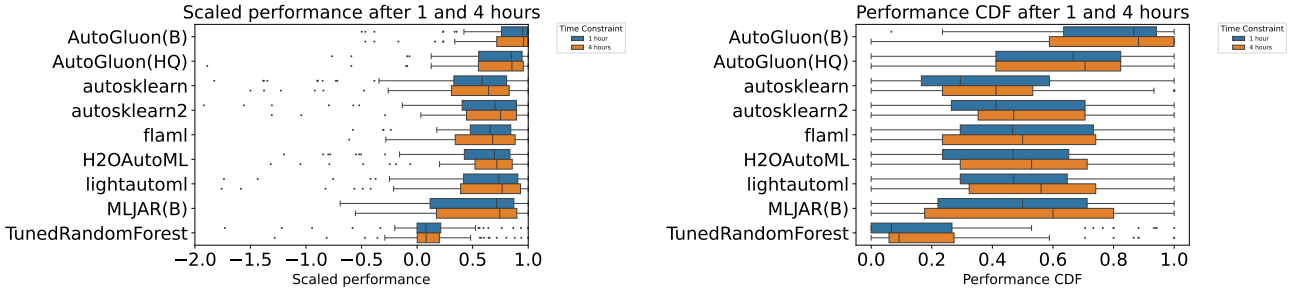
Finally, we also reproduce Figure 7 from the original publication in Fig. 14, where different Pareto plots are generated according to the type of tasks, showing the performance-speed trade-off, similar in spirit to Fig. 1 in this work. Here, we use COPA with $p = 2$ and equal weights. We can observe that, while some of the figures are quite similar, e.g., binary classification in the top row, some others differ significantly, e.g., regression in the bottom row, where COPA reports two less Pareto-optimal models. Beyond the differences in using scaled vs. CDF-transformed objectives, which we have extensively discussed during this paper, and showed the significant advantages of employing the latter, the differences in the number of Pareto-optimal models is due to the fact that the Pareto front is computed *after* aggregating the performance metrics. This is in stark contrast with the approach taken in this work (except for Fig. 1 for visualization purposes, see App. A.2), where we compute Pareto-optimal points on the space of all objectives. As we have been arguing during this work, COPA allows us to meaningfully navigate the Pareto front, enabling the creation of plots such as those reported in this work (e.g., Figs. 1, 4 and 5), which are significantly more informative than those reported before our work, as it can be clearly observed in Fig. 14.



(a) Using scaled performance.

(b) Using CCDF values.

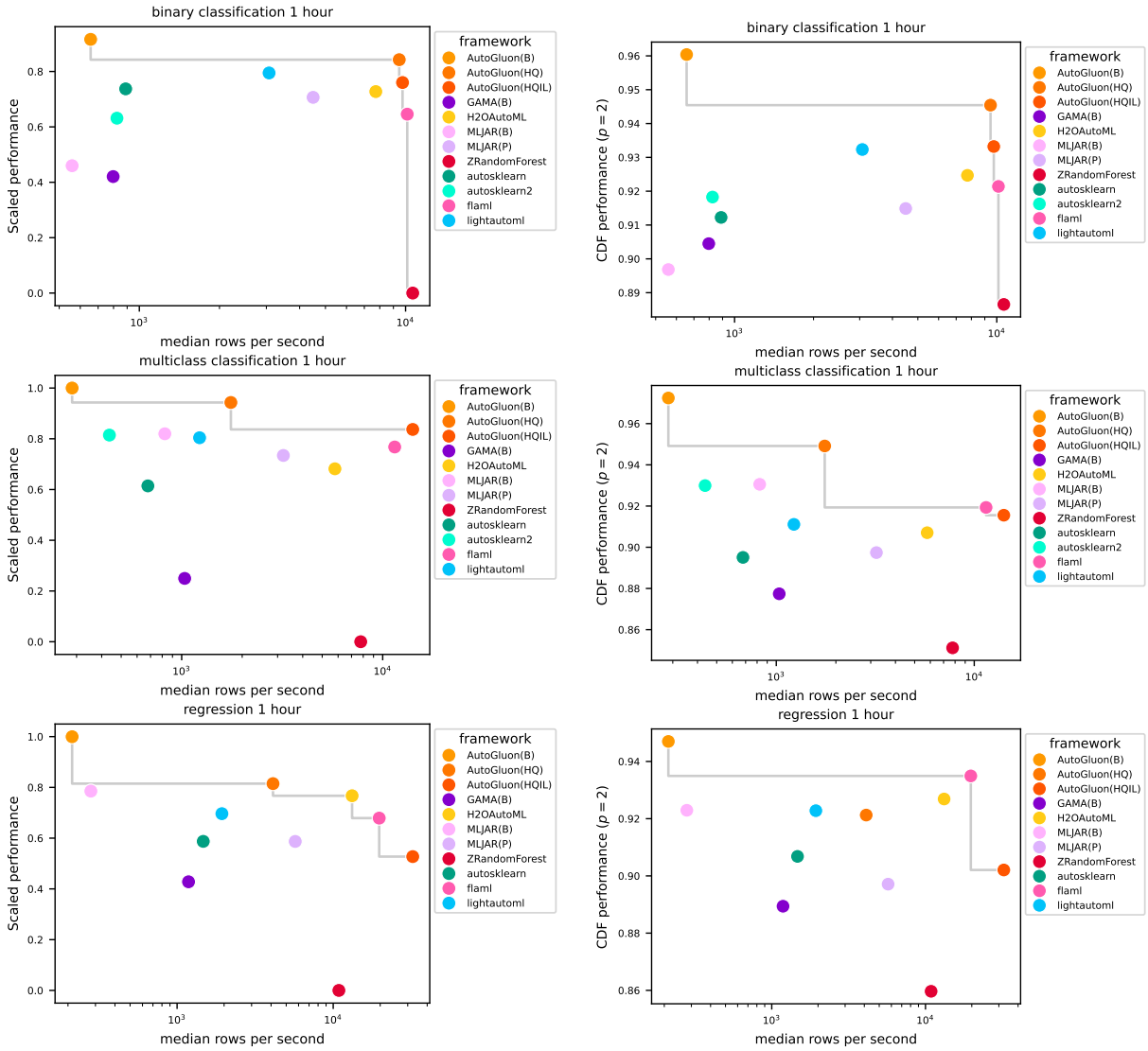
Figure 12: We reproduce Fig. 3 from Gijsbers et al. (2024) in (a) using their proposed scaled performance, and we show the same figure in (b) but using complementary CDF values (CCDF, one minus the CDF value). The same advantages as those discussed in §5.4 can be observed here.



(a) Using scaled performance.

(b) Using CCDF performance.

Figure 13: We reproduce Fig. 4 from Gijsbers et al. (2024) in (a) using their proposed scaled performance, and we show the same figure in (b) but using complementary CDF values (CCDF, one minus the CDF value). The same advantages as those discussed in §5.4 can be observed here.



(a) Using scaled performance.

(b) Using CCDF values.

Figure 14: We reproduce Fig. 7 from Gijsbers et al. (2024) in (a) using their proposed scaled performance, and we show the same figure in (b) but using complementary CDF values (CCDF, one minus the CDF value).