
COPA: Comparing the incomparable in multi-objective model evaluation

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 As machine learning (ML) practitioners, we often have hundreds of (trained) ML
2 models at hand from which we need to choose one, based on various objectives
3 such as accuracy, robustness, fairness, scalability, etc. However, how to *compare*,
4 *aggregate* and, ultimately, *trade-off* these objectives is usually a time-consuming
5 task that requires expert knowledge, as they may be measured in different units or
6 scales. In this work, we investigate *how* objectives can be automatically normalized
7 and aggregated to systematically navigate their Pareto front. To do so, we make
8 incomparable objectives comparable using their CDFs, approximated by their
9 relative rankings. As a result, we can aggregate them while matching user-specific
10 preferences, allowing practitioners to meaningfully navigate and search for models
11 in the Pareto front. We demonstrate the potential impact of our approach, named
12 COPA, in both model selection and benchmarking tasks across diverse ML areas
13 such as fair ML, domain generalization, AutoML and foundation models, where
14 classical ways to normalize and aggregate objectives fall short.

15 1 Introduction

16 In many phases of machine learning (ML), from model development to deployment, we often need
17 to *compare and select among a population of trained models according to multiple objectives*.
18 For example, even in the simple scenario of a single classification task, model selection involves
19 comparing and selecting among a population of trained classifiers with different hyperparameters
20 to find a specific compromise among objectives such as accuracy, sensitivity, or specificity [24]. A
21 common and more complex scenario these days involves benchmarking a large number of large deep
22 learning models in terms of how they perform with respect to many and diverse objectives that go
23 beyond accuracy, such as robustness [61], fairness [21], and CO₂ footprint [9, 35]. In both examples,
24 we encounter the following challenge: *how do we systematically compare and select among a large*
25 *number of ML models in terms of multiple objectives?*


26 Moreover, *different users and applications often have different needs and preferences*. For example, a
27 user may want to download a subset of trained large language models (LLMs) from the *Open LLM*
28 *Leaderboard* [12] to compare different prompt engineering approaches for a new task. The user
29 requires LLMs that perform relatively well without leaving unnecessarily large CO₂ footprints. To this
30 end, they need to compare the 2148 submitted LLMs in terms of 7 objectives, i.e., their performance
31 across 6 benchmarks and inference CO₂ cost. Among these models, 487 present non-trivial trade-offs,
32 i.e., for every pair, one is better in an objective but worse in another (see Fig. 1). How should they
33 compare the hundreds of models to decide what are acceptable performance-emission trade-offs?
34 Should they manually inspect all 487 LLMs? And what if another user required the most robust
35 model, rather than the most performant? Should they start from scratch?

36 Similar challenges can be easily found in the literature related to, e.g., multitask learning or domain
37 generalization [43, 48], where the selected model is expected to work ‘well’ on several tasks/domains;

fair classification [62], where it is often unclear what is an acceptable fairness-accuracy trade-off for deployment; or AutoML [15], where tens of frameworks are compared on hundreds of objectives. Crucially, all these works highlight two important limitations in multi-objective ML evaluation:

- L1. Objectives with different semantics and domains, such as average performance score and CO₂ 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 meters and grams.
- L2. When dealing with many objectives (7 in our LLM example), it is challenging for humans to translate their preferences into a concrete decision, as the number of plausible trade-offs quickly becomes overwhelming (487 in our example).

These challenges 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. The most common approach would be to perform a weighted combination of either the raw (Naive) or normalized objectives (Norm. and Delta, see §2 for their definitions). However, as we show in Fig. 1, both fail to address L1 and, thus, to evenly explore the Pareto front. In other words, they map most CO₂ importance values, α , to a small region of the front. To overcome these issues, prior works had to devise heuristic approaches tailored to their specific cases [5, 44]. For example, the authors of DecodingTrust [57] had to provide 8 ad hoc rules to normalize their objectives, one per objective. To date, we lack grounded approaches to compare, aggregate and, ultimately, trade-off objectives according to user preferences, that can be used out-of-the-box in multi-objective ML evaluation.

Contributions. We first motivate and *establish* the incomparability problem in multi-objective ML evaluation, shedding light on why previous approaches fail (§2). Next, we introduce **COPA** , a novel approach to allow practitioners to *meaningfully navigate the Pareto front*, and thus 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 their cumulative distribution functions, which we approximate using relative rankings; and **ii)** a criterion function with two easily interpretable parameters controlling the aggregation and importance of each objective. We then place COPA in the context of related work (§4), and finally demonstrate its potential impact (§5) in diverse and timely applications such as domain generalization, multitask learning, fair ML, AutoML benchmarking, and LLM selection. As we illustrate in Fig. 1, COPA enables thoroughly exploring the Pareto front as a function of the user preferences, here controlled by α . For instance, a deployer equally interested in the performance and CO₂ 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), ranked top-18 % for both objectives.

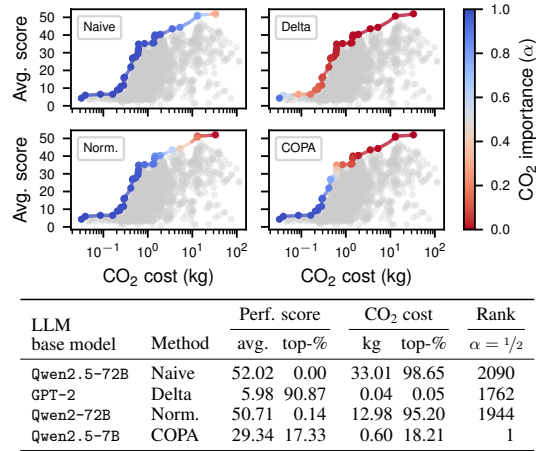


Figure 1: **COPA meaningfully navigates the performance-emissions trade-off of the Open LLM Leaderboard [12]**, evenly mapping the importance of CO₂ cost to the Pareto front. In contrast, existing approaches are biased toward one of the objectives. This is reflected in the retrieved LLMs where, e.g., COPA maps $\alpha = 1/2$ to a top-18 % model for both objectives, and all other approaches select either a high-performing but CO₂-demanding model, or vice versa.

2 Problem statement

We are given a population of already-trained models \mathcal{H} , typically obtained by changing hyperparameters, where each model $h \in \mathcal{H}$ is associated to a vector of K metrics assessing its performance with respect to 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:

$$\min_{h \in \mathcal{H}} \mathbf{y}(h) := [y_1(h), y_2(h), \dots, y_K(h)], \quad (1)$$

where $\mathbf{y}(h)$ is the objective vector of model h , and $y_k(h)$ its performance on the k -th objective. When it is clear from the context, we will omit the argument and write \mathbf{y} and y_k directly.

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 is not always better than the other for all objectives. In the MOO literature, the set of 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 $y_k \leq y_k^*$ for all $k \in \{1, 2, \dots, K\}$, and $y_k < y_k^*$ for at least one of the objectives.

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

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

One remarkable example of the latter is the *global-criterion method* [63] which maps DM preferences to the problem geometry by interpreting Eq. 2 as selecting the model closest to the *ideal* one, i.e.,

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

where $\mathbf{y}^{\text{ideal}}$ is the ideal solution, $\mathbf{y}^{\text{ideal}} := [\min_h y_1, \min_h y_2, \dots, \min_h y_K]$, and $\|\cdot\|_*$ is typically a p -norm. However, naively solving Eq. 3 (and, more generally, Eq. 2) is well-known in the MOO literature to be sensitive to the scaling of the objectives [4] (recall **L1** in §1), and thus prevents us from properly accounting for any DM preferences (**L2**). In this work, we argue that the criterion function C should fulfill the following desiderata:

D1. Reflect the DM preferences, translating their model expectations into an optimization problem.

D2. Provide a simple way to tune its parameters to meaningfully explore the Pareto front.

When are objectives incomparable? Similar to dimensional analysis in physics [2]—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 [3] of the form

$$\min_{h \in \mathcal{H}} \max \{ \alpha |y_1|, (1 - \alpha) |y_2| \}, \quad (4)$$

which solves Eq. 3 with C as the ∞ -norm weighted by $\alpha \in [0, 1]$. Intuitively, Eq. 4 looks for robust solutions that account for the importance of solving one objective over the other, seemingly satisfying our desiderata, **D1-2**. However, its naive application over the original objectives clearly shows how we can bias model selection in favor of Objective 2, as it can be seen in Fig. 2: for any given preference α smaller than 0.75, Eq. 4 yields a solution which *completely ignores Objective 1 performance*.

How can we make objectives comparable? As we just discussed, *even if we use a well-designed criterion function*, semantic incomparability can hinder our goal to meaningfully explore the Pareto

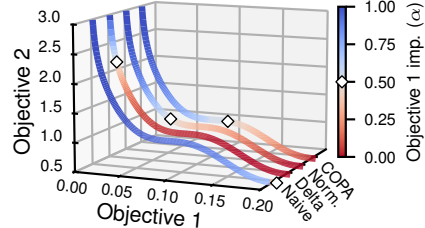


Figure 2: As we explore a synthetic Pareto front with different normalization functions to solve Eq. 3, only COPA meaningfully navigates it as we change α , and its min-max solution agrees with our expectations of a robust solution.

¹Note that, when we plot the Pareto front in 2D, e.g., in Fig. 1, the linear interpolation between models (dots) only serves visualization purposes, i.e., we cannot interpolate between models.

front. Historically, this has been addressed in the MOO literature by applying **component-wise transformations** to the objectives to normalize them [40], turning Eq. 2 into

$$\min_{h \in \mathcal{H}} C(\phi(\mathbf{y})) := C([\phi_1(y_1), \dots, \phi_K(y_K)]) . \quad (5)$$

Two classic examples of these transformations are

$$\Delta_k(y_k) := \frac{y_k - y_k^{\text{ideal}}}{y_k^{\text{ideal}}} \quad \text{and} \quad \text{norm}_k(y_k) := \frac{y_k - y_k^{\text{ideal}}}{y_k^{\text{nadir}} - y_k^{\text{ideal}}} , \quad (6)$$

where $y_k^{\text{nadir}} := [\max_h y_1, \max_h y_2, \dots, \max_h y_K]$ is the worst possible 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 y_k^{ideal} with a reference vector, as computing it can be challenging [33, 38, 40]. Back to our synthetic case, we now want to solve

$$\min_{h \in \mathcal{H}} \max \{ \alpha |\phi_1(y_1)|, (1 - \alpha) |\phi_2(y_2)| \} . \quad (7)$$

By testing different ϕ_k , we can understand why classic approaches fail to make objectives comparable. More specifically: **i)** using Δ_k now biases the problem toward the first objective, since $\min_h y_1 \approx 0$; 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 making a more meaningful use of α , spreading it uniformly along the curve.

The *main goal* of the functions $\phi_k: \mathbb{R} \rightarrow \mathbb{R}$ is thus to make the objectives semantically comparable, so that we can seamlessly aggregate them with the criterion function C . 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 (i.e., strictly increasing), so that it preserves Pareto-optimality.

In summary, to meaningfully explore the Pareto front, it is important to design a criterion function C that translates well DM preferences into an optimization problem (**D1-2**), and a normalization function ϕ that makes objectives semantically comparable (**D3-4**). These desiderata will blend in COPA, discussed in the next section. In the synthetic experiment above, COPA maps the value $\alpha = 1/2$, which turns Eq. 7 into a robust min-max problem [56], 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 fulfill desiderata **D3-4**, i.e., it should make any objectives semantically comparable while preserving their Pareto-optimality. Taking advantage of our probabilistic perspective (recall that y_k is a continuous random variable), 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 $\mathbf{u} := [u_1, u_2, \dots, u_K]$ instead of \mathbf{y} , where

$$u_k := F_k(y_k) \sim \mathcal{U}(0, 1) \quad \forall k \in \{1, 2, \dots, K\} , \quad (8)$$

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 [6, Example 5.6.3], and u_k is guaranteed to follow a standard uniform distribution if y_k is continuous.

Remarkably, Eq. 8 makes all criterion functions *marginal-distribution-free* in the sense of Kendall and Sundrum [29], i.e., it strips away all individual properties of the marginal distributions (e.g., the domain) of any given objective (**D3**). We note that normalizing random variables this way is one of the fundamental building stones of copulae in statistics [14, 51], ensuring that copula functions exclusively learn the relationship across random variables.

178 **How can we interpret the values of \mathbf{u} ?** One important advantage of using \mathbf{u} in place of \mathbf{y} in Eq. 5
 179 is that it provides a common framework to think about all objectives, since all their values are now
 180 framed as *elements within a population*. In practice, this means that the DM has a common language
 181 to express their expectations on the model. For example, $\mathbf{u} = 1/2$ corresponds for all objectives to the
 182 *the median value*, which divides \mathcal{H} into two *halves* comprising the best and worst performing models.
 183 However, there is still one caveat we need yet to address: we have no access to the marginal CDF of
 184 each objective, but only to samples of the joint distribution in \mathcal{H} .

185 3.2 Rankings as finite-sample approximations

186 As mentioned above, while we have no access to the CDFs themselves, we have samples from the
 187 joint distribution over the objectives, i.e., over, $p([y_1, y_2, \dots, y_K])$. Namely, we can consider each
 188 model $h \in \mathcal{H}$ as a sample from the joint distribution and, by looking at each objective individually,
 189 as a sample from the marginal distributions.

190 Let us now focus on the k -th objective, y_k , and drop the subindex in the following to ease notation. Say
 191 that we have $|\mathcal{H}| = N$ i.i.d. realizations of the objective, i.e., $\{y_1, y_2, \dots, y_N\} \stackrel{\text{i.i.d.}}{\sim} P_k$. Then, we can
 192 approximate Eq. 8 for the i -th sample, $u_i = F(y_i)$, by computing its order statistic, i.e., the random
 193 variable representing its relative ranking within the population, $R(i) := \sum_{j=1}^N [y_j < y_i]$, where
 194 Iverson brackets denote the indicator function, such that $y_{R(1)} \leq y_{R(2)} \leq \dots \leq y_{R(N)}$. Specifically,
 195 since the *empirical CDF* is the fraction of samples smaller than the input, it is direct to show that

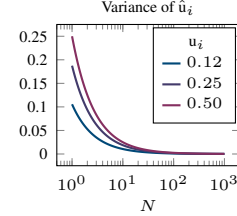
$$\hat{u}_i = \hat{F}(i) := \frac{1}{N} \sum_{j=1}^N [y_j < y_i] = \frac{1}{N} R(i) \quad (9)$$

196 enjoys the following properties [6]:

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

199 *Proof.* First, note that $[y_j < y_i] \sim \text{Bern}(u_i)$. Then, we have $R(i) \sim \text{Bin}(N, u_i)$ with mean Nu_i
 200 and variance $Nu_i(1 - u_i)$. Hence, \hat{u}_i has mean $\frac{1}{N} \mathbb{E}[R(i)] = u_i$, and variance $\frac{1}{N^2} \mathbb{V}[R(i)] =$
 201 $u_i(1 - u_i)/N$ which, by taking derivatives w.r.t. u_i , $\partial_{u_i} \mathbb{V}[\hat{u}_i] = 1 - 2u_i = 0 \Rightarrow u_i = 1/2$, which is a
 202 maximum since $\partial_{u_i}^2 \mathbb{V}[1/2] < 0$. \square

203 In other words, we can use the relative rankings of each objective to build
 204 an unbiased² estimator of the CDF, \hat{u}_i , whose variance rapidly decreases as
 205 we increase the size of \mathcal{H} , i.e., $\mathbb{V}[\hat{u}_i] \rightarrow 0$ as $N \rightarrow \infty$. Indeed, the inset
 206 figure shows the variance of \hat{u}_i as a function of the sample size for three
 207 different values of u_i . Note that the relative ranking is strictly increasing:
 208 if $y_i < y_j$, then $\hat{F}(y_i) < \hat{F}(y_j)$ for any \mathcal{H} containing both samples (D4).
 209 While this is an approximation of the true CDF, which would retain instead
 210 all the information about the joint distribution, it works egregiously well in our experiments (§5).
 211 Furthermore, note that this transformation is meant to ease inter-objective computations, we can (and
 212 should) use the original values of y_k to perform intra-objective comparisons or decisions.



213 3.3 Incorporating preferences into the optimization

214 Now that we can effectively approximate our normalization function, we introduce a criterion function
 215 to translate DM preferences into an optimization problem (D1). To do so, we start by looking back
 216 at global criterion methods, since plugging in our transformation $\mathbf{u} = \phi(\mathbf{y})$ simplifies the problem
 217 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
 218 approximation described in §3.2, the problem becomes a simple finite search of the form

$$\min_{i \in \{1, 2, \dots, N\}} \|\hat{\mathbf{u}}_i\|_* \quad (10)$$

219 That is, we have reduced our problem to finding the model whose ranking vector has the smallest
 220 norm. Using this new *marginal-free global-criterion method*, mapping the DM preferences now boils
 221 down to selecting an appropriate norm for the problem in Eq. 10. To this end, we propose to use as
 222 criterion function C a norm with parameters $p \geq 1$ and $\omega \in \mathbb{R}_+^K$ defined as

²In fact, it is known to be a consistent estimator [55].

$$\|\mathbf{u}\|_{p,\omega} := \left(\sum_{k=1}^K |\omega_k \mathbf{u}_k|^p \right)^{1/p}, \quad (11)$$

where $\sum_k \omega_k = 1$. This norm can be interpreted as a regular p -norm on a space with coordinates scaled by ω . More remarkably, 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 \mathbf{u}_k lie in the unit interval, and the power would often make them vanish too quickly, as we demonstrate in Fig. 10.

How can we interpret the parameters? Fortunately, the parameters of the proposed criterion function, p and ω , provide an easy and interpretable way for the DM to navigate the Pareto front (D2). Regarding ω , as we apply them in Eq. 11 *before* taking the power, we can provide a clear interpretation of ω in terms of ratio trade-offs. For example, if we had two objectives with $\omega = [0.75, 0.25]$, then we can see by equating the weighted objectives that minimizing the first objective to a value of \mathbf{u}_1 is worth the same as minimizing the second objective to a value of $\mathbf{u}_2 = \omega_1/\omega_2 \mathbf{u}_1 = 3\mathbf{u}_1$, i.e., \mathbf{u}_1 is three times more important than \mathbf{u}_2 . If we combine this interpretation with that 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.

We can interpret p using the same intuition as in ML regularization [16]: the models selected in Eq. 10 are those first intersecting an ever-expanding p -ball centered at the origin, whose shape depends on p . Higher values of p lead to denser objective vectors, while smaller values lead instead to sparser ones. Additionally, some values of p have clear interpretations: $p = 1$ is the average rank; $p = 2$ is the Euclidean distance; and $p = \infty$ turns Eq. 10 into a min-max problem, typically used to formulate robust optimization problems [56].

Does Eq. 11 enjoy theoretical guarantees? Given the similarity with commonly-used norms, it is natural to ask 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. 11 with $1 \leq p < \infty$ are always Pareto-optimal [40, Thm. 3.4.1]. However, it might not reach all optima. Similarly, note that $p = \infty$ reduces Eq. 10 to a weighted Tchebycheff problem which reaches any Pareto-optimal solution [40, Thm. 3.4.5], but also weakly optimal ones.

In practice, using a weighted Tchebycheff problem ($p = \infty$) is a good practice when we have few objectives and a large budget for the weights ω to test. 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 (as lower values of p lead to higher tolerance to bad performance on individual objectives), and ω based on the importance of solving each objective given by the DM.

4 Related work

Our work draws connections with other scientific domains, e.g., the notion of semantically incomparability is akin to that of incommensurability in dimensional analysis [2]. Similarly, using relative rankings to make better comparisons has been previously explored in microeconomics [47], MOO [23, 31], and statistics, designing methods that avoid the normality assumption, e.g., the Friedman test [13], Wilcoxon signed-rank test [60], or Kendall’s τ coefficient [28]. Finally, as mentioned in §2, copulas exploit the probability integral transform to become marginal-distribution-free [14], and the proposed criterion functions share similarities with weighted L_p -problems in MOO [40].

In ML, the closest work to ours is Park et al. [46], which learns the joint CDF, approximated with a copula, to recover a partial order for multi-objective Bayesian optimization. In contrast, we employ marginal CDFs and provide a principled way to translate DM preferences to an optimization problem. Another line of related works are those that attempt to learn the Pareto front either for model merging [7, 32] or a posteriori MOO methods [64]. Unfortunately, these methods fail to address semantic incomparability as they use the raw objectives. ROC curves [11] provide an interesting connection, since their axes can be understood as the CDFs of the target classes [18]. In practice, many prior works proposed ad hoc approaches to normalize and aggregate objectives using, e.g., normalized RMSEs [44, 57]—we refer to §8.3 of Japkowicz and Shah [24] for other references. Notoriously, some works in multitask learning [33, 43] and domain generalization [48] 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 [39], federated learning [27], probabilistic ML [26], and multimodal learning [1].

5 COPA in action

In this section, we motivate the use of COPA by showing a range of practical scenarios which would benefit from adopting the proposed methodology. We defer additional details and results to §A.

5.1 Synthetic evaluation

To qualitative assess COPA, we consider a synthetic Pareto front of the form $y_2 = 0.25 \cos(39y_1^{0.85}) - \log(y_1) - 0.46$ where $y_1 \sim \mathcal{U}(0.02, 0.2)$. We obtain as a result 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 the front is *strictly* increasing except in $[0.083, 0.091]$, we have that $u_1 \approx 1 - u_2$. As a result, we see that $p = 1$ almost exclusively finds solutions on the extrema. When we increase p , the distribution of solutions better spreads along the front and, as the p -balls become more square-like, we gain finer control on the solution found by tuning α . It is important to stress, however, that the finer control of $p = \infty$ comes at cost: as we increase K , finding a proper ω could prove challenging.

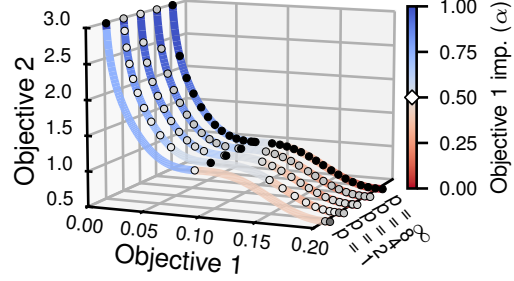


Figure 3: Distribution of solutions (circles) found for different values of p as we sweep over values of α . The darkness of the circles represents the number of times they were selected.

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 [42], their CO₂ footprint can be concerning and needs to be taken into account [9]. 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 [12] and take as objectives their inference CO₂ cost and performance on 6 different datasets: IFEval [65], BBH [54], MATH [20], GPQA [49], MuSR [52], and MMLU-Pro [58]. Then, we use COPA with $p = \infty$ to select an LLM, changing ω as we vary the importance given to their CO₂ footprint, denoted by α , as $\omega := [\alpha, (1-\alpha)/6, \dots, (1-\alpha)/6]$.

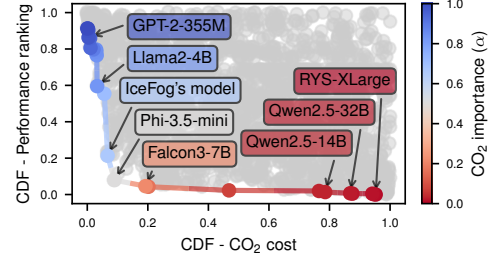


Figure 4: We can meaningfully explore the Pareto-optimal models of the Open LLM Leaderboard [12] with COPA. We use $p = \infty$ on the 7 objectives and highlight some of the selected models as we change the value of α .

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 CO₂ 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 [62].

We reproduce the CelebA [34] experiment from Maheshwari and Perrot [37] 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 directly solving Eq. 3 leads to the solution with maximum accuracy, as in §5.1. Instead, using COPA we can uniformly 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.

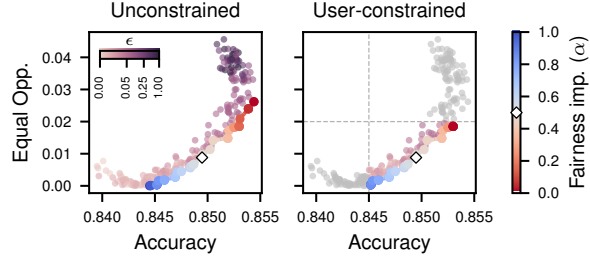


Figure 5: COPA can be used to meaningfully explore the Pareto front between accuracy and fairness (equal opportunity) in the CelebA experiment from Maheshwari and Perrot [37] in unconstrained (left) as well as user-constrained scenarios (right).

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 [36]. Despite constraining the Pareto front to consider only valid solutions (we still use invalid ones to approximate the CDF), 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 the conclusions we draw from comparative analyses in ML research?*

1. Incomparable objectives. First, we consider a multitask 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, Δ , as discussed in §4. To clearly showcase the issue, we look at the multi-SVHN experiment from Javaloy and Valera [25], which uses a modified version of SVHN [45] 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 and Valera [25], if we were to use different criterion functions, namely: COPA with different values of 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 .

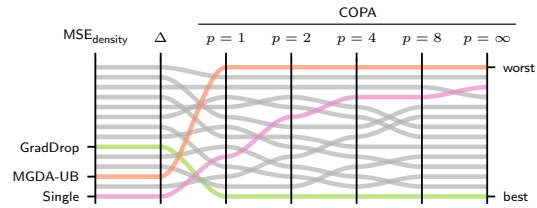


Figure 6: Ranking of MTL methods using different criteria to evaluate them. Methods whose rankings drastically change with Δ are colored.

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 [50], becomes the 3rd best method w.r.t. Δ ; or **ii)** the best one for every COPA, GradDrop [8], 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 look less at individual performances ($p = 1$). 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 [17] experiment from Hemati et al. [19]. 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, RSC [22] and SagNet

[41], 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 [19] and Mixup [59], 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 (see Tab 2) and differences in domains with less variance are less important in the average computation. 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.

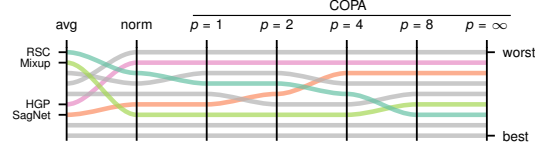


Figure 7: Ranking of domain generalization methods as we change the criterion function. Average accuracy is inconsistent with every COPA instance.

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. Additional plots can be found in §A.5.

We take the AutoML Benchmark (AMLB) [15] for the use-case as it “follows best practices and avoids common mistakes when comparing frameworks.” We reproduce all figures from the original work, comparing 15 AutoML methods evaluated on 104 different objectives. Since objectives are incomparable, the authors scale them using norm_k (Eq. 6) with a random forest as reference model, providing a number of analyses from these objectives. Remarkably, the authors also encourage the use of CD diagrams and Friedman tests, two methods that based on relative rankings.

A natural step is therefore to use CDF-normalized objectives. Fig. 8 shows the same AMLB boxplot using scaled and CCDF performance, i.e., $1 - F_k(y_k)$. We find that using CCDFs comes with several benefits: *i)* there are no outliers to report, unlike in the original plot (all values lie in $[0, 1]$); *ii)* there is no need for an arbitrary reference model; and *iii)* we can provide clear population-based interpretations, e.g., “on average, AutoGluon(B) [10] yields over top-10 % performance on the considered objectives.” These benefits extend to all AMLB plots, demonstrating that the proposed CDF transformation is a sensible way of normalizing objectives in general.

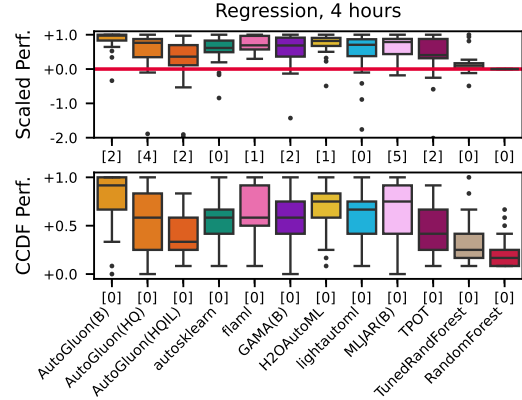


Figure 8: Comparison of AutoML methods on AMLB [15] using scaled performance, norm , with a random forest as reference method (red line); and using a CCDF-transformation (bottom). Brackets indicate the number of off-view outliers.

6 Concluding remarks

In this work, we have shown the importance of meaningfully navigating the Pareto front in multi-objective ML evaluation, allowing users to perform better-informed decisions. To this end, we have highlighted how crucial is to properly normalize all objectives and to have a criterion function that sensibly reflects DM preferences into an optimization problem. Finally, we have implemented these insights in COPA, and extensively demonstrated the impact that it can have in areas as fundamental and timely 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 settings with humans-in-the-loop, criterion functions that parametrize other preference types, a formal systematization of model selection enabled by COPA, or its adoption in public portals such as the Open LLM Leaderboard [12] or the DecodingTrust benchmark [57].

Bibliography

- [1] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018. (Cited in page 6.)
- [2] Grigory Isaakovich Barenblatt. *Dimensional analysis*. CRC Press, 1987. (Cited in pages 3 and 6.)
- [3] V. Joseph Bowman. On the Relationship of the Tchebycheff Norm and the Efficient Frontier of Multiple-Criteria Objectives. In Hervé Thiriez and Stanley Zionts, editors, *Multiple Criteria Decision Making*, pages 76–86, Berlin, Heidelberg, 1976. Springer Berlin Heidelberg. ISBN 978-3-642-87563-2. (Cited in page 3.)
- [4] Juergen Branke, Kalyan Deb, Kaisa Miettinen, and Slowinski Roman. Multiobjective Optimization, Interactive and Evolutionary Approaches [outcome of Dagstuhl seminars]. 2008. (Cited in page 3.)
- [5] Rich Caruana and Alexandru Niculescu-Mizil. 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*, pages 69–78, 2004. (Cited in page 2.)
- [6] George Casella and Roger L Berger. *Statistical inference*. Cengage Learning, 2021. (Cited in pages 4 and 5.)
- [7] Weiyu Chen and James T. Kwok. Pareto Merging: Multi-Objective Optimization for Preference-Aware Model Merging. 2024. URL <https://api.semanticscholar.org/CorpusID:271924011>. (Cited in page 6.)
- [8] Zhao Chen, Jiquan Ngiam, Yanping Huang, Thang Luong, Henrik Kretzschmar, Yuning Chai, and Dragomir Anguelov. Just Pick a Sign: Optimizing Deep Multitask Models with Gradient Sign Dropout. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *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 8.)
- [9] Tristan Coignion, Clément Quinton, and Romain Rouvoy. 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 7.)
- [10] Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. 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 9.)
- [11] Peter A. Flach. *ROC Analysis*, pages 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 6.)
- [12] Clémentine Fourier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. Open LLM Leaderboard v2, 2024. URL https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard. (Cited in pages 1, 2, 7, 9, 16, 17, and 23.)
- [13] Milton Friedman. 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 6.)
- [14] Gerry Geenens. (Re-)Reading Sklar (1959)—A Personal View on Sklar’s Theorem. *Mathematics*, 12(3):380, 2024. (Cited in pages 4 and 6.)
- [15] Pieter Gijsbers, Marcos L. P. Bueno, Stefan Coors, Erin LeDell, Sébastien Poirier, Janek Thomas, Bernd Bischl, and Joaquin Vanschoren. 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 2, 9, 20, 21, and 22.)

- [16] Ian J. Goodfellow, Yoshua Bengio, and Aaron C. Courville. *Deep Learning*. Adaptive computation and machine learning. MIT Press, 2016. ISBN 978-0-262-03561-3. URL <http://www.deeplearningbook.org/>. (Cited in page 6.)
- [17] Ishaan Gulrajani and David Lopez-Paz. 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 8.)
- [18] David J Hand. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine learning*, 77(1):103–123, 2009. (Cited in page 6.)
- [19] Sobhan Hemati, Guojun Zhang, Amir Hossein Estiri, and Xi Chen. Understanding Hessian Alignment for Domain Generalization. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 18958–18968. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01742. URL <https://doi.org/10.1109/ICCV51070.2023.01742>. (Cited in pages 8, 9, 18, 19, and 20.)
- [20] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset, 2021. URL <https://arxiv.org/abs/2103.03874>. (Cited in pages 7 and 17.)
- [21] Dong Huang, Qingwen Bu, Jie Zhang, Xiaofei Xie, Junjie Chen, and Heming Cui. 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.)
- [22] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *Computer vision—ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part II 16*, pages 124–140. Springer, 2020. (Cited in page 8.)
- [23] Amin Ibrahim, Azam Asilian Bidgoli, Shahryar Rahnamayan, and Kalyanmoy Deb. 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 6.)
- [24] Nathalie Japkowicz and Mohak Shah. *Evaluating learning algorithms: a classification perspective*. Cambridge University Press, 2011. (Cited in pages 1 and 6.)
- [25] Adrián Javaloy and Isabel Valera. 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 8 and 18.)
- [26] Adrián Javaloy, Maryam Meghdadi, and Isabel Valera. Mitigating Modality Collapse in Multimodal VAEs via Impartial Optimization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 9938–9964. PMLR, 2022. URL <https://proceedings.mlr.press/v162/javaloy22a.html>. (Cited in page 6.)
- [27] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D’Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badi Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and Open Problems in Federated Learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021. ISSN 1935-8237. doi: 10.1561/22000000083. URL <http://dx.doi.org/10.1561/22000000083>. (Cited in page 6.)
- [28] M. G. Kendall. 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 6.)

- [29] M. G. Kendall and R. M. Sundrum. 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.)
- [30] David Krueger, Ethan Caballero, Jörn-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Rémi Le Priol, and Aaron C. Courville. Out-of-Distribution Generalization via Risk Extrapolation (REx). In Marina Meila and Tong Zhang, editors, *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*, pages 5815–5826. PMLR, 2021. URL <http://proceedings.mlr.press/v139/krueger21a.html>. (Cited in page 19.)
- [31] Saku Kukkonen and Jouni Lampinen. Ranking-dominance and many-objective optimization. In *2007 IEEE Congress on Evolutionary Computation*, pages 3983–3990. IEEE, 2007. (Cited in page 6.)
- [32] Lu Li, Tianyu Zhang, Zhiqi Bu, Suyuchen Wang, Huan He, Jie Fu, Yonghui Wu, Jiang Bian, Yong Chen, and Yoshua Bengio. MAP: Low-compute model merging with amortized Pareto fronts via quadratic approximation. *ArXiv preprint*, abs/2406.07529, 2024. URL <https://arxiv.org/abs/2406.07529>. (Cited in page 6.)
- [33] Bo Liu, Yihao Feng, Peter Stone, and Qiang Liu. FAMO: Fast Adaptive Multitask Optimization. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *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 4 and 6.)
- [34] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 3730–3738. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.425. URL <https://doi.org/10.1109/ICCV.2015.425>. (Cited in pages 7 and 17.)
- [35] Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 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.)
- [36] Mark MacCarthy. Standards of fairness for disparate impact assessment of big data algorithms. *Cumb. L. Rev.*, 48:67, 2017. (Cited in page 8.)
- [37] Gaurav Maheshwari and Michaël Perrot. FairGrad: Fairness Aware Gradient Descent. *ArXiv preprint*, abs/2206.10923, 2022. URL <https://arxiv.org/abs/2206.10923>. (Cited in pages 7, 8, and 17.)
- [38] Kevis-Kokitsi Maninis, Ilija Radosavovic, and Iasonas Kokkinos. 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*, pages 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 4.)
- [39] Natalia Martínez, Martín Bertrán, and Guillermo Sapiro. 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*, pages 6755–6764. PMLR, 2020. URL <http://proceedings.mlr.press/v119/martinez20a.html>. (Cited in page 6.)
- [40] Kaisa Miettinen. *Nonlinear multiobjective optimization*, volume 12. Springer Science & Business Media, 1999. (Cited in pages 4 and 6.)
- [41] Jun Hyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from Failure: De-biasing Classifier from Biased Classifier. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *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 9.)

- [42] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *ArXiv preprint*, abs/2307.06435, 2023. URL <https://arxiv.org/abs/2307.06435>. (Cited in page 7.)
- [43] Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-Task Learning as a Bargaining Game. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 16428–16446. PMLR, 2022. URL <https://proceedings.mlr.press/v162/navon22a.html>. (Cited in pages 1 and 6.)
- [44] Alfredo Nazabal, Pablo M Olmos, Zoubin Ghahramani, and Isabel Valera. Handling incomplete heterogeneous data using VAEs. *Pattern Recognition*, 107:107501, 2020. (Cited in pages 2 and 6.)
- [45] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011. (Cited in page 8.)
- [46] Ji Won Park, Natasa Tagasovska, Michael Maser, Stephen Ra, and Kyunghyun Cho. BOTied: Multi-objective Bayesian optimization with tied multivariate ranks. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=cj5HbaX14p>. (Cited in page 6.)
- [47] Ashley Piggins. *Collective Choice and Social Welfare—Expanded Edition*, 2019. (Cited in page 6.)
- [48] Alexandre Ramé, Corentin Dancette, and Matthieu Cord. Fishr: Invariant Gradient Variances for Out-of-Distribution Generalization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 18347–18377. PMLR, 2022. URL <https://proceedings.mlr.press/v162/rame22a.html>. (Cited in pages 1 and 6.)
- [49] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A Graduate-Level Google-Proof Q&A Benchmark, 2023. URL <https://arxiv.org/abs/2311.12022>. (Cited in pages 7 and 17.)
- [50] Ozan Sener and Vladlen Koltun. Multi-Task Learning as Multi-Objective Optimization. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 525–536, 2018. URL <https://proceedings.neurips.cc/paper/2018/hash/432aca3a1e345e339f35a30c8f65edce-Abstract.html>. (Cited in page 8.)
- [51] M Sklar. Fonctions de répartition à n dimensions et leurs marges. In *Annales de l’ISUP*, volume 8, pages 229–231, 1959. (Cited in page 4.)
- [52] Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. MuSR: Testing the Limits of Chain-of-thought with Multistep Soft Reasoning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=jenyYQzue1>. (Cited in pages 7 and 17.)
- [53] Baochen Sun and Kate Saenko. 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*, pages 443–450. Springer, 2016. (Cited in page 19.)
- [54] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 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 7 and 17.)

- [55] Howard G Tucker. A generalization of the Glivenko-Cantelli theorem. *The Annals of Mathematical Statistics*, 30(3):828–830, 1959. (Cited in page 5.)
- [56] Sergio Verdu and H Poor. On minimax robustness: A general approach and applications. *IEEE transactions on Information Theory*, 30(2):328–340, 1984. (Cited in pages 4 and 6.)
- [57] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *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/63cb9921eecf51bfad27a99b2c53dd6d-Abstract-Datasets_and_Benchmarks.html. (Cited in pages 2, 6, 9, 23, and 24.)
- [58] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhui Chen. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/ad236edc564f3e3156e1b2feafb99a24-Abstract-Datasets_and_Benchmarks_Track.html. (Cited in pages 7 and 17.)
- [59] Yufei Wang, Haoliang Li, and Alex C. Kot. 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*, pages 3622–3626. IEEE, 2020. doi: 10.1109/ICASSP40776.2020.9053273. URL <https://doi.org/10.1109/ICASSP40776.2020.9053273>. (Cited in page 9.)
- [60] Frank Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6): 80–83, 1945. ISSN 00994987. URL <http://www.jstor.org/stable/3001968>. (Cited in page 6.)
- [61] Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, Fangyuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. Revisiting Out-of-distribution Robustness in NLP: Benchmarks, Analysis, and LLMs Evaluations. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *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-Dataset_s_and_Benchmarks.html. (Cited in page 1.)
- [62] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. In Aarti Singh and Xiaojin (Jerry) Zhu, editors, *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*, pages 962–970. PMLR, 2017. URL <http://proceedings.mlr.press/v54/zafar17a.html>. (Cited in pages 2 and 7.)
- [63] Milan Zeleny. Compromise programming. *Multiple criteria decision making*, 1973. (Cited in page 3.)
- [64] Yifan Zhong, Chengdong Ma, Xiaoyuan Zhang, Ziran Yang, Haojun Chen, Qingfu Zhang, Siyuan Qi, and Yaodong Yang. Panacea: Pareto Alignment via Preference Adaptation for LLMs. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/89f39d0b3d49a47606a165eefba2778c-Abstract-Conference.html. (Cited in page 6.)

710 [65] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny
711 Zhou, and Le Hou. Instruction-Following Evaluation for Large Language Models, 2023. URL
712 <https://arxiv.org/abs/2311.07911>. (Cited in pages 7 and 17.)

Appendix

Table of Contents

A Experimental details and additional results	16
A.1 Synthetic evaluation	16
A.2 Open LLM Leaderboard: Navigating the LLM performance-cost Pareto front . .	16
A.3 Navigating the fairness-accuracy trade-off	17
A.4 Comparative model analysis experiments	18
A.5 AutoML Benchmarking (AMLB) experiment	20
A.6 DecodingTrust: Navigating the LLM trustworthiness Pareto front	23

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:

$$y_2 = 0.25 \cos(39y_1^{0.85}) - \log(y_1) - 0.46, \quad (12)$$

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.

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 Open LLM Leaderboard: Navigating the LLM performance-cost Pareto front

Dataset details. In order to conduct our experiments, we retrieved the publicly available results from the Open LLM Leaderboard [12] using Huggingface’s dataset Python package and, for reproducibility purposes, saved a local copy with the state as of the 9th 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.

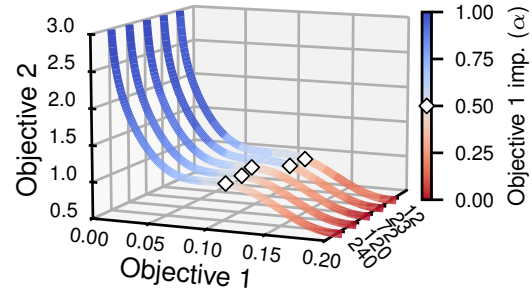


Figure 9: Synthetic Pareto front showing the Pareto front using COPA with $p = \infty$ as we change the number of sampled points. While it can be observed a deterioration on the estimated Pareto front (see quantized colors as we reduce N), COPA offers a robust estimator even with 12 datapoints.

Experimental details. As explained in the main text, we consider all reported values as objectives. Namely, we take as objectives the CO₂ emissions and all 6 benchmark performance scores computed on the following datasets: IFEval [65], BBH [54], MATH [20], GPQA [49], MuSR [52], and MMLU-Pro [58]. Then, we use COPA with $p = \infty$ to produce both Figs. 1 and 4, setting the values of ω according to the importance given to CO₂ emissions, α , as $\omega := [\alpha, \frac{1-\alpha}{6}, \dots, \frac{1-\alpha}{6}]$. To create these figures, we take 10 000 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 CO₂ 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 CO₂ 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 CO₂, 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 CO₂ emissions.

Table 1: Quantitative results of the LLMs highlighted in Fig. 4 from the Open LLM Leaderboard [12] using COPA with $p = \infty$, as we change the importance of CO₂ 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 CO₂ from the rest of objectives).

Full model name	Benchmarks scores						Summary		CDF values		
	IFEval (%)	BBH (%)	MATH (%)	GPQA (%)	MUSR (%)	MMLU-PRO (%)	Average (%)	CO ₂ cost (kg)	Perf. ($p = \infty$)	CO ₂ cost	α
<code>dfurman/CalmeRys-78B-Orpo-v0.1</code>	81.63	61.92	40.71	20.02	36.37	66.80	51.24	13.00	0.00	0.95	0.01
<code>maldy/Qwen2.5-32B-Instruct</code>	73.93	57.21	38.07	17.90	19.96	54.21	43.55	3.53	0.01	0.87	0.02
<code>sometimesanot/Qwen2.5-14B-Vimarcoso-v3</code>	72.57	48.58	34.44	17.34	19.39	48.26	40.10	1.93	0.01	0.79	0.03
<code>hotmailuser/FalconSlerp3-7B</code>	60.96	36.83	27.42	9.17	15.90	34.75	30.84	0.61	0.05	0.19	0.21
<code>unsloth/Phi-3-mini-4k-instruct</code>	54.40	36.73	15.41	9.73	13.12	33.68	27.18	0.47	0.08	0.09	0.50
<code>icefog72/Ice0.37-18.11-RP</code>	49.72	31.04	6.42	8.28	12.21	23.81	21.91	0.41	0.21	0.07	0.66
<code>h2oai/h2o-danube3.1-4b-chat</code>	50.21	10.94	2.11	4.70	10.20	19.10	16.21	0.30	0.60	0.03	0.82
<code>postbot/gpt2-medium-emailgen</code>	14.92	3.67	0.00	1.34	6.89	1.63	4.74	0.08	0.86	0.00	0.97

Differences in p -norms. To show the differences between using as criterion function the usual p -norm or the one proposed in this paper (Eq. 11), we have reproduced Fig. 1 from the manuscript but using the usual p -norm instead (still, using CDF-transformed objectives, so that semantic comparability is not an issue). We can observe the results in Fig. 10. Note that we do not use $p = \infty$ as in the original figure since, for the usual weighted p -norm, $\|\mathbf{u}\|_{\infty, \omega} = \|\mathbf{u}\|_{\infty}$, while for the proposed norm it corresponds to the weighted Tchebycheff problem, as discussed in §3.3.

Using piece-wise criterion functions. In the main pages, we exclusively consider criterion functions as weighted norms, as proposed in Eq. 11. To showcase that this is not a real restriction—in fact, we can use any sensible criterion function which we can evaluate—we show in Fig. 11 the same experiment as in Fig. 1, where we have replaced the criterion function such that it depends on the CO₂ footprint of the model we are evaluating. As we can see, it is still crucial to have semantically comparable objectives, and we can use any criterion function as long as it is sensible for the DM.

A.3 Navigating the fairness-accuracy trade-off

Experimental details. We reproduce the CelebA [34] experiment from [37] 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:

$$\epsilon \in \{0.001, 0.002, 0.003, 0.004, 0.005, 0.006, 0.007, 0.008, 0.009,$$

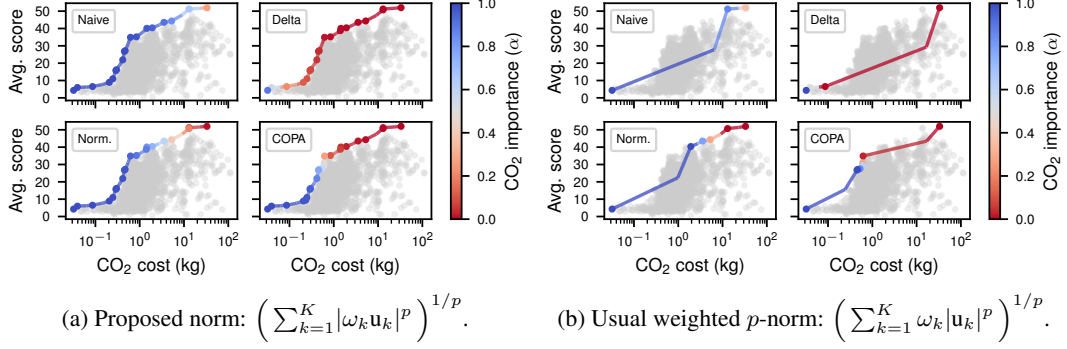


Figure 10: Reproduction of Fig. 1 using $p = 8$ and taking 10 000 evenly-spaced values for α . As discussed when introducing Eq. 11, the usual weighted p -norm is not well-suited for our purposes, as CDF-transformed objective lie in the unit interval and quickly vanish.

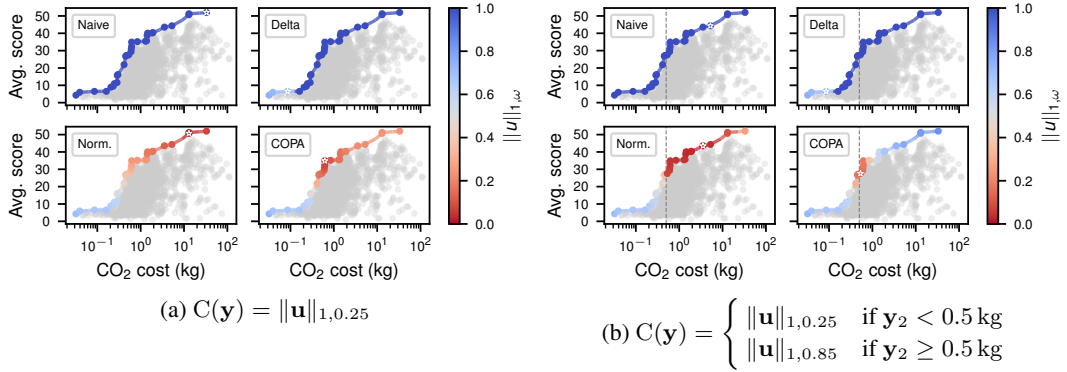


Figure 11: Reproduction of Fig. 1 where now colors represent the norm value for each model in the Pareto front (i.e., the value of the criterion function), and we look for the point with minimum norm (marked with a star): (a) $\alpha = 1/4$, (b) piece-wise criterion function, conditional on the CO₂ footprint. In both cases the importance of having comparable objectives is clear.

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.

808 A.3.1 Additional results

We show in Fig. 12 the same plot as in Fig. 5, 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.

812 A.4 Comparative model analysis experiments

Experimental details. For the figures shown in §5.3, we retrieved the results reported by two selected works. In particular, we took the values reported in the second half of Table 5 from the work of Javaloy and Valera [25] for the MTL experiment, and values reported in Table 4 of Hemati et al. [19] 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. To ease visualization, as ties are more frequent in $p = 1$ and $p = \infty$ —especially when we have only a handful

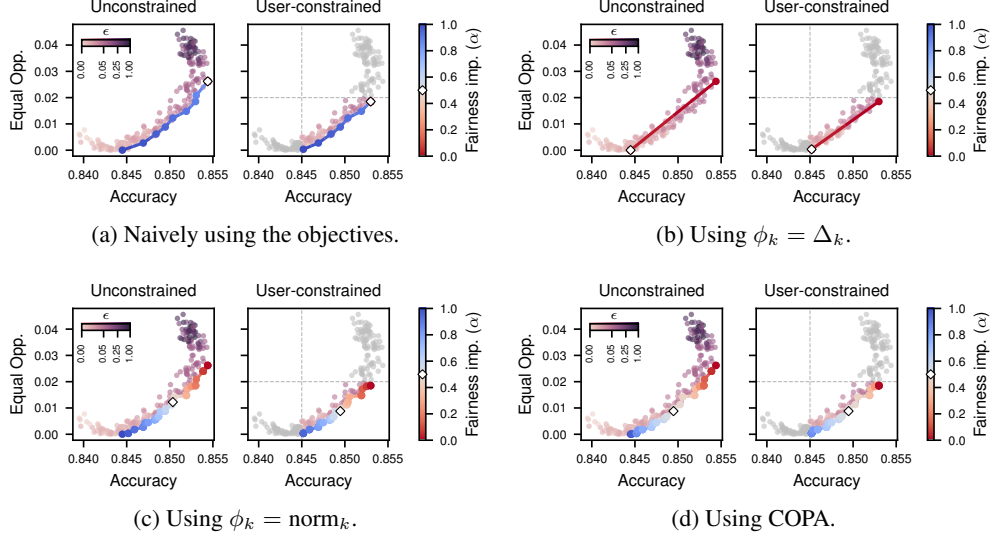


Figure 12: We reproduce the fair ML experiment from §5.2 using different normalization functions. We can observe that only COPA meaningfully navigates the Pareto front, with all other approaches being biased towards one of the extreme solutions. Indeed, Δ_k only reaches the two extreme solution despite sampling 50 evenly-spaced values for α .

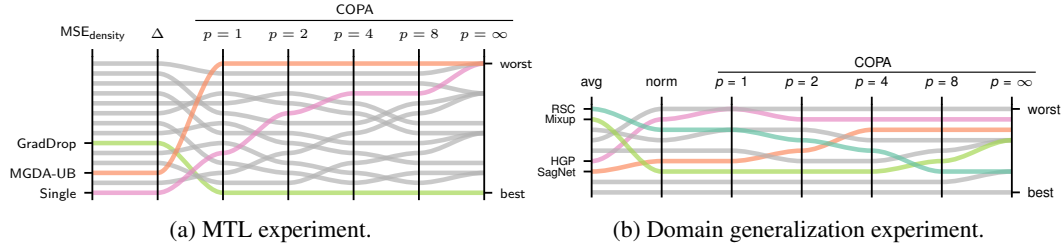


Figure 13: Reproductions of Figs. 6 and 7 where we do not untie methods. As it can be observed, conclusions drawn in the main paper do not change and untieing only serves aesthetic purposes here.

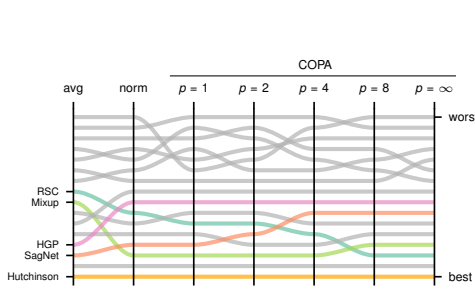
of models—we untie by using the ranking of $p = 8$ as a secondary criterion. That is, if two models tie, we rank those by their performance with $p = 8$. We plot the figures without untieing in Fig. 13 for the sake of transparency, showing that conclusions do not change. We use equal weights for all 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

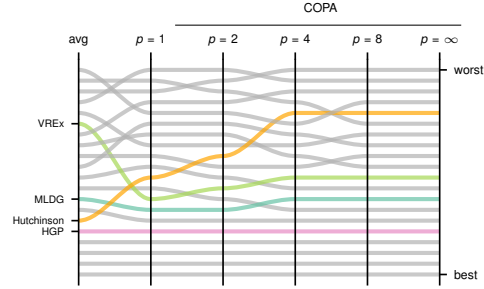
As mentioned just above, we discarded some methods in the domain generalization figure of the main text (i.e., Fig. 7). For completeness, we show in Fig. 14a the full figure with all methods included, and highlighting Hutchinson, the second method proposed by the authors, along HGP. Also, we show in Fig. 14b the same figure but using as data the one reported in Table 9 from Hemati et al. [19] (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. 14a, and the fourth to worst method in Fig. 14b. We can again observe how much our final conclusions can change in Fig. 14b, where the fourth to worst method in terms of average domain accuracy, VREx [30], 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 [53].

Table 2: 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 (norm_k , Eq. 6) 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



(a) Table 4 from Hemati et al. [19].



(b) Table 9 from Hemati et al. [19].

Figure 14: Ranking of the domain generalization methods considered by Hemati et al. [19] 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 Gijbbers et al. [15]. To achieve this, we simply modify the Jupyter notebook publicly available at github.com/PGijbbers/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 Gijbbers et al. [15], further reinforcing the argument of broadly adopting CDF-transformed objectives for general cases.

In particular, we show in Fig. 15 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. 16 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. 17, 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 §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

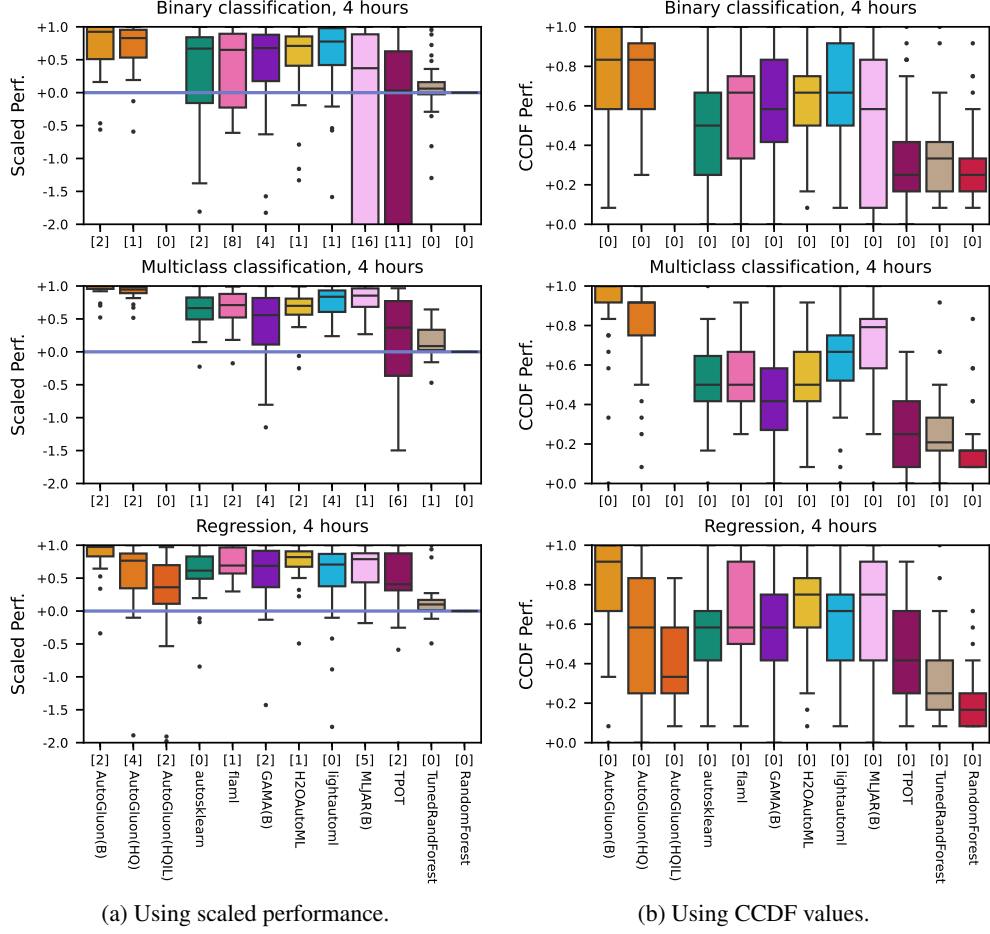


Figure 15: We reproduce Fig. 3 from Gijssbers et al. [15] in (a) using their proposed scaled performance, and we show the same figure in (b) but using complementary CDF values (CCDF, one minus the CDF value). The same advantages as those discussed in §5.4 can be observed here.

866 work (e.g., Figs. 1, 4 and 5), which are significantly more informative than those reported before our
 867 work, as it can be clearly observed in Fig. 17.

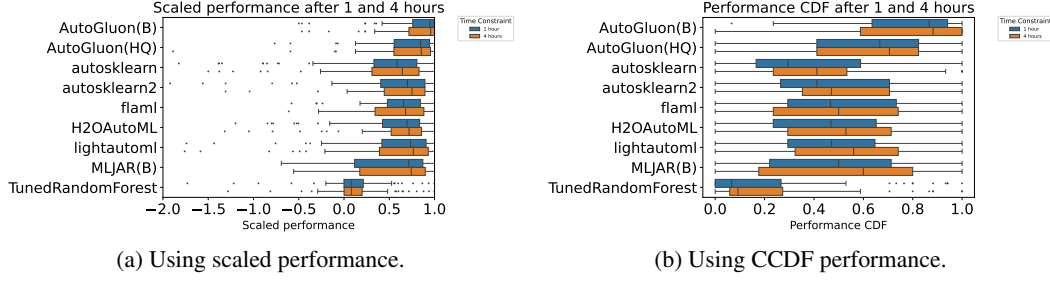


Figure 16: We reproduce Fig. 4 from Gijsbers et al. [15] in (a) using their proposed scaled performance, and we show the same figure in (b) but using complementary CDF values (CCDF, one minus the CDF value). The same advantages as those discussed in §5.4 can be observed here.

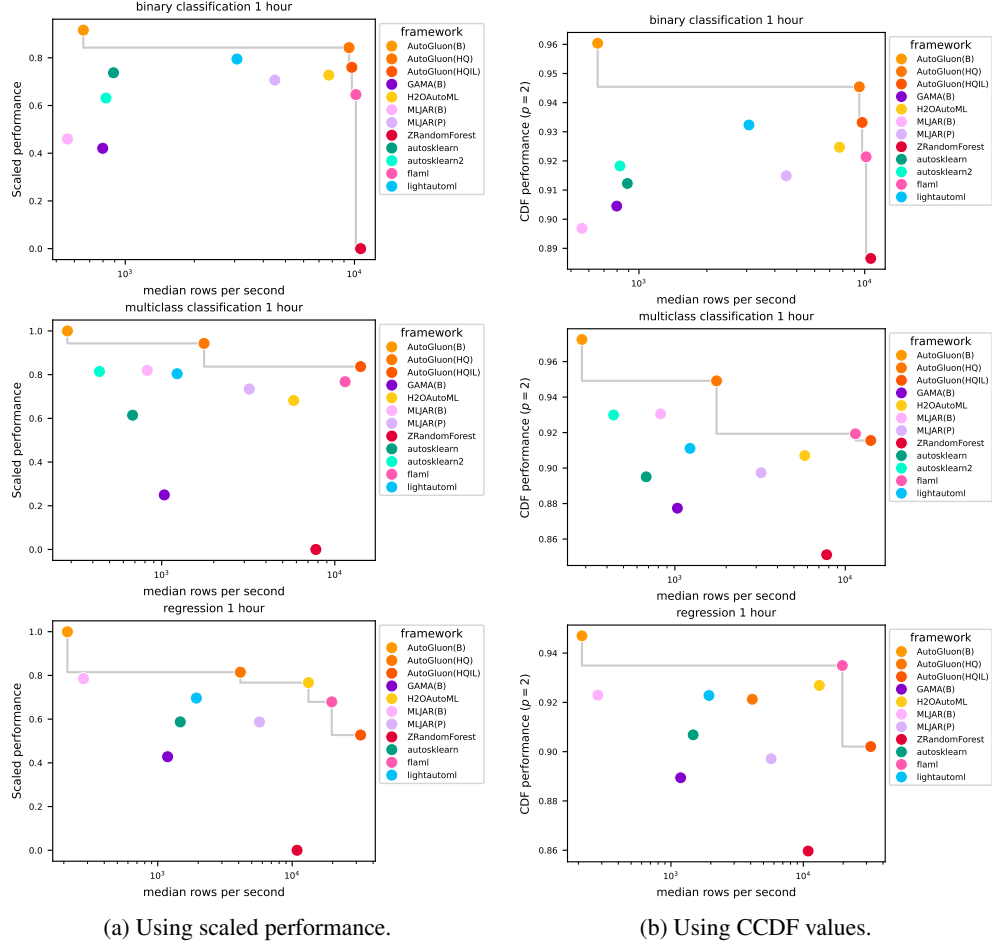


Figure 17: We reproduce Fig. 7 from Gijsbers et al. [15] in (a) using their proposed scaled performance, and we show the same figure in (b) but using complementary CDF values (CCDF, one minus the CDF value).

868 A.6 DecodingTrust: Navigating the LLM trustworthiness Pareto front

869 **Dataset details.** We look once more into the LLM space and, this time, we focus on the Decod-
 870 ingTrust leaderboard [57]. In contrast with the Open LLM leaderboard [12], DecodingTrust focuses
 871 in assessing the trustworthiness of LLM models, rather than on sheer performance. To this end, the
 872 authors design a comprehensive list of prompts that the model should successfully answer to. For the
 873 interest of this work, it suffices to say that DecodingTrust evaluates LLMs on 8 completely different
 874 aspects and the authors had to come up with different ad hoc normalization functions for each of
 875 the objectives so that they all lie in the $[0, 100]$ interval. Below there is a summary of the formulas
 876 employed by the authors and, while we do not give context for the variables shown in the equations,
 877 we want to stress the diversity of scores and normalizations that the authors had to propose to make
 878 objectives more comparable. For further details on the normalization functions and, more in general,
 879 DecodingTrust, refer to [57, App. I]:

880 • *Toxicity*: $1 - \frac{1}{2 \sum_i |D_i|} \sum_{i=1}^4 \sum_{x \in D_i} (f(x_{\text{adv}}^*; x) + f(x_{\text{benign}}^*; x))$.

• *Stereotype bias*:

$$\frac{S_{\text{benign}} + S_{\text{untargeted}} + S_{\text{targeted}}}{3} \quad \text{where} \quad S_{\text{scenario}} = 1 - \left(\sum_{i=1}^{n_{\text{ST}}} \sum_{j=1}^{n_{\text{DG}}} S_{ij} \right) / (n_{\text{ST}} n_{\text{DG}}).$$

881 • *Adversarial robustness*: $\frac{\sum_{i=1}^T \text{acc}_i * d_i}{\sum_{i=1}^T d_i}$.

882 • *Out-of-distribution robustness*: $(\text{ACC}_{\text{style}} + \text{Reliability}_{\text{OOD}} + \text{ACC}_{\text{style}}^{\text{icl}} + \text{ACC}_{\text{domain}}^{\text{icl}}) / 4$ where:

$$\text{ACC}_{\text{style}} = \frac{1}{S} \sum_{s=1}^S \text{acc}_s,$$

$$\text{Reliability}_{\text{OOD}} = \frac{\text{Reliability}_{2023} + \text{Reliability}_{2023\text{idk}}}{2}$$

$$\text{Reliability}_{\text{setting}} = \text{RR}_{\text{setting}} + (1 - \text{RR}_{\text{setting}}) * \text{macc}_{\text{setting}},$$

$$\text{ACC}_{\text{setting}}^{\text{icl}} = \frac{1}{D * N} \sum_{d=1}^D \sum_{i=1}^N \text{acc}_{di}^{\text{setting}}.$$

883 • *Robustness to adversarial demonstrations*: $(s^{(\text{cf})} + s^{(\text{sc})} + s^{(\text{bkd})}) / 3$ where:

$$s^{(\text{cf})} = \frac{1}{|D^{(\text{cf})}|} \sum_{i \in D^{(\text{cf})}} \text{acc}_i^{(\text{Demo+CF})},$$

$$s^{(\text{sc})} = \frac{1}{|D^{(\text{sc})}|} \sum_{i \in D^{(\text{sc})}} \frac{\text{acc}_i^{(\text{entail})} + \text{acc}_i^{(\text{non-entail})}}{2},$$

$$s^{(\text{bkd})} = 1 - \frac{1}{|M||B|} \sum_{i \in B} \sum_{j \in M} \text{ASR}_{ij}.$$

884 • *Privacy*: $1 - (0.4\text{LR}^{(\text{Enron})} + 0.3\text{LR}^{(\text{PII})} + 0.3\text{LR}^{(\text{Understand})})$ where

$$\text{LR}^{(\text{Enron})} = \frac{1}{T} \sum_{t=1}^T \frac{\text{LR}_t^{(\text{Email})} + \text{LR}_t^{(\text{Local})} + \text{LR}_t^{(\text{Domain})}}{3},$$

$$\text{LR}^{(\text{PII})} = \frac{1}{P} \sum_{p=1}^P \overline{\text{LR}}^p,$$

$$\text{LR}^{(\text{Understand})} = \frac{1}{WE} \sum_{w=1}^W \sum_{e=1}^E \overline{\text{LR}}_{we}.$$

- 885 • *Machine ethics*: $(\text{ACC}^{\text{zero}} + \text{ACC}^{\text{few}} + (1 - \overline{\text{FPR}}^{\text{jailbreak}}) + (1 - \overline{\text{FPR}}^{\text{evasive}}))/4$.
- 886 • *Fairness*: $100 \left(1 - \frac{M_{\text{dpd}}^{(\text{zero})} + M_{\text{dpd}}^{(\text{few-unfair})} + M_{\text{dpd}}^{(\text{few-fair})}}{3} \right)$.

887 **Additional results.** After computing all the
888 variables above, we end up with 8 objectives,
889 which are aggregated in the their official lead-
890 erboard by taking the average of all objectives
891 (i.e., using $p = 1$ in Eq. 11 since all values are
892 non-negative). One natural question then is how
893 do the rankings obtained with the average score
894 change if we plug in COPA on the objectives
895 given above. To this end, we take the results pub-
896 lished in the official site of DecodingTrust and
897 recompute their rankings using as aggregated
898 scores COPA with $p \in \{1, 2, 4, 8, \infty\}$. While
899 this can be easily done in the full leaderboard
900 with 55 models to this date (and it is indeed done
901 this way in the code accompanying this manu-
902 script), we show in Fig. 18 a bar plot showing
903 the differences in ranking of a small subset of
904 representative models w.r.t. average ranking. We
905 summarize the main takeaways as follows:

- 906 1. *Differences in normalization*: Even with
907 $p = 1$, we observe significant differences
908 in COPA w.r.t. the original ranking where,
909 e.g., Vicuna and MPT rank around 10 positions
910 better and the two GPT models close to 10 worse.
- 911 2. *Robustness with p* : As discussed in the manuscript, increasing (resp. decreasing) p can be
912 interpreted as adding (removing) importance to the performance of the models on individual
913 objectives. We observe a similar trend here, where those models that lack in one of the specific
914 objectives (e.g., GPT-4 which performs the worst in the Fairness objective) start losing ranks as
915 we increase p , and those which are more robust are rewarded instead (e.g., Vicuna and Tulu-2-7b).
- 916 3. *Robustness trend*. Similar to those experiments from §5.3, we observe a clear correlation between
917 the rankings and the values of p , that is, the aforementioned robustness criterion is stressed as we
918 increase p . While this is not always the case, e.g., Tulu-2-13b fluctuates by one or two rankings
919 as we change p , the trend is rather clear.
- 920 4. *Quasi-dominant models*. While there is no a dominant model (i.e., one which is better for all
921 objectives), some models like Claude and Gemini-Pro-1.0 rank top-10 for all but, respectively,
922 one and two objectives. As a result, we see that they consistently rank first and second for almost
923 all values of p , just like if we took the average. This is, in part, a result of the CDF estimator
924 having less variance for more extreme samples (see Prop. 3.1). In layman’s terms, a model that
925 does almost everything great is easy to spot.
- 926 5. *Diversity of solutions*. While not shown in Fig. 18, it is interesting to remark that in this setting,
927 with 8 diverse objectives, we find that out of the 55 available models, the best-worst performing
928 model (that is, the most robust model according to COPA with $p = \infty$) achieves a worst-ranking
929 of 22, with the two second-best LLMs obtaining a worst-ranking of 40. That is, only with 8
930 objectives we already find that any model performs relatively bad in at least one of them.

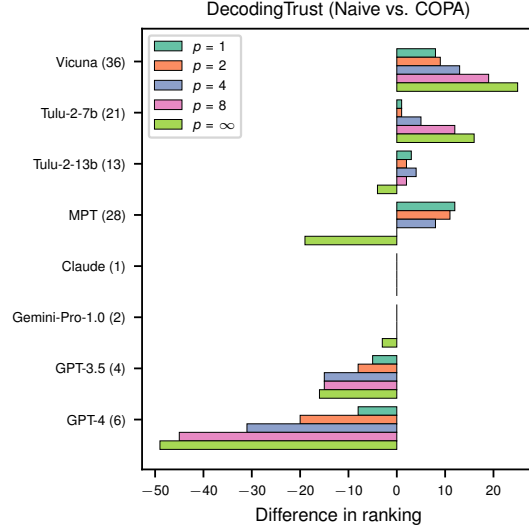


Figure 18: Difference in ranking for a subset of models on the DecodingTrust leaderboard [57] for different p values of COPA, where the number within parenthesis indicate their original ranking.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [\[Yes\]](#)

Justification: The claims made in the abstract and introduction (the need of normalizing our objectives and of having a meaningful criterion function to map the user preferences) are appropriately reflected and demonstrated through all experiments and discussions in the main manuscript.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We do not dedicate a specific section for limitations due to COPA simplicity and space constraints. However, we make really clear throughout the paper what are our main assumptions and thus limitations. Mainly, we assume the existence of a population of models \mathcal{H} with continuous random variables. If these do not hold, COPA cannot be directly applied.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate “Limitations” section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren’t acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

982 Question: For each theoretical result, does the paper provide the full set of assumptions and a
983 complete (and correct) proof?

984 Answer: [Yes]

985 Justification: The main theoretical claims have to do with the properties of the rank estimator,
986 which we fully proof in Prop. 3.1. While we point to other results without writing down their
987 proofs, e.g. the properties of the proposed norm Eq. 11, the demonstration require minimal
988 changes from those proofs we point to in the main paper.

989 Guidelines:

- 990 • The answer NA means that the paper does not include theoretical results.
- 991 • All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- 992 • All assumptions should be clearly stated or referenced in the statement of any theorems.
- 993 • The proofs can either appear in the main paper or the supplemental material, but if they appear
994 in the supplemental material, the authors are encouraged to provide a short proof sketch to
995 provide intuition.
- 996 • Inversely, any informal proof provided in the core of the paper should be complemented by
997 formal proofs provided in appendix or supplemental material.
- 998 • Theorems and Lemmas that the proof relies upon should be properly referenced.

999 4. Experimental result reproducibility

1000 Question: Does the paper fully disclose all the information needed to reproduce the main
1001 experimental results of the paper to the extent that it affects the main claims and/or conclusions
1002 of the paper (regardless of whether the code and data are provided or not)?

1003 Answer: [Yes]

1004 Justification: Yes, all details are fully specified in the main paper and in §A.

1005 Guidelines:

- 1006 • The answer NA means that the paper does not include experiments.
- 1007 • If the paper includes experiments, a No answer to this question will not be perceived well by
1008 the reviewers: Making the paper reproducible is important, regardless of whether the code
1009 and data are provided or not.
- 1010 • If the contribution is a dataset and/or model, the authors should describe the steps taken to
1011 make their results reproducible or verifiable.
- 1012 • Depending on the contribution, reproducibility can be accomplished in various ways. For
1013 example, if the contribution is a novel architecture, describing the architecture fully might
1014 suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary
1015 to either make it possible for others to replicate the model with the same dataset, or provide
1016 access to the model. In general, releasing code and data is often one good way to accomplish
1017 this, but reproducibility can also be provided via detailed instructions for how to replicate the
1018 results, access to a hosted model (e.g., in the case of a large language model), releasing of a
1019 model checkpoint, or other means that are appropriate to the research performed.
- 1020 • While NeurIPS does not require releasing code, the conference does require all submissions
1021 to provide some reasonable avenue for reproducibility, which may depend on the nature of
1022 the contribution. For example
 - 1023 (a) If the contribution is primarily a new algorithm, the paper should make it clear how to
1024 reproduce that algorithm.
 - 1025 (b) If the contribution is primarily a new model architecture, the paper should describe the
1026 architecture clearly and fully.
 - 1027 (c) If the contribution is a new model (e.g., a large language model), then there should either
1028 be a way to access this model for reproducing the results or a way to reproduce the model
1029 (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - 1030 (d) We recognize that reproducibility may be tricky in some cases, in which case authors are
1031 welcome to describe the particular way they provide for reproducibility. In the case of
1032 closed-source models, it may be that access to the model is limited in some way (e.g.,

1033 to registered users), but it should be possible for other researchers to have some path to
1034 reproducing or verifying the results.

1035 5. Open access to data and code

1036 Question: Does the paper provide open access to the data and code, with sufficient instructions to
1037 faithfully reproduce the main experimental results, as described in supplemental material?

1038 Answer: [Yes]

1039 Justification: All experiments use publicly-available data (except for the fairness use-case data,
1040 which is attached), and the code to reproduce the experiments are attached too.

1041 Guidelines:

- 1042 • The answer NA means that paper does not include experiments requiring code.
- 1043 • Please see the NeurIPS code and data submission guidelines ([https://nips.cc/public](https://nips.cc/public/guides/CodeSubmissionPolicy)
1044 [/guides/CodeSubmissionPolicy](https://nips.cc/public/guides/CodeSubmissionPolicy)) for more details.
- 1045 • While we encourage the release of code and data, we understand that this might not be
1046 possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including
1047 code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- 1048 • The instructions should contain the exact command and environment needed to run to
1049 reproduce the results. See the NeurIPS code and data submission guidelines ([https:](https://nips.cc/public/guides/CodeSubmissionPolicy)
1050 [//nips.cc/public/guides/CodeSubmissionPolicy](https://nips.cc/public/guides/CodeSubmissionPolicy)) for more details.
- 1051 • The authors should provide instructions on data access and preparation, including how to
1052 access the raw data, preprocessed data, intermediate data, and generated data, etc.
- 1053 • The authors should provide scripts to reproduce all experimental results for the new proposed
1054 method and baselines. If only a subset of experiments are reproducible, they should state
1055 which ones are omitted from the script and why.
- 1056 • At submission time, to preserve anonymity, the authors should release anonymized versions
1057 (if applicable).
- 1058 • Providing as much information as possible in supplemental material (appended to the paper)
1059 is recommended, but including URLs to data and code is permitted.

1060 6. Experimental setting/details

1061 Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters,
1062 how they were chosen, type of optimizer, etc.) necessary to understand the results?

1063 Answer: [Yes]

1064 Justification: Yes, all details are provided in §A.

1065 Guidelines:

- 1066 • The answer NA means that the paper does not include experiments.
- 1067 • The experimental setting should be presented in the core of the paper to a level of detail that
1068 is necessary to appreciate the results and make sense of them.
- 1069 • The full details can be provided either with the code, in appendix, or as supplemental material.

1070 7. Experiment statistical significance

1071 Question: Does the paper report error bars suitably and correctly defined or other appropriate
1072 information about the statistical significance of the experiments?

1073 Answer: [Yes]

1074 Justification: While we assume no stochasticity in the classical ML sense (i.e., via the datasets),
1075 we do so through \mathcal{H} and clearly describe the properties of the ranking estimator in Prop. 3.1.

1076 Guidelines:

- 1077 • The answer NA means that the paper does not include experiments.
- 1078 • The authors should answer “Yes” if the results are accompanied by error bars, confidence
1079 intervals, or statistical significance tests, at least for the experiments that support the main
1080 claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: Experiments are extremely lightweight and can be run in any modern device.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: Data does not involve human participants and is publicly-available.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: While we do dedicate a specific paragraph, the entire discussion within the manuscript concerns the need of properly mapping the preferences of users into the Pareto front and their possible misuses (e.g., by not normalizing objectives), and we thus believe that the impacts of our work are clear.

Guidelines:

- 1131 • The answer NA means that there is no societal impact of the work performed.
- 1132 • If the authors answer NA or No, they should explain why their work has no societal impact or
- 1133 why the paper does not address societal impact.
- 1134 • Examples of negative societal impacts include potential malicious or unintended uses (e.g.,
- 1135 disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deploy-
- 1136 ment of technologies that could make decisions that unfairly impact specific groups), privacy
- 1137 considerations, and security considerations.
- 1138 • The conference expects that many papers will be foundational research and not tied to
- 1139 particular applications, let alone deployments. However, if there is a direct path to any
- 1140 negative applications, the authors should point it out. For example, it is legitimate to point out
- 1141 that an improvement in the quality of generative models could be used to generate deepfakes
- 1142 for disinformation. On the other hand, it is not needed to point out that a generic algorithm
- 1143 for optimizing neural networks could enable people to train models that generate Deepfakes
- 1144 faster.
- 1145 • The authors should consider possible harms that could arise when the technology is being
- 1146 used as intended and functioning correctly, harms that could arise when the technology is
- 1147 being used as intended but gives incorrect results, and harms following from (intentional or
- 1148 unintentional) misuse of the technology.
- 1149 • If there are negative societal impacts, the authors could also discuss possible mitigation
- 1150 strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms
- 1151 for monitoring misuse, mechanisms to monitor how a system learns from feedback over time,
- 1152 improving the efficiency and accessibility of ML).

1133 11. Safeguards

1154 Question: Does the paper describe safeguards that have been put in place for responsible release
1155 of data or models that have a high risk for misuse (e.g., pretrained language models, image
1156 generators, or scraped datasets)?

1157 Answer: [NA]

1158 Justification: We do not provide any of the above.

1159 Guidelines:

- 1160 • The answer NA means that the paper poses no such risks.
- 1161 • Released models that have a high risk for misuse or dual-use should be released with necessary
- 1162 safeguards to allow for controlled use of the model, for example by requiring that users adhere
- 1163 to usage guidelines or restrictions to access the model or implementing safety filters.
- 1164 • Datasets that have been scraped from the Internet could pose safety risks. The authors should
- 1165 describe how they avoided releasing unsafe images.
- 1166 • We recognize that providing effective safeguards is challenging, and many papers do not
- 1167 require this, but we encourage authors to take this into account and make a best faith effort.

1168 12. Licenses for existing assets

1169 Question: Are the creators or original owners of assets (e.g., code, data, models), used in the
1170 paper, properly credited and are the license and terms of use explicitly mentioned and properly
1171 respected?

1172 Answer: [Yes]

1173 Justification: We properly cite and point to every method and data we use.

1174 Guidelines:

- 1175 • The answer NA means that the paper does not use existing assets.
- 1176 • The authors should cite the original paper that produced the code package or dataset.
- 1177 • The authors should state which version of the asset is used and, if possible, include a URL.
- 1178 • The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- 1179 • For scraped data from a particular source (e.g., website), the copyright and terms of service of
- 1180 that source should be provided.

1181 • If assets are released, the license, copyright information, and terms of use in the package
 1182 should be provided. For popular datasets, paperswithcode.com/datasets has curated
 1183 licenses for some datasets. Their licensing guide can help determine the license of a dataset.
 1184 • For existing datasets that are re-packaged, both the original license and the license of the
 1185 derived asset (if it has changed) should be provided.
 1186 • If this information is not available online, the authors are encouraged to reach out to the
 1187 asset’s creators.

1188 **13. New assets**

1189 Question: Are new assets introduced in the paper well documented and is the documentation
 1190 provided alongside the assets?

1191 Answer: [NA]

1192 Justification: No new assets.

1193 Guidelines:

1194 • The answer NA means that the paper does not release new assets.
 1195 • Researchers should communicate the details of the dataset/code/model as part of their sub-
 1196 missions via structured templates. This includes details about training, license, limitations,
 1197 etc.
 1198 • The paper should discuss whether and how consent was obtained from people whose asset is
 1199 used.
 1200 • At submission time, remember to anonymize your assets (if applicable). You can either create
 1201 an anonymized URL or include an anonymized zip file.

1202 **14. Crowdsourcing and research with human subjects**

1203 Question: For crowdsourcing experiments and research with human subjects, does the paper
 1204 include the full text of instructions given to participants and screenshots, if applicable, as well as
 1205 details about compensation (if any)?

1206 Answer: [NA]

1207 Justification: Not applicable.

1208 Guidelines:

1209 • The answer NA means that the paper does not involve crowdsourcing nor research with human
 1210 subjects.
 1211 • Including this information in the supplemental material is fine, but if the main contribution of
 1212 the paper involves human subjects, then as much detail as possible should be included in the
 1213 main paper.
 1214 • According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or
 1215 other labor should be paid at least the minimum wage in the country of the data collector.

1216 **15. Institutional review board (IRB) approvals or equivalent for research with human subjects**

1217 Question: Does the paper describe potential risks incurred by study participants, whether such
 1218 risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals
 1219 (or an equivalent approval/review based on the requirements of your country or institution) were
 1220 obtained?

1221 Answer: [NA]

1222 Justification: Not applicable.

1223 Guidelines:

1224 • The answer NA means that the paper does not involve crowdsourcing nor research with human
 1225 subjects.
 1226 • Depending on the country in which research is conducted, IRB approval (or equivalent) may
 1227 be required for any human subjects research. If you obtained IRB approval, you should clearly
 1228 state this in the paper.
 1229 • We recognize that the procedures for this may vary significantly between institutions and
 1230 locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines
 1231 for their institution.

1232 • For initial submissions, do not include any information that would break anonymity (if
1233 applicable), such as the institution conducting the review.

1234 **16. Declaration of LLM usage**

1235 Question: Does the paper describe the usage of LLMs if it is an important, original, or non-
1236 standard component of the core methods in this research? Note that if the LLM is used only for
1237 writing, editing, or formatting purposes and does not impact the core methodology, scientific
1238 rigorousness, or originality of the research, declaration is not required.

1239 Answer: [NA]

1240 Justification: We do not use any LLM.

1241 Guidelines:

1242 • The answer NA means that the core method development in this research does not involve
1243 LLMs as any important, original, or non-standard components.

1244 • Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for
1245 what should or should not be described.