Generalizability of experimental studies

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Experimental studies are a cornerstone of machine learning (ML) research. A common, but often implicit, assumption is that the results of a study will generalize beyond the study itself, e.g. to new data. That is, there is a high probability that repeating the study under different conditions will yield similar results. Despite the importance of the concept, the problem of measuring generalizability remains open. This is probably due to the lack of a mathematical formalization of experimental studies. In this paper, we propose such a formalization and develop a quantifiable notion of generalizability. This notion allows to explore the generalizability of existing studies and to estimate the number of experiments needed to achieve the generalizability of new studies. To demonstrate its usefulness, we apply it to two recently published benchmarks to discern generalizable and non-generalizable results. We also publish a Python module that allows our analysis to be repeated for other experimental studies.

1 Introduction

Due to the importance of experimental studies, the machine learning (ML) community advocates for high methodological standards [20, 12, 13, 17, 8, 31, 32, 44]. Failure to meet these standards can have significant consequences, such as the ongoing reproducibility crisis [6, 47, 50, 51, 30].

Reproducibility is not the only desirable property of a study. For example, the reader expects that the best encoders of categorical features identified in [41] will not only remain the best when the study is reproduced, but will also outperform their competitors on new datasets. This property of getting the same results from different data is known as *replicability* [46, 48]. Replicability is a special case of *generalizability*, the property of obtaining the same results with any change in the inputs. The assumption of generalizability is arguably the main motivation for extensive experimental studies and benchmarks. However, existing definitions of generalizability do not *quantify* how well the results of a study can be transferred to other contexts. This hinders the usefulness of such studies and leads to confusion. For example, articles [38, 41, 49, 42] and [19, 8, 11, 13, 29, 43] report that the results of experimental studies are often contradictory.

Quantifying generalizability can also help determine the appropriate size of experimental studies. For example, one dataset is unlikely to be sufficient to draw far-reaching conclusions, but 10^6 datasets are likely enough. Of course, such large studies are usually not practical: it is crucial to determine the minimum amount of data needed to achieve generalizability. This principle also applies to decisions other than the number of datasets, such as the choice of quality metric and the initialization seed.

A notion similar to generalizability is model replicability [1, 22, 23, 24, 33, 36, 37]. A model is ρ-replicable if, given i.i.d. samples from the same data distribution, the trained models are the same with probability $1 - \rho$ [33]. Adapting this definition to quantify generalizability is not trivial, as it requires formalizing experimental studies. The latter must take into account several aspects: the research question, the results of a study, and how to compare the results. Regarding the problem of defining the size of experimental studies, the current literature addresses the (crucial, but orthogonal) problem of choosing appropriate experimental factors [20, 12, 13, 17, 8, 31, 32, 44]. While these studies recommend varying the factors, they do not help decide how many of the factor levels are enough.

Our contributions are as follows:

1. we formalize experimental studies and their results;
2. we propose a quantifiable definition of the generalizability of experimental studies;
3. we develop an algorithm to estimate the size of a study to obtain generalizable results;
4. we consider two recent experimental studies on categorical encoders [41] and Large Language Models [55] and show how their results may or may not be generalizable.
5. we will publish the GENEXPY[1] Python module to repeat our analysis in other studies.

Paper outline: Section 2 is related work, Section 3 formalizes experimental studies, Section 4 defines generalizability and provides the algorithm to estimate the required size of a study for generalizability, Section 5 contains the case studies, Section 6 describes limitations and concludes.

2 Related work

We first discuss the literature related to the motivation we are tackling, i.e., why experimental studies may not generalize. Second, we overview the existing concept of model replicability, closely related to our work. Finally, we show other meanings that these words can assume in other domains.

Non-generalizable results. It is well known that experimental results can significantly vary based on design choices [38, 41, 49, 42]. Possible reasons include an insufficient number of datasets [19, 41, 3, 12] as well as differences in hyperparameter tuning [13, 41], initialization seed [30], and hardware [56]. As a result, the statistical benchmarking literature advocates for experimenters to motivate their design choices [7, 43, 11, 13, 44] and clearly state the conclusions they are attempting to draw from their study [7, 45].

Replicability and generalizability in ML. Our work formalizes the definitions of replicability and generalizability given in [48, 46]. Intuitively, replicable work consists of repeating an experiment on different data, while generalizable work varies other factors as well — e.g., quality metric, implementation. A recent line of work, initiated by [33], has linked replicability to model stability: a ρ-replicable model learns (with probability $1 - \rho$) the same parameters from different i.i.d. samples. This definition has later been adapted and applied to other learning algorithms [23], clustering [24], reinforcement learning [22, 37], convex optimization [1], and learning rules [36]. Recent efforts have been bridging the gap between replicability, differential privacy, generalization error, and global stability [15, 16, 26, 45, 21]. However, these applications remain limited to model replicability.

Replicability and generalizability in Science. In other fields of Science, generalizability and replicability take different meanings. In social sciences, generalizability theory is a tool to quantify the effect of different factors on numerical responses [14]. In medicine, the replicability proposed in [34] is the probability of observing a positive treatment effect in a meta-study. Although these concepts are related to generalizability of experimental studies, they are limited to purely numerical responses or specific study designs.

3 Experiments and experimental studies

An *experimental study* is a set of *experiments* comparing the same *alternatives* under different *experimental conditions*. An experimental condition is a tuple of *levels* of *experimental factors*, the parameters defining the experiments. Different factors play different roles in the study: the *design* and *held-constant* factors are fixed by design, while the generalizability of a study is defined in terms of the *allowed-to-vary* factors. The study aims at answering a *research question*, which defines its *scope* and *goals*.

[1] `https://anonymous.4open.science/r/genexpy-B94D`

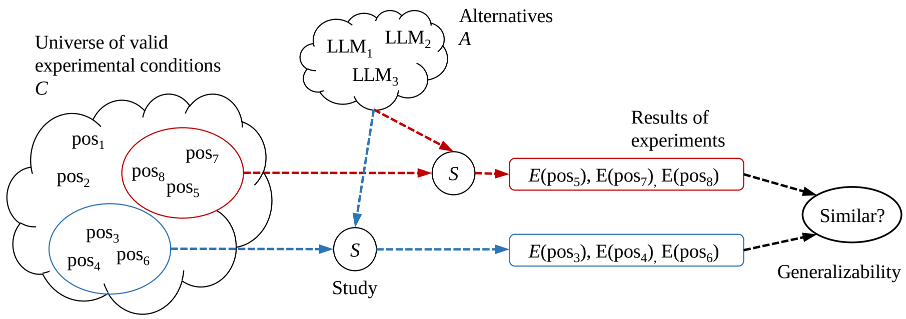


Figure 1: Two empirical studies on the checkmate-in-one task, cf. Example 3.1.

Example 3.1. (The "checkmate-in-one" task, cf. Figure 1) An experimenter wants to compare three Large Language Models (LLMs), the *alternatives*, on the "checkmate-in-one" task [55, 2, 5, 4, 18]. The assignment is to find the unique checkmating move from a position of pieces on a chessboard: an LLM succeeds if and only if it outputs the correct move. The experimenter considers two *experimental factors*: the number of shots, n, and the initial position on the chessboard, pos_l. The number of shots is a *design factor*, while the initial position is an *allowed-to-vary* factor. The experimenter wants to find if LLM_1 ranks consistently against the other two LLMs when changing the initial position, for a fixed number of shots.

The rest of this section defines the terms introduced above.

3.1 Experiments

An experiment evaluates all the considered *alternatives* under a *valid experimental condition*.

Alternatives. An alternative $a \in A$ is an object compared in the study, like an LLM in Example 3.1. Here, A is the set of alternatives considered in the study, with cardinality n_a.

Experimental factors. An experimental factor is anything that could, in principle, affect the result of an experiment. i denotes a factor, C_i the (possibly infinite) set of *levels* i can take, $c \in C_i$ a level of i, and I the set of all factors. We adapt Montgomery's classification of experimental factors [44, Chapter 1] and discern between *design factors*, *held-constant factors*, and *allowed-to-vary factors*.

- *Design factors*, e.g., whether and how to tune the hyperparameters, quality metrics, number of shots, are chosen by the experimenter.
- *Held-constant factors*, e.g., implementation, initialization seed, number of cross-validated folds, may affect the outcome but are not in the scope of the experiment and are fixed by the experimenter.
- *Allowed-to-vary factors*, e.g., "dataset" or "chessboard position" in Example 3.1, may affect the outcome but cannot be held constant: the experimenter expects results to generalize w.r.t. these factors; I_{atv} denotes them.

Experimental conditions. An *experimental condition* $\mathbf{c}$ is a tuple of levels of experimental factors, $\mathbf{c} = (c_i)_{i \in I} \in C \subseteq \prod_{i \in I} C_i$. We endow C with a probability μ, as we will need to sample from it to define the result of a study in Section **??**. The probability space $(C, \mathcal{F}, \mu)$ is the *universe of valid experimental conditions*. C may not coincide with $\prod_{i \in I} C_i$ as some experimental conditions may be *invalid*, i.e., illegal or not of interest. Validity has to be assessed on a case-by-case basis. For instance, in Example 3.1, $C = \{(\mathrm{pos}_l, n)\}_{l,n}$, where pos_l is a legal configuration of pieces on a chessboard and m is the non-negative number of shots.

Experimental results. The *experiment function* E evaluates the alternatives A under a valid experimental condition $\mathbf{c} \in C$. Unless necessary, we consider A fixed and omit it in our notation. We require that $E : C \to \mathcal{R}_{n_a}$ is a measurable function, for some fixed A. Finally, the *result* of an experiment $E(A, \mathbf{c})$ is a ranking on A.

Definition 3.1 (Ranking (with ties))**.** A ranking r on A is a transitive and reflexive binary endorelation on A. Equivalently, r is a totally ordered partition of A into *tiers* of equivalent alternatives. $r(a)$ denotes the *rank* of $a \in A$, i.e., the position of the tier of a in the ordering. W.l.o.g. $(\mathcal{R}_{n_a}, \mathcal{P}(\mathcal{R}_{n_a}))$ denotes the measure space of all rankings of n_a objects, where $\mathcal{P}$ indicates the power set.

Example 3.1 (Continued). The result of an experiment on (pos_l, n) is a ranking of the three LLMs, according to whether or not they output the checkmating move. Suppose that only LLM_1 and LLM_2 output the correct move. Then $E(\text{pos}_l, n)$ ranks LLM_1 and LLM_2 tied as best and LLM_3 as worst.

3.2 Experimental studies

A study is defined by its *research question* $\mathcal{Q}$, i.e., its *scope* and *goals*. The *scope* consists of the alternatives A, the valid experimental conditions C, and the allowed-to-vary factors I_{atv}. The *goal* is the kind of conclusions one is attempting to draw from the study. For now, the goal is a statement of interests, i.e., a set of strings.

Definition 3.2 (Research question)**.** The research question $\mathcal{Q} = (A, C, I_{\text{atv}}, \text{goals})$ is a tuple containing the set of alternatives A, the experimental conditions C, the set of allowed-to-vary-factors I_{atv}, and the goals of the study.

Example 3.1 (Continued). The research question of the "checkmate-in-one" study is as follows. The *scope* is $\left(A = \{\text{LLM}_a\}_{a=1,2,3}, C = \{(\text{pos}_l, n)\}_{l,n}, I_{\text{atv}} = \{\text{“position”}\}\right)$. The *goal* is "Does LLM_1 rank consistently against the other LLMs?"

A crucial element of our formalization is the distinction between *ideal* and *empirical* studies. An ideal study exhausts its research question; however, its result is not observable. An empirical study is an observable sample of an ideal study.

3.2.1 Ideal studies

The *ideal study* on a research question $\mathcal{Q} = (A, C, I_{\text{atv}}, \text{goals})$ is the experimental study consisting of an experiment for each valid experimental condition $\mathbf{c} \in C$. We say that such a study exhausts $\mathcal{Q}$. Hence, there exists exactly one ideal study on $\mathcal{Q}$. The *result* of an ideal study is the probability distribution of the results of its experiments. Recall that the experiment function $E : (C, \mathcal{F}, \mu) \to (\mathcal{R}_{n_a}, \mathcal{P}(\mathcal{R}_{n_a}))$ is measurable.

Definition 3.3 (Result of an ideal study)**.** The *result of an ideal study* with research question $\mathcal{Q} = (A, C, I_{\text{atv}}, \text{goals})$ is

$$\begin{aligned} S(\mathcal{Q}) = \mathbb{P} : \mathcal{R}_{n_a} &\to [0,1] \\ r &\mapsto \mathbb{P}(r) := \mu\left(E^{-1}(r)\right), \end{aligned}$$

where $E^{-1}(r) = \{\mathbf{c} : E(\mathbf{c}) = r\} \subseteq C$ is the preimage of r through E.

In general, multiple experiments of a study may yield identical results. Definition 3.3 supports this by assigning a higher probability mass to results that occur more often.

3.2.2 Empirical studies

Consider again a research question $\mathcal{Q} = (A, C, I_{\text{atv}}, \text{goals})$. In practice, as C might be infinite or too large, one can only run experiments on a sample of valid experimental conditions $\{\mathbf{c}_j\}_{j=1}^{N} \overset{\text{iid}}{\sim} (C, \mu)$. The study performed on $\{\mathbf{c}_j\}_{j=1}^{N}$ is *an empirical study* on $\mathcal{Q}$, of *size* N. As for ideal studies, the result of an empirical study is the probability distribution of the results of its experiments.

Definition 3.4 (Result of an empirical study)**.** The *result of an empirical study* on $\mathcal{Q}$ is

$$\begin{aligned} \hat{S}_N(\mathcal{Q}) : \mathcal{R}_{n_a} &\to [0,1] \\ r &\mapsto \#\left\{j \in \{\mathbf{c}_j\}_{j=1}^{N} : E(A, \mathbf{c}_j) = r\right\}. \end{aligned}$$

Where $\mathcal{Q}$, $\{\mathbf{c}_j\}_{j=1}^{N}$ is a research question and a set of valid experimental conditions as above.

The result of an empirical study can be thought of as the empirical distribution of a sample following the distribution of the result of the corresponding ideal study. With a slight abuse of notation, indicating both the sample and its empirical distribution as $\hat{S}_N(\mathcal{Q})$, we write

$$\hat{S}_N(\mathcal{Q}) \overset{\text{iid}}{\sim} S(\mathcal{Q}).$$

4 Generalizability of experimental studies

The currently accepted definition of generalizability is the property of two independent studies with the same research question to yield similar results [46, 48]. Although intuitive, this notion is not directly applicable as it does not provide a way to measure the generalizability of a study. We now introduce a quantifiable notion of generalizability of experimental studies, as the probability that any two empirical studies approximating the same ideal study yield similar results.

Definition 4.1 (Generalizability)**.** Let $\mathcal{Q} = (A, C, I_{\text{atv}}, \kappa)$ be the research question of an ideal study, let $\mathbb{P} = S(\mathcal{Q})$ be the result of that study, and let d be some distance between probability distributions. The *generalizability of the ideal study on* $\mathcal{Q}$ is

$$\text{Gen}(\mathcal{Q}; \varepsilon, n) := \mathbb{P}^n \otimes \mathbb{P}^n \left((X_j, Y_j)_{j=1}^n : d(X, Y) \leq \varepsilon \right), \tag{1}$$

where $\varepsilon \in \mathbb{R}^+$ is a *similarity threshold.*

As the result of an ideal study is usually unobservable (cf. Section 3.2), we do not know the true distribution $\mathbb{P}$. However, we can observe the result of an empirical study, $\hat{\mathbb{P}}_N = \hat{S}_N(\mathcal{Q})$, which approximates $\mathbb{P}$ under the assumption that the experimental conditions are i.i.d. samples from C. As the sample size N increases (the empirical study becomes larger), $\hat{\mathbb{P}}_N$ converges in distribution to $\mathbb{P}$.

Definition (1) requires a distance d between probability distributions. In the next sections, we propose to use a generalizability based on kernels and Maximum Mean Discrepancy (MMD) [27], as it allows to compute generalizability w.r.t. different research questions. The underlying idea is that we can capture the goal of a study with an appropriate kernel. We conclude this section with an algorithm to estimate the number of experimental conditions required to obtain generalizable results.

4.1 Similarity between rankings — kernels

Whether two experimental results (i.e., rankings) are similar or not ultimately depends on the goal of the study. For instance, consider two rankings on $A = \{a_1, a_2, a_3\}$, $\mathbf{r} = (1, 2, 3)$ and $\mathbf{r}' = (1, 3, 2)$, where r_i is the tier of alternative a_i. The conclusions drawn from r and r' are identical if one's goal is to find the best alternative, but very different if one's goal is to obtain an ordering of the alternatives. One can use kernels to quantify the similarity between experimental results. Kernels are suitable to formalize the aspects of the result of a study one wants to generalize, i.e., the goals of the study. For instance, one kernel is suitable to identify the best tier while another kernel focuses on the position of a specific alternative. In the following, we describe three representative kernels that cover a wide spectrum of possible goals.

Borda kernel. The Borda kernel is suitable for goals in the form "Is the alternative a^* consistently ranked the same?". It uses the Borda count: the number of alternatives (weakly) dominated by a given one [9]. For a pair of rankings, we compute the Borda counts of a^*, and then take their difference.

$$\kappa_b^{a^*,\nu}(r_1, r_2) = e^{-\nu|b_1 - b_2|},$$

where $b_l = \{a \in A : r_l(a) \geq r_l(a^*)\}$ is the number of alternatives dominated by a^* in r_l and $\nu \in \mathbb{R}$ is the kernel bandwidth. The Borda kernel takes values in $\left[e^{(-\nu n_a)}, 1\right]$. If ν is too large compared to $1/|b_1 - b_2|$, the kernel is oversensitive and will penalize every deviation too much. On the contrary, if ν is too small, the kernel is undersensitive and will not penalize deviations unless they are very large. As $|b_1 - b_2| \in [0, n_a]$, we recommend $\nu = 1/n_a$.

Jaccard kernel. The Jaccard kernel is suitable for goals in the form "Are the best alternatives consistently the same ones?". As it measures the similarity between sets [25, 10], we use it to compare the top-k tiers of two rankings.

$$\kappa_j^k(r_1, r_2) = \frac{\left|r_1^{-1}([k]) \cap r_2^{-1}([k])\right|}{\left|r_1^{-1}([k]) \cup r_2^{-1}([k])\right|},$$

where $r^{-1}([k]) = \{a \in A : r_1(a) \leq k\}$ is the set of alternatives whose rank is better than or equal to k. The Jaccard kernel takes values in $[0, 1]$.

Mallows kernel. The Mallows kernel is suitable for goals in the form "Are the alternatives ranked consistently?". It measures the overall similarity between rankings [35, 40, 39]. We adapt the original definition in [39] for ties,

$$\kappa_m^\nu(r_1, r_2) = e^{-\nu n_d},$$

where $n_d = \sum_{a_1, a_2 \in A} \left|\text{sign}\left(r_1(a_1) - r_1(a_2)\right) - \text{sign}\left(r_2(a_1) - r_2(a_2)\right)\right|$ is the number of discordant pairs and $\nu \in \mathbb{R}$ is the kernel bandwidth. If a pair is tied in one ranking but not in the other, one counts it as half a discordant pair. The Mallows kernel takes values in $\left[\exp\left(-2\nu\binom{n_a}{2}\right), 1\right]$. If ν is too large compared to $1/n_d$, the kernel is oversensitive and it will penalize every deviation too much. On the contrary, if ν is too small, the kernel is undersensitive and will not penalize deviations unless they are very large. As $n_d \in \left[0, \binom{n_a}{2}\right]$, we recommend $\nu = 1/\binom{n_a}{2}$.

4.2 Distance between distributions — Maximum Mean Discrepancy

Having sorted out how to measure the similarity between the results of experiments, we now discuss how to measure the distance between the results of studies. We chose the Maximum Mean Discrepancy (MMD) [27], for the following reasons. First, MMD is compatible with the kernels described in Section 4.1, i.e., it takes into consideration the goal of the studies. Second, it handles sparse distributions well; this is needed as empirical studies are typically small compared to the number of all possible rankings, which grows exponentially in the number of alternatives. [2] Finally, it comes with bounds and theoretical guarantees, which we will use in Section 4.3.

Definition 4.2 (MMD (empirical distributions))**.** Let X be a set with a kernel κ, and let $\mathbb{Q}_1$ and $\mathbb{Q}_2$ be two probability distributions on $\mathcal{R}_{n_a}$. Let $\mathbf{x} = (x_i)_{i=1}^n, \mathbf{y} = (y_i)_{i=1}^m$ be two i.i.d. samples from $\mathbb{Q}_1$ and $\mathbb{Q}_2$ respectively. Then,

$$\text{MMD}\left(\mathbf{x}, \mathbf{y}\right)^2 := \frac{1}{n^2}\sum_{i,j=1}^{n} \kappa(x_i, x_j) + \frac{1}{m^2}\sum_{i,j=1}^{m} \kappa(y_i, y_j) - \frac{2}{mn}\sum_{\substack{i=1\dots n\\ j=1\dots m}} \kappa(x_i, y_j).$$

Proposition 4.1. *MMD takes values in* $\left[0, \sqrt{2 \cdot (\kappa_{sup} - \kappa_{inf})}\right]$*, where* $\kappa_{sup} = \sup_{x,y \in X} \kappa(x, y)$ *and* $\kappa_{inf} = \inf_{x,y \in X} \kappa(x, y)$*.*

4.3 How many experiments ensure generalizability?

When designing a study, an experimenter has to decide how many experiments to run in order to obtain generalizable results. In other words, they need to choose a (minimum) sample size n^* that achieves the desired generalizability α^* and the desired similarity ε^*.

$$n^* = \min\left\{n \in \mathbb{N}_0 : \text{Gen}\left(\mathbb{P}; \varepsilon^*, n\right) \geq \alpha^*\right\}. \tag{2}$$

To estimate n^* we make use of a linear dependency between the logarithms of the sample size n and the logarithm of the α^*-quantile of MMD $\varepsilon_n^{\alpha^*}$ that we have observed in our experiments.

Proposition 4.2. $\forall \alpha^*$*, there exist* $\beta_0 \geq 0, \beta_1 \leq 0$ *s.t.*

$$\log(n) \approx \beta_1 \log\left(\varepsilon_n^{\alpha^*}\right) + \beta_0 \tag{3}$$

Appendix A.3.2 provides a proof for a simplified case. Proposition 4.2 suggests that one can use a small set of N preliminary experiments to estimate n^*. One can then iteratively improve that estimate with the results of additional experiments.

Our algorithm, shown in detail in Appendix A.3.3, requires specifying the desired generalizability, α^*, and the similarity threshold between the studies results, ε^*. Then, it performs the following steps:

[2] Fubini or ordered Bell numbers, OEIS sequence A000670.

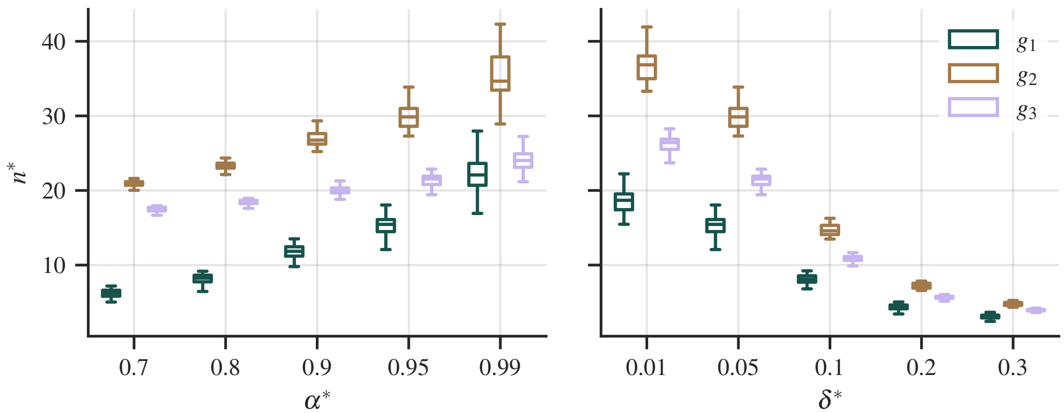


Figure 2: Predicted n^* for categorical encoders.

1. it estimates the α^*-quantile of MMD for all n less than some budget $n_{\max}$. If there exists an n less than $n_{\max}$ that satisfies the condition in (2), we return it as n^*;
2. it then fits the linear model in (3), computing the coefficients β_0 and β_1;
3. finally, it outputs $n^* = \exp\left(\beta_1 \log\left(\varepsilon_n^{\alpha^*}\right) + \beta_0\right)$, which satisfies the condition in (2) thanks to Proposition 4.2.

In practice, choosing ε^* is hardly interpretable as it is a threshold on MMD. To solve this, we propose choosing ε^* as a function of another parameter δ^*, such that

$$\varepsilon^*(\delta^*) = \sqrt{2(\kappa_{\text{sup}} - f_\kappa(\delta^*))}.$$

Here, δ^* represents the distance between two rankings as computed by the kernel and f_κ is the function linking the distance to the kernel value. For instance, for the Jaccard kernel, δ^* is simply the Jaccard coefficient between the top-k tiers of two rankings, $f_\kappa(\delta^*) = 1 - \delta^*$, and $\varepsilon^*(\delta^*) = \sqrt{2(1 - (1 - \delta^*))}$. For the Mallows kernel (with our recommendation for ν), δ^* is the fraction of discordant pairs, $f_\kappa(x) = e^{-x}$, and $\varepsilon^*(\delta^*) = \sqrt{2(1 - e^{-\delta^*})}$. As a concrete example, achieving $(\alpha^* = 0.99, \delta^* = 0.05)$-generalizable results for the Jaccard kernel means that, with probability 0.99, the average Jaccard coefficient between two rankings drawn from the results is 0.95.

5 Case studies

5.1 Case Study 1: A benchmark of categorical encoders

We now evaluate the generalizability of a recent study [41] that analyzes the performance of encoders for categorical data. The performance of an encoder is approximated by the quality of a model trained on the encoded data. The *design factors* are the model, the tuning strategy for the pipeline, and the quality metric for the model, while the only *allowed-to-vary factor* is the dataset. We impute missing values in the results of the study by assigning the worst rank. We evaluate how well the results of the study generalize w.r.t. three goals:

(g_1) Find out if One-Hot encoder (a popular encoder) ranks consistently amongst its competitors, using the Borda kernel with $\nu = {}^1\!/_{n_a}$.
(g_2) Investigate if some encoders outperform all the others using the Jaccard kernel with $k = 1$.
(g_3) Evaluate whether the encoders are typically ranked in a similar order, using the Mallows kernel with $\nu = 1/\binom{n_a}{2}$.

Figure 2 shows the predicted n^* for different choices of α^* and δ^*, the other one fixed at 0.95 and 0.05 respectively. The variance in the boxes comes from variance in the design factors. For example, the results for the design factors "decision tree, full tuning, accuracy" have a different (α^*, δ^*)-generalizability than the results for "SVM, no tuning, accuracy". We observe on the left that — as expected — obtaining generalizable results requires more experiments as the desired

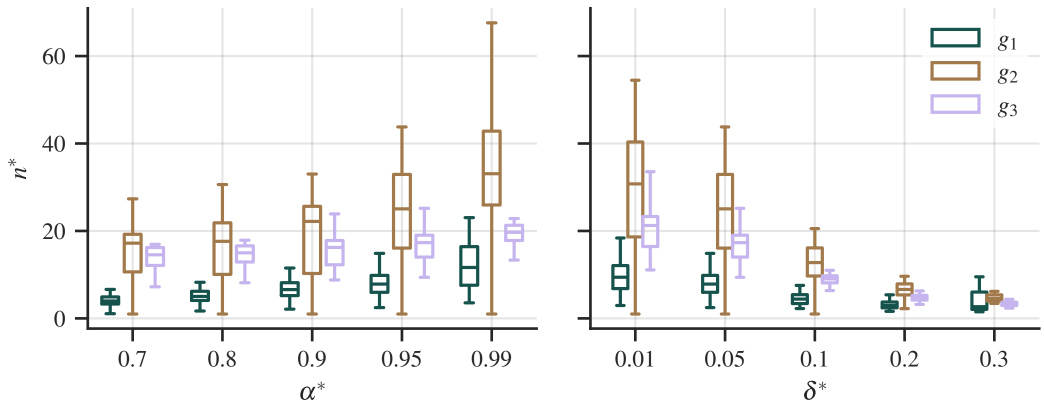


Figure 3: Predicted n^* for LLMs.

generalizability α^* increases. We can also see that the variance of the boxes increases with α^*. This means that the choice of the design factors has a larger influence on the achieved generalizability. We observe the same when decreasing δ^*, as it corresponds to a stricter similarity condition on the rankings. In the rather extreme cases of $\alpha^* = 0.7$ or $\delta^* = 0.3$, even less than 10 datasets are enough to achieve (α^*, δ^*)-generalizability.

Consider now goal g_2 for two different choices of design factors: (A): "decision tree, full tuning, accuracy" and (B): "SVM, full tuning, balanced accuracy". Furthermore, let $(\alpha^*, \delta^*) = (0.95, 0.05)$: we estimate $n^* = 28$ for (A) and $n^* = 34$ for (B), corresponding to the bottom and top whiskers of the corresponding box in Figure 2. As both (A) and (B) were evaluated using $n = 30$ experiments, we conclude that the results of (A) are (barely) $(0.95, 0.05)$-generalizable, while those of (B) are not. Hence, one should run more experiments with fixed factors (B) to make the study generalizable.

5.2 Case study 2: BIG-bench — A benchmark of Large Language Models

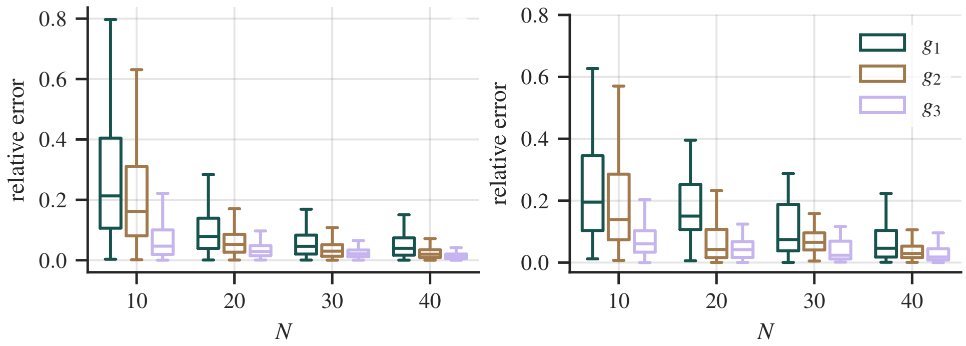
We now evaluate the generalizability of BIG-bench [55], a collaborative benchmark of Large Language Models (LLMs). The benchmark compares LLMs on different tasks, such as the checkmate-in-one task (cf. Example 3.1), and for different numbers of shots. Task and number of shots are the *design factors*. Every task has a number of subtasks, which is the *allowed-to-vary factor*. We stick to the preferred scoring for each subtask. As the results have too many missing values to impute them, we only consider the experimental conditions where at least 80% of the LLMs had results, and to the LLMs whose results cover at least 80% of the conditions.

Similar to before, we define the three goals as follows:

- (g_1) Find out if GPT3 (to date, one of the most popular LLMs) ranks consistently amongst its competitors, using the Borda kernel with $\nu = {}^1/_{n_a}$.
- (g_2) Investigate if some encoders outperform all the others using the Jaccard kernel with $k = 1$.
- (g_3) Evaluate whether the LLMs are typically ranked in a similar order, using the Mallows kernel with $\nu = {}^1/\binom{n_a}{2}$.

Figure 3 shows the predicted n^* for different choices of α^* and δ^*, the other one fixed at 0.95 and 0.05 respectively. Again, the variance in the boxes comes from variance in the design factors, i.e., the task and the number of shots. As before, increasing α^* or decreasing δ^* leads to higher n^*. Unlike in the previous section, n^* for g_2 greatly depends on the combination of fixed factors, as we now detail.

Consider now goal g_2 for two different choices of design factors: (A): "conlang_translation, 0 shots", and (B): "arithmetic, 2 shots". Furthermore, let $(\alpha^*, \delta^*) = (0.95, 0.05)$. For this choice of of parameters, we estimate $n^* = 44$ for (A), corresponding to the top whisker of the corresponding box in Figure 2. As the study evaluates (A) on 10 subtasks, it is therefore not $(0.95, 0.05)$-generalizable. In fact, we estimate that this would require 34 more subtasks. For (B), on the other hand, we estimate $n^* = 1$: the best 2-shot LLM for the observed subtasks is always PaLM 535B. Hence, the result of a single experiment is enough to achieve $(0.95, 0.05)$-generalizability.



(a) Categorical encoders

(b) LLMs

Figure 4: Relative error in the estimate of n^* against n^*_{50}.

Note that, although we correctly estimated $n^* = 1$ for (B), this estimate relies on 10 preliminary experiments. In other words, our algorithm was able to quantify *in hindsight* that a single experiment would have been enough to obtain generalizable results. Of course, however, one cannot trust an estimate of n^* based on only one experiment. The next section thus investigates how the number of preliminary experiments influences the estimate of n^*.

5.3 How many preliminary experiments?

This section evaluates the influence of the number of preliminary experiments N on n^*. For each study, we consider the design factor combinations for which we have at least 50 experiments. This results in 23 out of 48 combinations for the categorical encoders and 9 out of 24 combinations for the LLMs. For each of those combinations, we consider the estimate n^*_{50} made at $N = 50$ as the ground truth and observe how the estimates of n^* for $N < 50$ differ. Figure 4 shows the relative error $|n^*_N - n^*_{50}|/n^*_{50}$, for different goals: the relative errors behave very differently. For goal g_3 (Mallows kernel), even n^*_{10} is close to n^*_{50} for a majority of the design factor combinations. On the contrary, one needs 20 to 30 preliminary experiments for goal g_1 (Borda kernel). This means that knowing the goals of a study when performing preliminary experiments can help understand how trustworthy the estimate of n^* is.

6 Conclusion

Limitations & future work. First, we dealt with experimental results as rankings. Other forms of results, e.g., the absolute performance of alternatives according to some quality measure, will require the development of appropriate kernels. Second, our approach uses kernels to compute the similarity of experimental results and MMD the distance between the results of studies. There are, however. other possible choices. Third, we processed missing evaluations by either dropping them or imputing them. One could analyze different solutions, for instance by adapting the kernels to missing values. Fourth, we estimate the distribution of the MMD by sampling multiple times from the results. A non-asymptotic theory of MMD, at least for some kernels, might yield more insights in improving this procedure. Fifth, we plan to investigate the possibility of actively selecting experiments to obtain good estimates of the required size n^* with less preliminary experiments. Sixth and related to the previous one, we intend to obtain some guarantees on the convergence of n^* to the true value.

Conclusions. An experimental study is generalizable if, with high probability, its findings will hold under different experimental conditions, e.g., on unseen datasets. Non-generalizable studies might be of limited use or even misleading. This study is, to our knowledge, the first to develop a quantifiable notion for the generalizability of experimental studies. To achieve this, we formalize experiments, experimental studies and their results — rankings and distributions over rankings. Our approach allows us to estimate the number of experiments needed to achieve a desired level of generalizability in new experimental studies. We demonstrate its utility showing generalizable and non-generalizable results in two recent experimental studies.

Acknowledgments

...

References

[1] Kwangjun Ahn et al. "Reproducibility in Optimization: Theoretical Framework and Limits". In: *NeurIPS*. 2022.

[2] Scott Alexander. "A very unlikely chess game, 2020". In: *URL https://slatestarcodex.com/2020/01/06/a-very-unlikely-chessgame/.(cited on pp. 29 and 30)* ().

[3] Maxime Alvarez et al. "A Revealing Large-Scale Evaluation of Unsupervised Anomaly Detection Algorithms". In: *CoRR* abs/2204.09825 (2022).

[4] Prithviraj Ammanabrolu et al. "Bringing stories alive: Generating interactive fiction worlds". In: *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*. Vol. 16. 1. 2020, pp. 3–9.

[5] Prithviraj Ammanabrolu et al. "Toward automated quest generation in text-adventure games". In: *arXiv preprint arXiv:1909.06283* (2019).

[6] Monya Baker. "1,500 scientists lift the lid on reproducibility". In: *Nature* 533.7604 (2016).

[7] Thomas Bartz-Beielstein et al. "Benchmarking in Optimization: Best Practice and Open Issues". In: *CoRR* abs/2007.03488 (2020).

[8] Alessio Benavoli et al. "Time for a Change: a Tutorial for Comparing Multiple Classifiers Through Bayesian Analysis". In: *J. Mach. Learn. Res.* 18 (2017), 77:1–77:36.

[9] JC de Borda. "M'emoire sur les' elections au scrutin". In: *Histoire de l'Acad'emie Royale des Sciences* (1781).

[10] Mathieu Bouchard, Anne-Laure Jousselme, and Pierre-Emmanuel Doré. "A proof for the positive definiteness of the Jaccard index matrix". In: *International Journal of Approximate Reasoning* 54.5 (2013), pp. 615–626.

[11] Anne-Laure Boulesteix, Rory Wilson, and Alexander Hapfelmeier. "Towards evidence-based computational statistics: lessons from clinical research on the role and design of real-data benchmark studies". In: *BMC Medical Research Methodology* 17 (2017), pp. 1–12.

[12] Anne-Laure Boulesteix et al. "A statistical framework for hypothesis testing in real data comparison studies". In: *The American Statistician* 69.3 (2015), pp. 201–212.

[13] Xavier Bouthillier et al. "Accounting for Variance in Machine Learning Benchmarks". In: *MLSys*. mlsys.org, 2021.

[14] Robert L Brennan. "Generalizability theory". In: *Educational Measurement: Issues and Practice* 11.4 (1992), pp. 27–34.

[15] Mark Bun et al. "Stability Is Stable: Connections between Replicability, Privacy, and Adaptive Generalization". In: *STOC*. ACM, 2023, pp. 520–527.

[16] Zachary Chase, Shay Moran, and Amir Yehudayoff. "Stability and Replicability in Learning". In: *FOCS*. IEEE, 2023, pp. 2430–2439.

[17] Giorgio Corani et al. "Statistical comparison of classifiers through Bayesian hierarchical modelling". In: *Mach. Learn.* 106.11 (2017), pp. 1817–1837.

[18] Sahith Dambekodi et al. "Playing text-based games with common sense". In: *arXiv preprint arXiv:2012.02757* (2020).

[19] Mostafa Dehghani et al. "The Benchmark Lottery". In: *CoRR* abs/2107.07002 (2021).

[20] Janez Demsar. "Statistical Comparisons of Classifiers over Multiple Data Sets". In: *J. Mach. Learn. Res.* 7 (2006), pp. 1–30.

[21] Peter Dixon et al. "List and Certificate Complexities in Replicable Learning". In: *NeurIPS*. 2023.

[22] Eric Eaton et al. "Replicable Reinforcement Learning". In: *NeurIPS*. 2023.

[23] Hossein Esfandiari et al. "Replicable Bandits". In: *ICLR*. OpenReview.net, 2023.

[24] Hossein Esfandiari et al. "Replicable Clustering". In: *NeurIPS*. 2023.

[25] Thomas Gärtner, Quoc Viet Le, and Alex J Smola. "A short tour of kernel methods for graphs". In: *Under Preparation* (2006).

[26] Badih Ghazi et al. "On User-Level Private Convex Optimization". In: *ICML*. Vol. 202. Proceedings of Machine Learning Research. PMLR, 2023, pp. 11283–11299.

[27] Arthur Gretton et al. "A Kernel Method for the Two-Sample-Problem". In: *NIPS*. MIT Press, 2006, pp. 513–520.

[28] Arthur Gretton et al. "A Kernel Two-Sample Test". In: *J. Mach. Learn. Res.* 13 (2012), pp. 723–773.

[29] Odd Erik Gundersen, Kevin L. Coakley, and Christine R. Kirkpatrick. "Sources of Irreproducibility in Machine Learning: A Review". In: *CoRR* abs/2204.07610 (2022).

[30] Odd Erik Gundersen et al. "On Reporting Robust and Trustworthy Conclusions from Model Comparison Studies Involving Neural Networks and Randomness". In: *ACM-REP*. ACM, 2023, pp. 37–61.

[31] Torsten Hothorn et al. "The design and analysis of benchmark experiments". In: *Journal of Computational and Graphical Statistics* 14.3 (2005), pp. 675–699.

[32] Karl Huppler. "The Art of Building a Good Benchmark". In: *TPCTC*. Vol. 5895. Lecture Notes in Computer Science. Springer, 2009, pp. 18–30.

[33] Russell Impagliazzo et al. "Reproducibility in learning". In: *STOC*. ACM, 2022, pp. 818–831.

[34] Iman Jaljuli et al. "Quantifying replicability and consistency in systematic reviews". In: *Statistics in Biopharmaceutical Research* 15.2 (2023), pp. 372–385.

[35] Yunlong Jiao and Jean-Philippe Vert. "The Kendall and Mallows Kernels for Permutations". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 40.7 (2018), pp. 1755–1769.

[36] Alkis Kalavasis et al. "Statistical Indistinguishability of Learning Algorithms". In: *ICML*. Vol. 202. Proceedings of Machine Learning Research. PMLR, 2023, pp. 15586–15622.

[37] Amin Karbasi et al. "Replicability in Reinforcement Learning". In: *NeurIPS*. 2023.

[38] Fred Lu, Edward Raff, and James Holt. "A Coreset Learning Reality Check". In: *AAAI*. AAAI Press, 2023, pp. 8940–8948.

[39] Colin L Mallows. "Non-null ranking models. I". In: *Biometrika* 44.1/2 (1957), pp. 114–130.

[40] Horia Mania et al. "On kernel methods for covariates that are rankings". In: (2018).

[41] Federico Matteucci, Vadim Arzamasov, and Klemens Böhm. "A benchmark of categorical encoders for binary classification". In: *NeurIPS*. 2023.

[42] Duncan C. McElfresh et al. "On the Generalizability and Predictability of Recommender Systems". In: *NeurIPS*. 2022.

[43] Iven Van Mechelen et al. "A white paper on good research practices in benchmarking: The case of cluster analysis". In: *WIREs Data. Mining. Knowl. Discov.* 13.6 (2023).

[44] Douglas C Montgomery. *Design and analysis of experiments*. John wiley & sons, 2017.

[45] Shay Moran, Hilla Schefler, and Jonathan Shafer. "The Bayesian Stability Zoo". In: *NeurIPS*. 2023.

[46] Engineering National Academies of Sciences, Medicine, et al. *Reproducibility and replicability in science*. National Academies Press, 2019.

[47] Roger D Peng. "Reproducible research in computational science". In: *Science* 334.6060 (2011), pp. 1226–1227.

[48] Joelle Pineau et al. "Improving Reproducibility in Machine Learning Research(A Report from the NeurIPS 2019 Reproducibility Program)". In: *J. Mach. Learn. Res.* 22 (2021), 164:1–164:20.

[49] Zhen Qin et al. "RD-Suite: A Benchmark for Ranking Distillation". In: *NeurIPS*. 2023.

[50] Edward Raff. "Does the Market of Citations Reward Reproducible Work?" In: *ACM-REP*. ACM, 2023, pp. 89–96.

[51] Edward Raff. "Research Reproducibility as a Survival Analysis". In: *AAAI*. AAAI Press, 2021, pp. 469–478.

[52] Isaac J Schoenberg. "Metric spaces and positive definite functions". In: *Transactions of the American Mathematical Society* 44.3 (1938), pp. 522–536.

[53] Bernhard Schölkopf. "The kernel trick for distances". In: *Advances in neural information processing systems* 13 (2000).

[54] J Laurie Snell and John G Kemeny. "Mathematical models in the social sciences". In: *(No Title)* (1962).

[55] Aarohi Srivastava et al. “Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models”. In: *CoRR* abs/2206.04615 (2022).

[56] Donglin Zhuang et al. “Randomness in Neural Network Training: Characterizing the Impact of Tooling”. In: *MLSys*. mlsys.org, 2022.

A Details for Section 4

A.1 Details for Section 4.1

This section contains the proofs to show that the similarities introduced in Section 4.1 are kernels, i.e., symmetric and positive definite functions. As symmetry is a clear property of all of them, we only discuss their positive definiteness. Our proofs for the Borda and Mallows kernels follow that in [35]: we define a distance d on the set of rankings $\mathcal{R}_{n_a}$ and show that $(\mathcal{R}_{n_a}, d)$ is isometric to an L_2 space. This ensures that d is a conditionally positive definite (c.p.d.) function and, thus, that $e^{-\nu d}$ is positive definite [52, 53]. Our proof for the Jaccard kernel, instead, follows without much effort from previous results. For ease of reading, we restate the definitions as well.

Definition A.1 (Borda kernel)**.**

$$\kappa_b^{a^*,\nu}(r_1, r_2) = e^{-\nu|d_1 - d_2|}, \tag{4}$$

where $d_l = \{a \in A : r_l(a) \geq r_l(a^*)\}$ is the number of alternatives dominated by a^* in r_l and $\nu \in \mathbb{R}$.

Proposition A.1. *The Borda kernel as defined in* (4) *is a kernel.*

Proof. Define a distance

$$\begin{aligned} d : \mathcal{R}_{n_a} \times \mathcal{R}_{n_a} &\to \mathbb{R}^+ \\ (r_1, r_2) &\mapsto |d_1, d_2|, \end{aligned}$$

where $d_l = \{a \in A : r_l(a) \geq r_l(a^*)\}$ is the number of alternatives dominated by a^* in r_l. Now, $(\mathcal{R}_{n_a}, d)$ is isometric to $(\mathbb{R}, \|\cdot\|_2)$ via the map $r_l \mapsto d_l$. Hence, d is c.p.d. and κ_b is a kernel. ☐

Definition A.2 (Jaccard kernel)**.**

$$\kappa_j^k(r_1, r_2) = \frac{\left|r_1^{-1}([k]) \cap r_2^{-1}([k])\right|}{\left|r_1^{-1}([k]) \cup r_2^{-1}([k])\right|}, \tag{5}$$

where $r^{-1}([k]) = \{a \in A : r_1(a) \leq k\}$ is the set of alternatives whose rank is better than or equal to k.

Proposition A.2. *The Jaccard kernel as defined in* (5) *is a kernel.*

Proof. It is already know that the Jaccard coefficients for sets is a kernel [25, 10]. As the Jaccard kernel for rankings is equivalent to the Jaccard coefficient for the k-best tiers of said rankings, the former is also a kernel. ☐

Definition A.3 (Mallows kernel)**.**

$$\kappa_m^\nu(r_1, r_2) = e^{-\nu n_d}, \tag{6}$$

where $n_d = \sum_{a_1, a_2 \in A} |\text{sign}\left(r_1(a_1) - r_1(a_2)\right) - \text{sign}\left(r_2(a_1) - r_2(a_2)\right)|$ is the number of discordant pairs and $\nu \in \mathbb{R}$ is the kernel bandwidth.

Proposition A.3. *The Mallows kernel as defined in* (6) *is a kernel.*

Proof. The number of discordant pairs n_d is a distance on $\mathcal{R}_{n_a}$ [54]. Consider now the mapping of a ranking into its adjacency matrix,

$$\begin{aligned} \Phi : \mathcal{R}_{n_a} &\to \{0, 1\}^{n_a \times n_a} \\ r &\mapsto \left(\text{sign}\left(r(i) - r(j)\right)\right)_{i,j=1}^{n_a}. \end{aligned}$$

Then,

$$n_d = \|\Phi(r_1) - \Phi(r_2)\|_1 = \|\Phi(r_1) - \Phi(r_2)\|_2^2$$

where $\|\cdot\|_p$ indicates the entry-wise matrix p-norm and the equality holds because the entries of the matrices are either 0 or 1. As a consequence, $(\mathcal{R}_{n_a}, n_d)$ is isometric to $(\mathbb{R}^{n_a \times n_a}, \|\cdot\|_2)$ via Φ. Hence, n_d is c.p.d. and κ_m is a kernel. ☐

A.2 Details for Section 4.2

Proposition 4.1. *MMD takes values in* $\left[0, \sqrt{2 \cdot (\kappa_{sup} - \kappa_{inf})}\right]$, *where* $\kappa_{sup} = \sup_{x,y \in X} \kappa(x, y)$ *and* $\kappa_{inf} = \inf_{x,y \in X} \kappa(x, y)$.

Proof.

$$0 \leq \mathrm{MMD}_\kappa(\mathbf{x}, \mathbf{y})^2 = \frac{1}{n^2} \sum_{i,j=1}^{n} \kappa(x_i, x_j) + \frac{1}{m^2} \sum_{i,j=1}^{m} \kappa(y_i, y_j) - \frac{2}{mn} \sum_{\substack{i=1\ldots n \\ j=1\ldots m}} \kappa(x_i, y_j) \tag{7}$$

$$\leq \frac{1}{n^2} \sum_{i,j=1}^{n} \kappa_{\text{sup}} + \frac{1}{m^2} \sum_{i,j=1}^{n} \kappa_{\text{sup}} - \frac{2}{mn} \sum_{\substack{i=1\ldots n \\ j=1\ldots m}} \kappa_{\text{inf}}$$

$$= 2(\kappa_{\text{sup}} - \kappa_{\text{inf}})$$

□

A.3 Details for Section 4.3

A.3.1 Choice of α^*, ε^*, and δ^*

Consider a research question $\mathcal{Q} = (A, C, I_{\text{atv}}, \kappa)$ and the corresponding ideal study with result $\mathbb{P}$. The algorithm introduced in Section 4.3 aims at finding the minimum n^* such that, given two independent empirical studies on $\mathcal{Q}$, they achieve similar results. It has two hyperparameters, α^* and ε^*. $\alpha^* \in [0, 1]$ is the generalizability that one wants to achieve from the study, i.e., the probability that two independent realizations of the same ideal study will yield similar results. $\varepsilon^* \in \mathbb{R}^+$ is a similarity threshold: the results of two empirical studies $\mathbf{x}, \mathbf{y} \overset{\text{iid}}{\sim} \mathbb{P}$ are similar if $\mathrm{MMD}_\kappa(\mathbf{x}, \mathbf{y}) \leq \varepsilon^*$. However, as it is, ε^* is not interpretable. Instead, adapting the proof of Proposition 4.1, we can bound MMD by imposing a condition on the kernel, as we'll now illustrate. The key remark is that we are looking for a condition in the form

$$\mathrm{MMD}_\kappa(\mathbf{x}, \mathbf{y}) \leq \varepsilon^* = \sqrt{2(\kappa_{\text{sup}} - \delta')},$$

where $\delta' \in [0, \kappa_{\text{sup}}]$ replaces the third summatory in (7). In other terms, we can interpret δ' as the minimum acceptable value for the average of the kernel, $\mathbb{E}_{\mathbb{P}^2}\left[\kappa(x, y)\right]$. We now go a step further and compute δ' (a condition on the kernel) from $\delta^* \in [0, 1]$ (a condition on the rankings). The relation between δ' and δ^* changes with the kernel, and so does the interpretation of δ^*. For the three kernels we discuss in Section 4.1:

- *Mallows kernel with* $\nu = 1/\binom{n}{2}$: δ^* is the fraction of discordant pairs, $\delta' = e^{-\delta^*}$.
- *Jaccard kernel*: δ^* is the intersection over union of the top k tiers, $\delta' = 1 - \delta^*$.
- *Borda kernel with* $\nu = 1/n_a$: δ^* is the difference in relative position of a^* in the rankings, normalized to the length of the rankings, $\delta' = e^{-\delta^*}$

A.3.2 Proof of proposition 4.2

Proposition 4.2. $\forall \alpha^*$, *there exist* $\beta_0 \geq 0, \beta_1 \leq 0$ *s.t.*

$$\log(n) \approx \beta_1 \log\left(\varepsilon_n^{\alpha^*}\right) + \beta_0 \tag{3}$$

Proof. We provide a proof replacing MMD with the distribution-free bound defined in [28].

$$
\begin{aligned}
&\mathbb{P}^n \otimes \mathbb{P}^n \left((X_j, Y_j)_{j=1}^n : \text{MMD}(X,Y) - \left(\frac{2\kappa_{\text{sup}}}{n} \right) > \varepsilon \right) < \exp\left(-\frac{n\varepsilon^2}{4\kappa_{\text{sup}}} \right) \\
\overset{(1)}{\Longrightarrow} &\mathbb{P}^n \otimes \mathbb{P}^n \left((X_j, Y_j)_{j=1}^n : \text{MMD}(X,Y) > \varepsilon` \right) < \exp\left(-\frac{n\left(\varepsilon' - \left(\frac{2\kappa_{\text{sup}}}{n}\right)\right)^2}{4\kappa_{\text{sup}}} \right) \\
\overset{(2)}{\Longrightarrow} &\mathbb{P}^n \otimes \mathbb{P}^n \left((X_j, Y_j)_{j=1}^n : \text{MMD}(X,Y) > n^{-\frac{1}{2}} \left(\sqrt{-\log\left(1-\alpha\right) 4\kappa_{\text{sup}}} \right) + \sqrt{2\kappa_{\text{sup}}} \right) < 1 - \alpha \\
\overset{(3)}{\Longrightarrow} &\mathbb{P}^n \otimes \mathbb{P}^n \left((X_j, Y_j)_{j=1}^n : \text{MMD}(X,Y) \leq n^{-\frac{1}{2}} \left(\sqrt{-\log\left(1-\alpha\right) 4\kappa_{\text{sup}}} \right) + \sqrt{2\kappa_{\text{sup}}} \right) \geq \alpha
\end{aligned}
$$

where:

(1) $\varepsilon' = \varepsilon + \sqrt{2\kappa_{\text{sup}}/n}$.

(2) $1 - \alpha = \exp\left(-\frac{n\left(\varepsilon' - \left(\frac{2\kappa_{\text{sup}}}{n}\right)\right)^2}{4\kappa_{\text{sup}}} \right)$ and $\varepsilon` = n^{-\frac{1}{2}} \left(\sqrt{-\log\left(1-\alpha\right) 4\kappa_{\text{sup}}} + \sqrt{2\kappa_{\text{sup}}} \right)$.

(3) Take the complementary event.

Now,

$$
\begin{aligned}
&q_n^\alpha = n^{-\frac{1}{2}} \left(\sqrt{-\log\left(1-\alpha\right) 4\kappa_{\text{sup}}} \right) + \sqrt{2\kappa_{\text{sup}}} \\
\Rightarrow &n = (q_n^\alpha)^{-2} \left(\sqrt{-4\kappa_{\text{sup}} \log\left(1-\alpha\right)} + \sqrt{2\kappa_{\text{sup}}} \right)^2 \\
\Rightarrow &\log(n) = -2\log(q_n^\alpha) + 2\log\left(\sqrt{-4\kappa_{\text{sup}} \log\left(1-\alpha\right)} + \sqrt{2\kappa_{\text{sup}}} \right).
\end{aligned}
$$

concluding the proof. □

Remark. Altohugh theoretically sound, using the abovementioned bound instead of MMD leads to excessively conservative estimates for n^*, roughly one order of magnitude greater than the empirical estimate.

A.3.3 Pseudocode for the algorithm

Algorithm 1 Compute n_N^* from preliminary study

Require: α^* ▷ desired generalizability
Require: δ^* ▷ similarity threshold on rankings
Require: $\mathcal{Q}$ ▷ research question, $\mathcal{Q} = (A, C, I_{\text{atv}}, \kappa)$
Require: N ▷ size of preliminary study
Require: n_{max} ▷ maximum sample size to compute MMD
Require: n_{rep} ▷ number of repetitions to compute MMD

procedure ESTIMATENSTAR($\alpha^*, \delta^*, \mathcal{Q}, N, n_{\text{max}}, n_{\text{rep}}$)
  $\varepsilon^* \leftarrow$ compute ε^* from δ^* ▷ cf. Appendix A.3
  sample $\{\mathbf{c}_j\}_{j=1}^N \overset{\text{iid}}{\sim} C$
  $n_{\text{max}} \leftarrow \min\{n_{\text{max}}, [N/2]\}$ ▷ we need two disjoint samples of size n_{max} from $\{\mathbf{c}_j\}_{j=1}^N$
  for $n = 1 \ldots n_{\text{max}}$ **do**
    mmds $\leftarrow$ empty list
    for $n = 1 \ldots n_{\text{rep}}$ **do**
      sample without replacement $(\mathbf{c}_j)_{j=1}^{2n_{\text{max}}} \sim \{\mathbf{c}_j\}_{j=1}^N$
      $\mathbf{x} \leftarrow (\mathbf{c}_j)_{j=1}^{n_{\text{max}}}$ ▷ split the disjoint samples
      $\mathbf{y} \leftarrow (\mathbf{c}_j)_{j=n_{\text{max}}}^{2n_{\text{max}}}$
      append MMD $(\mathbf{x}, \mathbf{y})$ to mmds
    end for
    $\varepsilon_n^{\alpha^*} \leftarrow \alpha^*$-quantile of mmds
  end for
  fit a linear regression $\log(n) = \beta_1 \log\left(\varepsilon_n^{\alpha^*}\right) + \beta_0$
  $n_N^* \leftarrow \beta_1 \log(\varepsilon^*) + \beta_0$
  return n_N^*
end procedure

procedure RUNEXPERIMENTS($\alpha^*, \delta^*, \mathcal{Q}, n_{\text{max}}, n_{\text{rep}}$, step)
  $N \leftarrow$ step
  while $n^* > N$ **do**
    sample $\{\mathbf{c}_j\}_{j=1}^N \overset{\text{iid}}{\sim} C$
    $n^* \leftarrow$ ESTIMATENSTAR($\alpha^*, \delta^*, \mathcal{Q}, N, n_{\text{max}}, n_{\text{rep}}$)
    $N \leftarrow N +$ step
  end while
end procedure

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: In order, our claims are: the formalization in Section 3; the definition generalizability in Section 4; the algorithm for study size in Section 4.3, the case studies in Section 5, and we provide a link to the anonymized GitHub repository for the module.

   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: In Section 6.

   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**

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

Answer: [Yes]

Justification: The proofs are in the Appendix.

Guidelines:

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

4. **Experimental Result Reproducibility**

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

Answer: [Yes]

Justification: On GitHub.

Guidelines:

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

5. **Open access to data and code**

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

Answer: [Yes]

Justification: On GitHub.

Guidelines:

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

6. **Experimental Setting/Details**

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

Answer: [Yes]

Justification: In Section 5.

Guidelines:

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

7. **Experiment Statistical Significance**

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

Answer: [Yes]

Justification: The boxplots in Section 5 show the variability for the choice of fixed factors.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main 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: The analysis we showcase in Section 5 executes very fast, requiring in total less than 4 hours on a standard office laptop.

   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:

   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: [NA]

    Justification:

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [NA]

    Justification:

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: To the best of our knowledge, we referenced all sources in the appropriate way.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

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

    Answer: [Yes]

    Justification: Our Python module is documented on GitHub.

    Guidelines:

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

14. **Crowdsourcing and Research with Human Subjects**

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

    Answer: [NA]

    Justification:

    Guidelines:

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

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

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

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.