TOWARDS INFERENTIAL REPRODUCIBILITY OF MACHINE LEARNING RESEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Non-determinism in deep learning and the consequential randomness and variability in performance evaluation has spawned attempts to foster reproducibility of SOTA benchmark results by sharing data, code, and meta-parameter settings. In this paper, we propose to shift from the goal of duplicating a SOTA training result without any changes to a new type of reproducibility called inferential reproducibility that treats performance variation depending on data characteristics, meta-parameter settings, and their interactions as an inherent and interesting feature of non-deterministic deep learning, not as a bug that needs to be resolved. We propose to answer questions of inferential reproducibility by classical statistical methods: We show how to design a linear mixed effects model (LMEM) to analyze performance evaluation scores of machine learning algorithms, and to conduct statistical inference on the interpretable parameters of this model with a generalized likelihood ratio test (GLRT). This approach allows us to efficiently assess statistical significance of performance differences between models by simultaneously acknowledging for variability in meta-parameters and data. Furthermore, performance differences conditional on data properties can be assessed, and a variance component analysis (VCA) can be performed to reveal the contribution of meta-parameters to overall variance. Lastly, a reliability coefficient can be computed to assess the general robustness of the model. Code (R and Python) and sample applications of our tools are publicly available.[1]

1 INTRODUCTION

A typical research project in machine learning starts with optimizing a model on given training data, tuning meta-parameters on development data, and ends with evaluating the model using a standard automatic evaluation metric on benchmark test data. For example, a neural machine translation project could use the train-dev-test split of a benchmark parallel dataset provided at `paperswithcode.com` and claim a new SOTA result if a performance difference in BLEU score of a commonly accepted magnitude is achieved over the previous best score published on the leaderboard. By adding the program code (and, if necessary, new data) and publicly sharing meta-parameter settings (following reproducibility checklists[2]), the game can be re-opened to new competitors. This train-dev-test paradigm has greatly fostered research progress in many areas, and allows the researcher to happily focus on improving model performance.

Unfortunately, the party is spoiled by the inherent non-determinism of deep learning that lurks behind randomness in weight initialization, dropout, data shuffling and batching in non-convex optimization (Dauphin et al., 2014; D'Amour et al., 2020), non-determinism due to variations in model architecture and meta-parameter settings (Lucic et al., 2018; Henderson et al., 2018), non-determinism due to pre-processing variants and data splits (Gorman & Bedrick, 2019; Søgaard et al., 2021), and non-determinism due to differences in available computational budget (Strubell et al., 2019; Dodge et al., 2019). The problem caused by non-determinism is that slight changes in training settings can reverse relations between baseline and SOTA (Reimers & Gurevych, 2017; Melis

[1] `https://www.anonymized.com`

[2] See, for example, the checklists used at recent issues of NeurIPS (`https://neurips.cc/public/guides/PaperChecklist`) or ACL (`https://aclrollingreview.org/responsibleNLPresearch/`).

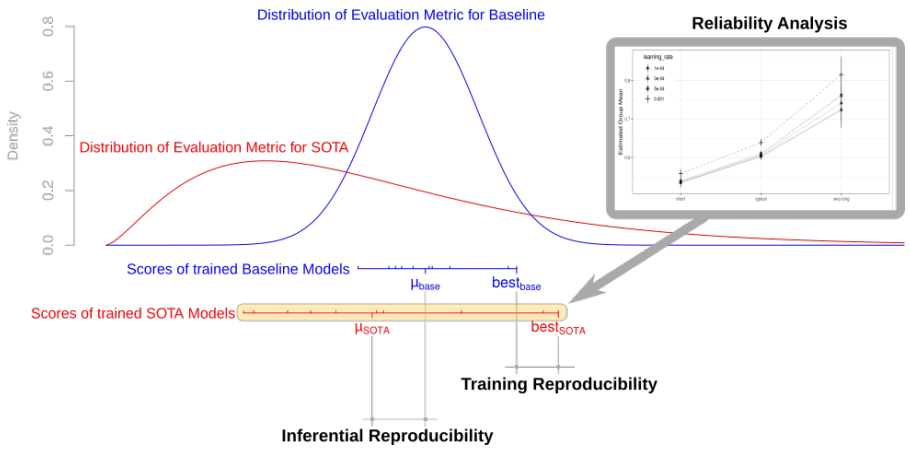

Figure 1: Performance comparison between baseline (blue curve) and SOTA (red curve) with respect to training reproducibility (best baseline versus best SOTA result) and inferential reproducibility (comparison of estimated expected performance under meta-parameter variation). Here relations are reversed if meta-parameter variation is taken into account, requiring a closed look with the magnifying glass of a reliability analysis.

et al., 2018). This has led to claims of a "reproducibility crisis" in AI (Hutson, 2018), and consequently to considerable efforts to foster reproducibility of SOTA benchmark results in various areas of machine learning by raised standards of sharing data, code, and meta-parameter settings (Heil et al., 2021; Pineau et al., 2021; Lucic et al., 2022).

In this paper, we take a different stance. We raise the heretical question of whether reproducing the best training result under exactly the same circumstances is actually an attempt at solving the wrong problem. We believe that no researcher is genuinely interested in reproducing the top score of the leaderboard found by extensive meta-parameter search on a particular dataset. Instead, the real interest is in understanding the strengths and weaknesses of a SOTA model, across a distribution of datasets, under potentially different meta-parameter settings, with the ultimate goal of an application of that model to new data. Duplication of a SOTA result under exactly the same circumstances is neither a necessary nor a sufficient condition for this goal. On the one hand, a successful duplication of a SOTA result on a benchmark does not guarantee generalization to new data. On the other hand, a model might generalize well to new data even if the exact SOTA result cannot be duplicated because of differences in computational budget.

We would like to promote a new type of reproducibility called *inferential reproducibility* that acknowledges variability due to meta-parameter settings and data properties as an inherent and interesting feature of non-deterministic machine learning, not as a bug that needs to be resolved. It asks the following question: Can qualitatively similar conclusions be drawn from an independent replication of a study? Translated to the area of machine learning, inferential reproducibility corresponds to a comparative evaluation of machine learning algorithms themselves, by asking the question which conclusions about the advantages of a SOTA model over a baseline can be drawn when applying them to new datasets under potentially required new settings of meta-parameters. We propose statistical methods to analyze inferential reproducibility. These methods are based on *linear mixed effects models (LMEMs)* fitted to performance evaluation scores of machine learning algorithms. First, we conduct a *generalized likelihood ratio test (GLRT)* of the likelihood ratio of nested LMEMs trained on the performance scores of competing machine learning algorithms on the same test data. This allows us to efficiently assess statistical significance of performance differences between algorithms, while simultaneously acknowledging for randomness in meta-parameters and data. This principle can easily be extended to assess performance differences conditional on data properties. Furthermore, a *variance component analysis (VCA)* facilitates a nuanced assessment of variation in performance estimates. Lastly, a *reliability coefficient* can be computed to assess the general robustness of the model by the ratio of substantial variance out of total variance.

2 A SCHEME FOR ANALYZING INFERENTIAL REPRODUCIBILITY

Figure 1 shows the distributions of performance evaluation scores of a baseline system (blue curve) and a SOTA system (red curve). The depicted scenario is one where the best evaluation result for the SOTA model is located in the long tail of the score distribution and thus only found by extensive meta-parameter search. We define the assessment of statistical significance of the difference of the best SOTA results against that of the best score of the baseline system as the task of **training reproducibility**.[3] However, such a test fails to assess the systematic aspect of comparing the machine learning algorithms themselves. We define this to be the task of **inferential reproducibility**.[4] This can be achieved by a significance test on the system effect parameter of an LMEM that estimates the means of the distributions of performance scores obtained under meta-parameter variation of SOTA and baseline model, and can be conditioned on data properties. In the depicted scenario, the relations between SOTA and baseline are reversed if meta-parameter variation is taken into account. LMEMs allow us to conduct a deeper analysis to quantify the sources of randomness and variability in performance evaluation. This is done by a **reliability analysis** (shown by the magnifying glass) that analyzes the variance contributed by meta-parameters and their interaction with data properties.

3 LINEAR MIXED EFFECTS MODELS

A linear mixed effects model (LMEM) is an extension of a standard linear model that allows a rich linear structure in the random component of the model, where effects other than those that can be observed exhaustively (so-called *fixed effects*) are treated as a random samples from a larger population of normally distributed random variables (so-called *random effects*).

Given a dataset of N input-output pairs $\{(\mathbf{x}^n, y^n)\}_{n=1}^N$, the general form of an LMEM is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{X}$ is an $(N \times k)$-matrix and $\mathbf{Z}$ is an $(N \times m)$-matrix, called model- or design-matrices (both are known), which relate the unobserved vectors $\boldsymbol{\beta}$ and $\mathbf{b}$ to $\mathbf{Y}$. $\boldsymbol{\beta}$ is a k-vector of fixed effects and $\mathbf{b}$ is an m-dimensional random vector called the random effects vector. $\boldsymbol{\epsilon}$ is an N-dimensional vector called the error component. The random vectors are assumed to have the following distributions:

$$\mathbf{b} \sim \mathcal{N}(0, \psi_\theta),\ \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{\Lambda}_\theta), \tag{2}$$

where ψ_θ and $\boldsymbol{\Lambda}_\theta$ are covariance matrices parameterized by the vector θ.

The most common application of LMEMs is to model complex covariance structures in the data when the usual i.i.d. assumptions fail to be applicable. This is the case for repeated or grouped, and thus non-independent, measurements such as multiple ratings of same items and same subjects in psycho-linguistic experiments. LMEMs have become popular in this area due to their flexibility (Baayen et al., 2008; Bates et al., 2015), and have even been credited as candidates to replace ANOVA (Barr et al., 2013). The price for this flexibility is an elaborate estimation methodology for which we refer the reader to further literature (Pinheiro & Bates, 2000; McCulloch & Searle, 2001; West et al., 2007; Demidenko, 2013; Wood, 2017).

4 GENERALIZED LIKELIHOOD RATIO TESTS W/ AND W/O MEASUREMENT VARIATION

Let us assume our goal is to test the statistical significance of an observed performance difference between a baseline and a SOTA system. For the sake of concreteness, let us assume we are comparing Natural Language Processing (NLP) models on a benchmark test set of gold standard sentences. In order to conduct a generalized likelihood ratio test (GLRT) for this purpose, we need to fit two

[3] The term was coined by Leventi-Peetz & Östreich (2022) and corresponds to Drummond (2009)'s replicability, i.e., a duplication of results without any changes.

[4] This term was coined by Goodman et al. (2016) and corresponds to Drummond (2009)'s reproducibility, i.e., an independent reproduction of a study that does require changes. In order to avoid to contribute to the terminological confusion in this area (see Plesser (2018)), we stick to the terms training reproducibility and inferential reproducibility in the following.

LMEMs on the performance evaluation data of baseline and SOTA system which analyze the data differently, and compare their likelihood ratio. Let us further assume an experimental design where variants of the baseline and SOTA models, corresponding to different meta-parameter configurations during training, are evaluated on the benchmark data. Simple linear models are a suboptimal choice to analyze this experiment since they are based on the assumption that each system was evaluated once on a disjoint set of sentences. This would force us to average over variants, thereby losing useful information contained in the clusters of repeated measurements of the same test input.

LMEMs allow us to better reflect this design and to leverage its statistical benefits by adding a random effect b_s for each sentence in our evaluation model. Such a model decomposes the total variance of the evaluation score into three blocks: systematic variance due to the fixed effects of the model, variance due to sentence heterogeneity, and unexplained residual variance. This allows us to reduce the as of yet unaccounted residual variance by attributing a variance component σ_s^2 to variance between sentences. If we think of the residual error as noise that masks the signal of measured performance scores, we can effectively perform a noise reduction that increases the power of our tests to detect significant differences.

A straightforward technique to implement statistical significance tests using LMEMs is the so-called *nested models* setup (Pinheiro & Bates, 2000). First we train an LMEM that doesn't distinguish between systems. This restricted model

$$m_0 : Y = \beta + b_s + \epsilon_{res} \tag{3}$$

specifies a common mean β for both systems as fixed effect, and a sentence-specific deviation b_s as random effect with variance σ_s^2, and a residual error ϵ_{res} with variance σ_{res}^2 for the performance scores Y. It represents the null hypothesis that there is no difference between systems. This model is compared to a more general model that allows different means for baseline and SOTA scores:

$$m_1 : Y = \beta + \beta_c \cdot \mathbb{I}_c + b_s + \epsilon_{res} \tag{4}$$

This model includes an indicator function $\mathbb{I}_c$ to activate a fixed effect β_c that represents the deviation of the competing SOTA model from the baseline mean β when the data point was obtained by a SOTA evaluation. The restricted model is a special case of this model (thus "nested" within the more general model) since it can be obtained by setting β_c to zero. Let ℓ_0 be the likelihood of the restricted model, and ℓ_1 be the likelihood of the more general model, the intuition of the likelihood ratio test is to reject the null hypothesis of no difference between systems if the statistic

$$\lambda = \frac{\ell_o}{\ell_1} \tag{5}$$

yields values close to zero.

The incorporation of a random sentence effect b_s introduces a pairing of systems on the sentence level that corresponds to standard pairwise significance tests. However, clustering at the sentence level allows accounting for arbitrary kinds of uncertainty introduced by the random nature of the training process. This setup is thus not only suitable for pairwise comparisons of best baseline and best SOTA model in order to test training reproducibility, but it also allows incorporating broader variations induced by meta-parameter settings of baseline and SOTA systems, thus making it suitable to test inferential reproducibility.

A further distinctive advantage of GLRTs based on LMEMs is that this framework allows analyzing significance of system differences conditional on data properties. For example, we could extend models m_0 and m_1 by a fixed effect β_d modeling a test data property d like readability of an NLP input sequence, or rarity of the words in an input sequence, and by an interaction effect β_{cd} allowing to assess the expected system performance for different levels of d. The enhanced model

$$m_1' : Y = \beta + \beta_d \cdot d + (\beta_c + \beta_{cd} \cdot d) \cdot \mathbb{I}_c + b_s + \epsilon_{res} \tag{6}$$

would then be compared to a null hypothesis model of the form

$$m_0' : Y = \beta + \beta_d \cdot d + b_s + \epsilon_{res}. \tag{7}$$

For more information on extensions of GLRTs for multiple comparisons and on their asymptotic statistics we refer the reader to further literature (van der Vaart, 1998; Pinheiro & Bates, 2000; Pawitan, 2001; Davison, 2003; Larsen & Marx, 2012).

5 VARIANCE COMPONENT ANALYSIS AND RELIABILITY COEFFICIENTS

The main goal of a reliability analysis in the context of a reproducibility study is to quantify and analyze the sources of randomness and variability in performance evaluation, and to quantify the robustness of a model in a way that allows to draw conclusions beyond the concrete experiment. The first goal can be achieved by performing a variance component analysis (VCA). For example, let us assume we want to specify a model for performance evaluation scores that besides a global mean μ specifies random effects to account for variations in the outcome Y specific to different sentences s and specific to different settings of a regularization parameter r. A tautological decomposition of the response variable into the following four components can be motivated by classical ANOVA theory (Searle et al., 1992; Brennan, 2001):

$$Y = \mu + (\mu_s - \mu) + (\mu_r - \mu) + (Y - \mu_s - \mu_r + \mu). \tag{8}$$

The components of the observed score Y for a particular regularization setting r on a single sentence s are the grand mean μ of the observed evaluation score across all levels of regularization and sentences; the deviation $\nu_s = (\mu_s - \mu)$ of the mean score μ_s for a sentence s from the grand mean μ; the deviation $\nu_r = (\mu_r - \mu)$ of the mean score μ_r for a regularization setting r from the grand mean μ; and the residual error, reflecting the deviation of the observed score Y from what would be expected given the first three terms. Except for μ, each of the components of the observed score varies from one sentence to another, from one regularization setting to another, and from one regularization-sentence combination to another. Since these components are uncorrelated with each other, the total variance $\sigma^2(Y - \mu)$ can be decomposed into the following *variance components*:

$$\sigma^2(Y - \mu) = \sigma_s^2 + \sigma_r^2 + \sigma_{res}^2, \tag{9}$$

where σ_s^2 and σ_r^2 denote the variance due to sentences and regularization settings, and σ_{res}^2 denotes the residual variance component including the variance due to interaction of s and r.

Let $\nu_f = \mu_f - \mu$ denote a deviation from the mean for a facet[5] f whose contribution to variance we are interested in. Instead of estimating the corresponding variance components σ_f by ANOVA expected mean square equations, we use LMEMs to model each ν_f as a component of the random effects vector $\mathbf{b}$ in equation 2, and model each corresponding variance component σ_f^2 as an entry of the diagonal variance-covariance matrix ψ_θ in equation 2. Besides greater flexibility in estimation[6], LMEMs also allow analyzing the interaction of meta-parameters and data properties. This can be achieved, for example, by changing the random effect b_r to a fixed effect β_r, and by adding a fixed effect β_d modeling test data characteristics, and an interaction effect β_{rd} modeling the interaction between data property d and choice of meta-parameter r.

The final ingredient of a reliability analysis is the definition of a coefficient that relates variance components to each other, instead of inspecting them in isolation. The key concept is the so-called *intra-class correlation coefficient (ICC)*, dating back to Fisher (1925). A fundamental interpretation of the ICC is as a measure of the proportion of variance that is attributable to substantial variance, i.e., to variance between the objects of measurement. The name of the coefficient is derived from the goal of measuring how strongly objects in the same class are grouped together in a measurement. Following Brennan (2001), we can define a concrete reliability coefficient, denoted by φ, for our application scenario. In our case, objects of interest are test sentences s, and substantial variance is variance σ_s^2 between sentences. Assume facets $f_1, f_2, \ldots$ and selected interactions $sf_1, sf_2, f_1f_2, \ldots$. Then the reliability coefficient φ is computed by the ratio of substantial variance σ_s^2 to the total variance, i.e., to itself and the error variance σ_Δ^2 that includes variance components for all random effects and selected interactions of random effects:

$$\varphi = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_\Delta^2}, \text{ where } \sigma_\Delta^2 = \sigma_{f_1}^2 + \sigma_{f_2}^2 + \ldots + \sigma_{sf_1}^2 + \sigma_{sf_2}^2 + \ldots + \sigma_{f_1f_2}^2 + \cdots + \sigma_{res}^2. \tag{10}$$

[5]In the psychometric approach to reliability of Brennan (2001), the conditions of measurement that contribute to variance in the measurement besides the objects of interest are called *facets* of measurement. In our running NLP example, the objects of interest in our measurement procedure are the sentences. They are the essential conditions of measurement. The only facet of measurement in this example are the regularization settings, while the objects of interest are not usually called a facet.

[6]Among the many advantages of using LMEMs to estimate variance components is that the same model structure can be used for designs that are special cases of the fully crossed design, and the elegant handling of missing data situations. See Baayen et al. (2008); Barr et al. (2013); Bates et al. (2015) for further discussions on the advantages of LMEMs over mixed-model ANOVA estimators.

Table 1: Text summarization results (Rouge-1/2/L) for baseline (BART) and SOTA (BART+R3F) reported in Aghajanyan et al. (2021).

	CNN/DailyMail	RedditTIFU
baseline	44.16/21.28/40.90	24.19/8.12/21.31
SOTA	44.38/21.53/41.17	30.31/10.98/24.74

Based on this definition, reliability of a performance evaluation across replicated measurements is assessed as the ratio by which the amount of substantial variance outweighs the total error variance. That is, a performance evaluation is deemed reliable if most of the variance is explained by variance between sentences and not by variance within a sentence, such as variance caused by random regularization settings or by residual variance due to unspecified facets of the measurement procedure. Naturally, different assumptions on thresholds on this ratio will lead to different assessments of reliability. A threshold of 80% is used, for example, by Jiang (2018). Values less than 50%, between 50% and 75%, between 75% and 90%, and above 90%, are indicative of poor, moderate, good, and excellent reliability, respectively, according to Koo & Li (2016).

More information on VCA and ICCs can be found in Shrout & Fleiss (1979); Searle et al. (1992); McGraw & Wong (1996); Brennan (2001); Webb et al. (2006).

6 A Worked-Through Example

We exemplify the methods introduced above on an NLP example from the `paperswithcode.com` open resource, namely the BART+R3F fine-tuning algorithm presented by Aghajanyan et al. (2021) for the task of text summarization, evaluated on the CNN/DailyMail (Hermann et al., 2015) and RedditTIFU (Kim et al., 2019) datasets.

BART+R3F was listed as SOTA on these datasets at the time of writing. It uses an approximate trust region method to constrain updates on embeddings f and classifier g during fine-tuning in order not to forget the original pre-trained representations. This is done by minimizing a task loss $\mathcal{L}(\theta)$ regularized by the Kullback-Leibler distance on normally or uniformly distributed parameters:

$$\mathcal{L}(\theta) + \lambda KL(g \cdot f(x) || g \cdot f(x+z)) \text{ s.t. } z \sim \mathcal{N}(0, \sigma^2 I) \text{ or } z \sim \mathcal{U}(-\sigma, \sigma). \tag{11}$$

The first question we want to answer is that of training reproducibility – is the result difference between baseline and new SOTA reproducible on the data[7] and the code[8] linked on the repository, and under the meta-parameter and preprocessing setup reported in the paper. As baseline we take a pre-trained BART-large[9] model (Lewis et al., 2020). The Rouge-1/2/L[10] (Lin & Hovy, 2003) results for the text summarization task reported in Aghajanyan et al. (2021) are shown in Table 1.

Let us first look at the results on the CNN/DailyMail dataset. The paper gives detailed meta-parameter settings for the text summarization experiments, but reports final results as maxima over training runs started from 10 unknown random seeds. Furthermore, the regularization parameter is specified as a choice of $\lambda \in [0.001, 0.01, 0.1]$, and the noise type as a choice from $[\mathcal{U}, \mathcal{N}]$. Using the given settings, we started the BART+R3F code from 5 new random seeds and the BART-large baseline from 18 random seeds on 4 Nvidia Tesla V100 GPUs each with 32 GB RAM and a update frequency of 8. All models were trained for 20-30 epochs using a loss-based stopping criterion. Searching over the given meta-parameter choices, we obtained a training reproducibility result given in Table 2: We find significant improvements of the best SOTA model over the best baseline with respect to all Rouge-X metrics (the difference baseline - SOTA is negative). However, the effect sizes (standardized mean difference between evaluation scores) are small.

Let us next inspect significance conditional on data properties. We quantify properties of summarization inputs by word rarity (Platanios et al., 2019), i.e., the negative logarithm of the empiri-

[7] `https://github.com/abisee/cnn-dailymail`, `https://github.com/ctr4si/MMN`
[8] `https://github.com/facebookresearch/fairseq/tree/main/examples/rxf`
[9] `https://github.com/facebookresearch/fairseq/tree/main/examples/bart`
[10] `https://github.com/google-research/google-research/tree/master/rouge`

Table 2: Significance of result difference baseline-SOTA on CNN/DailyNews.

	baseline	SOTA	p-value	effect size
Rouge-1	44.09	44.41	< 0.0001	-0.101
Rouge-2	21.13	21.44	< 0.0001	-0.080
Rouge-L	40.81	41.16	< 0.0001	-0.105

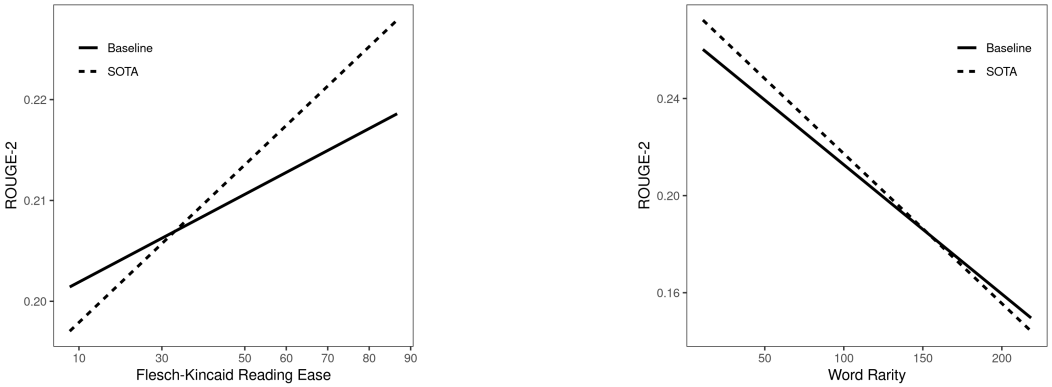

Figure 2: Interaction of Rouge-2 of baseline (solid) and SOTA (dashed) with readability (left) and word rarity (right).

Table 3: Significance of baseline-SOTA on CNN/DailyNews under meta-parameter variation.

	baseline	SOTA	p-value	effect size
Rouge-1	44.15	42.21	< 0.0001	0.390
Rouge-2	21.26	19.64	< 0.0001	0.301
Rouge-L	40.84	38.53	< 0.0001	0.531

cal probabilities of words in summary inputs, where higher values mean higher rarity. Furthermore, we calculate readability (Kincaid et al., 1975) of summary inputs by calculating the ratio of words/sentences and syllables/word. Readability scores are in principle unbounded, however, an interpretion scheme exists for the range from 0 (difficult) to 100 (easy). An analysis of significance conditional on data properties can be seen as first step of inferential reproducibility. The interaction plots given in Figure 2 show a significant difference in performance slope for Rouge-2 with respect to ease of readability, where the performance of the best SOTA system increases faster than that of the best baseline for easier inputs (left plot). Also, a significant difference in Rouge-2 with respect to word rarity is seen where the best SOTA model is better than the best baseline for inputs with lower word rarity (right plot).

The next question of inferential reproducibility is whether the results given above are robust against meta-parameter variations, and which meta-parameters are most important in order to achieve the best result. We inspect the original grid of meta-parameter configurations of the SOTA model, given by crossing the given choices of meta-parameters with each other, yielding 3 $\lambda \times$ 2 noise distributions $\times$ 5 random seeds = 30 configurations. As shown in Table 3, the relations between SOTA and baseline are turned around (the difference baseline - SOTA is positive) showing significant wins of baseline over SOTA at medium effect size.

Since the performance variation of the baseline model over 18 random seeds was negligible (standard deviations $< 0.2\%$ for Rouge-X scores), we conduct a reliability analysis of the SOTA model in order to reveal the culprit for this performance loss. The variance component analysis in Table 4 shows that the variance contributions due to variation in random seeds or choice of noise distribution are negligible. However, in all three cases the largest contribution to variance is due to the regularization parameter λ. The percentage of variance due to objects of interest, here summaries,

Table 4: Variance component analysis for Rouge-1 (top), Rouge-2 (middle), and Rouge-L (bottom) estimates.

Variance component v	Variance σ_v^2	Percent
summary_id	0.00923	**55.8**
lambda	0.00254	15.0
random_seed	0.00012	0.7
noise_distribution	0.00005	0.3
residual	0.00464	27.1
summary_id	0.00992	**62.7**
lambda	0.00131	8.3
random_seed	0.00008	0.5
noise_distribution	0.00003	0.2
residual	0.00449	28.3
summary_id	0.00875	**47.9**
lambda	0.00519	28.4
random_seed	0.00004	0.2
noise_distribution	0.00001	0.1
residual	0.00428	23.4

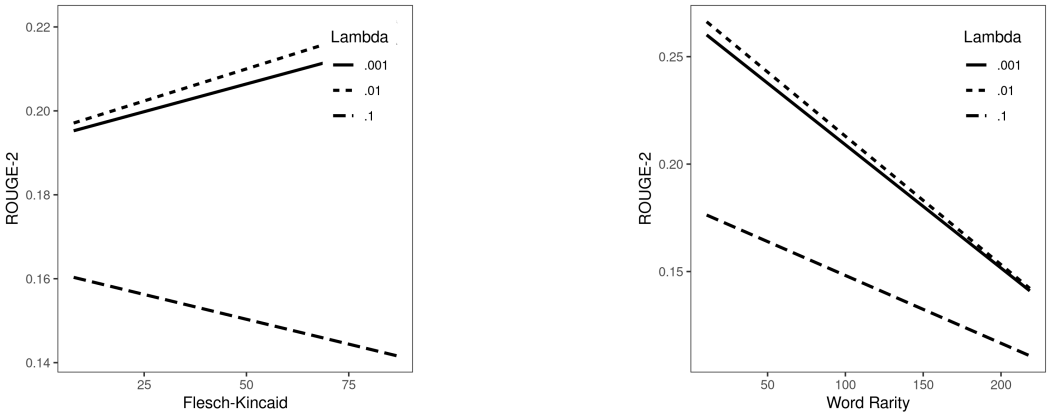

Figure 3: Interaction of Rouge-2 of SOTA for different values of regularization parameter λ with readability (left) and word rarity (right).

can readily be interpreted as reliability coefficient φ, yielding moderate reliability for performance evaluation under Rouge-1 and Rouge-2 (φ between 50% and 75%) and poor reliability for evaluation under Rouge-L (φ below 50%).

An inspection of the interaction of data properties with the regularization parameter is given in Figure 3. The interaction plots show a significant drop in Rouge-2 performance of the SOTA model for the regularization parameter $\lambda = 0.1$ across all levels of reading ease (top plot) and for rare words (bottom plot).

Let us inspect the results on the RedditTIFU dataset next. These data are interesting since they are much harder to read (mean readability score of -348.9), however, a reproducibility analysis on the RedditTIFU dataset was hampered by the fact that the train/dev/test split for RedditTIFU data (long version) was not given on `paperswithcode.com` nor reported in the paper or the code. We used the split[11] provided by Zhong et al. (2020) instead. Under this data split, we found a significant improvement of the best SOTA over the best baseline at a small effect size (-0.155) only for Rouge-2. If meta-parameter variation was taken into account, the effect size was even smaller

[11] `https://paperswithcode.com/sota/text-summarization-on-reddit-tifu`

(-0.0617). There were no significant interaction effects and neglible variance contributions from meta-parameters.

In sum, this small study allows a nuanced assessment of the strengths and weaknesses of the BART+R3F model: Losing or winning a new SOTA score strongly depends on finding the sweet spot of one meta-parameter (here: λ), while the paper's goal was explicitly to reduce instability across meta-parameter settings. Performance improvements by fine-tuning are achieved mostly on easy-to-read and frequent-word inputs – these comprise less than one quarter of the CNN/Dailynews data. Lastly, the model does not seem to be robust against variations in data – under a new random split on RedditTIFU the large gains reported for the split used in the paper can no longer be achieved.

7 Related Work

One key feature of the proposed LMEM-based approach is that it unifies special-purpose significance tests for particular evaluation metrics (Dror et al. (2020), Chapter 3), for meta-parameter variations (Dror et al., 2019), and for multiple test data (Dror et al., 2017) into a single framework for hypothesis testing. A distinctive advantage of our framework compared to extensions of bootstrap (Sellam et al., 2021; Bouthillier et al., 2021) or permutation (Clark et al., 2011) tests to incorporate meta-parameter variation is that it enables analyzing significance of result differences conditional on data properties. These can be generic data properties like readability as above, or properties of combined datasets obtained from different sources like data splits, bootstrapped data, or different-domain data sets.

Variance component analysis based on ANOVA techniques has been applied to information retrieval models (Ferro & Silvello, 2016; Voorhees et al., 2017) and machine learning models in general (Hutter et al., 2014; Zimmer et al., 2020). These approaches focus on meta-parameter importance and ignore an incorporation of data variability into their analysis. We replace ANOVA methods by LMEMs for modeling and estimation (Wood, 2017) and promote the ICC-based idea of quantifying reliability by the proportion of variance attributable to the objects of interest, which to our knowledge has not been applied to machine learning before.

From a broader perspective, our work can be seen as a contribution to trustworthiness (Huang et al., 2020) and interpretability (Zhang et al., 2021) of deep neural networks, with a focus on understanding variability in the performance of a neural network depending on data characteristics, meta-parameter settings, and their interactions.

From yet another perspective, a reliability coefficient can serve as criterion for model selection, adding a novel aspect of robustness that signals that an algorithm's performance does not (or only marginally) depend on random settings of meta-parameters or arbitrary characteristics of data. Following Brennan (2001)'s generalizability theory, on which our reliability coefficient is based on, the goal is to select models with maximal reliability and minimal variation in performance measurement.

8 Discussion

In the recent discussion of a reproducibility crisis of machine learning, Drummond (2009) and Belz et al. (2021) posed the question whether training reproducibility is actually interesting and worth having. Following Goodman et al. (2016), we proposed to pursue inferential reproducibility instead. This new paradigm of reproducibility embraces the inherent variability of deep learning due to data properties and meta-parameter settings and makes it the prime focus of the analysis.

We propose to reinvigorate the capabilities of some of the oldest analysis techniques in the history of statistics – VCA and ICCs (Fisher, 1925) and GLRTs (Neyman & Pearson, 1933; Wilks, 1938) – with more modern LMEM modeling techniques and apply them to address questions of genuine interest to researchers and users like "Will the SOTA algorithm's stellar performance on the benchmark testdata lead to top performance the kinds of datasets that my customers will bring?", or more specifically "How will individual test example characteristics or particular meta-parameter settings, and their interaction with data properties, affect performance?" Our methods are readily applicable to already existing performance evaluation data obtained during meta-parameter optimization. We believe that they will be especially useful for large-scale experiments where a manual inspection of variance due to interactions of large numbers of meta-parameters and data properties is prohibitive.

REFERENCES

Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. Better fine-tuning by reducing representational collapse. In *International Conference on Learning Representations (ICLR)*, 2021.

R.H. Baayen, D.J. Davidson, and D.M. Bates. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59:390–412, 2008.

Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tilly. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3): 255–278, 2013.

Douglas Bates, Martin Mächler, Benjamin M. Bolker, and Steven C. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. A systematic review of reproducibility research in natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Online, 2021.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gael Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. *Proceedings of Machine Learning and Systems (MLSys)*, 3, 2021.

Robert L. Brennan. *Generalizability theory*. Springer, 2001.

Jonathan Clark, Chris Dyer, Alon Lavie, and Noah Smith. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, Portland, OR, 2011.

Alexander D'Amour, Katherine A. Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yi-An Ma, Cory Y. McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *CoRR*, abs/2011.03395, 2020.

Yann N. Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS)*, Montreal, Canada, 2014.

A. C. Davison. *Statistical Models*. Cambridge University Press, 2003.

Eugene Demidenko. *Mixed Models: Theory and Applications with R*. Wiley, 2013.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, 2019.

Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. Replicability analysis for natural language processing: Testing significance with multiple datasets. In *Transactions of the Association for Computational Linguistics (TACL)*, volume 5, pp. 471–486, 2017.

Rotem Dror, Segev Shlomov, and Roi Reichart. Deep dominance - how to properly compare deep neural models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, Florence, Italy, 2019.

Rotem Dror, Lotem Peled, Segev Shlomov, and Roi Reichart. *Statistical Significance Testing for Natural Language Processing*. Morgan & Claypool, 2020.

Chris Drummond. Replicability is not reproducibility: Nor is it good science. In *Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th ICML*, Montreal, Canada, 2009.

Nicola Ferro and Gianmaria Silvello. A general linear mixed models approach to study system component effects. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pisa, Italy, 2016.

Ronald A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, 1925.

Steven N. Goodman, Daniele Fanelli, and John P. A. Ioannidis. What does research reproducibility mean? *Sci Transl Med*, 8(341):1–6, 2016.

Kyle Gorman and Steven Bedrick. We need to talk about standard splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, Florence, Italy, 2019.

B.J. Heil, M.M. Hoffman, F. Markowetz, S. Lee, C.S. Greene, and S.C. Hicks. Reproducibility standards for machine learning in the life sciences. *Nature Methods*, 18:1122–1144, 2021.

Peter Henderson, Riashat Islam, Philip Bachmann, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, New Orleans, LA, USA, 2018.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS)*, Montreal, Canada, 2015.

Xiaowei Huang, Daniel Kroening, Wenjie Ruan, James Sharp, Youcheng Sun, Emese Thamo, Min Wu, and Xinping Yi. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review*, 37, 2020.

Matthew Hutson. Artificial intelligence faces reproducibility crisis. *Science*, 359(6377):725–726, 2018. doi: 10.1126/science.359.6377.725.

Frank Hutter, Holger Hoss, and Kevin Leyton-Brown. An efficient approach for assessing hyperparameter importance. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, Beijing, China, 2014.

Zhehan Jiang. Using the linear mixed-effect model framework to estimate generalizability variance components in R. *Methodology*, 14(3):133–142, 2018.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. Abstractive summarization of Reddit posts with multi-level memory networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL:HLT)*, Minneapolis, Minnesota, 2019.

J. P. Kincaid, R. P. Fishburn, R. L. Rogers, and B. S. Chissom. Derivation of new readability formulas for navy enlisted personnel. Technical report, Technical Report, Naval Air Station, Millington, TN, 1975.

Terry K. Koo and Mae Y. Li. A guideline of selecting and reporting intraclass correlations coefficients for reliability research. *Journal of Chiropratic Medicine*, 15:155–163, 2016.

Richard J. Larsen and Morris L. Marx. *Mathematical Statistics and its Applications*. Prentice Hall, fifth edition, 2012.

A. M. Leventi-Peetz and T. Östreich. Deep learning reproducibility and explainable AI (XAI). *CoRR*, abs/2202.11452, 2022.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, Online, 2020.

Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, Edmonton, Canada, 2003.

Ana Lucic, Maurits Bleeker, Samarth Bhargav, Jessica Forde, Koustuv Sinha, Jesse Dodge, Sasha Luccioni, and Robert Stojnic. Towards reproducible machine learning research in natural language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, Dublin, Ireland, 2022.

Mario Lucic, Karol Kurach, Marcin Michalski, Olivier Bousquet, and Sylvain Gelly. Are GANs created equal? A large-scale study. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS)*, Montréal, Canada, 2018.

Charles E. McCulloch and Shayle R. Searle. *Generalized, Linear, and Mixed Models*. Wiley, 2001.

Kenneth O. McGraw and S. P. Wong. Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1):30–46, 1996.

Gabor Melis, Chris Dyer, and Phil Blunsom. On the state of the art of evaluation in neural language models. In *Proceedings of the 6th Conference on Learning Representations (ICLR)*, Vancouver, BC, Canada, 2018.

J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A*, 231:289–337, 1933.

Yudi Pawitan. *In All Likelihood. Statistical Modelling and Inference Using Likelihood*. Clarendon Press, 2001.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *Journal of Machine Learning Research (JMLR)*, 22:1–20, 2021.

José C. Pinheiro and Douglas M. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer, 2000.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. Competence-based curriculum learning for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL:HLT)*, Minneapolis, Minnesota, 2019. doi: 10.18653/v1/N19-1119.

Hans E. Plesser. Reproducibility vs. replicability: A brief history of a confused terminology. *Frontiers in Neuroinformatics*, 11(76):1–4, 2018.

Nils Reimers and Iryna Gurevych. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Copenhagen, Denmark, 2017.

Shayle R. Searle, George Casella, and Charles E. McCulloch. *Variance Components*. Wiley, 1992.

Thibault Sellam, Steve Yadlowsky, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Turc, Jacob Eisenstein, Dipanjan Das, Ian Tenney, and Ellie Pavlick. The multiberts: BERT reproductions for robustness analysis. *CoRR*, abs/2106.16163, 2021.

Patrick E. Shrout and Joseph L. Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979.

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. We need to talk about random splits. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Online, 2021.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, Florence, Italy, 2019.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Ellen M. Voorhees, Daniel Samarov, and Ian Soboroff. Using replicates in information retrieval evaluation. *ACM Transactions on Information Systems*, 36(2):1–31, 2017.

Noreen M. Webb, Richard J. Shavelson, and Edward H. Haertel. Reliability coefficients and generalizability theory. *Handbook of Statistics*, 26:81–214, 2006.

Brady T. West, Kathleen B. Welch, and Andrzej T. Galecki. *Linear Mixed Models: A Practical Guide Using Statistical Software*. Chapman & Hall/CRC, 2007.

S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics*, 19:60–92, 1938.

Simon N. Wood. *Generalized Additive Models. An Introduction with R*. Chapman & Hall/CRC, second edition, 2017.

Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2021.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, Online, 2020.

Lucas Zimmer, Marius Lindauer, and Frank Hutter. Auto-pytorch tabular: Multi-fidelity metalearning for efficient and robust autodl. *CoRR*, abs/2006.13799, 2020.