# TOWARDS INFERENTIAL REPRODUCIBILITY OF MACHINE LEARNING RESEARCH

**Michael Hagmann**
Computational Linguistics
Heidelberg University, Germany
hagmann@cl.uni-heidelberg.de

**Stefan Riezler**
Computational Linguistics & IWR
Heidelberg University, Germany
riezler@cl.uni-heidelberg.de

## ABSTRACT

Reliability of machine learning evaluation — the consistency of observed evaluation scores across replicated model training runs — is affected by several sources of nondeterminism which can be regarded as measurement noise. Current tendencies to remove noise in order to enforce reproducibility of research results neglects inherent nondeterminism at the implementation level and disregards crucial interaction effects between algorithmic noise factors and data properties. This limits the scope of conclusions that can be drawn from such experiments. Instead of removing noise, we propose to incorporate several sources of variance, including their interaction with data properties, into an analysis of significance and reliability of machine learning evaluation, with the aim to draw inferences beyond particular instances of trained models. We show how to use linear mixed effects models (LMEMs) to analyze performance evaluation scores, and to conduct statistical inference with a generalized likelihood ratio test (GLRT). This allows us to incorporate arbitrary sources of noise like meta-parameter variations into statistical significance testing, and to assess performance differences conditional on data properties. Furthermore, a variance component analysis (VCA) enables the analysis of the contribution of noise sources to overall variance and the computation of a reliability coefficient by the ratio of substantial to total variance.

## 1 INTRODUCTION

Training of deep learning models utilizes randomness to improve generalization and training efficiency, thus causing an inherent nondeterminism that hampers the reliability of machine learning evaluation — the consistency of the measurement of evaluation scores across replicated training runs. Gundersen et al. (2022) list several sources of nondeterminism, e.g., implementation-level nondeterminism such as random ordering in floating-point accumulation in parallel GPU threads (Pham et al., 2021), algorithmic factors such as variations in meta-parameters and model architecture (Lucic et al., 2018; Henderson et al., 2018; D'Amour et al., 2020), or data-level factors such as variations in pre-processing and evaluation metrics (Post, 2018; Chen et al., 2022) or varying characteristics of data in different splits (Gorman & Bedrick, 2019; Søgaard et al., 2021). Zhuang et al. (2022) show that implementation-level nondeterminism is partly irreducible, leading to variability in evaluation scores even for training runs on identical data, algorithmic settings and infrastructure. Furthermore, they point out strong effects of certain types of algorithm-level nondeterminism on certain subsets of the data.

Regarding the comparison of machine learning models, minor variations in these sources of nondeterminism can have huge impact on the resulting evaluation scores and sometimes even reverse the relation between optimal results for baseline and state-of-the-art (SOTA) model (Reimers & Gurevych, 2017; Melis et al., 2018). This fact questions what can be validly learned from a typical machine learning experiment. One current answer is to foster training reproducibility[1] in the sense of an exact duplication of a state-of-the-art (SOTA) training result under exactly the same conditions. In this view, all sources of nondeterminism are regarded as noise or nuisance factors

[1]The term was coined by Leventi-Peetz & Östreich (2022) and corresponds to Drummond (2009)'s replicability.

(Forde & Paganini, 2019) that are independent of the learning signal and need to be removed (or at least reduced, even if incurring a cost in efficiency (Ahn et al., 2022)). This goal is pursued by enforcing open-source program code, publicly available data, and explicit descriptions of experimental settings, following reproducibility checklists (Heil et al., 2021; Pineau et al., 2021; Lucic et al., 2022). An unintended side effect of this approach is that the conclusions that can be drawn from such experiments are restricted to statements about a single training configuration on a single test set.

Another viewpoint is to embrace certain types of nondeterminism as inherent and irreducible conditions of measurement that contribute to variance in performance evaluation in an interesting way. Instead of attempting to remove them, we propose to analyze the various components of measurement noise, and especially their interactions with certain properties of data. Such a study can be seen to fall under the umbrella of inferential reproducibility.[2] Goodman et al. (2016) define it to refer to the drawing of qualitatively similar conclusions from either an independent replication of a study or a reanalysis of the original study. For the case of machine learning evaluation, we focus on algorithmic-level factors such as variations in meta-parameters and model architecture and on data-level factors as the main sources of non-determinism in training replications. These are usually described independently of each other. Our goal is to answer the question whether a competitor model yields improvements over a baseline across different meta-parameter settings and across different characteristics of input data, and how variations in algorithmic settings interact with varying data characteristics.

The main contribution of our paper is to show how to apply well-known statistical methods to analyze machine learning evaluations under variability in meta-parameter settings and dependent on data characteristics, with a special focus on the detection of sources of variance and their interaction with data properties. 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)* to assess statistical significance of performance differences between algorithms, while simultaneously acknowledging for variation in nondeterministic factors. While applicable to any source of nondeterminism, in this paper, we focus on meta-parameter variations as main source of variability. A key feature of our approach is the possibility to assess the significance performance differences under meta-parameter variation conditional on data properties. Second, we show how to use *variance component analysis (VCA)* to facilitate a nuanced quantitative assessment of the sources of variation in performance estimates. Lastly, we compute a *reliability coefficient* to assess the general robustness of the model by the ratio of substantial variance out of total variance. Reliability is also intimately related to the power of the significance test.

Code (R and Python) for the toolkit and sample applications are publicly available.[3]

## 2 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*).

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, 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:

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

[2]This term corresponds to Drummond (2009)'s reproducibility and was coined by Goodman et al. (2016). Instead of contributing further to the terminological confusion in this area, we refer to the brief history of this discussion in Plesser (2018).

[3]https://www.cl.uni-heidelberg.de/statnlpgroup/empirical_methods_tutorial/

where $\psi_\theta$ and $\mathbf{\Lambda}_\theta$ are covariance matrices parameterized by the vector $\theta$.

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 by 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) to analyze experimental data. The price for this flexibility is an elaborate estimation methodology for which we refer the reader to Appendix A2 of Riezler & Hagmann (2022) and to further literature (Pinheiro & Bates, 2000; McCulloch & Searle, 2001; West et al., 2007; Demidenko, 2013; Wood, 2017).

## 3 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. Furthermore, 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 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 $\sigma_s^2$ 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 $\beta$ for both systems as fixed effect, and a sentence-specific deviation $b_s$ as random effect with variance $\sigma_s^2$, and a residual error $\epsilon_{res}$ with variance $\sigma_{res}^2$ 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 $\beta_c$ that represents the deviation of the competing SOTA model from the baseline mean $\beta$ 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 $\beta_c$ to zero. Let $\ell_0$ be the likelihood of the restricted model, and $\ell_1$ 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 $\beta_d$ 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 $\beta_{cd}$ 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

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

GLRTs belong to the oldest techniques in statistics, dating back to Neyman & Pearson (1933); Wilks (1938). For more information on extensions of GLRTs for multiple comparisons and on their asymptotic statistics we refer the reader to Chapter 4 and Appendix A3 of Riezler & Hagmann (2022) and to further literature (van der Vaart, 1998; Pinheiro & Bates, 2000; Pawitan, 2001; Davison, 2003; Larsen & Marx, 2012).

## 4 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 $\mu$ 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 $\mu$ of the observed evaluation score across all levels of regularization and sentences; the deviation $\nu_s = (\mu_s - \mu)$ of the mean score $\mu_s$ for a sentence $s$ from the grand mean $\mu$; the deviation $\nu_r = (\mu_r - \mu)$ of the mean score $\mu_r$ for a regularization setting $r$ from the grand mean $\mu$; and the residual error, reflecting the deviation of the observed score $Y$ from what would be expected given the first three terms. Except for $\mu$, 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 $\sigma_s^2$ and $\sigma_r^2$ denote the variance due to sentences and regularization settings, and $\sigma_{res}^2$ 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[4] $f$ whose contribution to variance we are interested in. Instead of estimating the corresponding variance components $\sigma_f$ by ANOVA expected mean square equations, we use LMEMs to model each $\nu_f$ as a component of the random

[4]In the psychometric approach of Brennan (2001), the conditions of measurement that contribute to variance in the measurement besides the objects of interest are called *facets* of measurement. Facets comprise what we called measurement noise above. 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.

effects vector $\mathbf{b}$ in equation 2, and model each corresponding variance component $\sigma_f^2$ as an entry of the diagonal variance-covariance matrix $\psi_\theta$ in equation 2.

Besides greater flexibility in estimation[5], 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 $\beta_r$, and by adding a fixed effect $\beta_d$ modeling test data characteristics, and an interaction effect $\beta_{rd}$ modeling the interaction between data property $d$ and 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 $\varphi$, for our application scenario. In our case, objects of interest are test sentences $s$, and substantial variance is variance $\sigma_s^2$ between sentences. Assume facets $f_1, f_2, \ldots$ and selected interactions $sf_1, sf_2, f_1f_2, \ldots$. Then the reliability coefficient $\varphi$ is computed by the ratio of substantial variance $\sigma_s^2$ to the total variance, i.e., to itself and the error variance $\sigma_\Delta^2$ that includes variance components for all random effects and selected interactions of random effects:

$$\varphi = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_\Delta^2}, \tag{10}$$

where $\sigma_\Delta^2 = \sigma_{f_1}^2 + \sigma_{f_2}^2 + \ldots + \sigma_{sf_1}^2 + \sigma_{sf_2}^2 + \ldots + \sigma_{f_1f_2}^2 + \cdots + \sigma_{res}^2$. 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).

VCA and ICCs date back to the works of Fisher (1925). More information can be found in to Chapter 3 and Appendix A2 of Riezler & Hagmann (2022) and Shrout & Fleiss (1979); Searle et al. (1992); McGraw & Wong (1996); Brennan (2001); Webb et al. (2006).

## 5 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 for text summarization on these datasets on `paperswithcode.com` at the time of paper publication. 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:

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

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[6] and the code[7] linked on the repository,

[5] 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.

[6] `https://github.com/abisee/cnn-dailymail`, `https://github.com/ctr4si/MMN`

[7] `https://github.com/facebookresearch/fairseq/tree/main/examples/rxf`

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

| | CNN/DailyMail | RedditTIFU |
|---|---|---|
| bl | 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 |

Table 2: Significance of result difference baseline-SOTA on CNN/DailyNews under Rouge-1 (R1), Rouge-2 (R2) and Rouge-L (RL) evaluation.

| | bl | SOTA | $p$-value | effect size |
|---|---|---|---|---|
| R1 | $44.09$ | $44.41$ | $< 0.0001$ | $-0.101$ |
| R2 | $21.13$ | $21.44$ | $< 0.0001$ | $-0.080$ |
| RL | $40.81$ | $41.16$ | $< 0.0001$ | $-0.105$ |

and under the meta-parameter and preprocessing setup reported in the paper. As baseline we take a pre-trained BART-large[8] model (Lewis et al., 2020). The Rouge-1/2/L[9] (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 empirical 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 1 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

[8] https://github.com/facebookresearch/fairseq/tree/main/examples/bart

[9] https://github.com/google-research/google-research/tree/master/rouge

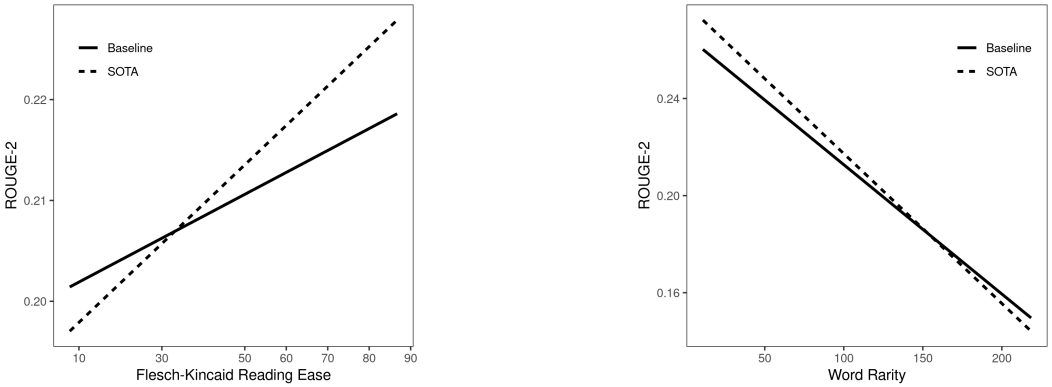


Figure 1: Interaction of Rouge-2 of baseline (solid) and SOTA (dashed) with readability (left) and word rarity (right).

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

| | baseline | SOTA | $p$-value | effect size |
|---|---|---|---|---|
| R1 | 44.15 | 42.21 | $< 0.0001$ | 0.390 |
| R2 | 21.26 | 19.64 | $< 0.0001$ | 0.301 |
| RL | 40.84 | 38.53 | $< 0.0001$ | 0.531 |

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

| Variance component $v$ | Variance $\sigma_v^2$ | 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 |

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 $\lambda$. The percentage of variance due to objects of interest, here summaries, can readily be interpreted as reliability coefficient $\varphi$, yielding moderate reliability for performance evaluation under Rouge-1 and Rouge-2 ($\varphi$ between $50\%$ and $75\%$) and poor reliability for evaluation under Rouge-L ($\varphi$ below $50\%$).

An inspection of the interaction of data properties with the regularization parameter is given in Figure 2. The interaction plots show a significant difference in Rouge-2 performance of the SOTA model between regularization parameters, where performance for $\lambda = 0.1$ is lower and decreases with increasing reading ease (top plot) and increasing word rarity (bottom plot).

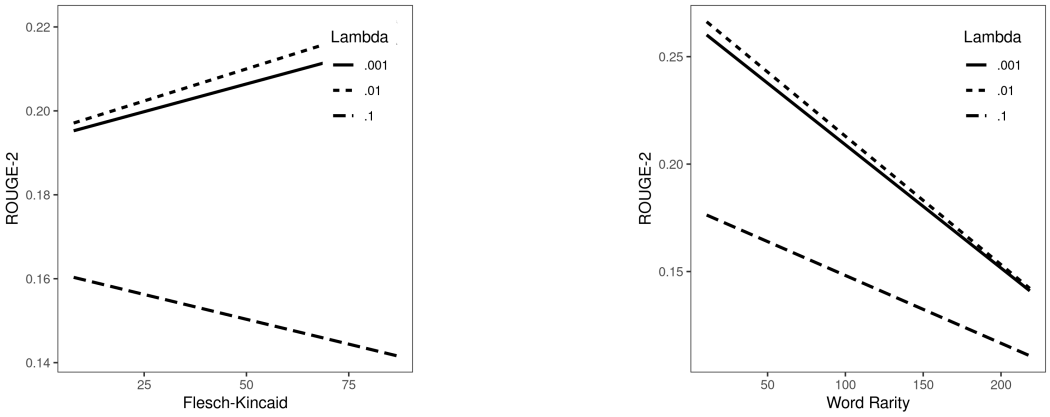


Figure 2: Interaction of Rouge-2 of SOTA for different values of regularization parameter $\lambda$ with readability (left) and word rarity (right).

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[10] 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 ($-0.0617$). 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: $\lambda$), 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. Furthermore, the optimal choice of main variance-introducing meta-parameter interacts strongly with the investigated data characteristics. 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.

## 6 Related Work

Our work must not be confused with approaches to automatic machine learning (AutoML) Zimmer et al. (2021); Habelitz & Keuper (2020) or neural architecture search (NAS) (Zoph & Le, 2017; Ebrahimi et al., 2017). While AutoML and NAS focus on efficient search of a space of meta-parameter configurations for an optimal result on a particular dataset, our interest is in analyzing the variance contributions of meta-parameters for a given meta-parameter search experiment, and especially in the interactions of meta-parameter variations with characteristics of data, with the goal of gaining insights about possible applications of a model to new data. ANOVA-like techniques have been used to analyze meta-parameter importance in the context of AutoML (Hutter et al., 2014), however, ignoring the crucial aspect of interactions of meta-parameters with data properties. Furthermore, or work is not restricted to meta-parameter variation as source of non-determinism, but can in principle be applied to any source of randomness in machine learning training.

Discussions of reproducibility problems in research date back at least to Ioannidis (2005), and for the area of machine learning at least to Hand (2006). Since then, a multitude of papers has been published on various sources of irreprocucibility in various machine learning areas (see Gundersen et al. (2022) for an overview), however, much less work has been invested in concrete techniques to

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

quantify reproducibility. Recent works try to capture reproducibility of evaluations under replicated training runs by a single statistical measure (e.g., coefficient of variation in Belz et al. (2021) or mean and confidence bounds in Lucic et al. (2018); Henderson et al. (2018)), in difference to our goal of decomposing measurement noise into different components and analyzing the interaction of noise factors with data properties. Variance component analysis based on ANOVA techniques has been applied to information retrieval models (Ferro & Silvello, 2016; Voorhees et al., 2017), however, these approaches again 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.

Special-purpose significance tests have been proposed 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). One advantage of the proposed LMEM-based approach is that it unifies these special-purpose techniques into a single framework for hypothesis testing. Furthermore, extensions of bootstrap (Sellam et al., 2022; Bouthillier et al., 2021) or permutation (Clark et al., 2011) tests have been proposed to incorporate meta-parameter variation. The distinctive advantage of our approach 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. The idea of treating test data as random effects and thus increasing the power of statistical significance testing has already proposed by Robertson & Kanoulas (2012). However, the general applicability of LMEMs and GLRTs as a unified framework for significance testing conditional on data properties and simultaneously incorporating arbitrary sources of variation has not yet been fully recognized in the wider community of machine learning research.

## 7 Discussion

Widely recognized work by applied statisticians has proposed to abandon non-confirmatory statistical significance testing, at least its role in screening of thresholds and as guarantor of reproducibility, but instead to report continuous $p$-values, along with other factors such as prior evidence (if available) (Gelman & Loken, 2014; Colquhoun, 2017; McShane et al., 2019). Our proposed use of GLRTs, VCA and ICCs aligns with these recommendations. Our focus is to use them as analysis tools of assess performance under different meta-parameter settings, dependent on characteristics of data, and to detect important sources of variance and their interactions with data properties. This allows us 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 on 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?"

Like related work that analyzes model performance under given meta-parameter configurations (Dodge et al., 2019; Strubell et al., 2019), our work is limited by the lack of standardization in the notion of a meta-parameter search. This includes human factors regarding an unclear differentiation between meta-parameters and fixed design choices for particular models. Furthermore, formal criteria on how proper ranges across model families should be defined, are lacking. Thus our work should be seen as a contribution to interpretability of deep learning experiments, not as the provision of a new decisive criterion to rank machine learning models with respect to reliability or a $p$-value under variability of meta-parameters and data. Nonetheless, our methods are readily applicable to performance evaluation data already obtained during meta-parameter optimization. They allow us to transform this usually unused data into new findings about algorithm behavior. 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.

### Acknowledgements

This research has been conducted in project SCIDATOS (Scientific Computing for Improved Detection and Therapy of Sepsis), funded by the Klaus Tschira Foundation, Germany (Grant number 00.0277.2015).

### ETHICS STATEMENT

The experiments reported in this paper are replications of published results and are not expected to raise any ethical concerns.

### REPRODUCIBILITY STATEMENT

The data, code, and meta-parameter settings for the experiments reported in this paper are documented therein and are publicly available.

## 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. URL https://openreview.net/forum?id=OQ08SN70M1V.

Kwangjun Ahn, Prateek Jain, Ziwei Ji, Satyen Kale, Praneeth Netrapalli, and Gil I. Shamir. Reproducibility in optimization: Theoretical framework and limits. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, USA, 2022.

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. doi: https://doi.org/10.1016/j.jml.2007.12.005.

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. doi: https://doi.org/10.1016/j.jml.2012.11.001.

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. doi: https://doi.org/10.18637/jss.v067.i01.

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. doi: http://dx.doi.org/10.18653/v1/2021.eacl-main.29.

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. URL https://proceedings.mlsys.org/paper/2021/file/cfecdb276f634854f3ef915e2e980c31-Paper.pdf.

Robert L. Brennan. *Generalizability theory*. Springer, 2001. doi: https://doi.org/10.1007/978-1-4757-3456-0.

Yanran Chen, Jonas Belouadi, and Steffen Eger. Reproducibility issues for bert-based evaluation metrics. *CoRR*, abs/2204.00004, 2022. doi: 10.48550/ARXIV.2204.00004.

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. URL https://aclanthology.org/P11-2031.

David Colquhoun. The reproducibility of research and the misinterpretation of $p$-values. *Royal Society Open Science*, 4(12), 2017. doi: 10.1098/rsos.171085.

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.

A. C. Davison. *Statistical Models*. Cambridge University Press, 2003. doi: https://doi.org/10.1017/cbo9780511815850.

Eugene Demidenko. *Mixed Models: Theory and Applications with R*. Wiley, 2013. doi: https://doi.org/10.1002/0471728438.

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. doi: http://dx.doi.org/10.18653/v1/D19-1224.

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. doi: http://dx.doi.org/10.1162/tacl_a_00074.

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. doi: http://dx.doi.org/10.18653/v1/P19-1266.

Rotem Dror, Lotem Peled, Segev Shlomov, and Roi Reichart. *Statistical Significance Testing for Natural Language Processing*. Morgan & Claypool, 2020. doi: https://doi.org/10.1007/978-3-031-02174-9.

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.

Sayna Ebrahimi, Anna Rohrbach, and Trevor Darrell. Gradient-free policy architecture search and adaptation. In *Proceedings of the Conference on Robot Learning (CoRL)*, Mountain View, CA, USA, 2017.

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. doi: https://doi.org/10.1145/2911451.2911530.

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

Jessica Zosa Forde and Michela Paganini. The scientific method in the science of machine learning. In *Proceedings of the ICLR 2019 Debugging Machine Learning Models Workshop*, New Orleans, LA, USA, 2019.

Andrew Gelman and Eric Loken. The statistical crisis in science. *American Scientist*, 102(6):460–465, 2014. doi: https://doi.org/10.1511/2014.111.460.

Steven N. Goodman, Daniele Fanelli, and John P. A. Ioannidis. What does research reproducibility mean? *Sci Transl Med*, 8(341):1–6, 2016. doi: https://doi.org/10.1126/scitranslmed.aaf5027.

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. doi: http://dx.doi.org/10.18653/v1/P19-1267.

Odd Erik Gundersen, Kevin Coakley, and Christine Kirkpatrick. Sources of irreproducibility in machine learning: A review. *CoRR*, abs/2204.07610, 2022. doi: 10.48550/ARXIV.2204.07610.

P. Habelitz and J. Keuper. PHS: A toolbox for parallel hyperparameter search. *CoRR*, abs/2002.11429, 2020. doi: 10.48550/ARXIV.2002.11429.

David J. Hand. Classifier Technology and the Illusion of Progress. *Statistical Science*, 21(1):1 – 14, 2006. doi: https://doi.org/10.1214/088342306000000060.

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. doi: https://doi.org/10.1038/s41592-021-01256-7.

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. doi: https://doi.org/10.1609/aaai.v32i1.11694.

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. URL https://proceedings.neurips.cc/paper/2015/file/afdec7005cc9f14302cd0474fd0f3c96-Paper.pdf.

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. URL https://proceedings.mlr.press/v32/hutter14.html.

John P. A. Ioannidis. Why most published research findings are false. *PLOS Medicine*, 2(8), 2005. doi: 10.1371/journal.pmed.0020124.

Zhehan Jiang. Using the linear mixed-effect model framework to estimate generalizability variance components in R. *Methodology*, 14(3):133–142, 2018. doi: https://doi.org/10.1027/1614-2241/a000149.

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. doi: doi={http://dx.doi.org/10.18653/v1/N19-1260},.

J. P. Kincaid, R. P. Fishburn, R. L. Rogers, and B. S. Chissom. Derivation of new readability formulas for navy enlisted personnel. 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. doi: https://doi.org/10.1016/j.jcm.2016.02.012.

Richard J. Larsen and Morris L. Marx. *Mathematical Statistics and its Applications*. Prentice Hall, fifth edition, 2012. doi: https://doi.org/10.1080/00031305.2011.645758.

A. M. Leventi-Peetz and T. Östreich. Deep learning reproducibility and explainable AI (XAI). *CoRR*, abs/2202.11452, 2022. doi: doi={10.48550/ARXIV.2202.11452},.

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. doi: http://dx.doi.org/10.18653/v1/2020.acl-main.703.

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. URL https://aclanthology.org/N03-1020.

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. doi: http://dx.doi.org/10.18653/v1/2022.acl-tutorials.2.

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. URL `https://proceedings.neurips.cc/paper/2018/file/e46de7e1bcaaced9a54f1e9d0d2f800d-Paper.pdf`.

Charles E. McCulloch and Shayle R. Searle. *Generalized, Linear, and Mixed Models*. Wiley, 2001. doi: https://doi.org/10.1002/0471722073.

Kenneth O. McGraw and S. P. Wong. Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1):30–46, 1996. doi: https://doi.org/10.1037/1082-989x.1.1.30.

Blakeley B. McShane, David Gal, Andrew Gelman, Christian Robert, and Jennifer L. Tackett. Abandon statistical significance. *The American Statistician*, 73(sup1):235–245, 2019. doi: https://doi.org/10.1080/00031305.2018.1527253.

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. URL `https://openreview.net/forum?id=ByJHuTgA-`.

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. doi: https://doi.org/10.1098/rsta.1933.0009.

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

Hung Viet Pham, Shangshu Qian, Jiannan Wang, Thibaud Lutellier, Jonathan Rosenthal, Lin Tan, Yaoliang Yu, and Nachiappan Nagappan. Problems and opportunities in training deep learning software systems: An analysis of variance. In *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, Virtual, 2021. doi: https://doi.org/10.1145/3324884.3416545.

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. doi: https://doi.org/10.1007/b98882.

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. doi: https://doi.org/10.3389/fninf.2017.00076.

Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the 3rd Conference on Machine Translation (WMT)*, Brussels, Belgium, 2018. doi: http://dx.doi.org/10.18653/v1/W18-6319.

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. doi: http://dx.doi.org/10.18653/v1/D17-1035.

Stefan Riezler and Michael Hagmann. *Validity, Reliability, and Significance: Empirical Methods for NLP and Data Science*. Springer, 2022. doi: https://doi.org/10.1007/978-3-031-02183-1.

Stephen E. Robertson and Evangelos Kanoulas. On per-topic variance in IR evaluation. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Portland, OR, USA, 2012. doi: https://doi.org/10.1145/2348283.2348402.

Shayle R. Searle, George Casella, and Charles E. McCulloch. *Variance Components*. Wiley, 1992. doi: https://doi.org/10.1002/9780470316856.

Thibault Sellam, Steve Yadlowsky, Ian Tenney, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Raluca Turc, Jacob Eisenstein, Dipanjan Das, and Ellie Pavlick. The multiBERTs: BERT reproductions for robustness analysis. In *International Conference on Learning Representations (ICLR)*, 2022. URL `https://openreview.net/forum?id=K0E_F0gFDgA`.

Patrick E. Shrout and Joseph L. Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979. doi: https://doi.org/10.1037/0033-2909.86.2.420.

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. doi: http://dx.doi.org/10.18653/v1/2021.eacl-main.156.

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. doi: http://dx.doi.org/10.18653/v1/P19-1355.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998. doi: https://doi.org/10.1017/cbo9780511802256.

Ellen M. Voorhees, Daniel Samarov, and Ian Soboroff. Using replicates in information retrieval evaluation. *ACM Transactions on Information Systems*, 36(2):1–31, 2017. doi: https://doi.org/10.1145/3086701.

Noreen M. Webb, Richard J. Shavelson, and Edward H. Haertel. Reliability coefficients and generalizability theory. *Handbook of Statistics*, 26:81–214, 2006. doi: https://doi.org/10.1016/s0169-7161(06)26004-8.

Brady T. West, Kathleen B. Welch, and Andrzej T. Galecki. *Linear Mixed Models: A Practical Guide Using Statistical Software*. Chapman & Hall/CRC, 2007. doi: https://doi.org/10.1201/9781420010435.

S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics*, 19:60–92, 1938. doi: https://doi.org/10.1214/aoms/1177732360.

Simon N. Wood. *Generalized Additive Models. An Introduction with R*. Chapman & Hall/CRC, second edition, 2017. doi: https://doi.org/10.1201/9781315370279.

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. doi: http://dx.doi.org/10.18653/v1/2020.acl-main.552.

Donglin Zhuang, Xingyao Zhang, Shuaiwen Leon Song, and Sara Hooker. Randomness in neural network training: Characterizing the impact of tooling. In *Proceedings of the 5th MLSys Conference*, Santa Clara, CA, USA, 2022. URL `https://proceedings.mlsys.org/paper/2022/file/757b505cfd34c64c85ca5b5690ee5293-Paper.pdf`.

Lucas Zimmer, Marius Lindauer, and Frank Hutter. Auto-pytorch: Multi-fidelity metalearning for efficient and robust autodl. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(9):3079–3090, 2021. doi: 10.1109/TPAMI.2021.3067763.

Barret Zoph and Quoc V. Le. Neural architecture search with reinforcement learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Toulon, France, 2017. URL https://openreview.net/forum?id=r1Ue8Hcxg.