Validity, Reliability, and Significance.

A Tutorial on Statistical Methods for Reproducible Machine Learning

Stefan Riezler and Michael Hagmann

Computational Lingustics & IWR
Heidelberg University, Germany

riezler@cl.uni-heidelberg.de

Introduction

Methodology of Machine Learning

Theory of machine learning

- Goal:
  - Learn a mathematical function to make predictions on unseen test data, based on given training data of inputs and outputs, without explicit programmed instructions on how to perform the task.
- Learning functional relationships between inputs and outputs builds on **methods of mathematical optimization**. [Bottou et al., 2018]
- Important twist: **Optimize prediction performance in expectation**, thus enabling **generalization to unseen data**.

  [von Luxburg and Schölkopf, 2011, Kawaguchi et al., 2020, Shen et al., 2021]

Practical workflow of machine learning experiments

- The **train-dev-test** paradigm:
  - Optimize a model on given training data,
  - tune meta-parameters on development data,
  - evaluate the model using a standard automatic evaluation metric on benchmark test data.
- Assume data splits to represent i.i.d. samples from a representative data population.
- Define SOTA by best achieved result, publish code, and report corresponding meta-parameter settings.

Digging Deeper: Methodology of Deep Learning

Inherent non-determinism of deep learning

- Non-convex optimization under randomness in weight initialization, dropout, data shuffling and batching.

  [Clark et al., 2011, Dauphin et al., 2014, D'Amour et al., 2020]

- Non-determinism due to variations in architecture, meta-parameter settings, pre-processing and data splits.

  [Lucic et al., 2018, Henderson et al., 2018, Post, 2018, Gorman and Bedrick, 2019, Søgaard et al., 2021]

- Non-determinism due to differences in available computational budget. [Strubell et al., 2019, Dodge et al., 2019]

The Quest for Replicability

Replicability = reproducibility of SOTA results under exactly same circumstances

- Quest for replicability fostered by sharing data, code, meta-parameter settings, e.g., on `paperswithcode.com`

  [Pineau et al., 2021, Heil et al., 2021, Lucic et al., 2022]
- **Non-determinism in deep learning is spoiling the party**
  - Slight changes in training settings can reverse relations between baseline and SOTA. [Reimers and Gurevych, 2017, Melis et al., 2018]
  - Large-scale SOTA results may be impossible to replicate, even if code and data are shared [Kaplan et al., 2020, Chowdhery et al., 2022].

Crisis Management

- Does AI face a **replicability crisis?** [Hutson, 2018]
- Or is **replicability uninteresting and not worth having**?

  [Drummond, 2009, Belz et al., 2021]
- ➡ Quest for replicability of SOTA result under exactly same circumstances is **asking the wrong question!**

An Alternative: Inferential Reproducibility

Inferential reproducibility

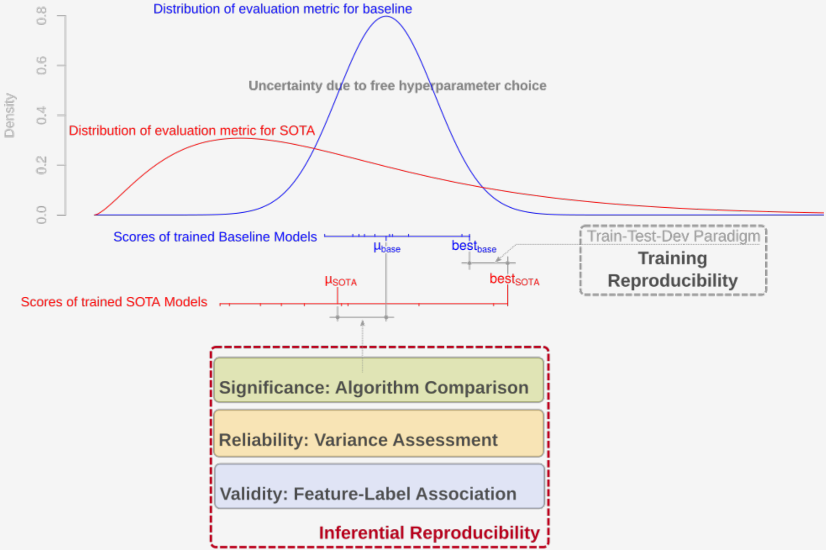
- Question: Can qualitatively similar conclusions be drawn from an independent replication of a study? [Goodman et al., 2016]
- **Inferential reproducibility in machine learning**:
    - Which conclusions about comparison SOTA-baseline can be drawn **across data properties** under **variability of meta-parameters**?
    - Inferential reproducibility is **interesting feature** of non-deterministic machine learning, **not a bug** that needs to be resolved.
    - **:: Training reproducibility ::** Ability to **duplicate prior results** using the same means as used in the original work.

      [Leventi-Peetz and Östreich, 2022]

Questions of theory of science to analyze inferential reproducibility

- **Significance** – how likely is it that a result difference between two models (incorporating sources of variation) is due to chance?
- **Reliability** – how consistent is a performance evaluation if replicated under variations of meta-parameters (or varying data properties)?
- **Validity** – does a machine learning model predict what it purports to predict?

Towards Inferential Reproducibility


Distribution of evaluation metric for baseline
Uncertainty due to free hyperparameter choice
Distribution of evaluation metric for SOTA
Density
0.0
0.2
0.4
0.6
0.8
Scores of trained Baseline Models
μ_{base}
$best_{base}$
μ_{SOTA}
$best_{SOTA}$
Scores of trained SOTA Models
Train-Test-Dev Paradigm

Training Reproducibility

Significance: Algorithm Comparison

Reliability: Variance Assessment

Validity: Feature-Label Association

Inferential Reproducibility

Towards Inferential Reproducibility

Statistical methods as analysis tools

- **Significance**:
  - **Training reproducibility**: Replicability of best SOTA result on benchmark testset.
  - **Inferential reproducibility**: Reproducibility of experiment under variations of meta-parameters and varying data properties.
- **Reliability**:
  - **Variance decomposition**: Decompose variance into components due to variations in meta-parameters and data properties.
  - **Reliability coefficient**: Calculate amount of variance attributable to objects of interest.
- **Validity**: Further reproducibility problems caused by dataset biases.

Tutorial Outline

Statistical models for significance, reliability, and validity

- Interpretable statistical models linear mixed effects models (**LMEMs**), generalized additive models (**GAMs**), trained on predictions of machine learning models. [Wood, 2017]
- **Significance testing under data/meta-parameter variation** by likelihood ratio test on nested LMEM models.
- **Reliability coefficient** and **variance component analysis** of meta-parameter and data effect of LMEM models.
- **Validity** test exposing circularity by **GAM feature shape analysis**.

Textbook

Stefan Riezler & Michael Hagmann (2021). Validity, Reliability, and Significance: Empirical Methods for NLP and Data Science. Morgan & Claypool/Springer.

- Data, code, and preprint available at `https://www.cl.uni-heidelberg.de/statnlpgroup/empirical_methods/`.

Linear Mixed Effect Models & Generalized Likelihood Ratio Test

Outline

- LMEMs in a nut shell
- Estimating LMEMs
- Asymptotic Results for Maximum Likelihood Estimators
- Principles of hypothesis testing
- Generalized Likelihood Ratio Test

Linear Mixed Effect Models (LMEMs)

- LMEM is a regression model for a random variable Y from the exponential family (*Gaussian*, gamma, Bernoulli, categorial, exponential, beta, ...)
- Extension of a standard linear regression model that allows to model *non-iid variance-covariance patterns*
- Tabular dataset: $[[\mathrm{x}^n, \mathrm{z}^n, y^n]_{n=1}^N]^\top$ where $x \in \mathbb{R}^p$, $z \in \mathbb{R}^q$ and $y \in \mathbb{R}$

General form of LMEM for a single observation

$$y^n = \mathrm{x}^{\mathrm{n}}\boldsymbol{\beta} + \mathrm{z}^{\mathrm{n}}\mathrm{b} + \boldsymbol{\epsilon}^n$$

- $\boldsymbol{\beta}$ fixed effect parameters
- Random effect parameters b and errors ϵ modeled by Gaussian variables
- $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are estimated from the data
- See [McCulloch and Searle, 2001, Demidenko, 2013, Wood, 2017, Riezler and Hagmann, 2021].

Linear Mixed Effect Models (LMEMs)

Matrix notation for the whole dataset

$$\mathrm{y} = \mathrm{X}\boldsymbol{\beta} + \mathrm{Z}\mathrm{b} + \boldsymbol{\epsilon}$$

- $\mathrm{y} \in \mathbb{R}^N$, $\mathrm{X} \in \mathbb{R}^{N \times p}$ and $\mathrm{Z} \in \mathbb{R}^{N \times q}$ denote the stacked $y^n/x^n/z^n$
- $\mathrm{b} \sim \mathcal{N}(0, \psi_{\boldsymbol{\theta}})$, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \Lambda_{\boldsymbol{\theta}})$ where $\psi_{\boldsymbol{\theta}}$ and $\Lambda_{\boldsymbol{\theta}}$ are positive definite
- $\mathbb{E}[\mathrm{y}|\mathrm{X}] = \mathrm{X}\boldsymbol{\beta}$ and $\mathbb{V}[\mathrm{y}] = \mathrm{Z}\psi_{\boldsymbol{\theta}}\mathrm{Z}^\top + \Lambda_{\boldsymbol{\theta}}$

Data distribution for Gaussian Y

$$\mathrm{Y} \sim \mathcal{N}(\mathrm{X}\boldsymbol{\beta}, \mathrm{Z}\psi_\theta\mathrm{Z}^\top + \Lambda_{\boldsymbol{\theta}})$$

REMARK I: $f_{\mathrm{y}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\mathrm{y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathrm{y}-\boldsymbol{\mu})\right)$
REMARK II: Inversion of $\mathrm{Z}\psi_\theta\mathrm{Z}^\top + \Lambda_\theta$ is computationally expensive

Profile Likelihood based ML Estimator

AIM: Find an expression for f_{y} that avoids the inversion of $\mathsf{Z}\boldsymbol{\psi}_{\theta}\mathsf{Z}^{\top} + \Lambda_{\boldsymbol{\theta}}$

Elementary facts

$$\text{Marginal Distribution: } f_{y} = \int f_{y,\mathsf{b}}\, d\mathsf{b}$$

$$\text{Conditional Distribution: } f_{\mathsf{y},\mathsf{b}} = f_{\mathsf{y}|\mathsf{b}} f_{\mathsf{b}}$$

Conditional distribution of Y given b

$$\mathsf{y}|\mathsf{b} \sim \mathcal{N}(\mathsf{X}\boldsymbol{\beta} + \mathsf{Z}\mathsf{b}, \Lambda_{\boldsymbol{\theta}})$$

$$\mathsf{b} \sim \mathcal{N}(0, \psi_{\boldsymbol{\theta}})$$

$\Rightarrow$ Just need to invert $\Lambda_{\boldsymbol{\theta}}$ and $\psi_{\boldsymbol{\theta}}$ which are typically simple and sparse

Profile Likelihood based ML Estimator

Result: Marginal distribution

$$f_{\mathrm{y}}(\boldsymbol{\beta},\boldsymbol{\theta}) \propto |\mathrm{Z}^\top \Lambda_{\boldsymbol{\theta}}^{-1}\mathrm{Z} + \psi_{\boldsymbol{\theta}}^{-1}|^{-1/2} f_{\mathrm{y}|\hat{\mathrm{b}}}(\boldsymbol{\beta},\boldsymbol{\theta}) f_{\hat{\mathrm{b}}}(\boldsymbol{\theta})$$

$$\text{where} \quad \hat{\mathrm{b}} := \underset{\mathrm{b}\in\mathbb{R}^q}{\operatorname{argmax}} \log(f_{\mathrm{y},\mathrm{b}}(\boldsymbol{\beta},\boldsymbol{\theta}))$$

PROOF: $f_y = \int \exp(\log(f_{y,\mathrm{b}}))\, d\mathrm{b}$ & Taylor Series of $\log(f_{y,\mathrm{b}})$ around $\hat{\mathrm{b}}$

ML Objective based on the marginal distribution

$$\begin{aligned} L(\boldsymbol{\beta},\boldsymbol{\theta},\mathrm{b}) = & -(\mathrm{y}-\mathrm{X}\boldsymbol{\beta}+\mathrm{Zb})^\top \Lambda_{\boldsymbol{\theta}}^{-1}(\mathrm{y}-\mathrm{X}\boldsymbol{\beta}+\mathrm{Zb}) - \mathrm{b}^\top \psi_{\boldsymbol{\theta}}^{-1}\mathrm{b} \\ & - \log(|\Lambda_{\boldsymbol{\theta}}|) - \log(|\psi_{\boldsymbol{\theta}}|) - \log(|\mathrm{Z}^\top \Lambda_{\boldsymbol{\theta}}^{-1}\mathrm{Z} + \psi_{\boldsymbol{\theta}}^{-1}|) \end{aligned}$$

What happened to $\hat{\mathrm{b}}$?

Profile Likelihood based ML Estimator

Assume we know $\boldsymbol{\theta}$: MLE for β and b (Henderson equations)

$$\begin{bmatrix} X^\top \Lambda_{\boldsymbol{\theta}}^{-1} X & X^\top \Lambda_{\boldsymbol{\theta}}^{-1} Z \\ Z^\top \Lambda_{\boldsymbol{\theta}}^{-1} X & Z^\top \Lambda_{\boldsymbol{\theta}}^{-1} Z + \psi_{\boldsymbol{\theta}}^{-1} \end{bmatrix} \cdot \begin{bmatrix} \hat{\beta} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} X^\top \Lambda_{\boldsymbol{\theta}}^{-1} y \\ Z^\top \Lambda_{\boldsymbol{\theta}}^{-1} y \end{bmatrix}$$

How to estimate $\boldsymbol{\theta}$

- ML: Find $\hat{\boldsymbol{\theta}}$ by optimizing the profile likelihood $L(\hat{\beta}, \hat{b}, \boldsymbol{\theta})$
  - No closed form solution
  - Computations can be sped up due to simple and sparse matrices
  - Convergence can be sped up combining EM and Newton-Raphson methods
- REML (Restricted Maximum Likelihood)

Important Large Sample Results for MLE

Consistency [Nie, 2006]

$$\forall \epsilon > 0: \quad P\left(d(\hat{\boldsymbol{\beta}}_N, \boldsymbol{\beta}) > \epsilon\right) \xrightarrow{N \to \infty} 0$$

- $\hat{\boldsymbol{\beta}}$ is asymptotically unbiased
- $\mathbb{V}[\hat{\boldsymbol{\beta}}]$ decreases with N

Distribution of $\hat{\boldsymbol{\beta}}$ given $\boldsymbol{\theta}$

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{app}{\sim} \mathcal{N}(0, (\mathsf{X}^\top(\mathsf{Z}\psi_\theta \mathsf{Z}^\top + \Lambda_{\boldsymbol{\theta}})^{-1}\mathsf{X})^{-1})$$

- Allows statistical inference for $\hat{\boldsymbol{\beta}}$
- Results still holds when $\boldsymbol{\theta}$ is replaced by $\hat{\boldsymbol{\theta}}$

Principles of Hypothesis Testing

Fundamental goal

Decide between two mutually exclusive and exhaustive sets of hypotheses, called *null hypothesis* H_0 and *alternative hypothesis* H_1 about the data generating probability measures by evidence obtained from observed random samples.

$\Rightarrow$ The test decision is a random event!

Important probabilities of a hypothesis test

$$\begin{aligned} \text{Type-I error probability:} \quad & P(\text{reject } H_0 \text{ while } H_0 \text{ is true}) \\ \text{Power:} \quad & P(\text{reject } H_0 \text{ while } H_1 \text{ is true}) \end{aligned}$$

Conducting a hypothesis test

1. Define a test statistic T which allows to discriminate between H_0 and H_1
2. Assume H_0 is true and derive the distribution of T
3. Set the Type-I error probability to a predefined level α
4. Reject H_0 when $P(|T| > t_{obs}) \leq \alpha$

Notes

- It is important that the actual α_{act} equals the nominal α, otherwise the test either wastes power ($\alpha_{act} < \alpha$) or is not admissible ($\alpha_{act} > \alpha$)
- Usually 2 is the difficult step. If one resorts to resampling based methods one has to be careful to implement an appropriate resampling mechanism [Canty et al., 2006].

Generalized Likelihood Ratio Test

Hypothesis

Suppose we have two *nested models* describing the same data $f(\Theta_0)$ and $f(\Theta_1)$ where $\Theta_0 \subseteq \Theta_1$ with $df_0 := \dim(\Theta_0) < \dim(\Theta_1) =: df_1$ are the parameter spaces of the models. We want to test if $m(\Theta_1)$ is more appropriate.

$$\begin{aligned} H_0&: \quad \theta \in \Theta_0 \\ H_1&: \quad \theta \in \Theta_1 \setminus \Theta_0 \end{aligned}$$

Likelihood ratio

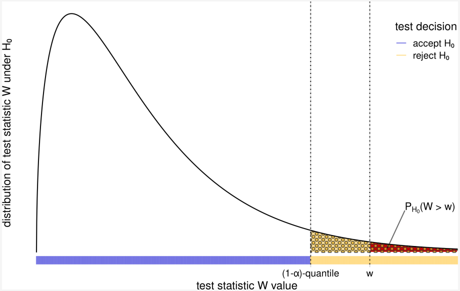
$$\lambda := \frac{f(\hat{\theta}_0^{ML})}{f(\hat{\theta}_1^{ML})} = \frac{\ell_0^*}{\ell_1^*}$$

Generalized Likelihood Ratio Test

Interpretation of $0 < \lambda \leq 1$:

- Values of λ close to 1 suggest that restricted model (H_0) explains the data as well as more complex model (H_1)
- H_0 should be accepted for large values of λ
- Conversely, values close to 0 suggest that the data are not very compatible with the parameter values in the restricted model
- H_0 should be rejected for small values of λ

Generalized Likelihood Ratio Test



Test statistic [Wilks, 1938, van der Vaart, 1998]

$$W = -2\log\lambda = 2(\log\ell_1^* - \log\ell_0^*) \overset{H_0}{\sim} \chi^2_{df_1-df_0}$$

Reject H_0 if $p := P_{H_0}(W > w_{obs}) \leq \alpha$

Significance

Significance Testing under Measurement Variations

- **State-of-the-art:** Statistical significance testing is mostly ignored in NLP and ML in general. [Marie et al., 2021, Ulmer et al., 2022]
- **Goal:** Start reproducibility analysis by significance testing, w/ and w/o incorporation of variability in meta-parameters and data.
- **Method:**
  - Train **LMEM** on performance scores of baseline and SOTA models, obtained w/ or w/o meta-parameter variation during training.
  - Apply **GLRT** to system effect parameter of LMEM.
  - Analyze **significance w/ and w/o meta-parameter variation** and **conditional on data properties**.

Reminder: General Form of LMEMs

- For given dataset of N input-output pairs $\{(\mathrm{x}^n, y^n)\}_{n=1}^N$, general form of an LMEM is

$$\mathrm{Y} = \mathrm{X}\boldsymbol{\beta} + \mathrm{Z}\mathrm{b} + \boldsymbol{\epsilon}.$$

  - Y are N stacked response variables,
  - X and Z known design matrices,
  - $\boldsymbol{\beta}$ fixed effects,
  - b random effects,
  - $\boldsymbol{\epsilon}$ residual errors,
  - where $\mathrm{b} \sim \mathcal{N}(0, \psi_\theta)$, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \Lambda_\theta)$.

Significance Testing with LMEMs

GLRTs based on LMEMS

- **Response variables** Y for LMEM training: **Performance evaluation scores** of meta-parameter variants of baseline and SOTA systems.
- **GLRT**: Train **LMEM with fixed effect** β_c **accounting for competing systems** on performance scores of baseline and SOTA systems, and compare their likelihood ratio.
- **Pairing of systems on the sentence level**: Incorporation of **random sentence effect** b_s allows incorporation of meta-parameter variations and reduces residual variance.

The nested models setup [Pinheiro and Bates, 2000]

- **Restricted null hypothesis model** not distinguishing between systems:

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

where β is fixed effect for common mean for both systems, b_s is random effect for sentence-specific deviation with variance σ_s^2, and residual error ϵ_{res} with variance σ_{res}^2.

- **General model with different means** for baseline and SOTA:

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

where indicator function $\mathbb{I}_c$ activates fixed effect β_c for deviation of competing SOTA model from the baseline mean β when data point was obtained by a SOTA evaluation.

GLRTs in the nested models setup

- Restricted model m_0 is special case ("nested") of more general model m_1 since it restricts factor β_c to zero.
- Let ℓ_0 be likelihood of restricted model m_0, ℓ_1 be likelihood of more general model m_1, intuition of GLRT is to reject the null hypothesis if the **test statistic of likelihood ratio**

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

yields values close to zero.

Significance Testing with LMEMs

Analyzing significance conditional on data properties

- Extend models m_0 and m_1 by a **fixed effect β_d modeling a test data property** d like segment length, readability, or word rarity.
- Add **interaction effect** $\beta_{c:d}$ to assess expected system performance for different levels of d.
- Perform GLRT comparing

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

to null hypothesis model

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

Example: Interactive Machine Translation

Fine-Tuning Neural Machine Translation (NMT) from human feedback [Kreutzer et al., 2020]

- Baseline: NMT system pre-trained on large out-of-domain data.
- SOTA: Fine-tuning on in-domain data annotated with human error markings or error corrections.
- Response variables for LMEM training: TER scores on test data. [Snover et al., 2006]
- Data properties: Sentence lengths, binned into short (< 15 words), typical ($15 - 55$ words), very long (> 55 words).

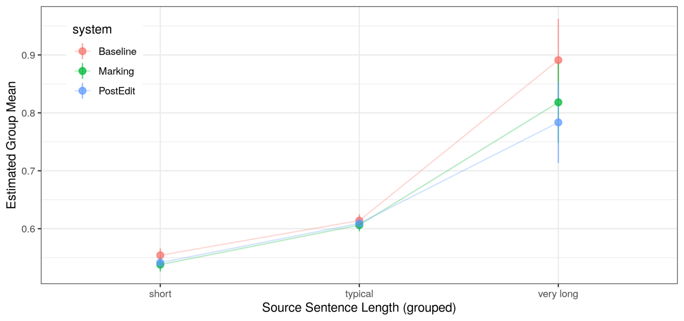
Example: Interactive Machine Translation

Meta-parameter	Grid values			
learning_rate	0.0001	0.0003	0.0005	0.003
random_seed	42	43	44	
encoder_dropout	0	0.2	0.4	0.6
decoder_dropout	0	0.2	0.4	0.6
decoder_dropout_hidden	0	0.2	0.4	0.6

- Meta-parameter grid of attention-based RNN for interactive NMT.

  [Kreutzer et al., 2020]

Significance Testing of Difference Baseline - SOTA



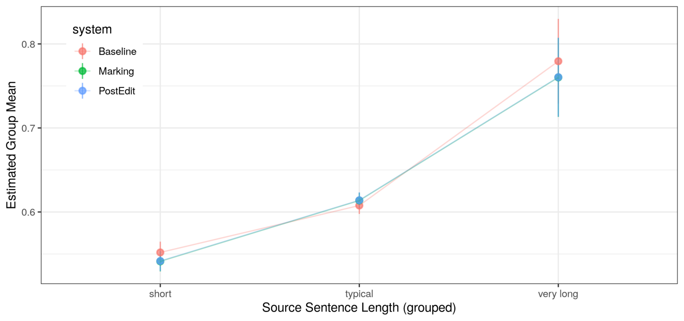
- TER scores for fine-tuning on human error markings or human post-edits compared to baseline, evaluated on test sentences of 3 length classes.
    - SOTA systems trained under three different random seeds, thus one replication for each of three random seeds in LMEM input data.

Significance Testing of Difference Baseline - SOTA

	p-value	effect size
baseline - marking	0.000332	1.24
baseline - post-edit	0.0000000358	1.28
marking - post-edit	0.0252	0.589

- *p*-values and effect sizes (standardized mean difference) for comparison of fine-tuning on human error markings or human post-edits to baseline on very long test sentences.
  - *p*-values < 0.05, medium to very large effect sizes

Significance Testing under Meta-Parameter Variation



- Extended meta-parameter configuration space by grid search over $4 \times 4 \times 4 \times 4 \times 3 = 768$ trained models for each of the fine-tuning runs.

Significance Testing under Meta-Parameter Variation

- Meta-parameters:
  - initial learning rate (`learning_rate`),
  - probability of zeroing out connections during training of encoder (`enc_dropout`), decoder (`dec_dropout`), and hidden decoder layers (`dec_dropout_h`),
  - seed of random number generator (`random_seed`).
- *p*-values for all pairwise differences are above 0.05 across different classes of sentence length.
  - **Significance of result difference lost!**
  - Investigate reasons ➡ **reliability analysis!**

Summary: Significance Testing with LMEMs

Advantages of Model-Based Significance Testing with LMEMs

- One-stop approach to test statistical significance of performance differences between machine learning models:
  - Variance in evaluation scores due **meta-parameter variation is incorporated naturally** into training data for LMEM.
  - **No matching of evaluation metrics to significance tests required** [Dror et al., 2020] since test statistics is not based on evaluation metrics, but on MLE parameters of LMEM [van der Vaart, 1998].
  - Further key advantage is analysis of **significance of result difference conditional on data properties**.
  - Power of significance test is **intimately related to reliability** of model under analysis - next chapter!
  - Further reading: [van der Vaart, 1998, Pinheiro and Bates, 2000, Davison, 2003].

Alternative: Sampling-Based Significance Tests

- Goal:
  - Applicability to arbitrary and arbitrarily complex evaluation metrics (e.g., non-linear combinations of counts like F-score [Manning et al., 2008], BLEU [Papineni et al., 2002], ROUGE [Lin and Hovy, 2003]).
  - No restriction to "mean of samples" metrics which is requirement in parametric tests (t-test, Z-test).
  - More powerful than nonparametric tests (e.g. sign test).

Alternative: Sampling-Based Significance Tests

Examples

- **Bootstrap resampling:** [Efron and Tibshirani, 1993] Sample itself is a representative "proxy" for the population, sampling distribution of test statistic is estimated by repeatedly sampling (with replacement) from the sample itself.
- **Permutation test:** [Fisher, 1935] Principle of stratified shuffling [Noreen, 1989] allows generation of null-hypothesis conditions by shuffling (sampling without replacement) outputs between two systems at strata that partition the data.

Bootstrap Test

```
Given test set outputs $(A_0, B_0) = (a_i, b_i)_{i=1}^N$, where $a_i$ is the output
of system $\mathcal{A}$, and $b_i$ is the output of system $\mathcal{B}$, on test instance $i$.
Compute score difference $\Delta S_0 = S(A_0) - S(B_0)$ on test data.
For $k = 1, \ldots, K$:
    Generate bootstrap dataset $S_k = (A_k, B_k)$ by sampling $N$ examples
    from $(a_i, b_i)_{i=1}^N$ with replacement.
    Compute score difference $\Delta S_k = S(A_k) - S(B_k)$ on bootstrap data.
Compute $\overline{\Delta S_k} = \frac{1}{K} \sum_{k=1}^K \Delta S_k$.
Set $c = 0$.
For $k = 1, \ldots, K$:
    If $|\Delta S_k - \overline{\Delta S_k}| \geq |\Delta S_0|$
        $c++$
$p = c/K$.
Reject null hypothesis if $p$ is less than or equal to rejection level $\alpha$.
```

Permutation Test

```
Given test set outputs $(A_0, B_0) = (a_i, b_i)_{i=1}^N$, where the first element
in the ordered pair $(a_i, b_i)$ is the output of system $\mathcal{A}$, and the second
element is the output of system $\mathcal{B}$, on test instance $i$.
Compute score difference $\Delta S_0 = S(A_0) - S(B_0)$ on test data.
Set $c = 0$.
For $r = 1, \ldots, R$:
    Compute shuffled outputs $(A_r, B_r)$ where for each $i = 1, \ldots, N$:
    $\mathrm{swap}(a_i, b_i) = \begin{cases} (a_i, b_i) & \text{with probability } 0.5, \\ (b_i, a_i) & \text{with probability } 0.5. \end{cases}$
    Compute score difference $\Delta S_r = S(A_r) - S(B_r)$ on shuffled data.
    If $|\Delta S_r| \geq |\Delta S_0|$
        $c++$
$p = c/R$.
Reject null hypothesis if $p$ is less than or equal to rejection level $\alpha$.
```

Problems with Bootstrap

- Bootstrap test makes more Type I errors (i.e., rejecting H_0 when it is true) and more Type II errors (i.e., not rejecting H_0 when it is false) than the permutation test if **bootstrap consistency** is not given (i.e., if data from which is resampled are not representative of population). [Canty et al., 2006, Riezler and Maxwell, 2005, Berg-Kirkpatrick et al., 2012]
- Designed for comparing a pair of selected systems on a single test set, no easy incorporation of variability in meta-parameters or data!

Problems with Permutation

- Permutation test has great power (i.e., high probability of rejecting H_0 when it is false) for large samples [Hoeffding, 1952].
- Stratified shuffling principle needs to be applicable, which is not always the case.
- Designed for comparing a pair of selected systems on a single test set, no easy incorporation of variability in meta-parameters or data!

Further Alternatives

Significance testing across multiple meta-parameter and data settings

- Bootstrap and permutation tests are designed for comparing a pair of selected systems on a single test set - extensions apply this principle to sampling w/ and w/o replacement from system outputs under meta-parameter variations, but ignore variation of data properties. [Clark et al., 2011, Sellam et al., 2021, Bouthillier et al., 2021].
- Statistical significance test based on the stochastic order/dominance of performance score distributions allow incorporation of meta-parameter variation, but still ignore variation of data properties. [Dror et al., 2019, Ulmer et al., 2022]

Reliability

Reliability of Measurements

- **State-of-the-art:** Bootstrap confidence intervals ("error bars") around evaluation scores under meta-parameter variation.

  [Lucic et al., 2018, Henderson et al., 2018]

- **Goal:**
  - Analyze sources of variability in performance evaluation,
  - analyze interaction of meta-parameters variance with data properties,
  - compute coefficient to quantify general robustness of a model.
- **Method:**
  - **Variance component analysis (VCA)**: Untangle sources of variability in measurement.
  - **Reliability coefficient**: Assess general robustness of model by ratio of substantial variance out of total variance.

Variance Component Analysis

VCA in classical ANOVA [Fisher, 1925, Searle et al., 1992]

- Example: Specify model with random effects for variation in outcome Y between sentences s and between settings of meta-parameter r.
- **Tautological decomposition:**

$$Y = \mu + (\mu_s - \mu) + (\mu_r - \mu) + (Y - \mu_s - \mu_r + \mu),$$

  - grand mean μ of observed evaluation score across all levels of meta-parameter r and sentences s,
  - deviation $\nu_s = (\mu_s - \mu)$ of mean score μ_s for sentence s from μ,
  - deviation $\nu_r = (\mu_r - \mu)$ of mean score μ_r for meta-param. r from μ,
  - residual error, reflecting deviation of observed score Y from what would be expected given the first three terms.

VCA in classical ANOVA [Fisher, 1925, Searle et al., 1992]

- Components in decomposition are uncorrelated with each other.
- Total variance $\sigma^2(Y-\mu)$ can be decomposed into following **variance components**:

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

  - σ_s^2 and σ_r^2 denote variance due to sentences and meta-parameter settings,
  - σ_{res}^2 denotes residual variance including variance due to interaction of s and r.

Reminder: General Form of LMEMs

- For given dataset of N input-output pairs $\{(\mathrm{x}^n, y^n)\}_{n=1}^N$, general form of an LMEM is

$$\mathrm{Y} = \mathrm{X}\boldsymbol{\beta} + \mathrm{Z}\mathrm{b} + \boldsymbol{\epsilon}.$$

  - Y are N stacked response variables,
  - X and Z known design matrices,
  - $\boldsymbol{\beta}$ fixed effects,
  - b random effects,
  - $\boldsymbol{\epsilon}$ residual errors,
  - where $\mathrm{b} \sim \mathcal{N}(0, \psi_\theta),\ \boldsymbol{\epsilon} \sim \mathcal{N}(0, \Lambda_\theta)$.

Estimation of Variance Components by LMEMs

- Conditions of measurement that contribute to variance in the measurement besides the objects of interest (here: sentences) are called *facets* of measurement (example: meta-parameters).
  - Each **facet-specific component** $\nu_f = \mu_f - \mu$ modeled as component b_f of **random effects** vector b,
  - corresponding **variance component** σ_f^2 modeled as component of **variance-covariance matrix** ψ_θ.

Advantages LMEM over ANOVA

- **Flexibility!**
  - General estimation procedure that is not design-driven.
  - Elegant handling of missing data situations.
  - Flexible modeling, e.g., random-effects-only models.
- Further reading: [Baayen et al., 2008, Barr et al., 2013, Bates et al., 2015]

Modeling Interactions with Data Properties in LMEMs

- Identify facet f with large variance contribution σ_f^2 in VCA.
- Analyze interaction of facet f with data property d:
  - Change random effect b_f to fixed effect β_f,
  - Add fixed effect β_d modeling test data characteristics,
  - Add interaction effect $\beta_{f:d}$ modeling interaction between data property d and facet f.

Reliability Coefficient

Intra-class correlation coefficient (ICC) [Fisher, 1925]

- Fundamental interpretation as measure of proportion of variance that is attributable to objects of measurement.
- Ratio of variance between objects of interest σ_B^2 to the total variance σ_{total}^2, including variance within objects of interest σ_W^2.

$$ICC = \frac{\sigma_B^2}{\sigma_{total}^2} = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2}.$$

- Name of coefficient is derived from goal of measuring how strongly objects in the same class are grouped together: **Variance between objects of interest should outweigh variance within!**

Reliability Coefficient

General reliability coefficient φ [Brennan, 2001]

- Ratio of substantial variance σ_s^2 to the sum of itself and absolute error variance σ_Δ^2, defined for facets $f_1, f_2, \ldots$ and selected interactions $s : f_1, s : f_2, f_1 : f_2, \ldots,$ all modeled as random effects:

$$\varphi = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_\Delta^2}, \text{ where } \sigma_\Delta^2 = \sigma_{f_1}^2 + \sigma_{f_2}^2 + \ldots + \sigma_{s:f_1}^2 + \sigma_{s:f_2}^2 + \ldots + \sigma_{f_1:f_2}^2 + \cdots + \sigma_{res}^2.$$

Reliability coefficient φ applied to NLP/data science

- **Reliability of performance evaluation across replicated measurements** is assessed as the **ratio by which the amount of substantial variance outweighs the total error variance**.
  - Variance should explained by variance between test sentences, not by variance-inducing facets like meta-parameter settings or by unspecified facets of measurement procedure.
  - Interpretation of threshold on ratio:
    - Values less than 50%, between 50% and 75%, between 75% and 90%, and above 90%, indicative of poor, moderate, good, and excellent reliability [Koo and Li, 2016]

Example: Variance Component Analysis of Meta-Parameter Importance

Assessing importance of meta-parameters

- Goal: Assess importance of meta-parameters in automatic meta-parameter search. [Habelitz and Keuper, 2020]
- Method: VCA using LMEM with random effects for meta-parameters (and interactions)
  - LMEMs offer unified framework to assess importance of meta-parameter across all levels of other meta-parameters, not just in context of a single fixed instantiation of remaining meta-parameters.
  - Previous work used less flexible functional ANOVA for same purpose.

    [Hutter et al., 2014, Zimmer et al., 2020]

Example: VCA of Meta-Parameter Importance

Example: A neural model for disease score prediction

- Multi-layer perceptron (MLP) to predict Sequential Organ Failure Assessment (SOFA) score.
- Meta-parameters:
  - maximal number of neurons in hidden layer (`hidden_size_max`),
  - number of hidden layers (`hidden_number`),
  - values of initial learning rate (`learning_rate`),
  - number of training examples in each gradient computation (`batch_size`),
  - seed of random number generator (`random_seed`),
  - number of iterations over training set (`epochs`),
  - probability of zeroing out hidden connections during training (`dropout`).

Example: VCA of Meta-Parameter Importance

Meta-parameter	Grid values					
batch_size	1	4	8	16	32	64
dropout	0	0.05	0.1	0.15	0.2	
epochs	1	5	10			
hidden_number	3	5	7			
hidden_size_max	16	32	64	128	256	
learning_rate	0.001	0.01	0.1			
random_seed	−7712	6483	20777			

- Meta-parameter values in grid search for SOFA-score MLP.

Example: VCA of Meta-Parameter Importance

- Random-effects-only LMEM:

$$Y = \mu + b_{hidden\_size\_max} + b_{hidden\_number} + b_{learning\_rate} \\ + b_{batch\_size} + b_{random\_seed} + b_{epochs} + b_{dropout} + \epsilon_{res}.$$

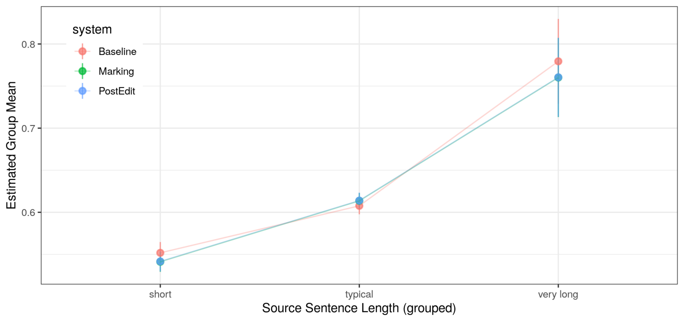
- Training data for LMEM:
  - Performance evaluations of summative evaluation metric, e.g., mean accuracy over test data instances.
  - Evaluations for fully crossed meta-parameter configuration space, yielding $6 \times 5 \times 3 \times 3 \times 5 \times 3 \times 3 = 12{,}150$ models.

Example: VCA of Meta-Parameter Importance

Variance component v	Variance σ_v^2	Percent
residual	0.0000314	61.2
hidden_number	0.0000159	31.0
learning_rate	0.00000318	6.2
batch_size	0.000000517	1.01
hidden_size_max	0.000000260	0.505
dropout	0.0000000599	0.117
random_seed	0.00000000405	0.00788

- Most variance induced by variation in number of hidden layers (31%),
- followed with a wide margin by learning rate (6.2% of total variance),
- all other meta-parameters introduce negligible variance of $\leq 1\%$.

Example: Reliability Analysis in Interactive Machine Translation



- Reminder: Significance between baseline and SOTA model was lost in extended meta-parameter grid search.
- Goal: Reliability analysis of SOTA model!
- Question: Which **meta-parameter setting is responsible** for performance drop, and what is **interaction with data** properties?

Reliability Analysis of SOTA model

- Response variable Y is TER score on test sentence, μ is grand mean, b_s is sentence-specific deviation, and b_{random_seed} is random effect modeling 3 random seeds:

$$Y = \mu + b_s + b_{random\_seed} + \epsilon_{res}.$$

- Excellent reliability $\varphi = 98.4\%$, essentially no contribution of variance due to replications under random seeds.

Variance component	Variance σ^2	Percent
sentence	0.984	98.4
residual	0.0163	1.63
random_seed	0	0

Reliability Analysis of SOTA under Meta-Parameter Variation

- Add random effect b_f for each meta-parameter f in grid search:

$$\begin{aligned} Y = \mu + b_s + b_{learning\_rate} + b_{random\_seed} + b_{enc\_dropout} \\ + b_{dec\_dropout} + b_{dec\_dropout\_h} + \epsilon_{res}. \end{aligned}$$

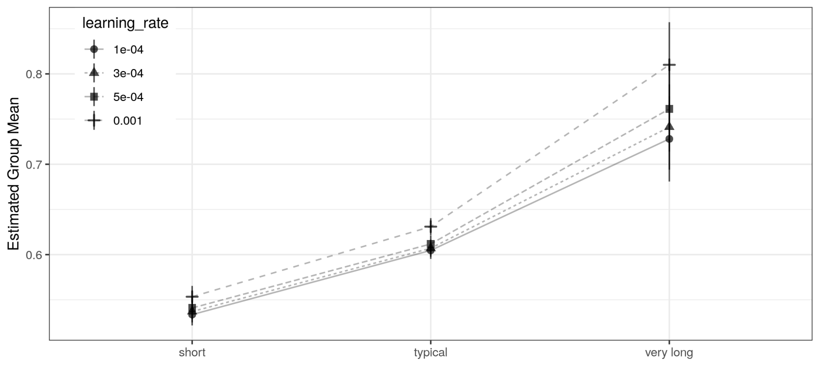
- Reliability coefficient drops below 90% with learning rate having largest contribution to variance.

Variance component	Variance σ^2	Percent
sentence	0.0574	88.4
residual	0.00737	11.3
learning_rate	0.000127	0.2
decoder_dropout	0.0000303	0.05
encoder_dropout	0.0000224	0.03
decoder_dropout_hidden	0.00000130	0
random_seed	0.000000578	0

Interaction between Meta-Parameters and Data Properties

- Add fixed effect β_{src_length} for source sentence length and interaction effect $\beta_{src\_length:learning\_rate}$.

$$Y = \mu + b_s + \beta_{src\_length} + \beta_{learning\_rate} + \beta_{src\_length:learning\_rate} + \epsilon_{res}.$$



Interpretation of Reliability Analysis

- Significant improvements by fine-tuning over baseline with large effect size only on very long sentences.
  - ➡ Such improvements are likely to be reproducible on very long sentences of new datasets.
- Strong dependency of consistency of evaluation results on initial learning rate settings.
  - ➡ Likely that the results will be reproducible only for small initial learning rates (< 0.0005), but not for large initial learning rates.
- Questionable reproducibility of result differences on short and medium length sentences, especially between fine-tuned systems.

Summary: LMEM-Based Reliability Analysis

- Distinctive idea:
  - Compute **reliability coefficient as proportion of substantial variance attributable to the objects of interest**, compared to insubstantial variance due to idiosyncrasies of measurement situation.
  - Ideas date back to [Fisher, 1925] and allow **interpretation of reasons for (un)reliability** and **understanding of interactions of variance components and data**.
  - Based on **well-understood statistical models (LMEMs)**.
  - Further reading: [Searle et al., 1992, Brennan, 2001, Webb et al., 2006].

Agreement coefficients for data annotation

- Scott's π [Scott, 1955], Cohen's κ [Cohen, 1960], or Krippendorff's α [Krippendorff, 2004] are commonly used descriptive statistics to measure agreement of raters in data annotation.
- Based on simple concept of percent agreement that is adjusted to include agreement by chance.
- Easily computable from experimental data by collecting relative count statistics.

Problems with agreement coefficients

- Convenience in computation is due to a fixed choice of a model for computing chance agreement:
    - Sampling with replacement (Scott's π and Cohen's κ) or without replacement (Krippendorff's α),
    - from distributions for individual raters (Cohen's κ) or from observed ratings averaged over raters (Scott's π and Krippendorff's α).

Problems with agreement coefficients

$$\text{chance-adjusted agreement} = \frac{\text{observed agreement - chance agreement}}{n - \text{chance agreement}}.$$

- Counter-intuitive principle of maximum randomness, leading to many paradoxes and abnormalities. [Zhao et al., 2013]
- Main disadvantages:
  - No generalization beyond concrete raters and concrete data points examined in a concrete experiment.
  - No explanation of reasons for high/low agreement by properties of raters or data, or by interactions between them.

Bootstrap confidence intervals for model evaluation

- Interest is in reliability of predictions of a machine learning algorithm itself, not just reliability of single concrete evaluation experiment.
- Bootstrap-inspired resampling to compute confidence bounds for evaluation scores on test data. [Henderson et al., 2018, Lucic et al., 2018]
    - Goal: Quantify variation in maximum out-of-sample performance with respect to meta-parameter choice and computational budget.
    - Method: Resample performance evaluation scores from pool of models trained under increasing budget for meta-parameter search.

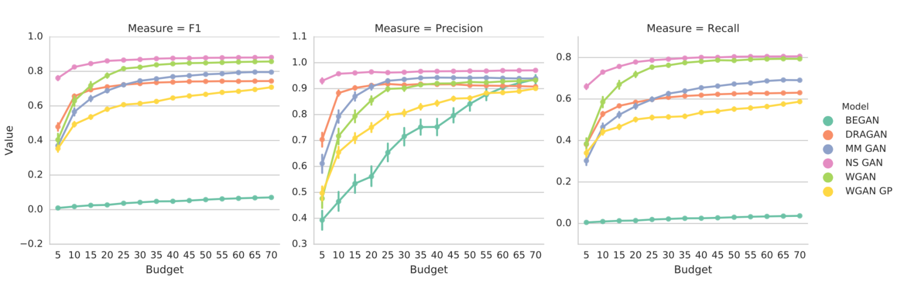
Bootstrap Confidence Interval for Evaluation Metric

1. Generate M meta-parameter configurations for the model class.
2. For each $m = 1, \dots, M$: Train model p_m and calculate the performance evaluation score $u_m = u(p_m)$.
3. For each $B \leq M$: Construct a bootstrap distribution by K times drawing B random samples with replacement from $\{u_m \colon m = 1, \dots, M\}$. For each sample select the maximum performance score.
4. Calculate the mean $\bar{x}$ and the standard deviation $\sigma_{\bar{x}}$ of this distribution. Plug both estimates into the standard normal 95% confidence interval of the population mean μ:

$$\bar{x} - 1.96\sigma_{\bar{x}} \leq \mu \leq \bar{x} + 1.96\sigma_{\bar{x}}.$$

Alternative Reliability Measures

- Mean and 95% confidence intervals for F1-score, precision, recall of GANs for different computational budgets. [Lucic et al., 2018]



Problems with bootstrap confidence intervals

- Idea: Use confidence bounds to directly signify reliability of an evaluation meta-parameter settings: At the same level of confidence, smaller confidence bounds indicate higher reliability.
- Problems:
  - Lacking bootstrap consistency, either if test set from which bootstrap samples are drawn is not representative of population [Canty et al., 2006], or if the parameter to be estimated is on the boundary of the parameter space [Andrews, 2000, Bickel and Freedman, 1981] as in calculations of expected maximum performance [Lucic et al., 2018, Dodge et al., 2019].
- Main shortcoming:
  - Confidence intervals do not tell us which meta-parameters have the most influence on variations in evaluation scores, and how meta-parameter settings interact with properties of test data.

Recap: Inferential Reproducibility - A Worked-Through Example

BART-RXF: Better Fine-Tuning by Reducing Representational Collapse [Aghajanyan et al., 2021]

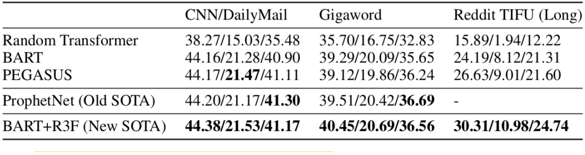
- SOTA on `paperswithcode.com` for text summarization task on CNN/Dailymail and RedditTIFU datasets.
- Baseline: BART [Lewis et al., 2019]
- SOTA Model: Approximate trust region method by constraining updates on embeddings f and classifier g during fine-tuning in order not to forget original pre-trained representations.

$$\begin{aligned}\mathcal{L}_{R3F}(f,g,\theta) &= \mathcal{L}(\theta) + \lambda KL(g \cdot f(x) || g \cdot f(x+z)) \\ &\text{s.t. } z \sim \mathcal{N}(0,\sigma^2 I) \text{ or } z \sim \mathcal{U}(-\sigma,\sigma).\end{aligned}$$

Recap: A Worked-Through Example

Experimental setup and SOTA results

- Datasets hosted on `paperwithcode.com`
  - train/dev/test split for Reddit not given, used split of [Zhong et al., 2020] instead.
- Reported meta-parameter ranges: $\lambda \in [0.001, 0.01, 0.1]$, noise distribution $\mathcal{N}$ or $\mathcal{U}$, maximum result of 10 random seeds .
  - Seeds of random number generator not given, used new 18 random seeds for baseline and 5 for SOTA.
- Results reported in [Aghajanyan et al., 2021]:



	CNN/DailyMail	Gigaword	Reddit TIFU (Long)
Random Transformer	38.27/15.03/35.48	35.70/16.75/32.83	15.89/1.94/12.22
BART	44.16/21.28/40.90	39.29/20.09/35.65	24.19/8.12/21.31
PEGASUS	44.17/**21.47**/41.11	39.12/19.86/36.24	26.63/9.01/21.60
ProphetNet (Old SOTA)	44.20/21.17/**41.30**	39.51/20.42/**36.69**	-
BART+R3F (New SOTA)	**44.38/21.53/41.17**	**40.45/20.69/36.56**	**30.31/10.98/24.74**

Significance Testing for Training Reproducibility

baseline - SOTA	p-value	effect size
Rouge1	$1.99e-14$	-0.101
Rouge2	0.00000000114	-0.0803
RougeL	$1.35e-15$	-0.105

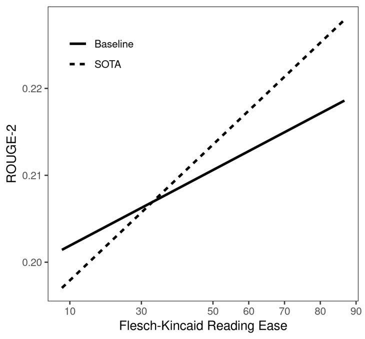
- Rouge [Lin and Hovy, 2003] evaluation of best baseline versus best SOTA model on CNN/DailyMail shows **significant improvements of best SOTA model over baseline** with small effect sizes.

A First Step towards Inferential Reproducibility: Significance Conditional on Data Properties

Measuring difficulty of summarization data

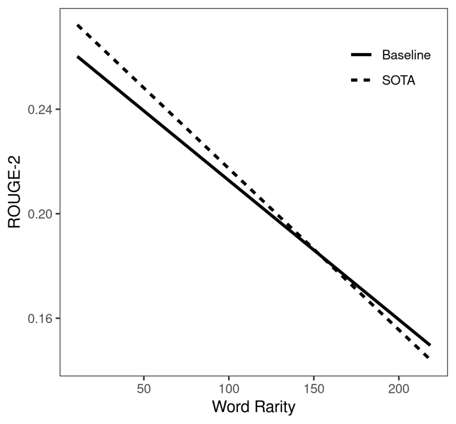
- **Word rarity** [Platanios et al., 2019]: Negative log of empirical probabilities of words in segment, higher value means higher rarity.
- **Flesch-Kincaid readability** [Kincaid et al., 1975]: Pro-rates words/sentences and syllables/word; in principle unbounded, but interpretation scheme exists for ranges from 0 (difficult) to 100 (easy).

Interaction of Performance with Data Properties



- Significant difference in performance slope with respect to ease of readability.
- Performance for SOTA system increases faster for easier inputs than for baseline.

Interaction of Performance with Data Properties



- Significant difference in performance with respect to word rarity.
- SOTA is better than baseline for inputs with lower word rarity.

Significance Testing for Inferential Reproducibility

Incorporating meta-parameter variation into significance testing

- Grid search over 18 random seeds for baseline, 30 SOTA models for 3 λ values × 2 noise distributions × 5 random seeds.

baseline - SOTA	p-value	effect size
Rouge1	0.0	0.390
Rouge2	0.0	0.301
RougeL	0.0	0.531

- **Relations turned around: Baseline significantly better than SOTA**, at medium effect size!
- Performance variation of baseline model over 18 random seeds negligible (standard deviations $< 0.2\%$ for Rouge-X scores)
- ➡ Reliability analysis of SOTA model!

Reliability Analysis for Inferential Reproducibility

Reliability coefficient and variance component analysis

Variance component v	Variance σ_v^2	Percent
summary_id	0.00923	55.7
lambda	0.00254	15.3
random_seed	0.000122	0.73
noise_distribution	0.0000473	0.29
residual	0.00464	28.0

- Only moderate value of reliability coefficient.
- Largest variance component for Rouge1 estimate due to regularization constant λ.

Reliability Analysis for Inferential Reproducibility

Reliability coefficient and variance component analysis

Variance component v	Variance σ_v^2	Percent
summary_id	0.00992	62.7
lambda	0.00131	8.31
random_seed	0.0000766	0.48
noise_distribution	0.0000318	0.2
residual	0.00449	28.3

- Only moderate value of reliability coefficient.
- Largest variance component for Rouge2 estimate due to regularization constant λ.

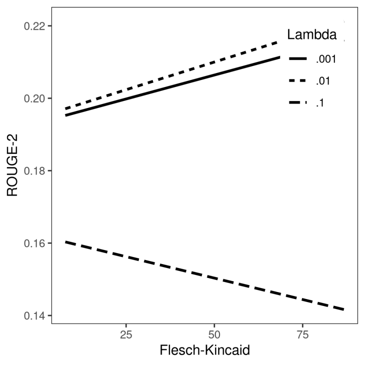
Reliability Analysis for Inferential Reproducibility

Reliability coefficient and variance component analysis

Variance component v	Variance σ_v^2	Percent
summary_id	0.00875	47.9
lambda	0.00519	28.4
random_seed	0.0000370	0.2
noise_distribution	0.0000144	0.08
residual	0.00428	23.4

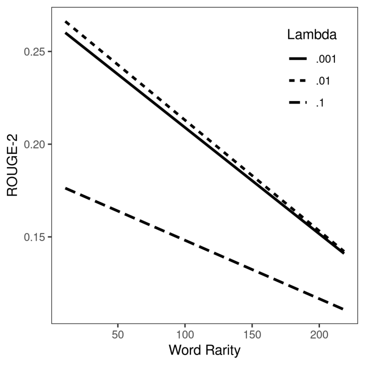
- Poor value of reliability coefficient.
- Largest variance component for RougeL estimate due to regularization constant λ.

Interaction of Meta-Parameters with Data Properties



- Significant drop in performance of SOTA model across levels of reading difficulty for regularization constant $\lambda = 0.1$.

Interaction of Meta-Parameters with Data Properties



- Significant drop in performance of SOTA model for regularization constant $\lambda = 0.1$, especially for rare words.

Reproducibility Results on RedditTIFU

- Interesting data since much harder to read (mean readability score of -348.9).
- Significant improvement of best SOTA over baseline only for Rouge2 at small effect size.
- No significant improvements of SOTA over baseline if meta-parameter variation is taken into account.
- Reliability coefficients of around 80% with negligible variance contributions from λ values.

Interpretation of Statistical Analysis

- Losing or winning a new SOTA score strongly depends on finding the **sweet spot of a single meta-parameter** (here: λ) – paper's goal was explicitly to reduce instability across meta-parameter settings!
- Performance improvements by fine-tuning **mostly on easy-to-read and frequent-word inputs** – less than one quarter of the CNN/Dailynews data.
- **Lacking robustness against data variability** – new random split on RedditTIFU negates gains reported for split used in paper.

Generalized Additive Models

Outline

- GAMs in a nut shell
- A brief sketch of splines
- Estimating spline based GAMs
- Two important measures
- Consistency

Generalized Additive Models (GAMS)

- GAMs are regression models for a random variable Y from the exponential family (*Gaussian*, gamma, Bernoulli, categorial, exponential, beta, ...)
- Extension of a standard linear regression model that allows to model *non-linear functions*
- Tabular dataset: $[[\mathrm{x}^n, y^n]_{n=1}^N]^\top$ where $x \in \mathbb{R}^p$ and $y \in \mathbb{R}$

Generalized Additive Models (GAMs)

General form of interpretable GAMs

$$\mathbb{E}[Y|x_1,\ldots,x_p] = \underbrace{\overbrace{f_{1,1}(x_1)+f_{1,2}(x_2)+\ldots}^{\text{univariate}} + \overbrace{f_{2,1}(x_1,x_2)+\ldots}^{\text{bivariate}}}_{\text{non-parametric}} + \underbrace{X\beta}_{\text{parametric}}$$

- $f(\cdot)$ called smoother (non-linear function)
  - non-parametric regression models
  - *splines*
  - deep neural networks
  - regression trees
- parametric part is typically used to model categorical variables
- $f(\cdot)$ and β are estimated from the data
- See [Wood, 2017, Hastie and Tibshirani, 1986, Hastie and Tibshirani, 1990, Wahba, 1990, Green and Silverman, 1993, Riezler and Hagmann, 2021].

What are splines?

- Well known technique from numerical mathematics for function interpolation
- Key Idea: Interpolation is done by piece-wise polynomial functions that connect smoothly at knots to model globally smooth functions

Definition: Spline

A function p:$[\tau_0, \tau_{n-1}) \mapsto \mathbb{R}$ that can be expressed by a polynomial with a degree of at most d for each sub-interval $[\tau_i, \tau_{i+1}]$ of a strictly increasing knot sequence $\tau := [\tau_i]_{i=0,\dots,n-1}$ is called a piece-wise polynomial function or *spline* on τ of maximum degree d.

The spline space $S_{d,\tau}$

$S_{d,\tau}$ denotes the vector space of all $(d-1)$-times continuously differentiable splines on τ.

A Basis for $S_{d,\tau}$

Truncated power function

$$(u)_+^d := \begin{cases} 0 & u < 0 \\ u^d & \text{otherwise} \end{cases} \quad \text{with} \quad d \in \mathbb{N}_0$$

Result

For every spline p on τ with maximum degree d exist a unique set of coefficients c_{ij} for $i = 0, \ldots, d$ and $j = 0, \ldots, (n-2)$ such that

$$p(x) = \sum_{j=0}^{n-2} \sum_{i=0}^{d} c_{ij} (x - \tau_j)_+^d$$

The most commonly used splines (natural splines, B-Spline, cubic splines, TP-splines, etc) differ mostly by the chosen base to represent $S_{d,\tau}$.

Functional minimization problem

Let $\mathfrak{H}$ be the class of twice differentiable univariate functions and assume N datapoints:

$$\min_{h \in \mathfrak{H}} \sum_{n=1}^{N} (y^n - h(x^n))^2 + \lambda \int (h''(x))^2 \, dx$$

where $\lambda \in \mathbb{R}^+$ and $\int (h''(x))^2 dx$ is a measure for the roughness of a function over its domain.

Solution: Natural cubic splines with knots at each input x^n

Idea for spline based GAMs

Fix a Basis for $S_{d,\tau}$, transform the input feature x by the base functions and estimate the $c_{i,j}$ from data

Matrix notation of a spline

$$f(\cdot) = \sum_{j=1}^{d} \beta_j b_j(\cdot) = \mathrm{b}(\cdot)\boldsymbol{\beta}$$

$$\text{where} \quad \mathrm{b}(\cdot) = [b_1(\cdot), b_2(\cdot), \dots, b_d(\cdot)]$$

$$\boldsymbol{\beta} = [\beta_1, \beta_2, \dots, \beta_d]^\top$$

Penalized Least Squares Parameter Estimation

Penalized least squares objective

$$\hat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^s} \left\| \mathsf{Y} - \mathsf{G}\boldsymbol{\beta} \right\|^2 + \sum_{k=1}^{p} \lambda_k \int (f_k''(x))^2 \, dx$$

where $s = \sum_{k=1}^{p} d_k$, $\lambda_k \in \mathbb{R}^+$ and G stores the base function values of the input features.

Useful fact about the roughness penalty

$$\int (f''(x))^2 \, dx = \boldsymbol{\beta}^\top \Omega \boldsymbol{\beta}$$

$$\text{where} \quad \Omega := [\int b_s''(x) b_t''(x) dx]_{s,t=1,\ldots,N}$$

Penalized Least Squares Parameter Estimation

PLSE objective (for one spline)

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^N} \left\| \mathrm{Y} - \mathrm{G}\boldsymbol{\beta} \right\|^2 + \lambda \boldsymbol{\beta}^\top \Omega \boldsymbol{\beta}$$

REMARK: Note similarity to OLS objective

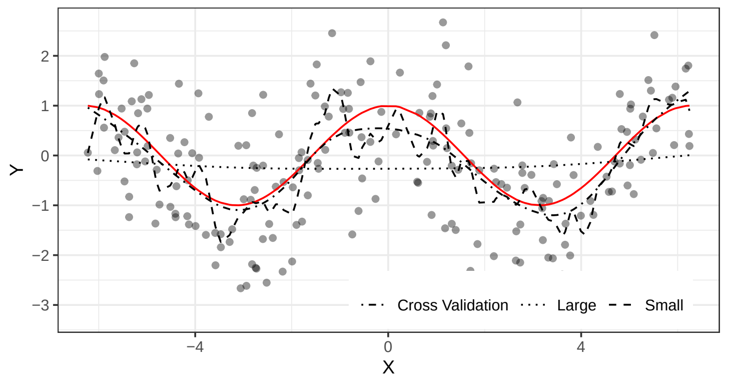
Estimators

$$\hat{\boldsymbol{\beta}} = (\mathrm{G}^\top \mathrm{G} + \lambda\Omega)^{-1} \mathrm{G}^\top \mathrm{y}$$

Thus, the estimated smoother is:

$$\hat{f}(\cdot) = \mathrm{b}(\cdot)(\mathrm{G}^\top \mathrm{G} + \lambda\Omega)^{-1} \mathrm{G}^\top \mathrm{y}$$

Choice of λ



Estimating λ

- cross validation [Wood, 2017]
- marginal likelihood estimation in tandem with β [Wood et al., 2016]

Consistency

Definition: Consistency

Let $M\colon = \{p_\theta\colon \theta \in \Theta\}$ be a parametric statistical model where $\theta \mapsto p_\theta$ is injective. Further, let $p_{\theta_0} \in M$ denote the true model of the data generating process for a dataset $D = \{(\mathrm{x}^n, y^n)\}_{n=1}^N$. Then an estimator θ_N is called *consistent* iff for all $\epsilon > 0$ holds

$$P\left(|\theta_N - \theta_0| > \epsilon\right) \xrightarrow{N\to\infty} 0.$$

Consistency has been shown for spline based GAMs by [Heckman, 1986].

Measuring Model Fit: Deviance

A likelihood based measure of model fit

Difference between the log-likelihood $\ell(\mu)$ of a model μ and the largest possible log-likelihood ℓ^*

$$D^*_\mu := 2(\ell^* - \ell(\mu))$$

ℓ^* corresponds to the likelihood of a model that perfectly reproduces the targets

Deviance explained

$$D^2(\mu) = 1 - \frac{D^*_\mu}{D_{\mu_0}} \in [0, 1]$$

where μ_0 denotes the intercept only model

Validity

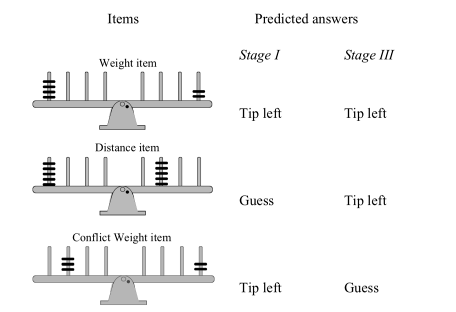
Validity in psychological measurement theory

"A test is valid for measuring an attribute if (a) the attribute exists and (b) variations in the attribute causally produce variation in the measurement outcomes."

[Borsboom et al., 2004, Borsboom, 2005, Borsboom and Mellenbergh, 2007]

- Measurement model explicates how the structure of theoretical attributes relates to the structure of observations
- Example: Measurement model for temperature stipulates how the level of mercury in a thermometer systematically relates to temperature

Example: Psychological Measurement of Developmental Stages



Items	Predicted answers	
	Stage I	*Stage III*
Weight item	Tip left	Tip left
Distance item	Guess	Tip left
Conflict Weight item	Tip left	Guess

- Example: Psychological test of developmental stages by Jean Piaget [Inhelder and Piaget, 1958]
  - Different positions in the attribute (e.g. children of age 3-5 versus 10-12) lead to different test outcomes
  - Observed test outcomes can be used to infer position of children in one of four discrete stages of cognitive development

Example: Measurement in the train-dev-test Paradigm

Machine learning models as measurement models

- Example: Multiclass classification
  - Variation in attribute = variation of test pairs (x, y) where x are inputs, y gold standard outputs
  - Measurement outcome = model prediction $\hat{y}$ for input x
  - Causal relation = variation in feature values and labels correlates invariantly across environments (here: test (re-)splits)

    [Peters et al., 2016, Arjovsky et al., 2019]

- Is accurate prediction across test sets all we need to claim validity?
- No! Further criterion of **absence of circularity** from philosophy of science. [Balzer and Brendel, 2019]

Circularity in Machine Learning

Circular features

- **Indirect measurements:** Target label is determined by indirect measurement, but fundamental measurements needed to determine this indirect measurement are part of input feature representation.
- **Circularity:** Circular feature (= fundamental measurement) will lead to an exact reconstruction of the known deterministic function (= indirect measurement) by machine learning, but achieves nothing else.

The Circularity Problem

Why is circularity a problem?

- Machine learning models trained on data including circular features will yield **nearly perfect predictions** on input data including the defining measurements, but they **cannot be transferred to unseen data** where the defining features are not or only incompletely available.
- Circular features in machine learning models will **nullify the contribution of all features except those defining the target**, thus such models will **not learn new predictive patterns** that involve features other than the known defining measurements.

Circularity Test

Goal: A Circularity Test for Black-Box Models

- Assume we know the functional definition of the target, but not the training data of the machine learning model
- Our data are model predictions on test data $T = \{(\mathrm{x}^m, \hat{y}^m)\}_{m=1}^{M}$
- Detect whether black-box model used circular features and remove them from dataset.

GAMs, Deviance, and Nullification

- **GAMs:** Expressive and yet interpretable model class [Wood, 2017]
  - Decompose complex function into sum of non-linear *feature shapes* $f_k(x_k)$, e.g., regression splines [Hastie and Tibshirani, 1990]

$$Y^n = \sum_{k=1}^{p} f_k(x_k^n) + \sum_{i \neq j} f_{ij}(x_i^n, x_j^n) + \epsilon^n, \text{ where } \epsilon^n \sim \mathcal{N}(0, \sigma^2).$$

- **Deviance:** $D^2(\mu) \in [0,1]$ measures proportion of log-likelihood of model μ out of maximal data fit.
- **Nullification:** Identifies circular features by non-null feature shapes, based on identifiability and consistency of maximum likelihood estimators for GAMs.

- **Dataset** $T = \{(\mathrm{x}^m, \hat{y}^m)\}_{m=1}^{M}$ where $\mathrm{x}^m \in \mathbb{R}^p$,
- **Candidate circular features** $C \subseteq \mathcal{P}(\{1, \ldots, p\})$,
- **Models** $\mathcal{M} := \{\mu_c \colon c \in C\}$ obtained by fitting a GAM based on feature set c to data T.

Two-step test to detect circular features c^*

1. **Deviance:** $c^* = \mathrm{argmax}_{c \in C} D^2(\mu_c)$ where $D^2(\mu_{c^*})$ is close to 1, and in case the maximizer is not unique, the maximizer is chosen whose associated GAM μ_{c^*} has the smallest degrees of freedom.
2. **Nullification:** The feature shape of every feature $x_j : j \in \{1, \ldots, p\} \setminus c^*$ added to the GAM μ_{c^*} is nullified in the resulting model.

Example: Circularity in Patent Prior-Art Search

Condition	Relevance Score
no citation	0
inventor citation	1
examiner citation	2
family patent	3

- Data: Construction of gold standard relevance judgements from citations of patents in other patents. [Graf and Azzopardi, 2008]
- Model: Non-linear combination of features by MLP, trained for logistic regression on binarized relevance ranks (level 1 for citations, 0 else).
- Teacher MLP trained on 1,500 patent queries, resulting in 318,375 observations of query-document pairs.

Example: Circularity in Patent Prior-Art Search

- **What if patent citations are included as features in ranking model** (e.g., KISS principle of [Magdy and Jones, 2010])?

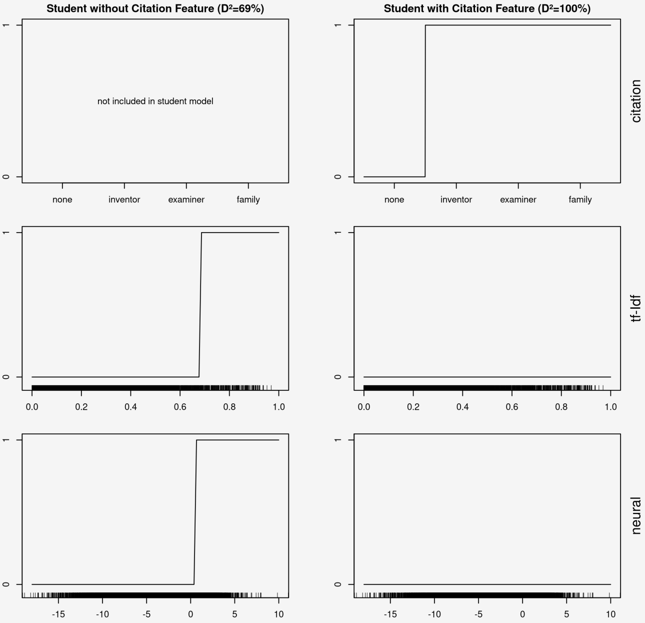
	Feature	Meaning	Range
(1)	`neural`	similarity score learned by neural network	$\mathbb{R}$
(2)	`tf-Idf`	cosine similarity of tf-Idf scores	$\mathbb{R}$
(3)	`inventor`	indicator for inventor citation	$\{0,1\}$
(4)	`examiner`	indicator for examiner citation	$\{0,1\}$
(5)	`family`	indicator for family patent	$\{0,1\}$

Deviance of Student GAMs

Rank	Included Features	D^2	Complexity
1	{inventor, examiner, family}	100%	5
2	{inventor, examiner, family, neural}	100%	6.33
3	{inventor, examiner, family, tf-Idf}	100%	7.95
4	{inventor, examiner, family, neural, tf-Idf}	100%	11.1
5	{examiner, family, neural, tf-Idf}	95%	22

- Top five models visited during circularity search for IR training data.

Nullification in Student GAMs



Example: Circularity in Medical Data Science

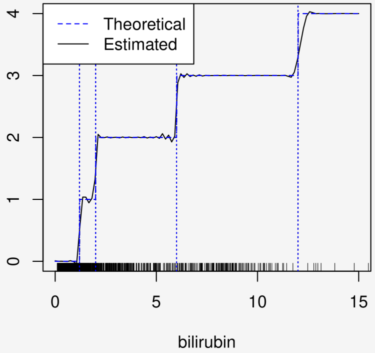
Circularity in SOFA/Sepsis Score Prediction

- Sepsis-3 consensus definition defines sepsis as a change in total SOFA (sequential organ failure assessment) score of at least 2 points consequent to an infection. [Singer et al., 2016, Seymour et al., 2016]
- SOFA scoring system itself is defined for 6 organ systems whose scores are defined by thresholds on measurable physiological quantities like heart rate, creatinin, bilirubine, urine output etc.

  [Vincent et al., 1996]
- Recent overview examined 22 studies on machine learning for (early) prediction of sepsis, with the exception of one, all studies define ground-truth sepsis labels using the deterministic rules of the consensus definition like Sepsis-3. [Moor et al., 2021]

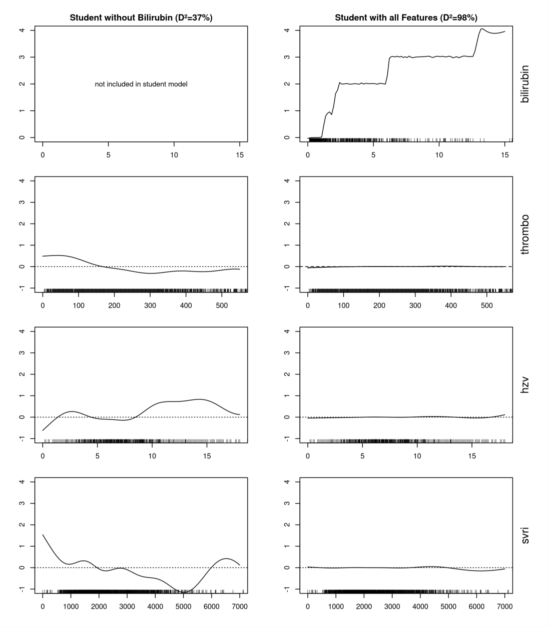
Circularity in Liver SOFA Datasets

Condition	Liver SOFA Score
$0 <$ bilirubin ≤ 1.2	0
$1.2 <$ bilirubin ≤ 1.9	1
$1.9 <$ bilirubin ≤ 5.9	2
$5.9 <$ bilirubin ≤ 11.9	3
bilirubin > 11.9	4



- Definition and reconstruction of liver SOFA score.

Circularity in Liver SOFA Prediction



Circularity in Medical Data Science

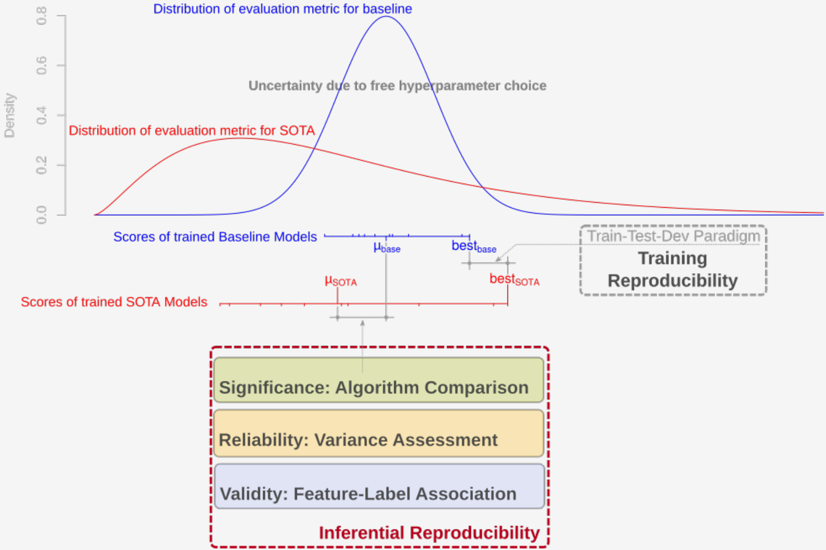
- How likely is it that other datasets and machine learning models exhibit a yet undetected circularity problem?
    - **Critical candidates** are machine learning applications in measurement-based sciences like medicine that define the objects of their research, e.g., diseases, by **rigid measurement procedures** (e.g., on physiological features, images, or text data).
    - **Circularity problems extend to a longitudinal design for prognosis** where feature measurements at a current point in time are used to forecast disease status at future points in time.
        - Circularity will be introduced by the auto-correlation of the time series, especially if data imputation methods like last-value carried forward are used in dataset creation.

Summary: Circularity

- **Circularity inhibits machine learning at its core**
  - If circular features are included in data/models, **nothing else but a reconstruction of the known functional definition of the target will be learned.**
  - If circular features are not or only incompletely available, **reproducibility is lost** in any case.
  - **No insights into new predictive patterns**, no transfer to data labeled in other ways.
- Remedy: Detect and remove circular features in data/models!

Conclusion

Conclusion: Inferential Reproducibility


Density
0.0
0.2
0.4
0.6
0.8
Distribution of evaluation metric for baseline
Uncertainty due to free hyperparameter choice
Distribution of evaluation metric for SOTA
Scores of trained Baseline Models
μ_{base}
$best_{base}$
μ_{SOTA}
$best_{SOTA}$
Scores of trained SOTA Models

Train-Test-Dev Paradigm

Training Reproducibility

Significance: Algorithm Comparison

Reliability: Variance Assessment

Validity: Feature-Label Association

Inferential Reproducibility

Inferential Reproducibility

- Validity, reliability, and significance are methodological pillars of empirical science.
- Easily neglected in race for improved state-of-the-art results on benchmark data.
- Old-fashioned statistical methods come the rescue to analyze inferential reproducibility!
    - Enter **interpretable GAMs and LMEMs** as analysis tools.
    - **Statistical tests like GLRT, VCA, or circularity test** are **justified by identifiability and consistency** of maximum likelihood estimators for GAMs and LMEMs.
    - **Wide applicability, well established software**.

Focus of our work

- **Significance:**
  - Related to partial conjunction testing for multiple datasets [Dror et al., 2017],
  - and to score distribution comparison for multiple models [Dror et al., 2019].
  - **Our focus**: **Unified approach** for significance testing under **meta-parameter and data variation**, using likelihood ratio tests.

Focus of our work

- **Reliability:**
  - Related to approaches that analyze meta-parameter importance in model prediction [Hutter et al., 2014, Bergstra and Bengio, 2012],
  - or report expected validation performance w.r.t. computational budget [Dodge et al., 2019, Tang et al., 2020].
  - **Our focus**: **Explain variability** by LMEM variance component analysis and **justify reliability** by ICC-like coefficient.

Conclusion

Focus of our work

- **Validity:**
  - Related to descriptive statistics to detect dataset bias [Poliak et al., 2018, Gururangan et al., 2018],
  - with goal of using machine learning to reduce influence of bias features[Clark et al., 2019, Kim et al., 2019].
  - **Our focus:** GAM-based test to **detect** validity-violating features and **remove** them from datasets.

Open Questions, Comments, Suggestions

- Towards **inferential reproducibility** as a **new standard in machine learning evaluation**?
  - How to get there?
  - Would you go the extra mile?
  - What did we forget?
- Please tell us in Q&A or by email to `{riezler,hagmann}@cl.uni-heidelberg.de`

Thank you!

Data, code, and preprint:

`https://www.cl.uni-heidelberg.de/statnlpgroup/empirical_methods/`

Aghajanyan, A., Shrivastava, A., Gupta, A., Goyal, N., Zettlemoyer, L., and Gupta, S. (2021).
Better fine-tuning by reducing representational collapse.
In *International Conference on Learning Representations (ICLR)*.

Andrews, D. W. (2000).
Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space.
Econometrica, 68(2):399–405.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2019).
Invariant risk minimization.
CoRR, abs/1907.02893.

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

Balzer, W. and Brendel, K. R. (2019).
Theorie der Wissenschaften.
Springer.

Barr, D. J., Levy, R., Scheepers, C., and Tilly, H. J. (2013).
Random effects structure for confirmatory hypothesis testing: Keep it maximal.
Journal of Memory and Language, 68(3):255–278.

Bates, D., Mächler, M., Bolker, B. M., and Walker, S. C. (2015).
Fitting linear mixed-effects models using lme4.
Journal of Statistical Software, 67(1):1–48.

Belz, A., Agarwal, S., Shimorina, A., and Reiter, E. (2021).
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.

Berg-Kirkpatrick, T., Burkett, D., and Klein, D. (2012).
An empirical investigation of statistical significance in NLP.
In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Jeju Island, Korea.

Bergstra, J. and Bengio, Y. (2012).
Random search for hyper-parameter optimization.
Journal of Machine Learning Research (JMLR), 13:281–305.

Bickel, P. J. and Freedman, D. A. (1981).
Some asymptotic theory for the bootstrap.
The Annals of Statistics, 9(6):1196–1217.

Borsboom, D. (2005).
Measuring the Mind. Conceptual Issues in Contemporary Psychometrics.
Cambridge University Press.

Borsboom, D. and Mellenbergh, G. J. (2007).
Test validity in cognitive assessment.
In Leighton, J. P. and Gierl, M. J., editors, *Cognitive Diagnostic Assessment for Education. Theory and Applications*, pages 85–115. Cambridge University Press.

Borsboom, D., Mellenbergh, G. J., and van Heerden, J. (2004).

The concept of validity.
Psychological Review, 111(4):1061–1071.

Bottou, L., Curtis, F. E., and Nocedal, J. (2018).
Optimization methods for large-scale machine learning.
SIAM Review, 60(2):223–311.

Bouthillier, X., Delaunay, P., Bronzi, M., Trofimov, A., Nichyporuk, B., Szeto, J., Mohammadi Sepahvand, N., Raff, E., Madan, K., Voleti, V., Ebrahimi Kahou, S., Michalski, V., Arbel, T., Pal, C., Varoquaux, G., and Vincent, P. (2021).
Accounting for variance in machine learning benchmarks.
Proceedings of Machine Learning and Systems (MLSys), 3.

Brennan, R. L. (2001).
Generalizability theory.
Springer.

Canty, A. J., Davison, A. C., Hinkley, D. V., and Ventura, V. (2006).
Bootstrap diagnostics and remedies.
The Canadian Journal of Statistics, 34(1):5–27.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M.,

Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N. (2022).
PaLM: Scaling language modeling with pathways.
CoRR, abs/2204.02311.

Clark, C., Yatskar, M., and Zettlemoyer, L. (2019).
Don't take the easy way out: Ensemble based methods for avoiding known dataset biases.
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.

Clark, J., Dyer, C., Lavie, A., and Smith, N. (2011).
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.

Cohen, J. (1960).
A coefficient of agreement for nominal scales.
Educational and Psychological Measurement, 20(1):37–46.

D'Amour, A., Heller, K. A., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., Hormozdiari, F., Houlsby, N., Hou, S., Jerfel, G., Karthikesalingam, A., Lucic, M., Ma, Y., McLean, C. Y., Mincu, D., Mitani, A., Montanari, A., Nado, Z., Natarajan, V., Nielson, C., Osborne, T. F., Raman, R., Ramasamy, K., Sayres, R., Schrouff, J., Seneviratne, M., Sequeira, S., Suresh, H., Veitch,

V., Vladymyrov, M., Wang, X., Webster, K., Yadlowsky, S., Yun, T., Zhai, X., and Sculley, D. (2020).
Underspecification presents challenges for credibility in modern machine learning.
CoRR, abs/2011.03395.

Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., and Bengio, Y. (2014).
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.

Davison, A. C. (2003).
Statistical Models.
Cambridge University Press.

Demidenko, E. (2013).
Mixed Models: Theory and Applications with R.
Wiley.

Dodge, J., Gururangan, S., Card, D., Schwartz, R., and Smith, N. A. (2019).
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.

Dror, R., Baumer, G., Bogomolov, M., and Reichart, R. (2017).
Replicability analysis for natural language processing: Testing significance with multiple datasets.

In *Transactions of the Association for Computational Linguistics (TACL)*, volume 5, pages 471–486.

Dror, R., Peled, L., Shlomov, S., and Reichart, R. (2020).
Statistical Significance Testing for Natural Language Processing.
Morgan & Claypool.

Dror, R., Shlomov, S., and Reichart, R. (2019).
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.

Drummond, C. (2009).
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.

Efron, B. and Tibshirani, R. J. (1993).
An Introduction to the Bootstrap.
Chapman and Hall.

Fisher, R. A. (1925).
Statistical Methods for Research Workers.
Oliver and Boyd.

Fisher, R. A. (1935).
The Design of Experiments.
Hafner.

Goodman, S. N., Fanelli, D., and Ioannidis, J. P. A. (2016).

What does research reproducibility mean?
Sci Transl Med, 8(341):1–6.

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

Graf, E. and Azzopardi, L. (2008).
A methodology for building a patent test collection for prior art search.
In *Proceedings of the 2nd International Workshop on Evaluating Information Access (EVIA)*, pages 60–71, Tokyo, Japan.

Green, P. J. and Silverman, B. W. (1993).
Nonparametric regression and generalized linear models: a roughness penalty approach.
Crc Press.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., and Smith, N. A. (2018).
Annotation artifacts in natural language inference data.
In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, New Orleans, Louisiana.

Habelitz, P. and Keuper, J. (2020).
PHS: A toolbox for parallel hyperparameter search.
CoRR, abs/2002.11429.

Hastie, T. and Tibshirani, R. (1986).

Generalized additive models.
Statistical Science, 1(3):297–318.

Hastie, T. and Tibshirani, R. (1990).
Generalized Additive Models.
Chapman and Hall.

Heckman, N. E. (1986).
Spline smoothing in a partly linear model.
Journal of the Royal Statistical Society B, 48(2):244–248.

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

Henderson, P., Islam, R., Bachmann, P., Pineau, J., Precup, D., and Meger, D. (2018).
Deep reinforcement learning that matters.
In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, New Orleans, LA, USA.

Hoeffding, W. (1952).
The large-sample power of tests based on permutations of observations.
Annals of Mathematical Statistics, 23:169–192.

Hutson, M. (2018).
Artificial intelligence faces reproducibility crisis.
Science, 359(6377):725–726.

Hutter, F., Hoss, H., and Leyton-Brown, K. (2014).

An efficient approach for assessing hyperparameter importance.
In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, Beijing, China.

Inhelder, B. and Piaget, J. (1958).
The Growth of Logical Thinking from Childhood to Adolescence.
Basic Books.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020).
Scaling laws for neural language models.
CoRR, abs/2001.08361.

Kawaguchi, K., Kaelbling, L. P., and Bengio, Y. (2020).
Generalization in deep learning.
CoRR, abs/1710.05468.

Kim, B., Kim, H., Kim, K., Kim, S., and Kim, J. (2019).
Learning not to learn: Training deep neural networks with biased data.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA.

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

Koo, T. K. and Li, M. Y. (2016).
A guideline of selecting and reporting intraclass correlations coefficients for reliability research.

Journal of Chiropratic Medicine, 15:155–163.

Kreutzer, J., Berger, N., and Riezler, S. (2020).
Correct me if you can: Learning from error corrections and markings.
In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Krippendorff, K. (2004).
Content Analysis. An Introduction to Its Methodology.
Sage.

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

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019).
BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Lin, C.-Y. and Hovy, E. (2003).
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.

Lucic, A., Bleeker, M., Bhargav, S., Forde, J., Sinha, K., Dodge, J., Luccioni, S., and Stojnic, R. (2022).
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.

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

Magdy, W. and Jones, G. J. F. (2010).
Applying the KISS principle for the CLEF- IP 2010 prior art candidate patent search task.
In *In Proceedings of the CLEF 2010 Workshop*, Padua, Italy.

Manning, C. D., Raghavan, P., and Schütze, H. (2008).
Introduction to Information Retrieval.
Cambridge University Press.

Marie, B., Fujita, A., and Rubino, R. (2021).
Scientific credibility of machine translation research: A meta-evaluation of 769 papers.
In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, Online.

McCulloch, C. E. and Searle, S. R. (2001).
Generalized, Linear, and Mixed Models.
Wiley.

Melis, G., Dyer, C., and Blunsom, P. (2018).
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.

Moor, M., Rieck, B., Horn, M., Jutzeler, C., and Borgwardt, K. (2021).
Early prediction of sepsis in the ICU using machine learning: A systematic review.
Frontiers in Medicine, 8.

Nie, L. (2006).
Strong consistency of the maximum likelihood estimator in generalized linear and nonlinear mixed-effects models.
Metrika, 63(2):123–143.

Noreen, E. W. (1989).
Computer Intensive Methods for Testing Hypotheses. An Introduction.
Wiley.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002).
Bleu: a method for automatic evaluation of machine translation.
In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*, Philadelphia, PA.

Peters, J., Bühlmann, P., and Meinshausen, N. (2016).
Causal inference using invariant prediction: identification and confidence intervals.
Journal of the Royal Statistical Society, Series B, 78(5):947–1012.

Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Larochelle, H. (2021).
Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program).

Journal of Machine Learning Research (JMLR), 22:1–20.

Pinheiro, J. C. and Bates, D. M. (2000).
Mixed-Effects Models in S and S-PLUS.
Springer.

Platanios, E. A., Stretcu, O., Neubig, G., Poczos, B., and Mitchell, T. (2019).
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.

Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., and Van Durme, B. (2018).
Hypothesis only baselines in natural language inference.
In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, New Orleans, Louisiana.

Post, M. (2018).
A call for clarity in reporting BLEU scores.
In *Proceedings of the 3rd Conference on Machine Translation (WMT)*, Brussels, Belgium.

Reimers, N. and Gurevych, I. (2017).
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.

Riezler, S. and Hagmann, M. (2021).
Validity, Reliability, and Significance: Empirical Methods for NLP and Data Science.

Morgan & Claypool Publishers.

Riezler, S. and Maxwell, J. (2005).
On some pitfalls in automatic evaluation and significance testing for MT.
In *Proceedings of the ACL-05 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, Ann Arbor, MI.

Scott, W. A. (1955).
Reliability of content analysis: The case of nominal scale coding.
Public Opinion Quarterly, 19:321–325.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992).
Variance Components.
Wiley.

Sellam, T., Yadlowsky, S., Wei, J., Saphra, N., D'Amour, A., Linzen, T., Bastings, J., Turc, I., Eisenstein, J., Das, D., Tenney, I., and Pavlick, E. (2021).
The multiberts: BERT reproductions for robustness analysis.
CoRR, abs/2106.16163.

Seymour, C. W., Liu, V. X., Iwashyna, T. J., Brunkhorst, F. M., Rea, T. D., Scherag, A., Rubenfeld, G., Kahn, J. M., Shankar-Hari, M., Singer, M., Deutschman, C. S., Escobar, G. J., and Angus, D. C. (2016).
Assessment of clinical criteria for sepsis for the third international consensus definitions for sepsis and septic shock (Sepsis-3).
JAMA, 315(8):762–774.

Shen, Z., Liu, J., He, Y., Zhang, X., Xu, R., Yu, H., and Cui, P. (2021).
Towards out-of-distribution generalization: A survey.

CoRR, abs/2108.13624.

Singer, M., Deutschman, C. S., and Seymour, C. W. (2016).
The third international consensus definitions for sepsis and septic shock (Sepsis-3).
JAMA, 315(8):801–810.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006).
A study of translation edit rate with targeted human annotation.
In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA'06)*, Cambridge, MA.

Søgaard, A., Ebert, S., Bastings, J., and Filippova, K. (2021).
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.

Strubell, E., Ganesh, A., and McCallum, A. (2019).
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.

Tang, R., Lee, J., Xin, J., Liu, X., Yu, Y., and Lin, J. (2020).
Showing your work doesn't always work.
In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, Online.

Ulmer, D., Hardmeier, C., and Frellsen, J. (2022).
deep-significance - easy and meaningful statistical significance testing in the age of neural networks.

CoRR, abs/2204.06815.

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

Vincent, J., Moreno, R., Takala, J., Willatts, S., Mendonça, A. D., Bruining, H., Reinhart, C., Suter, P., and Thijs, L. (1996).
The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure.
Intensive Care Medicine, 22(7):707–710.

von Luxburg, U. and Schölkopf, B. (2011).
Statistical learning theory: Models, concepts, and results.
In Gabbay, D., Hartmann, S., and Woods, J., editors, *Handbook of the History of Logic, vol. 10: Inductive Logic*, pages 651–706. Elsevier.

Wahba, G. (1990).
Spline models for observational data.
SIAM.

Webb, N. M., Shavelson, R. J., and Haertel, E. H. (2006).
Reliability coefficients and generalizability theory.
Handbook of Statistics, 26:81–214.

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

Wood, S. N. (2017).

Generalized Additive Models. An Introduction with R.
Chapman & Hall/CRC, second edition.

Wood, S. N., Pya, N., and Säfken, B. (2016).
Smoothing parameter and model selection for general smooth models.
Journal of the American Statistical Association, 111(516):1548–1575.

Zhao, X., Liu, J. S., and Deng, K. (2013).
Assumptions behind intercoder reliability indices.
Communication Yearbook, 36:419–480.

Zhong, M., Liu, P., Chen, Y., Wang, D., Qiu, X., and Huang, X. (2020).
Extractive summarization as text matching.
In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, Online.

Zimmer, L., Lindauer, M., and Hutter, F. (2020).
Auto-pytorch tabular: Multi-fidelity metalearning for efficient and robust autodl.
CoRR, abs/2006.13799.