Towards Inferential Reproducibility of Machine Learning Research

Stefan Riezler and Michael Hagmann

Computational Lingustics & Center for Scientific Computing (IWR)
Heidelberg University, Germany

riezler@cl.uni-heidelberg.de

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., 2022, Shen et al., 2021]

Practical workflow of supervised 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.
- Define SOTA by best achieved result, publish code and data, and report corresponding meta-parameter settings.

New paradigm: Practical workflow of in-context learning with LLMs

- The **pretrain-finetune/prompt** paradigm:
  - Access a **pretrained** LLM,
  - **finetune/prompt** model on task-specific data,
  - evaluate the model using a standard automatic evaluation metric on benchmark test data.
- Define SOTA by best achieved result, publish code and **fine-tuning data**, and report corresponding meta-parameters/**prompts**.

Sources of randomness and variability

- Non-convex optimization under **stochasticity** in weight initialization, dropout, data shuffling and batching. [Dauphin et al., 2014]
- **Implementation-level** nondeterminism in floating-point truncation error due to random accumulation ordering in parallel GPU threads.

  [Pham et al., 2021, Gundersen et al., 2022]
- **Algorithmic** factors of nondeterminism in choice of optimizers, meta-parameters and model architecture.

  [Henderson et al., 2018, Schmidt et al., 2021, D'Amour et al., 2022]
- **Data-level** variability in pre-processing, evaluation metrics, data splits [Post, 2018, Chen et al., 2022, Gorman and Bedrick, 2019, Søgaard et al., 2021].
- **Prompt-level** variability in number, ordering, and similarity metric of in-context examples. [Han et al., 2023]

Replicability = training reproducibility of SOTA results under exactly same circumstances

- **Nondeterminism in deep learning is spoiling the party**
  - Implementation-level nondeterminism is partly irreducible, leading to variability even for training runs in identical settings. [Zhuang et al., 2022]
  - Slight changes in training settings can reverse relations between baseline and SOTA. [Reimers and Gurevych, 2017, Melis et al., 2018]
  - Results on ever-growing data may be impossible to replicate, even if code and data are shared [Kaplan et al., 2020, Chowdhery et al., 2022].
  - For API-served black-box commercial LLMs, replicability of research is put in the hands of commercial providers.

- **Checklists:**
  - 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]
  - Unintended side effect: Conclusions that can be drawn from such experiments are restricted to statements about a single training configuration on a single test set.

- 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!**

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**:
  - Embrace certain types of **nondeterminism** as inherent and irreducible conditions of measurement that **contribute to variance in performance evaluation in an interesting way**.
  - Our focus: Which conclusions about comparison SOTA-baseline can be drawn **across data properties** under **variability of meta-parameters**?

Model-based statistical methods for significance and reliability testing

- Interpretable statistical model: Linear mixed effects models (**LMEM**), trained on predictions of machine learning models.
- Significance testing under data/meta-parameter variation by generalized likelihood ratio test (**GLRT**) on nested LMEM models.
- Reliability coefficient and variance component analysis (**VCA**) of meta-parameter and data effect of LMEM models.
- A **Worked-Through Example:** Inferential reproducibility of fine-tuning pre-trained LLMs [Aghajanyan et al., 2021]

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 **LMEMs** 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, conditional on data properties**.

GLRTs based on LMEMS

- **Response variables** Y for LMEM training: **Evaluation scores** for meta-parameter variations of baseline and SOTA.
- **GLRT**: Train **LMEMs 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.

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** β_{cd} to assess expected system performance for different levels of d.
- Perform GLRT comparing

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

to null hypothesis model

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

Reliability

- **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 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.

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 μ_s for sentence s from μ,
  - deviation $\nu_r = (\mu_r - \mu)$ of mean μ_r for meta-parameter 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 $\sigma^2(Y)$ can be decomposed into following independent **variance components**:

$$\sigma^2(Y) = \sigma^2_s + \sigma^2_r + \sigma^2_{res},$$

  - σ^2_s and σ^2_r denote variance due to sentences and meta-parameters,
  - σ^2_{res} denotes residual variance, including variance due to interactions of s and r.

- $$Y = \underbrace{\mu}_{=:\beta_0} + \underbrace{(\mu_s - \mu)}_{=:b_s} + \underbrace{(\mu_r - \mu)}_{=:b_r} + \underbrace{(Y - \mu_s - \mu_r + \mu)}_{=:\epsilon}.$$

  where

  $$\mathrm{b} = \begin{bmatrix} b_r \\ b_s \end{bmatrix} \sim \mathcal{N}(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_r^2 & 0 \\ 0 & \sigma_s^2 \end{bmatrix}), \epsilon \sim \mathcal{N}(0, \sigma^2_{residual}).$$

  - Each component $\mu_f - \mu$ 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.

- $$Y = \beta_0 + \boldsymbol{b_s} + \beta_f + \beta_d + \beta_{fd} + \epsilon.$$
- 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 β_{fd} modeling interaction between data property d and facet f.

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^2_B to the total variance σ^2_{total}, including variance within objects of interest σ^2_W.

$$ICC = \frac{\sigma^2_B}{\sigma^2_{total}} = \frac{\sigma^2_B}{\sigma^2_B + \sigma^2_W}.$$

- 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!**

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., 2020]
- 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.

$$\mathcal{L}_{R3F}(f, g, \theta) = \mathcal{L}(\theta) + \lambda KL_{\text{sym}}(g \cdot f(x) || g \cdot f(x+z))$$
$$\text{where } z \sim \mathcal{N}(0, \sigma^2 I) \text{ or } z \sim \mathcal{U}(-\sigma, \sigma).$$

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].
- 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]:

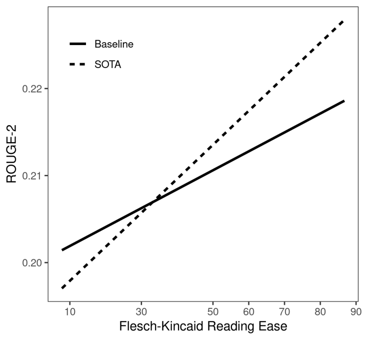
	CNN/DailyMail	Gigaword	Reddit TIFU (Long)
Random Transformer	38.27/15.03/35.48	35.70/16.75/32.83	15.89/1.94/12.22
BART	44.16/21.28/40.90	39.29/20.09/35.65	24.19/8.12/21.31
PEGASUS	44.17/**21.47**/41.11	39.12/19.86/36.24	26.63/9.01/21.60
ProphetNet (Old SOTA)	44.20/21.17/**41.30**	39.51/20.42/**36.69**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

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.

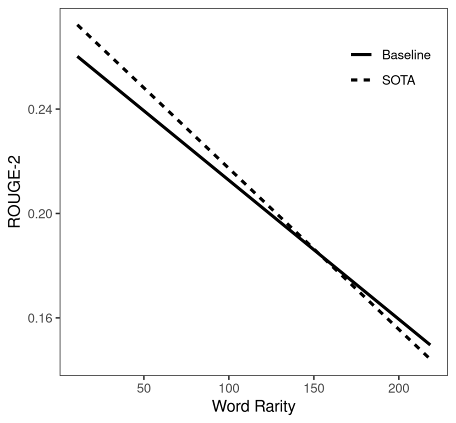
Measuring difficulty of summarization data

- **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, usually interpreted as ranging from 0 (difficult) to 100 (easy).



- Significant difference in performance slope regarding reading ease.
- Performance for SOTA system increases faster for easier inputs.

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.

Incorporating meta-parameter variation into significance testing

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

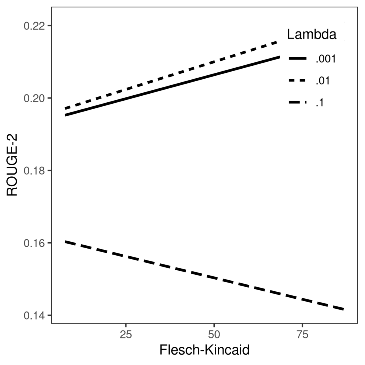
baseline - SOTA	p-value	effect size
Rouge1	0.0	0.390
Rouge2	0.0	0.301
RougeL	0.0	0.531

- **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 coefficient and variance component analysis

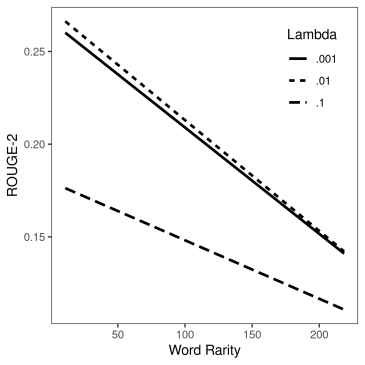
Variance component v	Variance σ_v^2	Percent
summary_id	0.00992	62.7
lambda	0.00131	8.31
random_seed	0.0000766	0.48
noise_distribution	0.0000318	0.2
residual	0.00449	28.3

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



- 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.

- 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.

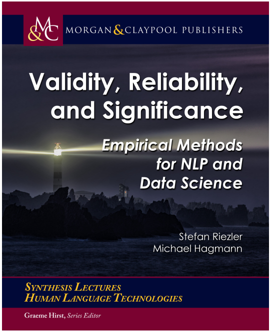
- Losing or winning a new SOTA score strongly depends on finding the **sweet spot of a single meta-parameter** (here: λ) – paper's goal was explicitly to reduce instability across meta-parameter settings!
- Performance improvements by fine-tuning **mostly on easy-to-read and frequent-word inputs** – 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.

Conclusion

Inferential Reproducibility

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

Thank you!

- Textbook: 

- Paper: *Towards Inferential Reproducibility of Machine Learning Research*, Michael Hagmann and Stefan Riezler, ICLR 2023.
- Data & code: `https://www.cl.uni-heidelberg.de/statnlpgroup/empirical_methods/`

Background

General Form of Model

- 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 $\mathrm{b} \sim \mathcal{N}(0, \psi_\theta)$, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \Lambda_\theta)$.
- Notation following [Wood, 2017].

Estimation

- **Fixed effects** can be observed exhaustively and are modeled as parameters of a **standard linear model**.
- **Random effects** are modeled as normally distributed random variables, and corresponding observations are treated as **random samples** from a larger population.
- LMEMs look like a linear model, however, linear combination of fixed effect predictor variables and normally distributed random components yields nonlinear objective.
- Several packages exist for efficient estimation.
- See [Pinheiro and Bates, 2000, Demidenko, 2013, Bates et al., 2015].

Comparison to ANOVA

- LMEMs offer **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:

  [McCulloch and Searle, 2001, West et al., 2007, Baayen et al., 2008, Barr et al., 2013]

Background: The Generalized Likelihood Ratio Test

- Let ℓ_0 be likelihood of restricted model (setting parameter for deviation of models to zero), and ℓ_1 likelihood of more general model.
- Null hypothesis H_0 is assumption that restricted model is adequate.

Generalized Likelihood Ratio Test (GLRT)

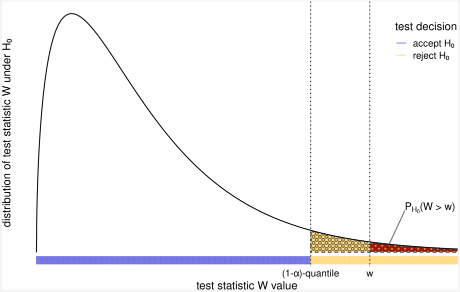
A GLRT computes the likelihood ratio

$$\lambda = \frac{\ell_0}{\ell_1},$$

and rejects H_0 if $0 < \lambda \leq \lambda^*$ where λ^* is chosen such that $P(0 < \lambda \leq \lambda^* | H_0 \text{ is true }) = \alpha$ for a significance level α.

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



χ^2 distribution of likelihood ratio statistic [Wilks, 1938]

- $W = -2\log\lambda = 2\log\frac{\ell_1}{\ell_0} = 2(\log\ell_1 - \log\ell_0) \sim \chi^2$, where χ^2 distribution has $k_1 - k_0$ degrees of freedom if general model has k_1 parameters and restricted model has k_0 parameters
- Reject H_0 if observed value w is greater than $(1-\alpha)$-quantile, i.e., if p-value $p := P_{H_0}(W > w)$ is smaller than rejection level α.
- Further reading: [Pawitan, 2001, Davison, 2003, van der Vaart, 1998].

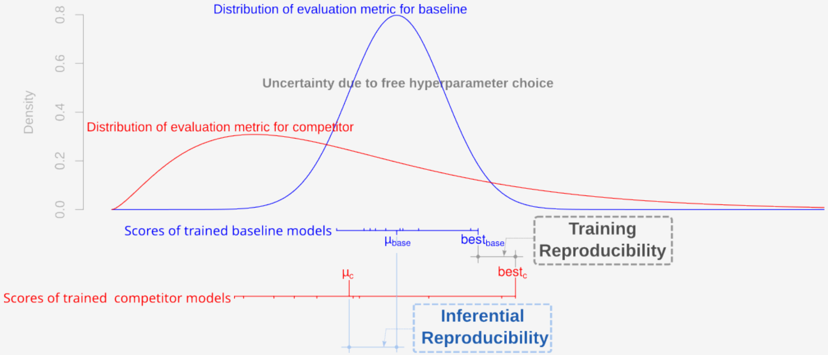
General reliability coefficient φ [Brennan, 2001]

- 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 $f_1 : f_2, \ldots,$ all modeled as random effects:

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

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]


Distribution of evaluation metric for baseline
Uncertainty due to free hyperparameter choice
Distribution of evaluation metric for competitor
Density
0.0
0.2
0.4
0.6
0.8
Scores of trained baseline models
μbase
bestbase
Training Reproducibility
μc
bestc
Scores of trained competitor models
Inferential Reproducibility

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)*.

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.

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.

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

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

Chen, Y., Belouadi, J., and Eger, S. (2022).
Reproducibility issues for BERT-based evaluation metrics.
arXiv, abs/2204.00004.

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.
arXiv, abs/2204.02311.

D'Amour, A., Heller, K., 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., 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. (2022).
Underspecification presents challenges for credibility in modern machine learning.
Journal of Machine Learning Research, 23(226):1–61.

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.

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.

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

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.

Gundersen, O. E., Coakley, K., and Kirkpatrick, C. (2022).
Sources of irreproducibility in machine learning: A review.
arXiv, abs/2204.07610.

Hagmann, M., Meier, P., and Riezler, S. (2023).
Towards inferential reproducibility of machine learning research.
In *The Eleventh International Conference on Learning Representations (ICLR)*.

Han, C., Wang, Z., Zhao, H., and Ji, H. (2023).
In-context learning of large language models explained as kernel regression.
arXiv, abs/2305.12766.

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.

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

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.
arXiv, abs/2001.08361.

Kawaguchi, K., Kaelbling, L. P., and Bengio, Y. (2022).
Generalization in deep learning.
In *Mathematical Aspects of Deep Learning*. Cambridge University Press.

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, 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.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020).
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.

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.

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.

Pawitan, Y. (2001).
In All Likelihood. Statistical Modelling and Inference Using Likelihood.
Clarendon Press.

Pham, H. V., Qian, S., Wang, J., Lutellier, T., Rosenthal, J., Tan, L., Yu, Y., and Nagappan, N. (2021).
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.

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.

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.

Schmidt, R. M., Schneider, F., and Hennig, P. (2021).
Descending through a crowded valley - benchmarking deep learning optimizers.
In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, virtual.

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

Shen, Z., Liu, J., He, Y., Zhang, X., Xu, R., Yu, H., and Cui, P. (2021).
Towards out-of-distribution generalization: A survey.
arXiv, abs/2108.13624.

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.

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

arXiv, abs/2204.06815.

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

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.

West, B. T., Welch, K. B., and Galecki, A. T. (2007).
Linear Mixed Models: A Practical Guide Using Statistical Software.
Chapman & Hall/CRC.

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.

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.

Zhuang, D., Zhang, X., Song, S. L., and Hooker, S. (2022).

Randomness in neural network training: Characterizing the impact of tooling.
In *Proceedings of the 5th MLSys Conference*, Santa Clara, CA, USA.