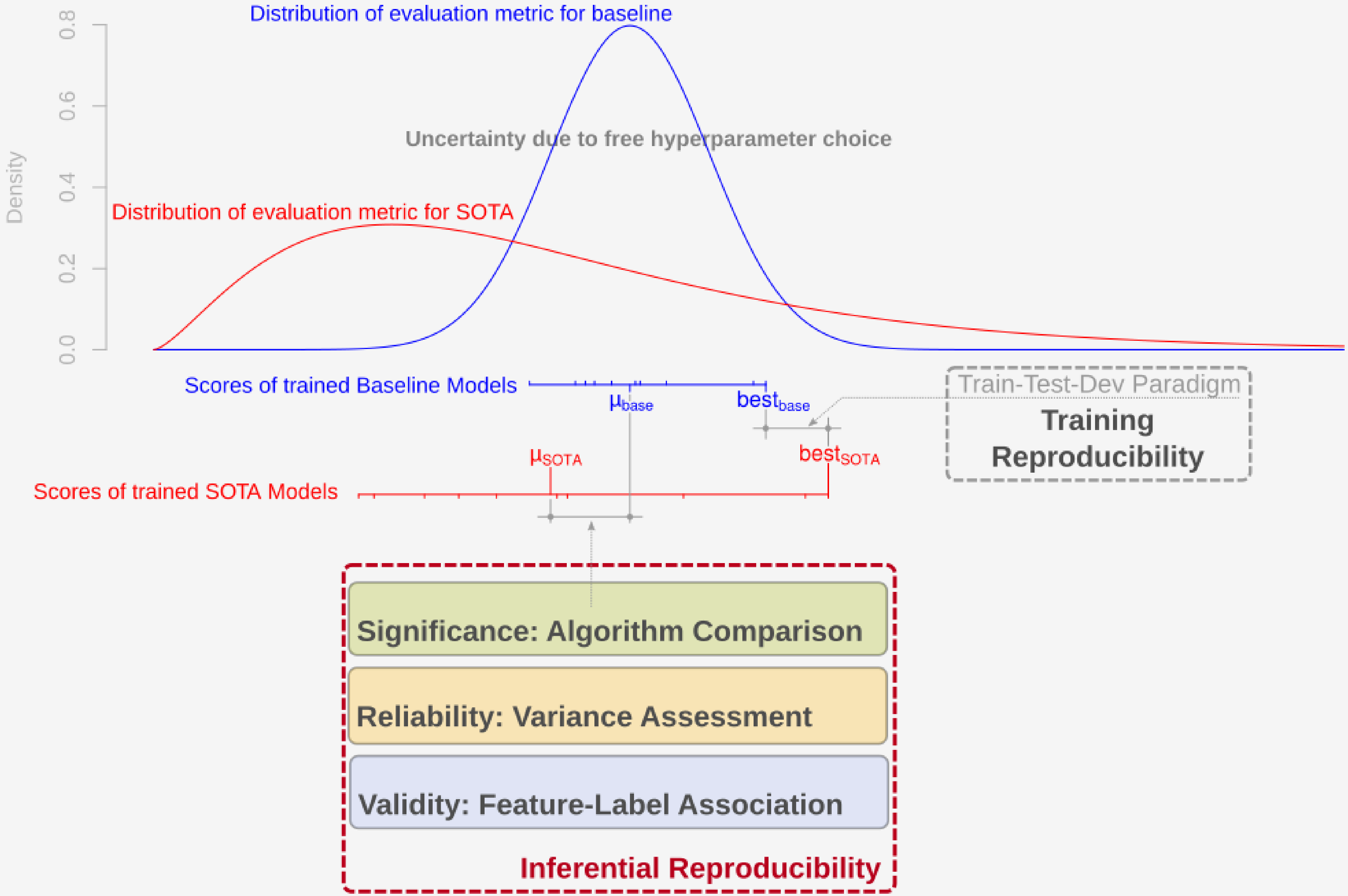
Conclusion


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
μbase
bestbase
Train-Test-Dev Paradigm
Training Reproducibility
μSOTA
bestSOTA
Scores of trained SOTA Models
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.

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.