Machine Learning and Open Science: On Risks and Challenges

Mohammad Arvan

B.Sc., Qazvin Islamic Azad University, 2016

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois at Chicago (UIC), 2024
Chicago, Illinois

Defense Committee:

Dr. Natalie Parde, Chair and Advisor, Department of Computer Science, UIC

Dr. Barbara Di Eugenio, Department of Computer Science, UIC

Dr. Xinhua Zhang, Department of Computer Science, UIC

Dr. Luis Gabriel Ganchinho de Pina, Department of Computer Science, UIC

Dr. Ehud Reiter, University of Aberdeen

COPYRIGHT

ACKNOWLEDGMENTS

I am incredibly proud the work I present in this dissertation, but I must acknowledge several factors and individuals that aided me in reaching this point.

Firstly, my PhD journey started with a letter from the department of Computer Science at the University of Illinois at Chicago (UIC), offering me a scholarship to pursue my PhD. I am grateful for being given this opportunity, which has allowed me to grow both personally and professionally. Furthermore, I believe there were equally equipped candidates who could have been selected, and I am thankful for being chosen.

I would like to express my deepest gratitude to my advisor, Dr. Natalie Parde, for her guidance, support, and mentorship throughout my PhD. Her patience, wisdom, and expertise have been invaluable to me. I am grateful for her unwavering support, encouragement, and belief in me. I am also thankful for her flexibility and understanding, which allowed me to explore my interests and grow as a researcher. I am fortunate to have had the opportunity to work with her, and I am grateful for the knowledge and skills I have gained under her guidance.

This dissertation has gone through several iterations, and I am thankful for the critical feedback and suggestions I received from UIC's Natural Language Processing group, and my committee members. I am grateful for the time and effort they invested in reviewing my work.

While I am the first author for this dissertation and the publications that stem from it, many have contributed to the work presented here. Therefore, throughout this dissertation, I have used ”we” to acknowledge the collective effort that went into this research.

Lastly, I would like to express my deepest gratitude to my friends and family, who have supported me throughout my PhD. I am grateful for their love, encouragement, and understanding. I am fortunate to have such a loving and supportive community, and I am thankful for everything they have done for me.

MA

Contribution of Authors

The following is a statement of the contributions made by each author for the publications included in this thesis:

- **Arvan et al. [1]**: *Reproducibility of Exploring Neural Text Simplification Models: A Review*
Mohammad Arvan was responsible for the primary data collection, analysis, conceptualization, and manuscript writing. Luís Pina and Natalie Parde contributed to the experimental design, technical setup, offered detailed feedback on multiple drafts and provided mentorship throughout the project.
- **Arvan et al. [2]**: *Reproducibility in Computational Linguistics: Is Source Code Enough?*
Mohammad Arvan managed the project, carrying out data collection, analysis, conceptualization, and writing. Luís Pina and Natalie Parde provided significant input on the methodological framework, and provided continuous feedback and guidance during the research process.
- **Arvan et al. [3]**: *Investigating Reproducibility at Interspeech Conferences: A Longitudinal and Comparative Perspective*
Mohammad Arvan led the project, performing data collection, analysis, conceptualization, and drafting of the manuscript. A. Seza Doğruöz and Natalie Parde provided technical and editorial feedback, contributed to the refinement of the research questions, and supervised the overall research direction.
- **Arvan et al. [4]**: *Human Evaluation Reproduction Report for Data-to-text Generation with Macro Planning*
Mohammad Arvan conducted the data collection, analysis, conceptualization, and manuscript writing. Natalie Parde provided guidance on the experimental design, offered substantial feedback on drafts, and supervised the overall direction of the research.
- **Arvan et al. [5]**: *ReproHum #0712-01: Human Evaluation Reproduction Report for "Hierarchical Sketch Induction for Paraphrase Generation"*
Mohammad Arvan was responsible for data collection, analysis, conceptualization, and writing of the manuscript. Natalie Parde provided advisory support, detailed feedback on the drafts, and contributed to shaping the research questions and methodology.

In all manuscripts, Mohammad Arvan was the primary and main contributor, leading the research and manuscript preparation efforts.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACL Association for Computational Linguistics

ACM Association for Computing Machinery

AMT Amazon Mechanical Turk

ANOVA Analysis of Variance

ASO Almost Stochastic Order

BWS Best-Worst Scaling

COLING International Conference on Computational Linguistics

CPU Central Processing Unit

CS Computer Science

CV Coefficient of Variation

CV* Unbiased Coefficient of Variation

EMNLP Empirical Methods in Natural Language Processing

EOL End of Life

FDR False Discovery Rate

FPO Floating-Point Operations

GDPR General Data Protection Regulation

GPU Graphics Processing Unit

GUM Guide to the expression of Uncertainty in Measurement

HEDS Human Evaluation Data Sheet

HITs Human Intelligence Tasks

IRB Institutional Review Board

LREC Language Resources and Evaluation Conference

LSTM Long Short-Term Memory

ML Machine Learning

MT Machine Translation

NAACL North American Chapter of the Association for Computational Linguistics

NER Named Entity Recognition

NeurIPS Neural Information Processing Systems

NHST Null Hypothesis Significance Testing

NLP Natural Language Processing

NTS Neural Text Simplification

NTS-w2v Neural Text Simplification with Word2Vec

ReLU Rectified Linear Unit

SGD Stochastic Gradient Descent

TACL Transactions of the Association for Computational Linguistics

UNESCO United Nations Educational, Scientific and Cultural Organization

VIM Vocabulary of Metrology

SUMMARY

Recent years have witnessed substantial growth in Machine Learning (ML) and Natural Language Processing (NLP), largely fueled by the accessibility and openness of data and models, which is a cornerstone of Open Science. This dissertation builds on this foundation by integrating additional principles of Open Science—transparency, scrutiny, critique, and reproducibility—into the study of these fields.

The dissertation extensively explores and tackles the challenges in reproducibility across both automatic and human evaluations in ML. It begins by unraveling the hidden complexities in evaluating uncertainty, emphasizing the necessity of rigorous statistical analysis, which include effect size and power analysis, and acknowledges the persistent risks of false discoveries despite careful considerations. This is complemented by a comprehensive guide to conducting and reporting uncertainties in evaluations, presenting a crucial resource for researchers to enhance the reliability of their findings.

Further dissecting reproducibility challenges, we investigate the trends in availability of research artifacts and examines the impact of community-driven initiatives aimed at improving reporting practices. Furthermore, we present reproducibility assessment of eight scientific papers. Despite certain improvements spurred by community-driven initiatives for better reporting practices, there remain major issues that hinder reproducibility. An in-depth case study on the reproducibility of a text simplification pipeline reveals several overlooked reproducibility challenges such as bugs and dependency issues. Reproducibility of human evaluations is also scrutinized through two case studies. After observing mixed results, we identify several factors that contribute to inconsistencies in human evaluations, including small sample sizes and dynamic conditions. Through these analyses, the dissertation underscores the ongoing challenges in achieving reproducibility in ML and NLP, offering insights to bolster the reliability of future research within these dynamic fields.

Preface: On Open Science and Machine Learning

The United Nations Educational, Scientific and Cultural Organization (UNESCO) has recently released its recommendation on Open Science [6]. These recommendations are based on the consideration that science's efficiency, effectiveness, and impact can be improved by making scientific knowledge, data, and information openly available, accessible, and reusable. Open science promotes the sharing of research data and materials, and the use of open licenses for publications. It highlights the importance of increased scrutiny and transparency in the research process. Furthermore, it underlines accessibility and inclusivity as key principles of science.

In delineating the framework of Open Science, UNESCO's report identifies quality and integrity, collective benefit, equity and fairness, as well as diversity and inclusiveness as its core values. These values serve as the foundation for the guiding principles outlined within the report, which include:

- Transparency, scrutiny, critique, and reproducibility
- Equality of opportunities
- Responsibility, respect, and accountability
- Collaboration, participation, and inclusion
- Flexibility
- Sustainability

UNESCO's report proposes a series of actionable steps aimed at fostering Open Science. These measures encompass the advocacy of Open Science values, collaboration, and the adoption of innovative technical methodologies. While the full implementation of these actions may span years, if not decades, initiating progress begins with undertaking smaller, tangible steps feasible in the present moment.

The pervasive culture of openness within the realm of machine learning facilitates the development of numerous open-source models, tools, and datasets, thereby significantly advancing research and practice in the field. This accessibility has catalyzed unprecedented growth over the past decade, underscoring the potential impact of machine learning on society. Given this potential, even minor adjustments in research practices hold promise for substantial long-term benefits.

Hence, the primary objective of this dissertation is to investigate the application of Open Science principles to research in ML and NLP. Specifically, we focus on center on guiding tenets of Open Science: transparency, scrutiny, critique, and reproducibility. By embracing openness and transparency, the community can aim to foster trust in scientific inquiry and facilitate rigorous scrutiny and critique of the research process. Ultimately, such efforts can serve as a self-correcting mechanism, engendering more dependable and reproducible research outcomes to the collective advantage of the scientific community.

Chapter 1

Measurement and Evaluation

1.1 Introduction

Evaluation is a fundamental part of any scientific endeavor. It is the process by which the quality, effectiveness, or value of a particular entity is assessed. In the context of ML and NLP, evaluation is crucial for determining the performance of models and algorithms. The results of these evaluations are often used to compare different approaches, validate hypotheses, and make decisions about the utility of a particular method or technique.

Evaluation in ML typically involves a combination of automatic and human assessments. Automatic evaluation employs metrics and algorithms to assess the performance of a model or system. Human evaluation, on the other hand, involves human annotators assessing the quality of a model's or system's output. Both types of evaluation have their strengths and weaknesses, each providing valuable insights into the performance of a model or system.

At its core, evaluation is a measurement process. It involves measuring the performance of a model or system against predefined criteria or benchmarks. Therefore, understanding the principles and limitations of measurements is essential for conducting rigorous and reliable evaluations.

1.2 Contents of This Chapter

This chapter serves as an introduction to the dissertation. It provides an overview of the key concepts and terms related to measurement and evaluation in ML and NLP. The chapter is organized as follows: Section 1.3 provides background information on key terms and definitions used throughout this dissertation. Section 1.4 discusses non-metric definitions of reproducibility and replicability. Section 1.5 provides a brief background on deep learning and its relevance to the reproducibility of ML research. Section 1.6 reviews related work on reproducibility and open science in the ML and NLP communities. With this background contextualized, we then discuss the challenges associated with reproducibility in ML research in Section 1.7 and outline the structure of the dissertation.

1.3 Terms and Definitions

Progress has been made in standardizing research practices for fostering reproducible and open science, but the approach to this topic has been diverse, resulting in a range of terms and definitions being used interchangeably. Given the novelty of this field, it is crucial to establish clear definitions to mitigate confusion and ensure effective communication. This section outlines key terms relevant to the discourse drawn from the framework outlined by Joint Committee for Guides in Metrology [7], JCGM [8], Plant et al. [9], and Belz et al. [10].

Consider a deep learning model that is trained to perform Named Entity Recognition (NER) on a dataset. The model is evaluated using precision, recall, and F1-score. The following terms are relevant to this scenario:

Measurand. The quantity or a property intended to be measured. The F1-score of the NER model is the measurand in this case.

True Value. True value of the measurand is the value consistent with the definition of the measurand, but it is **unknowable** in practice.

Value. Value of a quantity is the magnitude of the quantity expressed as a number and a reference to a particular unit.

Reference Value. Value used as a basis for comparison with the measured value of a quantity.

Measurement. Measurement is the set of operations performed with the goal of determining the value of the measurand. Measurements are imperfect due to errors or uncertainty. In the case of the NER model, all the process and operations involved in calculating the F1-score are considered the measurement.

Conditions of Measurement. Conditions of measurement are the set of variables that affect the observed result. These include but are not limited to source code, runtime environment, data, modeling algorithm, and hardware.

Error. The error of a measurement is the difference between the measured value and the reference value of the measurand. The error is often categorized into systematic and random. Systematic errors are consistent and predictable, while random errors are unpredictable and vary from one measurement to another. In the context of the NER model, floating-point arithmetic and rounding errors are examples of random errors. On the other hand, if an evaluation script has a bug that consistently overestimates the F1-score, it is a systematic error.

Uncertainty. The uncertainty of a measurement is a parameter that characterizes the dispersion of the measured values that could reasonably quantify the measurand. It reflects the lack of exact knowledge of the value of the measurand. Note that uncertainty and error are not synonyms. In the context of deep learning models, nonrepresentative sampling and variations in repeated observations of the measurand under apparently identical conditions could be attributed as the primary sources of uncertainty. Often a value of a measurand is determined by a set of quantities. Repeated observations are often considered more objective or statistically rigorous, but this consideration fails to account for the judgment and interpretation that is inherent in the process of measurement.

Reproducibility. Within the Vocabulary of Metrology (VIM) framework, reproducibility is defined as measurement precision under a set of conditions. Using precision enables the quantification of reproducibility by standard statistical measures such as the Coefficient of Variation (CV), computed as the ratio between the standard deviation σ and mean μ of measured scores: $\frac{\sigma}{\mu}$. Specifically, when computing these statistical measures the conditions are all the variables affecting an observed result, including but not limited to source code, runtime environment, data, modeling algorithm, and hardware. In practice, due to the small sample sizes of most experiments, the CV is adjusted to account for this bias as suggested by Belz [11] and is referred to as the Unbiased Coefficient of Variation (CV*). Aside from CV, correlation coefficients such as Pearson's r and Spearman's ρ can also be used to quantify reproducibility of two sets of scores.

Replicability. Following the definition of reproducibility, we define *replicability* (also sometimes referred to as *repeatability*) as reproducibility under the same conditions as the original measurement.

Research Artifacts. All of the materials and conditions recorded and released by authors to achieve reproducibility is referred to as research artifacts. These include but are not limited to source code, data, models, and hyperparameters. Undeniably, capturing all conditions involved in scientific experiments is an ever-evolving and challenging task. Currently, the best solution is to provide a *self-contained* docker container or a virtual machine image that can be used to reproduce the results. While self-containment might not seem necessary at first glance, it prolongs the life time of the artifacts by reducing the chance of reliance on unavailable dependencies or irreproducible setups. While this concept has not been fully adapted by the machine learning community, it is considered as the best practice in other computing research areas such as systems and software engineering [12, 13].

1.4 Non-Metric Definitions

Although not utilized in this work, it is important to note common definitions found in the literature. Reproducibility is often used interchangeably with correctness, but asserting so requires judgment and interpretation [7].

Reproducibility has also been defined as the ability to achieve identical results using the same methods and data. This definition is frequently applied in computational research, where replicating study results depends on employing the same code and data. In this context, reproducibility assessments often yield a binary outcome: either the results match or they do not. However, this approach overlooks factors beyond code and data that may influence study outcomes. For example, if reproduced F1 score of NER model is off by 1 point, it is considered as a fail reproduction. Thus, reducing reproducibility to a binary outcome is insufficient to capture the complexity of the reproducibility problem. A recent study [14] found that only 14.03% of 513 reproduction score pairs matched, underscoring the importance of employing more nuanced quantitative measures to comprehend the extent of reproducibility.

Prior to the above definitions being popularized, Rougier et al. [15] defined *reproducing* as running the same software on the same input data and obtaining the same results. *Replicating* was then limited to running new software and achieving results judged as similar enough by an expert in the field. The Association for Computing Machinery (ACM) [16] considers the team and the experimental setup as contributing factors to reproducibility and replicability; furthermore, they add another term, *repeatability*, to the glossary of definitions. Whitaker [17] and Schloss [18] introduced two additional concepts known as *robustness* and *generalizability* to cover other missing dimensions pertaining to the extent that externally-produced research findings can be verified. Although the creation of these varied and specific definitions was well-intended, they attempt to cover an open-ended number of dimensions; rendering them obsolete upon the arrival of a more comprehensive definition [10].

1.5 Deep Learning

To ensure all readers grasp the unique challenges of reproducing modern ML and NLP research, this section offers a straightforward explanation of deep learning fundamentals, highlighting how they differ from other scientific and computational areas.

Deep learning (or deep neural networks) is currently the most advanced method in machine learning. It uses a system modeled with hierarchical layers of computational units known as neurons, crucial elements of neural networks. These networks draw inspiration from the way biological brains function.

Neural Networks. Neural networks include multiple neurons arranged into layers. Each neuron calculates a simple math operation, usually a linear transformation, to produce an output. Although basic, arranging multiple neurons into layers, adding non-linear functions, and layering these groups together—known as "hidden layers"—enhances the network's capabilities significantly.

Hidden Layers. Located between the input layer (where data enters) and the output layer (where the network's results are outputted), hidden layers enable neural networks to learn more complex and abstract aspects of the data. These layers are called "hidden" because they don't directly interact with the input or outputs.

Non-linearity. Neural networks apply non-linear functions such as Rectified Linear Unit (ReLU) or Sigmoid to outputs. This allows the networks to pick up on complex patterns beyond what linear operations could achieve alone.

Softmax Layer. Commonly used in the last layer of a classification network, the softmax layer transforms the output scores or logits from the network into probabilities. This helps in understanding how likely each class is the correct classification based on the given inputs.

Parameters and Hyperparameters. Parameters are parts of the model adjusted during training (like weights and biases in neurons). Hyperparameters are settings determined before training starts and typically remain unchanged, like learning rate, number of training rounds, or number of hidden layers.

Training with Backpropagation. Neural networks refine their accuracy through a process called backpropagation. It calculates how much a function, measuring prediction error, changes concerning each parameter in the network. Adjustments are then made to these parameters through techniques like gradient descent.

Stochastic Gradient Descent (SGD) Training large models or using big datasets makes computing gradients for the entire dataset impractical due to memory limits. SGD helps by estimating the gradient from smaller data batches, which helps in efficiently updating parameters to handle large data volumes.

Modern deep learning models are often complex pipelines and are reaching trillion parameter sizes [19]. They utilize attention mechanisms [20, 21], custom learning curves and other techniques to achieve state-of-the-art results. Each configuration or setting adds another hyperparameter that may require tuning. Furthermore, there are at least three sources of randomness in deep learning. First, the initialization of the parameters is random. Secondly, there is often a random component in the optimization algorithm. For example, stochastic gradient descent uses a random subset of the training data to calculate the gradient. Since the order of the training data is shuffled, each run of the algorithm uses a different subset of the training data and therefore produces different results. Lastly, Dropout [22] or random transformation applied to the input data impacts the results.

1.6 Related Work

Many top-tier venues have encouraged researchers to share more implementation details regarding their work over the last few years [23, 24, 16, 25, 26]. Proposed checklists such as the ML Reproducibility Checklist [27] and the ML Completeness Checklist [28] ask authors to provide specifications of dependencies, training and evaluation code, pre-trained models, and proper documentation on how to run the provided code.

Subsequently, we provide a succinct overview of the progression of efforts towards reproducibility and Open Science within the ML and NLP communities. Numerous studies have assessed the reproducibility of scientific publications. This task often involves attempting to achieve results *close enough* to the ones reported in the paper with little to no reliance on the released software artifacts, if available. Raff [29] attempts to quantify the reproducibility ratio of 255 papers published from 1984 to 2017. He selects different thresholds for a minimal acceptable error for algorithmic and empirical claims, ultimately reporting a 63% reproducibility ratio. In a similar study, Wieling et al. [30] survey 395 papers presented at the Association for Computational Linguistics (ACL) 2011 and 2016 conferences and identify whether links to data and code were provided. Then, they attempt to reproduce the results of ten papers using provided code and data. They ultimately find results that are close to those reported for six papers.

Olorisade et al. [31] attempt to independently investigate the claims of six studies on a very specific topic: text mining for citation screening. In the authors' words, 27% of machine learning papers lack the necessary information required for achieving reproducible results; hence, they introduce a checklist to help mitigate this issue. The challenge of dealing with missing information has also been brought up by Gundersen et al. [32]. Utilizing checklists and guiding authors towards better standards during the paper submission process has become a common practice at several venues, including NeurIPS [33], Nature [23], AAAI [24], ACM [16], and

ACL [25]. Aside from guidelines, communities have organized reproducibility challenges [34, 35] that attempt to promote improved reproducibility across the field.

Pineau et al. [33] provided a report after conducting the NeurIPS 2019 Reproducibility Challenge, recapping their conclusions. They highlight that in the field of Computer Science (CS), empirical results have been a major driving factor for recent research advancements. However, the reproducibility of the experiments in the CS field, where experiments rely upon hardware and software designed by humans, is even worse than that observed in other scientific fields including biology, physics, and sociology [36]. They list a few possible roots for this problem, especially when it comes to ML research. These include:

- Lack of access to the same training data
- Misspecification or under-specification of the model or training procedure
- Lack of availability of the code necessary to run the experiments, or errors in the available code
- Under-specification of the metrics used to report results
- Improper use of statistics to analyze results, such as claiming significance without proper statistical testing or using the wrong statistical test
- Selective reporting of results and ignoring the danger of overfitting
- Over-claiming of the results, by drawing conclusions that go beyond the evidence presented (*e.g.,* , making claims based on an insufficient number of experiments, or making claims that are mismatched from the hypothesis)

To improve the standards of reproducibility across the community, Pineau et al. [33] present a reproducibility program with three components: 1) a code submission policy, 2) a community-wide reproducibility program, and 3) the inclusion of the ML Reproducibility Checklist as part of the paper submission process. Then they highlight three major challenges in the field of ML compared to other disciplines. First, ML research suffers from a uniquely insufficient exploration of variables that might affect the conclusions of a study. Moreover, ML research is often improperly documented and reported, which could complicate reproduction attempts. Lastly, statistical analysis has become less common in ML research, often resulting in unclear statistical significance.

In the end, they discuss four frequent objections to reproducibility standards, including dataset confidentiality, proprietary software, computation infrastructure, and replication of mistakes. They mention that they have provided enough flexibility in their own guidelines to account for these exceptions so that authors do not feel pressured when they submit their work.

1.7 Challenges in Reproducibility of Machine Learning Research

Every evaluation procedure conducted as part of ML research is a measurement reported. Each measurement is impacted by three main factors, the conditions of measurement, uncertainty, and error. Therefore, understanding the risks and challenges associated with reproducibility of ML research necessitates a comprehensive understanding of the factors affecting measurements performed as part of both automatic and human evaluations.

This dissertation is organized into three main chapters. Chapter 2 addresses uncertainty in evaluation. Here, we explore best practices for managing uncertainty in both automatic and human evaluations and emphasize the importance of conducting rigorous statistical analyses. We provide a detailed discussion on often overlooked elements such as effect size and power analysis. Moreover, through a case study, we illustrate that despite meticulous statistical analysis, the risk of encountering false discoveries still persists. We demonstrate using current evaluation practices when there is a marginal performance difference between two models can lead to false discoveries. Such discoveries in turn lead to irreproducible results. The primary contribution of this chapter is to provide a comprehensive guide for researchers to accurately address and report uncertainty in their evaluations.

Chapter 3 explores the challenges specific to the reproducibility of automatic evaluations. The chapter is structured into three sections: Research Artifacts Availability, Adequacy of Existing Research Artifacts, and In-depth Reproducibility Assessment of a Neural Text Simplification (NTS) Pipeline.

A primary challenge in replicating ML research is the specificity of reported conditions, which encompasses all variables influencing the observed results. These variables include, but are not limited to, source code, runtime environment, data, modeling algorithms, and hardware. The lack of detailed reporting on these conditions often necessitates guesswork to fill in missing details, thereby compromising the integrity of reproduction attempts. This chapter begins with an analysis of the availability trends of research artifacts and examines the impact of community-driven initiatives aimed at enhancing reporting practices (Section 3.3). Our findings indicate that such initiatives have indeed improved the availability of research artifacts, particularly in communities with a strong focus on reproducibility.

Further, this chapter presents a reproducibility assessment of eight papers from the Empirical Methods in Natural Language Processing (EMNLP) 2021 conference (Section 3.4). The assessment reveals that, despite improvements, reporting practices still fall short of enabling straightforward reproductions. We advocate for containerization as a useful technique to mitigate these issues, offering a more controlled environment for machine learning experiments.

Lastly, the chapter conducts a detailed case study on the reproducibility of a NTS pipeline (Section 3.5). Despite its significant citation count and adherence to recommended checklist items, the study exposes several overlooked reproducibility challenges. We identified three major bugs within the codebase, emphasizing the need for thorough evaluations of research artifacts. Such discoveries underscore the importance of rigorous scrutiny in research to ensure genuine reproducibility.

Chapter 4 shifts focus towards the reproducibility of human evaluations, an essential component alongside automatic evaluations in ML research. Particularly in NLP tasks, human evaluations are often regarded as the gold standard due to their ability to capture nuances that automated methods might miss. This chapter details our efforts to replicate human evaluation results from two NLP-focused research papers.

Our replication attempts yielded mixed results: we successfully reproduced the human evaluations from one paper, but encountered high variability in the results from the other. This variability prompts a critical examination of several aspects of experimental design that may contribute to such inconsistencies. One significant factor identified was the low statistical power resulting from the small sample sizes used in the studies. Furthermore, we draw attention to conditions with variable outcome. For example, the minimum number of approved tasks for each annotator results in a different annotator pool as time progresses.

We propose that in scenarios where outcomes of conditions are likely to change, researchers should consider equating the conditions to reflect the current state rather than strictly adhering to those used in the original experiment. This approach increases the relevance and adaptability of the research, potentially leading to more consistent and reproducible outcomes in human evaluations

Finally, in Chapter 5, we close this dissertation re-iterating key insights and take-home messages from the thesis as a whole, both for the ML and NLP community and the scientific community at large.

Chapter 2

Uncertainty in Evaluation

2.1 Introduction

Recent strides in machine learning are driven by empirical evidence. Deep learning models' performance is assessed across various automatic benchmarks (c.f., Wang et al. [37] or Srivastava et al. [38]) and human evaluations; however, measures of their performance are susceptible to uncertainties stemming from factors like random initialization, data shuffling, stochastic gradient descent, and dropout [39, 36, 40, 41, 42, 43, 44, 45, 1, 46]. Human evaluations are also affected by uncertainty. Participants of each evaluation are individuals that are sampled from the entire population; therefore, one cannot expect the exact same results from repeating the evaluation with different participants. What is worse is that human evaluations are costly to conduct, forcing researchers to rely on small sample sizes, which in turns increases this variability.

In this chapter, we attempt to examine uncertainty inherent in the automatic and human evaluation of machine learning models. While uncertainty cannot be eliminated, following best practices in reporting can reduce the chances of false discoveries and improve the reliability of research findings [9]. Adopting these practices also aligns with the broader goals of the Open Science movement, which advocates for transparency, reproducibility, and inclusivity in scientific research to foster improved scholarship [6]. At an application level, particularly in sensitive domains such as healthcare, following rigorous evaluation practices is essential to ensure the safety and efficacy of machine learning models.

Our work draws inspiration from Guide to the expression of Uncertainty in Measurement (GUM) [7]. This document is based on the recommendations of Joint Committee for Guides in Metrology [7] and provides a comprehensive guide on how to evaluate and report uncertainty in measurement. By using standard definitions utilized in other scientific fields, we aim to provide a common language for researchers to discuss uncertainty in the evaluation of machine learning learning models. Furthermore, this enables us to leverage lessons learned in other evidence-based fields, such as medicine, physics, and chemistry, to improve the reporting of uncertainty in machine learning model evaluation.

Additionally, we focus on best practices for data analysis and reporting in scientific research as outlined by Plant et al. [9]. We present a unique perspective on the evaluation of machine learning models that is often ignored in the literature. Evidence suggests that much of the ML and NLP community is frustrated by indiscriminate reliance on benchmark values as a measure of success [47], and our work here validates that frustration. Our contribution is threefold: (1) we provide a comprehensive guide on how to report uncertainty in the evaluation of machine learning models, (2) we present a case study on the automatic evaluation of deep learning models trained with different random seeds, and (3) we discuss the implications of our findings for the broader communities that utilize machine learning in their work. We hope that our work serves as a reference for future researchers interested in rigorously evaluating and reporting the results of their research on machine learning models.

2.2 Contents of this Chapter

In this chapter, we provide a comprehensive guide on how to report uncertainty in the evaluation of machine learning models. We begin by discussing the metrology perspective on uncertainty in evaluation, highlighting the importance of understanding and reporting uncertainty in the evaluation of machine learning models. We then delve into statistical significance testing, providing background on the concepts of Type I and Type II errors, statistical power, and effect size. We discuss the reliability of statistical inference and address issues in statistical significance testing. We present a case study to demonstrate the application of these concepts in practice, using two experiments that utilize deep learning models for text generation tasks. We conclude with a discussion of our results and their implications for the broader ML community. This chapter heavily relies upon definitions provided in the first chapter, section 1.3.

2.3 Metrology Perspective on Uncertainty in Evaluation

As previously discussed in Chapter 1 (Section 1.3), measurements are imperfect due to errors or uncertainty. The *uncertainty* of a measurement is a parameter that characterizes the dispersion of the measured values that could reasonably quantify the measurand. It reflects the lack of exact knowledge of the value of the measurand. Note that uncertainty and *error* are not synonyms. In the context of deep learning models, nonrepresentative sampling and variations in repeated observations of the measurand under apparently identical conditions could be attributed as the primary sources of uncertainty. Often a value of a measurand is determined by a set of quantities. Repeated observations are often considered more *objective* or *statistically rigorous*, but this consideration fails to account for the judgment and interpretation that is inherent in the process of measurement. GUM highlights two primary methods for evaluating uncertainty: Type A and Type B evaluations.

Type A evaluation of uncertainty relies on observed frequency distributions, deriving standard uncertainty from probability density functions based on these distributions. On the other hand, Type B evaluation is based on assumed probability density functions derived from beliefs about the likelihood of events occurring, often termed *subjective probability*. Both methods use probability interpretations, with Type B evaluations drawing from a pool of information including previous data, experience, manufacturer specifications, calibration certificates, and uncertainties from reference data. Type B evaluations can be as reliable as Type A, particularly when Type A evaluations are based on a limited number of statistically independent observations. Alas, in the context of deep learning where underlying data distributions and model behaviors can be complex and poorly understood, the lack of knowledge about the prior distribution limits the applicability of Type B evaluation methods.

Given the reliance of Type A evaluation on observed data, statistical analysis becomes a critical tool for quantifying uncertainty by extrapolating from observed frequency distributions to evaluate the reliability of these data-based assessments. One of the simplest ways for statistical analysis of a series of observations is to calculate the estimated standard deviation of the measurand. The estimated standard deviation is termed the *standard uncertainty* of the measurand. Statistical tests can also be utilized for this purpose. In general, GUM recommends that it is preferable to err on the side of providing **too much information** rather than too little when it comes to reporting uncertainty. We focus our investigation on the statistical analysis of a series of observations, and extend our study beyond what is covered in GUM.

	Pred H_0	**Pred** H_a
True H_0	True Negative (TN)	False Positive (Type I Error, α)
True H_a	False Negative (Type II Error, β)	True Positive (TP)

Table 2.1: Four possible outcomes of a hypothesis test.

2.4 Statistical Significance Testing

2.4.1 Background

Null Hypothesis Significance Testing (NHST) is one of the most well-known statistical methods for quantifying the likelihood that the observed data would occur if the null hypothesis were true [48]. Many statistics textbooks cover this process in depth (c.f., Warne [49] or Moore et al. [50]); briefly, the null hypothesis, H_0, is a statement that there is no *effect* or no difference between groups. The alternative hypothesis, H_a, is a statement that there is an effect or difference between groups. The goal of NHST is to determine whether the observed data provides enough evidence to reject the null hypothesis, and the *test statistic* is a value calculated from the data that is used to make this determination. The p-value is correspondingly the probability of obtaining a test statistic with a value that is at least as extreme as the one that was actually observed, assuming that the null hypothesis is true.

In the realm of hypothesis testing, several key concepts play pivotal roles in understanding the validity and reliability of statistical inferences. *Type I error* (false negative) refers to the probability of incorrectly rejecting the null hypothesis when it is true, denoted as α. *Type II error* (false positive) occurs when the null hypothesis is not rejected despite being false, signifying a failure to detect a true effect or difference when one exists. Beta (β) represents the probability of committing a Type II error. If the p-value is less than α, the results are considered statistically significant; hence, α is also known as the *significance level*. Dror et al. [51] provide a useful primer regarding the use of these concepts specifically in NLP settings. These two types of errors are crucial to consider in hypothesis testing as they influence the accuracy of research conclusions and the interpretation of study outcomes [52]. Table 2.1 illustrates the four possible outcomes of a hypothesis test.

Statistical power, often represented as $1 - \beta$, reflects the ability of a statistical test to detect a true effect or difference when it exists. It quantifies the likelihood of correctly rejecting the null hypothesis, thus minimizing the risk of Type II errors. Power is influenced by various factors, including the sample size, effect size, variability of the data, and chosen significance level (α). A larger sample size and effect size tend to increase power, as they provide greater precision and enhance the ability to detect differences [53]. Statistical power is known to be difficult to compute for many studies, and many studies have been found to be underpowered [54]; as in other fields, this contributes to the reproducibility crisis [53].

Finally, *effect size*, another critical factor, measures the magnitude of the observed effect or difference. A larger effect size indicates a more substantial impact of the independent variable on the dependent variable, making it easier to detect statistically significant results [55]. However, like other concepts, its importance has been under-recognized by researchers [56, 57].

2.4.2 Reliability of Statistical Inference

Overall, understanding the concepts defined in §2.4 is essential for researchers to design studies with sufficient power, interpret results accurately, and draw meaningful conclusions from their findings. However, this practice has been also under the receiving end of severe criticism [58, 55, 59]. In fact, Cohen [55] highlights several major issues with the NHST. First, the p-value is often misinterpreted as the probability that the null hypothesis is true, when in fact it is the probability of observing the data given that the null hypothesis is true ($P(D|H_0)$). Moreover, researchers often misinterpret the complement of the p-value, $1 - p$, as the probability that the results can be replicated. He underscores that rejecting the null hypothesis does not imply that the original theory is correct.

Ioannidis [60] drew attention to post-study probability that a statistically significant finding is true. Consider a simple example, given 1,000 possible hypotheses and assuming that 100 of such hypotheses are true. With 80% power, we would be able to detect 80 of the 100 true hypotheses. Simultaneously, there would be 45 false positives (5% significance level × 900). This would result in 36% (45 / (45 + 80)) statistically significant results that are false. This probability is referred to as false-positive report probability [60]. What is worse is that this probability increases when the experiment is underpowered. Forstmeier et al. [61] visually demonstrate this probability in various scenarios.

2.4.3 Addressing Issues in Statistical Significance Testing

Such reports paint a grim picture of the reliability of scientific research; nonetheless, it is essential to keep in mind that *not* conducting such analyses would be objectively worse [62]. Inclusion of statistical analyses highlights attention to and care for reporting uncertainty [63], and it enables other researchers to more confidently reproduce scientific results and assess the validity of reported findings [10]. Ultimately, similar to Simons [64], we believe in the *Trust but Verify* principle. The verification, in the form of reproductions conducted by independent researchers, would eliminate various sources of bias and errors.

There have been numerous proposals to address issues regarding statistical significance testing. McShane et al. [65] suggested abandoning the usage of thresholds for statistical significance; instead they propose using p-value as a continuous variable. Wasserstein et al. [66] highlighted six principles for proper use of p-values, and Benjamin et al. [67] suggested changing the p-value threshold from 0.05 to 0.005. While there is no consensus on the best approach, these proposals underscore the importance of understanding the limitations of statistical significance testing and the need for more nuanced and thoughtful approaches to evaluating scientific results.

Excluding statistical analyses would result in under-specified publications—this would be a step backward for the community, and it is against the recommendation of GUM [7]. Several other studies have also focused on utilizing the current best practices while simultaneously highlighting the importance of meta-analyses that mitigate the issues mentioned in §2.4.2 [68, 69, 70, 71, 72, 9, 64, 73, 74]. While "correctness" and "truth" are desired outcomes of a scientific method, they are not achievable without judgment, interpretation, and subjective evaluation.

2.5 Case Study

In the previous section, we discussed the importance of understanding and reporting uncertainty in the evaluation of machine learning models. We now present a case study to demonstrate the application of these concepts in practice, using two experiments that utilize deep learning models for text generation tasks. For each experiment, we repeat an identical training process with different random seeds. Then, we use appropriate statistical tests to evaluate the significance of the differences in the performance of the models trained with different random seeds. Since the null hypothesis, H_0, that the performance of the models is the same, is true,[1] we expect the statistical tests to have a high p-value. In other words, we expect the tests to fail to reject the null hypothesis.

Earlier studies have shown that the performance variation observed due to seemingly inconsequential factors can be statistically significant [75, 76, 77]. This is an expected outcome; however, we are interested in the rate of such significant differences. To achieve this, we randomly split the results into two groups and compare the performance of the models in each

[1] Models by definition are the same when they are trained under identical conditions, and random initialization should not have a noticeable influence on this.

group. We repeat this process multiple times to evaluate the reliability of the statistical tests. Essentially, this experimental design is intended to “stress test” the significance tests.

2.5.1 Statistical Tests

We selected two non-parametric tests and one parametric test for evaluating statistical significance. Non-parametric tests do not rely on any assumptions regarding score distribution, which is important when evaluating deep learning models since their results distributions are hard to define. The non-parametric tests we selected are the *paired bootstrap test* [78, 79, 51] and the Almost Stochastic Order (ASO) [80, 81] test. The paired bootstrap test is a common statistical significance test that makes little to no assumptions about the underlying distribution of the data [82]. It is especially suitable for translation tasks, and was implemented in the sacreBLEU Python package [83] as part of the effort to standardize the evaluation of machine translation models. The ASO test is a more recent test designed to compare the performance of two score distributions [81]. We discuss the use of these tests to assess significance in §2.5.3.

The selected parametric test is the two-sided *t-test* [84]. This test is widely used in across the literature to compare the means of two groups. It is based on the assumption that the data is normally distributed, which is often not the case for deep learning models. However, we include this test to compare the results with the non-parametric tests; we expect the results of the paired t-test to be less reliable due to the assumption of normality.

2.5.2 Selected Papers

For our case study we considered papers that: (1) were available in the ACL Anthology; (2) had a documented track record of straightforward reproducibility, such as through their inclusion in reproducibility studies or challenges; and (3) had anticipated resource costs within our compute budget. Specifically regarding the last criterion, the nature of our case study required training models many times to build a representative empirical score distribution; thus, we needed the central experiment in our selected papers to be repeatable enough times within a compute budget of 12 hours of training on a single Graphics Processing Unit (GPU)[2] to perform sound statistical tests. We selected two papers that met these criteria due to our limited budget. These papers are:

***Rethinking Perturbations in Encoder-Decoders for Fast Training* by Takase et al. [85].** This paper examines the problem of efficient training of sequence-to-sequence models. The authors compare simple regularization techniques to perturbations from an efficiency perspective. They report that word dropout and random input replacement can perform similarly to more complex perturbation techniques while being faster.

***Exploring Neural Text Simplification Models* by Nisioi et al. [86].** This paper explores the use of sequence-to-sequence models to generate simplified versions of input text. The authors find that these models yield impressive grammar and effectively preserve the source sentence content, overall achieving higher performance than other contemporary text simplification approaches.

2.5.3 Experiment Design

Changing the random seed during model training results in variations in the model's performance. Thus, the random seed was the independent variable in our experiments; which we varied at each run whereas all other conditions were kept the same. To report the rate of significant differences in statistical tests when varying the random seed during model training, we followed the experimental setup originally utilized by Colas et al. [87, 88]. We trained a model from each

[2] We used consumer-grade GPUs, specifically the NVIDIA RTX 2080TI and RTX Titan.

#	Reference	Task	Reps.
1	Takase et al. [85]	Machine Translation (MT)	77
2	Nisioi et al. [86]	Simplification	39

Table 2.2: Reproduced papers. The number of reproductions (*Reps.*) is the number of random seeds that were used to train the model in separate training runs.

selected paper n times using different random seeds, yielding n unique results for each paper.[3] Then, we used the selected statistical tests to compare the results of the same model across different seeds, and calculated the percentage of tests that rejected the null hypothesis.

The paired bootstrap test required selection of a pair of models to compare among all $\binom{n}{2}$ pairs of trained models. We repeated this process for each of the $\binom{n}{2}$ pairs to calculate the rate of significant differences. The t-test and ASO test accept two sets of results as input, and we repeated the process of selecting two sets of results from the n results for each paper. We calculated the rate of significant differences for each test. In contrast to the paired bootstrap test, the number of potential splits is far greater: $\binom{n}{\frac{n}{2}}$. We used a constant 10,000 random splits to calculate the rate of significant differences for these tests. We set the significance level α to 0.05 for all tests, to accept or reject the following null hypothesis:

***H_0*:** The performance of set A is the same as the performance of set B.

For the selected papers, we re-ran experiments using the authors' code and data to ensure consistency with the original experimental setup and methods. We summarize the experimental tasks and number of training runs included in our study in Table 2.2. We repeated Takase et al. [85]'s work 77 times and Nisioi et al. [86]'s work 39 times, respectively. We carefully followed the instructions in the papers and supplementary materials to set up the required

[3] We set n for each paper to the maximum number of experimental runs we could afford to allocate within our budget.

environment, including using necessary software dependencies, hardware configurations, and data preprocessing steps.

Performing multiple statistical tests on the same data increases the risk of false positives; thus, although our process can estimate the rate of false positives it should not be used to draw other conclusions. Some techniques can be used to control the False Discovery Rate (FDR) under arbitrary dependence assumptions, such as the Benjamini-Yekutieli method [89], but these techniques directly limit the rate of false positives to a predefined desired level across many tests. Since our goal is to provide an estimation of the chance of falsely rejecting the null hypothesis in a single test, we do not apply any FDR control techniques.

2.6 Results

2.6.1 Statistical Power Analysis

Performing a power analysis for the t-test is straightforward. We set $\alpha = 0.05$ and power to 0.8. We followed Cohen [90] to define the effect sizes as follows: small effect=0.2, medium effect=0.5, large effect=0.8, and very large effect=1.2. We calculated the number of samples required to achieve these effect sizes and found that we need 12, 26, 64, and 394 samples for achieving effect size of 1.2, 0.8, 0.5, and 0.2, respectively (smaller effect sizes means that the groups are more similar, therefore, more samples are needed to detect a difference). Note that the t-test is *not* the best test for our case, as the data is not normally distributed, but we included this test to compare the results with the non-parametric tests.

2.6.2 Statistical Significance Testing

We summarize our results for the two selected papers in Table 2.3 using the standard performance metric (BLEU score) for text generation reported in those papers, and provide additional details

	Takase et al. [85]	Nisioi et al. [86]
Reported	36.22[a]	84.51
Mean	36.18	87.90
Median	36.19	88.11
STD	0.20	1.16
Min	34.76	84.47
Max	36.46	89.59

Table 2.3: Summary of BLEU scores for each paper. The superscript [a] indicates the average of three models trained using different seeds by the original authors.

for each reproduced paper in the subsections below.

Paper 1. Takase et al. [85] evaluated their work on several datasets. We limit our experiments to those with the IWSLT 2014 German-English training set,[4] which contains 160k sentence pairs. This dataset is considered low resource and enables training smaller models. In Figure 2.2, we plot the distribution of results generated across all 77 of our experimental runs of their machine translation model using unique random seeds. The plot shows that the model's performance is distributed across a wide range of values. With BLEU=36.18 and BLEU=36.19 for mean and median, respectively, our average results are close to BLEU=36.22, the value originally reported by the authors. However, the best and especially the worst results show the extent of the variation in the model's performance.

Next, we ran the selected statistical tests to evaluate the significance using a set of 77 sample results. We calculated $\binom{77}{2} = 2926$ p-values using the paired bootstrap test. We observed $p < 0.05$ in 22% (655 out of 2926) of cases. For ASO, there were 1.36e+22 possible combinations (all possible combinations of 38[5] out of 77 results), but as mentioned previously we randomly sampled results from 10,000 combinations.[6] We observed $\epsilon < 0.5$ in 24% (2,459

[4] https://sites.google.com/site/iwsltevaluation2014/data-provided

[5] 0.5×77, rounded down to the nearest whole number.

[6] It was computationally intractable to perform the test on all combinations.

out of 10,000) of cases using ASO.[7] A more conservative threshold of $\epsilon < 0.2$ yielded 3% (345 out of 10,000).

Lastly, for the t-test we observed $p < 0.05$ in 4.61% (461 out of 10,000) of cases. This is surprising, as the t-test is considered to be less reliable when the data is not normally distributed. All 461 cases had an effect size larger than 0.5 (medium effect size). Of these, 179 of the 461 cases had an effect size larger than 0.8 (large effect size), and 0 had an effect size greater than 1.2 (very large effect size). The ratio of false positives of the t-test falls within the expected range of 5% for $\alpha = 0.05$.

Paper 2. Nisioi et al. [86] evaluated their work on English Wikipedia[8] and Simple English Wikipedia [91]. They proposed two model variants, one of which used pre-trained word vectors and the other of which did not. We used the latter and repeated the authors' experiment 39 times; Figure 2.1 shows the violin plot of our results. Similarly to Takase et al. [85], the authors used BLEU to evaluate their model output. Surprisingly, our results (see Table 2.3) suggest that their reported BLEU=84.51 is closer to the minimum than the mean and median, which are BLEU=87.90 and BLEU=88.11, respectively. Furthermore, the model with the best BLEU score (BLEU=89.59) outperforms all models in the original paper. A finding that does not align with the original work is that the model with pre-trained word vectors outperforms the rest.

Following the same procedure as for Paper 1, we calculate $\binom{39}{2} = 741$ p-values for the paired bootstrap test. We observe $p < 0.05$ in 29% (216 out of 741) of cases. For our sample of 10,000 sets of results for ASO, we observe $\epsilon < 0.5$ in 17% (1,760 out of 10,000) of cases. On the other hand, with $\epsilon < 0.2$, we observe 4% (438 out of 10,000) of cases with a significant difference. Lastly, for the t-test, only 2.16% (216 out of 10,000) of cases have $p < 0.05$. All 216 of these cases had at least a small effect size (greater than 0.2). Of these, 82 of the 216

[7] ASO returns ϵ, a confidence score, rather than the traditional p; both can reject H_0 when compared to a threshold.

[8] https://en.wikipedia.org/wiki/Main_Page

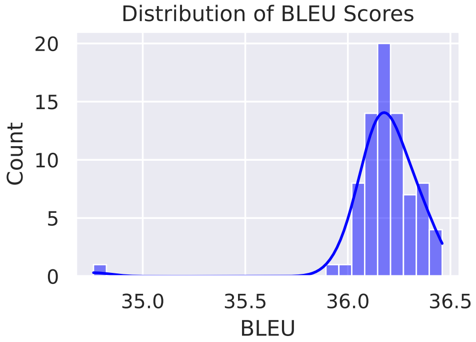


Figure 2.1: Histogram of BLEU scores for Nisioi et al. [86]. The x-axis represents the BLEU score, and the y-axis represents the number of occurrences. The sample size is 39.

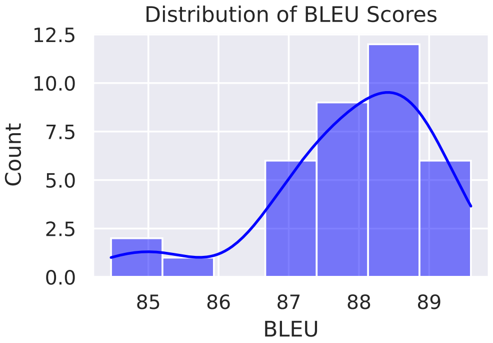


Figure 2.2: Histogram of BLEU scores for Takase et al. [85]. The x-axis represents the BLEU score, and the y-axis represents the number of occurrences. The sample size is 77.

cases had an effect size greater than 0.5 (medium effect size), and 0 had an effect size greater than 0.8 (large effect size). Once again, the ratio of false positives resulting from the t-test falls within the expected range of 5% for $\alpha = 0.05$.

2.7 Discussion

Our results demonstrate a large variation in the performance of models trained with different random seeds, for both papers. This variation is reflected in the wide range of BLEU scores obtained across the different runs, rendering evaluation of the models challenging [82, 51, 54]. We believe that without the additional statistical analysis conducted, one might easily draw incorrect conclusions about the performance of the models.

This empirical observation has important consequences, especially when considering all the hard work that researchers dedicate to designing models that beat the state-of-the-art in a specific task. Given the wide range of scores, it is hard to build confidence that a particular change to an existing model results in any performance improvement observed, specially since statistical significance tests are rarely included in the analysis. It is possible that trivial changes result in better performance, and that meaningful changes result in lower performance, all due to sheer luck.

We found that for a very large effect size, the t-test requires 8 samples to achieve 80% power. This number may be an underestimate, since the data is not necessarily normally distributed. Regardless, we recommend future studies to at least use 8 unique random seeds when reporting the performance of their models. Ideally, the desired effect size is determined before the experiment is conducted, and the number of samples is calculated accordingly. Only then can researchers confidently claim that the observed differences are not due to random chance.

Numerous publications report achieving marginal improvements in performance over the state-of-the-art. Marginal improvements most likely result in a small effect size, which sets the sample size required to achieve 80% power to 199. Currently, even at best scenario, researchers collect 5 to 10 samples, which is far from the required number of samples to achieve 80% power. Knowingly or unknowingly, researchers are likely to report false positives, which in turn result in failed reproductions in the future.

Out of the three tests, the t-test was the only one that had a false positive rate within the expected range of 5%. We ground this finding in the fact that the data is not normally distributed. Ultimately, due to ease of usage of statistical tests in publicly available libraries in python [92, 93, 94, 95], we encourage researchers to utilize all three tests highlighted here to evaluate the significance of their results.

Ultimately, uncertainty engraved in deep learning models is a reality that we must recognize and address to the best of our abilities. This stochasticity poses a risk to the reliability of model evaluation in our field. We hope that our work raises awareness of this and encourages researchers to conduct more rigorous evaluations of their models. Reviewers and community organizers should also take an active role in ensuring scientific publications are not underspecified. Given the limitations of significance testing, we do not recommend gatekeeping publications to only those reporting results that pass significance tests. Rather, we recommend for the inclusion of broad statistical significance analyses in those publications. We underline that false positives can and do occur. Such findings should be investigated to determine the root cause of the false positives. Particularly, if false positives are attributed to the experimental design, future studies can take steps to mitigate the issue.

Our work was limited in several ways. First, the claims that we make are purposely based on experiments that do not leverage pre-trained models. If additional experiments in the future reveal that the variance in results across training runs with different random seeds is not as strong with models utilizing pre-trained language models, then our claims would generalize less broadly. There are many advantages to reusing pre-trained models; one of them is to vastly minimize the impact of weight initialization since a portion of the model is shared everywhere. In turn, this could reduce the variability as a result of random seeds. We hope to investigate this in a future study.

Furthermore, while there is little evidence to suggest that risks pertaining to variation from random seeds are task-specific, we do only explore two tasks in this work, covering machine translation and text simplification. Thus, the number of domains explored in this chapter is limited. Addressing this is complicated due to widespread issues affecting the reproducibility of results. Although we originally hoped to include more tasks in this study, we ran into reproducibility issues of our own that prohibited this.

Finally, the models included in our study were evaluated on relatively small datasets compared to some datasets that are more commonly used. Our selection of models for inclusion in this study was guided by careful resource considerations, since a necessary component of our work was the repeated retraining of these models. Moreover, reproducing work in the first place often requires a lot of debugging and testing in itself. Rigorous evaluation often necessitates down-scaling since it requires many times more computation, and our study exemplified although was not unique in this. We leave it to others to scale this study as yet another future direction, given access to the necessary compute resources.

2.8 Conclusions

In this work, we have shown that different samples of results from the same model, varying only the random seed, can exhibit so much variability that the samples could be considered drawn from two different models. We also demonstrated that additional statistical analysis in the form of significance testing can be employed to evaluate the impact of uncertainty in deep learning models, and that some tests may do better than others at estimating the false positive rate. Surprisingly, in our experiments the t-test was the only significance test for which false positives fell within the anticipated range. Using a more conservative threshold for the ASO test also decreased the false positive rate to an acceptable level.

We underscore that this added scrutiny is beneficial to the community, and results in stronger, more reliable advances. While the risk of false discoveries cannot be eliminated, providing detailed statistical analyses implies a level of consideration for the uncertainty in the results. Furthermore, it enables future researchers to more easily investigate the findings and build upon them. Without a doubt, early detection of any false discovery would be beneficial to the community.

We encourage community organizers and reviewers to further emphasize and demand appropriate significance testing with adequate sample sizes. There does not seem to be a way to "fix" the stochasticity of deep learning models. Eliminating or optimizing the random seed are not viable options, as such actions cannot be transferred to other tasks or models. Instead, the only way to reliably support claims of performance is through inclusion of extensive statistical analyses.

This chapter provided a guide on how to handle uncertainty in evaluations. However, aside from uncertainty, conditions also impact the outcomes of measurements. The next chapter more fully examines these conditions in the context of automated evaluation.

Chapter 3

Automatic Evaluation Reproducibility[1]

3.1 Introduction

The reproducibility of research findings is a cornerstone of scientific integrity, yet in the realm of computational linguistics, challenges persist that can obstruct this fundamental principle. There have been several initiatives to address these challenges, including the introduction of reproducibility checklists and guidelines by major conferences in the field. However, the effectiveness of these measures in promoting reproducibility remains an open question.

This chapter investigates the availability of research artifacts in papers from key conferences such as the ACL, Neural Information Processing Systems (NeurIPS), and Interspeech, utilizing data from respective repositories such as the ACL Anthology. We aim to evaluate whether a conference's focus on reproducibility correlates with enhanced accessibility to research artifacts, reflecting on past and present practices. Moreover, issues such as "code dumps," characterized by the release of source code without adequate guidance or documentation, contribute significantly to reproducibility problems. We hypothesize that these issues, while prevalent, are more effectively addressed in conferences that demand for higher reproducibility standards.

[1] Parts of this chapter were previously published in Arvan et al. [3, 1, 2]

Through a close examination, including attempting to reproduce findings from select EMNLP 2021 papers and reviewing a well-documented study by Nisioi et al. [86], this research aims to not only highlight existing gaps but also to suggest ways in which the community could fortify its commitment to open and reproducible science. This analysis is particularly pertinent in light of recent discussions within the scientific community regarding the reliability of published research and aims to foster a deeper understanding that could shape future conference policies and practices.

3.2 Contents of this Chapter

This chapter is devoted to exploring the nuanced topic of reproducibility within computational linguistics and related scientific disciplines. Recognizing reproducibility as a foundation of scientific integrity, our analysis aims to provide a multi-faceted examination of how reproducibility is currently addressed in various academic forums. Below are the main contents and contributions of this chapter:

1. **Trends in Research Artifact Availability:** Initially, we conduct a comprehensive analysis of the availability of research artifacts, including data and source code, from papers published at top-tier conferences such as those hosted by the ACL, NeurIPS, and Interspeech. The purpose is to identify trends in artifact sharing and assess reproducibility norms across these forums. We compare these practices to understand variances among conferences, investigating whether stringent reproducibility guidelines correlate with better research artifact dissemination.

2. **Adequacy of Existing Research Artifacts:** Next, we address the prevalent issue where the availability of source code does not necessarily equate to reproducibility. By attempting to replicate the results of selected papers from the 2021 Conference on EMNLP, we highlight

the common barriers encountered, such as ”code dumps” and missing dependencies. These challenges illustrate why merely having access to source code often falls short of ensuring that a study can be effectively reproduced.

3. **Detailed Case Study Evaluation:** Lastly, the chapter includes an in-depth evaluation of a specific study from Nisioi et al. [86], chosen due to its exemplary status in terms of reproducibility. We dissect how this particular paper's materials aid in replication efforts, using it as a model to understand what constitutes effective sharing and documentation of research artifacts. This case study helps underline the practices that enhance reproducibility and can serve as a guide for future scientific publications.

By examining these aspects, this chapter contributes to the ongoing discussion on how to improve scientific reproducibility. Each section builds upon the lessons learned from current practices, aiming to suggest actionable improvements for future research endeavors. Contents of this chapter are based on our publications [3, 1, 2].

3.3 Trends in Artifact Availability

3.3.1 Method

To analyze trends in artifact availability, we select five major NLP conferences (ACL, EMNLP, Language Resources and Evaluation Conference (LREC), North American Chapter of the Association for Computational Linguistics (NAACL), and International Conference on Computational Linguistics (COLING)), and scrape the ACL Anthology to obtain data associated with all papers published at those venues. The `aclanthology.org` portal contains “code” and “data” fields that indicate whether the paper contains a link to the source code and the data. We use these fields to determine the availability of research artifacts for these conferences.

We also download papers from Interspeech and NeurIPS for additional comparison. The papers published in Interspeech are available through the isca-speech.org website. This portal provides basic information regarding speech processing papers listed in the conference proceedings, as well as the papers themselves. At the time of writing, the website provides information about 35,050 papers published at 344 conferences. While the ISCA portal is useful for researchers who are searching for papers, it does not provide any information about the availability of research artifacts; hence, retrieving this information originally required us to manually search each paper. To expedite and streamline this process, we wrote a python script to download the PDF files. Then, we used the PyPDF2[2] library to extract the text from the PDFs. Following this, we performed a simple keyword search ("github.com") to determine whether the paper contained any information regarding released software artifacts. GitHub is by and large the most popular website for hosting open source code, making its presence in a paper a reasonable first clue towards source code availability. The references section of the paper was excluded from the search to reduce false positives.

We attempted to follow a similar process to that used for the ACL Anthology for papers published at NeurIPS. This conference utilizes the openreview.net website to host their proceedings. The website provides an API to download accepted papers published at conferences. NeurIPS also provides a paper portal[3] that provides a similar experience to the ACL Anthology. Once this data is collected, we analyze the trends in artifact availability for each conference in terms of ratio and number of papers with research artifacts.

Our research on the availability of artifacts in academic papers has several limitations, primarily due to reliance on basic keyword searches across various databases, including Interspeech and NeurIPS, which could generate false positives or negatives.

[2] https://pypi.org/project/PyPDF2

[3] https://papers.nips.cc

3.3.2 Results

We present our primary results in Figure 3.1. The figure shows the percentage of papers with research artifacts for each conference. On a positive note, all of the conferences have an upward trend for research artifact submission. In the case of NeurIPS, ACL, EMNLP, and NAACL, the percentage has surpassed 50%. The gap between research artifact availability in NeurIPS and Interspeech is nearly 40%.

Out of these conferences, LREC [96], COLING [97], and Interspeech have not yet formally emphasized *reproducibility* in their call for papers. On the other hand, besides NeurIPS which has been the frontrunner, EMNLP [98], ACL [99], and NAACL [100] highlight several reproducibility guidelines for authors to consider during the submission process. Figure 3.1 illustrates percentage of papers with research artifacts at the selected conferences over the years.

We demonstrate the total number of papers accepted at each conference in Figure 3.2. Despite having the most papers accepted, NeurIPS is the frontrunner in terms of research artifact availability. The rest of the conferences are relatively close in terms of total number of papers accepted. We have included this figure to provide context for the percentage of papers with research artifacts.

3.3.3 Discussion

Our results suggests that the percentage of papers with research artifacts in COLING, LREC, and Interspeech was lower than that observed in other conferences. We believe this could create unnecessary barriers for future researchers to adapt and build upon the work presented at these conferences. This problem is not easily solved, as it requires a change in established research practices and acceptance from the community. Certainly there are cases that cannot share their work due to privacy or other concerns. However, we believe that the community and its leadership should strive to make research artifacts available and promote the value of openness.

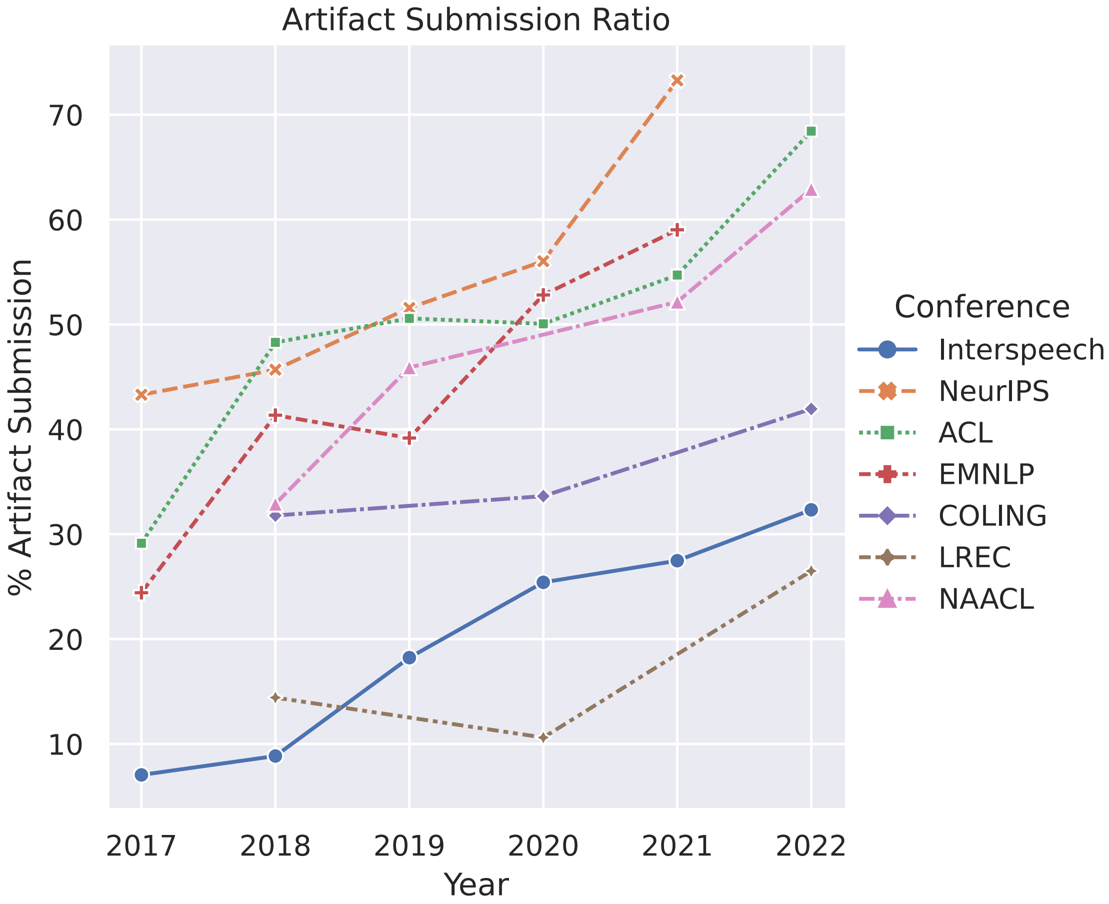


Figure 3.1: Percentage of papers with research artifacts at the selected conferences over the years.

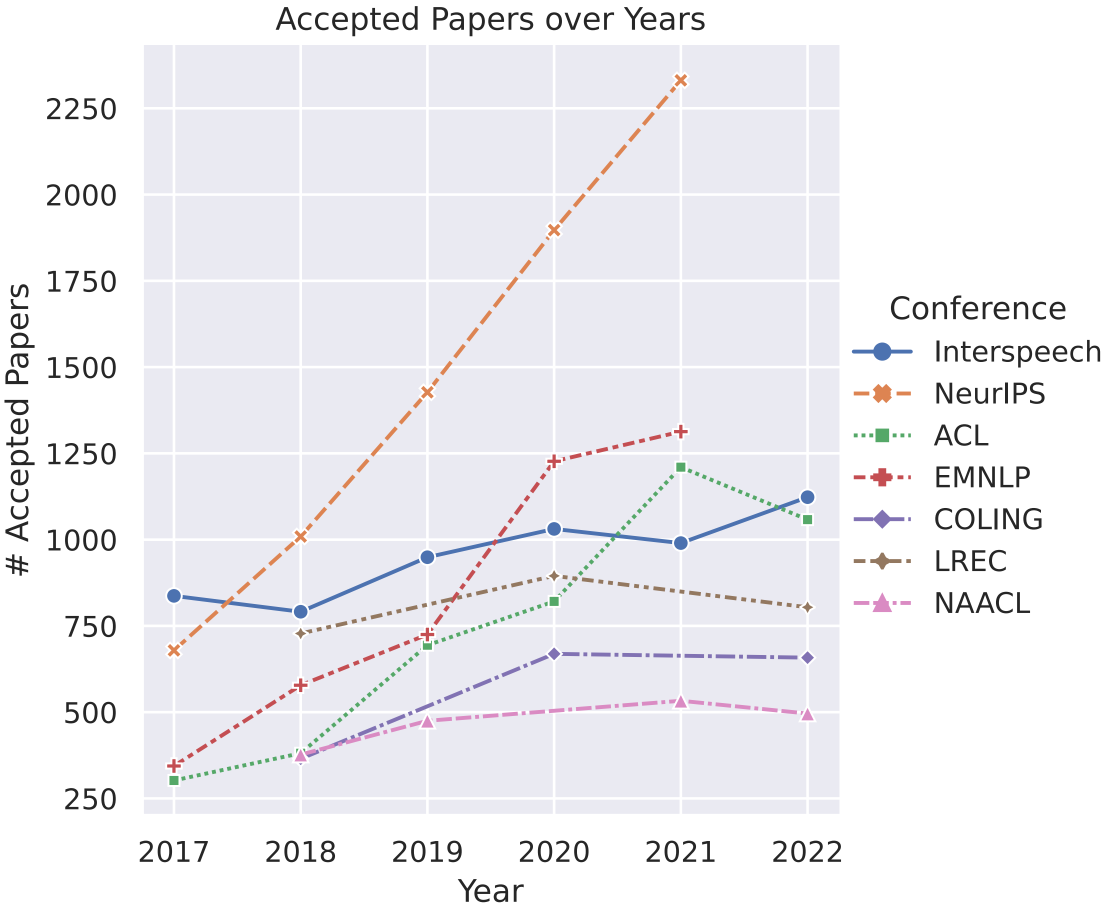


Figure 3.2: Total number of papers accepted at the selected conferences over the years.

3.4 Eight Paper Case Study

3.4.1 Method

Out of over 1300 accepted papers at EMNLP 2021, 723 had URLs to a repository containing the code required to run their experiments. We randomly select eight papers from the *2021 Conference on Empirical Methods in Natural Language Processing* [101]. We attempt to perform necessary steps to measure the performance of the models reported in these papers. Then, we compare the results to those reported in the papers. We use the CV* metric to quantify the reproducibility of the results. CV* was discussed earlier in Section 1.3. We also provide a checklist of the availability of instructions, dependencies, and scripts used for training and evaluation. We summarize the results of our reproducibility assessments in Table 3.1. Our selected papers are provided below. We refer to each paper by its associated number in this list.

1. A Massively Multilingual Analysis of Cross-linguality in Shared Embedding Space [102]
2. Automatically Exposing Problems with Neural Dialog Models [103]
3. Frustratingly Simple but Surprisingly Strong: Using Language-Independent Features for Zero-shot Cross-lingual Semantic Parsing [104]
4. Weakly-supervised Text Classification Based on Keyword Graph [105]
5. ReasonBERT: Pre-trained to Reason with Distant Supervision [106]
6. StreamHover: Livestream Transcript Summarization and Annotation [107]
7. ValNorm Quantifies Semantics to Reveal Consistent Valence Biases Across Languages and Over Centuries [108]
8. Measuring Association Between Labels and Free-Text Rationales [109]

While random selection of the papers avoids inserting selection bias into our findings, it may increase the difficulty of achieving reproducibility due to lower familiarity with certain concepts. We allotted fixed, limited time and computation resources for each paper, and report whether we were able to reproduce the findings of the paper within our time and resource budget. Certain works may require more time and resources than others, and we acknowledge that our results may not be comprehensive. The next case study presented in this chapter (Section 3.5) provides in-depth analysis of a single paper, which provides a contrast from the case study presented here.

For each selected paper, we attempted to follow the instructions provided by the authors to achieve the reported results. In case of errors or missing instructions, we reached out to the authors for clarification. We also reported the issues we faced during the reproducibility process. We provide a checklist of the availability of instructions, dependencies, and scripts used for training and evaluation. We summarize the results of our reproducibility assessments in Table 3.1.

Although we quantify reproducibility using CV*, we consider an attempt successful if we can run the code and record any results. This is in parts due to the limited time, resources, and the complexity of every paper. We also provide a detailed account of the reproducibility process for each paper in Section 3.4.2.

We understand that this process is not comprehensive; regardless, we believe it provides a good indicator for adequacy of released research artifacts. We also acknowledge that the reproducibility of results is not the only measure of the quality of a scientific publication.

3.4.2 Results

Table 3.1 provides an overview of the availability of instructions, dependency specifications, and scripts used for training and evaluation. It also indicates whether we were able to reproduce the results for each paper. This checklist highlights the availability of materials that ultimately aid the reproducibility process. Given that one of the biggest reproducibility hurdles is getting the provided code to a running state, we expected this table to provide an estimate of the difficulty in reproducing the results of each selected paper. We also document the reproduction process for each paper below.

Paper 1. The released source code for Paper 1 [102] contains a list of dependencies and full information on how to run the scripts provided. We found several syntax errors within the code released; fixing these errors took little effort. Unfortunately, the released code then terminated

Title	Instructions	Dependencies	Scripts	Reproducibility
Paper 1 [102]	Yes	Yes	Yes	No
Paper 2 [103]	No	No	No	No
Paper 3 [104]	Yes	Yes	Yes	No
Paper 4 [105]	Yes	Yes	Yes	No
Paper 5 [106]	Yes	Yes	Yes	Yes
Paper 6 [107]	Yes	Yes	Yes	No
Paper 7 [108]	No	No	No	No
Paper 8 [109]	Yes	Yes	Yes	Yes

Table 3.1: Overall results of the reproducibility attempts discussed in this work.

with a runtime error after executing for approximately one hour. We reported this issue to the authors through GitHub's tracking issue,[4] and followed up via email to the first author. While the authors responded to the follow-up email, they were unable to provide a solution to the issue. Thus, our efforts failed to get the released source code to a running state.

Paper 2. The source code released for Paper 2 [103] contains no information or documentation describing how to achieve the reported results. We were unable to understand which files were used to achieve the results reported in the paper. Communications with the authors through GitHub[5] and email were unsuccessful. We were unable to run the code provided for this paper.

Paper 3. In order to reproduce the results presented in Paper3 [104], it is necessary to access specific pre-trained embeddings. Unfortunately, these embeddings are no longer accessible, as indicated by the unavailable link.[6] Concerns also arise regarding the availability of the preprocessing scripts for the dataset, which appear to be similarly inaccessible. Efforts to resolve these issues included direct communication with the authors through a GitHub issue[7] and a follow-up email to the first author. However, these attempts have not elicited any response.

[4] https://github.com/AlexJonesNLP/XLAnalysis5K/issues/1

[5] https://github.com/DianDYu/trigger/issues/1

[6] http://www.let.rug.nl/rikvannoord/DRS/embeddings/

[7] https://github.com/SALT-NLP/Multilingual-DRS-Semantic-Parsing/issues/1

Paper 4. According to the authors of Paper 4 [105], their released source code requires multiple GPUs, and it is not possible to run it on a single GPU. This requirement raises the entry barrier for assessing the reproducibility of this work. Fortunately, we had access to a multi-GPU workstation, so we were able to continue with our reproducibility analysis only to find two errors. First, there was a missing dependency, which was straightforward to fix. Second, we ran into a mismatched device error during the source code runtime, which originates from mishandling the device (Central Processing Unit (CPU) or GPU) used for the data or the model. Regardless of what device is used, if there is an operation between two tensors, they have to be on the same device. We reported this issue through GitHub.[8] Although we did not receive a response prior to the initial submission, we did hear back later after we followed up with the first author via email. The authors responded with a solution. Other users reported experiencing the same issue and that the solution appears to be working. This case offers a good example of effective communication and collaboration to improve the reproducibility of the publication.

Paper 5. We were able to run the source code released for Paper 5 [106] without any issues. The source code included quick experiments with smaller data samples, which made the reproducibility assessment of this work easy and straightforward. Additionally, the authors provided clear and concise instructions on how to achieve the results reported in the paper. The training took minutes for each model, which would have made debugging easier if we had encountered issues.

Since we were able to successfully reproduce this paper, we measured the CV* for the results of several models on the SQuAD dataset reported in Table 4 of Paper 5 [106]. We present the results in the top portion of Table 3.2. We note that the reported results are from an average of five runs; however, to the best of our knowledge, the authors have not released the full results. We observe CV* values ranging from 17.04 to 139.92. This is less than ideal, but it can be

[8] https://github.com/zhanglu-cst/ClassKG/issues/5

justified by the small size of the dataset and the random seed affecting the data order. We believe that running more experiments to increase the sample size would yield lower CV*. We mark this reproducibility attempt as successful.

Paper	Model	Reported	Ours	Mean	Unbiased St. Dev.	CV* ↓
5	BERT	9.9	5.78	7.84	3.64	52.25
	ReasonBERT$_R$	41.3	34.79	38.04	5.76	17.04
	ReasonBERT$_B$	33.2	5.81	19.50	24.26	139.92
	SSPT	10.8	4.92	7.86	5.20	74.51
	SpanBERT	15.7	10.07	12.88	4.98	43.53
8	E-SNLI	90.52	87.72	89.12	2.48	3.13

Table 3.2: Partial reproducibility results for Papers 5 [106, Table 4] and 8 [109, Table 3]. Performance for Paper 5 [106] was measured as SQuAD dataset F$_1$ using a sample size of 16, and CV* scores are averaged across five runs. Performance for Paper 8 [109] was measured as accuracy of the trained self-rationalizing model (I→OR). Lower CV* is better.

Paper 6. The source code for Paper 6 [107] contains a full list of dependencies and instructions on how to run the code. As part of their evaluation, the source code used *pyrouge* to calculate the ROUGE metric [110]. Even though we were able to train a model according to the instructions, the final step of evaluation failed with a runtime error due to the missing installation of ROUGE. Following instructions provided by the *pyrouge* Python release package was not possible due to an unavailable (dead) URL that was supposed to explain how to install ROUGE.[9] Even after finding the instructions included in the main GitHub repository for *pyrouge*, we were not able to get it to a working state. We suspect this package is no longer being maintained as it had not been updated for more than three years at the time of writing. Furthermore, this issue was already reported by others in April 2021.[10] We contacted the authors via email offering our collaboration to migrate the source code to use SacreROUGE [111] but did not receive a response.

[9] https://pypi.org/project/pyrouge/

[10] https://github.com/bheinzerling/pyrouge/issues/38

Paper 7. The source code released with Paper 7 [108] is a collection of functions within a Python script. We were unable to determine which function(s) should be run for which experiment by inspecting the script. Without having access to the specific scripts used to run the experiments reported in the paper (or more documentation), we could not continue our work. We reached out to the authors by submitting an issue over GitHub,[11] and sent a follow-up email to the first author. We did not receive a response.

Paper 8. Paper 8 [109] raised our concerns about the hardware requirements for training and evaluation, as it uses T5 [112]—one of the largest available pre-trained Transformer models. Fortunately, the paper used T5 base, one of the smaller variants, and we were able to train and evaluate this model using a GPU with 24GB memory available. Unlike Paper 5 [106], this work did not contain a set of experiments using only a portion of the dataset. Given that training for 200 epochs (the authors' instructions) was beyond our allotted computing budget, we resorted to reducing the number of training epochs to one. This reduced the training time to less than an hour. Despite this reduction, our results were still close to those originally reported. We present these results in the bottom portion of Table 3.2.

3.4.3 Discussion

The results of our 8-paper reproduction case study show that code availability is not enough for reproducing the results present in published literature. Out of eight papers with released source code, we were only able to run two without issues. Furthermore, even though we made our best attempts to fix the issues with the others (including contacting the original authors), we were not successful in doing so.

To determine what new guidelines can be introduced to improve the state of reproducibility in the field, we first categorize the issues we found and check whether the existing guidelines

[11] https://github.com/autumntoney/ValNorm/issues/1

cover them. The primary problem of Paper 2 [103] and Paper 7 [108] was missing files, scripts, and instructions used to generate the reported results. Current reproducibility guidelines already address this problem. The ML Code Completeness checklist [28] highlights the importance of dependencies, code used for training and evaluation, and a README file accompanied by the instructions. We found that papers present training and evaluation scripts in many unique ways. This hinders the understandability of the code (*e.g.,* which script achieves which result, or what is the correct order of operations). We believe this problem could be addressed by recommending that authors include explicit scripts to generate each result reported in the paper.

Dependency on external resources was the main issue with Paper 3 [104], Paper 4 [105], and Paper 6 [107]. Paper 3 [104] used a pre-trained embedding file that was no longer available for download. Paper 4 [105] and Paper 6 [107] had missing and broken dependencies, respectively. Dependencies introduce variability over time, and may become broken as packages cease to be maintained. Simply listing dependencies, even with exact versions (which may become broken or inaccessible in the future), is not adequate to ensure long term reproducibility. Instead, *self-contained artifacts* such as Docker containers and Virtual Machine (VM) images can recreate an executing environment with high fidelity without relying on external resources (*e.g.,* files or URLs). The use of virtual environments is already mentioned in the ML Code Completeness Checklist [28]. The NAACL reproducibility track [26] also focuses on model verification using Docker containers, however, the containers are not publicly available.

Outside of NLP, the NeurIPS Code and Data Submission Guidelines [113] suggest the submitted codes should be self-contained and executable. Regardless, we believe reproducibility standards need to prioritize releasing self-contained environments. This shift would reduce the workload near submission deadlines while helping authors to document and record their work throughout development.

Except for Paper 5 [106] and Paper 8 [109], we encountered issues requiring the authors' assistance (*e.g.,* syntax and runtime errors). Aside from one case, our attempts to communicate with the authors were not successful. We understand that authors may not be available after their work is published, and that there are no guidelines regarding support of published research. Instead, this further strengthens our recommendation that future conferences require self-contained artifacts. We also recommend that they provide a venue to evaluate such artifacts at the time of publication, as performed in other fields of CS [12, 114, 115, 13].

On the positive side, Paper 5 [106] eased our reproducibility attempt through the inclusion of small-scale experiments. Often, the resources available for assessing reproducibility are limited compared to the original study. Therefore, unique hardware requirements and compute-intensive methods raise the barrier for reproducibility assessments. We recommend including limited experiments that are able to run on commodity hardware and with modest time requirements.

Given the empirical results provided in the previous section, we believe the following guidelines would help future reproducibility:

1. Include small scale experiments.
2. Include and document explicit scripts to generate each result in the paper.
3. Release executable self-contained artifacts.
4. Require (and evaluate) artifacts, not source code.

3.5 Case Study: Neural Text Simplification

While a high-level and fast reproducibility assessment provides good overview of the state of reproducibility in the field, a more detailed case study can provide a deeper understanding of the challenges faced by researchers in reproducing the results of a paper. In this section, we present a detailed case study of the work of Nisioi et al. [86] on the task of text simplification. We selected this work because it is one of the first to explore neural sequence to sequence models for automatic text simplification. We aim to reproduce the results of this work and evaluate the reproducibility of the released artifacts. We also provide a detailed account of the reproducibility process for this paper.

Nisioi et al. [86]'s work explores the task of *neural text simplification*. In this task, the goal is to transform a given text into a simpler version while retaining its meaning. What constitutes simplicity itself raises complicated questions since simplicity could be observed in the form of lexical simplification, content reduction, and grammatical or structural modification. Data-driven techniques attempt to achieve simplicity through automated metrics and human evaluation. The task holds many parallels with Machine Translation (MT), and this framing allows models studied in the context of neural MT (*e.g.,* , neural sequence to sequence models) to be adapted and deployed for neural text simplification.

Nisioi et al. [86] is one of the first investigations of neural sequence to sequence models for automatic text simplification. In particular, they use Long Short-Term Memory (LSTM) networks [116, 117] in an encoder-decoder architecture that has demonstrated success in similar sequence to sequence problems [118]. The encoder LSTM computes a representation for each source sentence, and the decoder LSTM generates an output given the encoded representation and previously generated tokens. The authors also employ a global attention mechanism that provides a more dynamic information flow and increases the representation bandwidth. To avoid overfitting, they use dropout [22], a technique that injects noise into the input during training by masking out certain features.

Nisioi et al. [86] experiment with two variants of networks: one with random embedding weight initialization (*NTS*), and another with pre-trained embeddings. The latter is built by concatenating pre-trained word2vec embeddings from the Google News corpus [119] with a locally trained skip-gram model [120] with hierarchical softmax and a window size of 10. The concatenation process involves utilizing a unique dictionary associated with the source and target embeddings. The authors refer to this variant as *Neural Text Simplification with Word2Vec (NTS-w2v)*.

3.5.1 Method

Data, source code, and training process are some of the conditions that affect the reproducibility of a study. Therefore, in order to assess the reproducibility of the work of Nisioi et al. [86], we first examine the data used for this work, and then shift our attention to the released software artifacts. We perform these steps to identify potential obstacles to reproducibility and to test the adequacy of the existing standards. Later, we assess the reproducibility of the reported automatic evaluations. Although we do not fill out any checklists [33], as they are not created for the purpose of third-party evaluation, we cover nearly all the concerns they attempt to address.

Data The data utilized for training and evaluation is a critical component of any machine learning study. Unsurprisingly, all reproducibility checklists emphasize the importance of data transparency. Nisioi et al. [86] used a corpus of parallel English Wikipedia and Simple English Wikipedia (EW-SEW) articles [121] when developing and evaluating their text simplification model. EW-SEW includes both manually and automatically aligned sentence pairs and was one of the largest publicly available datasets for text simplification at the time Nisioi et al. [86]'s paper was published. Sentences in EW-SEW were filtered based on Wiktionary-based word-level semantic similarity scores included in the dataset, with a retention threshold set at 0.45. This resulted in a final set of 280K+ aligned sentences. EW-SEW does not have standard validation and test splits; thus, although Nisioi et al. [86] used EW-SEW for training, they used TurkCorpus [122] for validation and testing. TurkCorpus is considerably smaller than EW-SEW and consists of 2000 validation and 359 test sentences.

Split	**Sentences**
train (EW-SEW)	284,677
validation (TurkCorpus)	2,000
test (TurkCorpus)	359

Table 3.3: Distribution of sentence pairs across different data splits, with data sources in parentheses.

The final distribution of training, validation, and test data is shown in Table 3.3. To preprocess the data, Nisioi et al. [86] used the Stanford NER system [123] to automatically tag the locations, persons, organizations, and miscellaneous entities in the dataset. We check the reproducibility of the preprocessing steps in Subsection 3.5.2 by reviewing the original dataset, as well as steps taken to filter and process the data.

Research Artifacts Authors may omit purportedly trivial details from research publications due to strict length limits. Such details may be crucial for later successful replication. Fortunately, released software artifacts often provide these details and other necessary engineering steps. The ML Completeness Checklist [28] underlines the inclusion of five items in software artifacts that facilitate reproducibility and are expected to result in easier adaptability for future researchers: (1) specification of dependencies, (2) training code, (3) evaluation code, (4) pre-trained models, and (5) a README file including a table of results accompanied by precise commands to run and produce those results. Given that Nisioi et al. [86] provided all of these items, we investigate the quality and the functionality of the released artifacts within this context by reviewing the aforementioned checklist items, testing out the provided commands, and rebuilding the environment using provided materials.

Automatic Evaluation Reproducibility and reporting quality are complementary to one another, and improvements to one often accompany improvements to the other. We include task-agnostic metrics and details commonly used in training and evaluations of neural networks [124, 25] in our assessment of the reproducibility of Nisioi et al. [86]'s automated evaluations. Namely, we check the number of parameters in the model, the computing infrastructure used to achieve results, and the total GPU hours required to train the model. We also report the model's total Floating-Point Operations (FPO) [125], providing an estimate of the amount of computational work performed irrespective of the hardware setting. In neural networks, the dominant floating-point operations are ADD and MUL operations performed by a GPU.

Nisioi et al. [86] evaluated the performance of their neural text simplification approach using two automated metrics as well as a human performance assessment. Their automated metrics included BLEU [126, 127], a precision-based metric commonly used for machine translation and text simplification; and SARI [122], a metric designed specifically for text simplification that compares the system output against reference output and the input sentence. The evaluation scripts for these metrics are included in the source code released by the authors. In addition to calculating the BLEU score using the script provided, we also calculate it using sacreBLEU v2.1,[12] a Python library that aims to unify standards for calculating the BLEU score [83].

Ultimately, these metrics are used on various output files generated by different variants. At first, we evaluate the original outputs provided by Nisioi et al. [86]. Then, we use the trained model released by the authors to generate a new output and evaluate it using the mentioned metrics. Lastly, we use the code and the configuration provided by the authors in their publication and in their source code to train new models. Using these newly trained models, we generate yet another set of outputs. During this process, we are reducing the set of controlled conditions affecting the final results. We expected variation to increase as fewer conditions are controlled.

We used these metrics to evaluate the performance of our reproduced model, facilitating a

[12] https://github.com/mjpost/sacrebleu

direct comparison with the originally reported performance. Instead of viewing the reproducibility of the automatic evaluations as a binary state, *e.g.,* , reproducible or not reproducible, we use CV* (defined in Section 1.3) to quantify reproducibility as a measurement precision.

We take one additional step to verify the claims based on the empirical results. We use paired bootstrap resampling [79] with 1000 samples to compare the performance of the two main variants on the output files released by Nisioi et al. [86].

3.5.2 Results

In this section, we describe the outcomes of our reproducibility assessment split into three main categories: data, software artifacts, and automatic evaluation.

Data We were unable to analyze the original unfiltered version of the EW-SEW dataset [121] as planned because the webpage containing the dataset no longer exists,[13] nor could earlier versions be retrieved using web archival tools (*e.g.,* , the Wayback Machine[14]). The released code repository for the selected paper also does not include scripts for filtering the dataset. As such, we could not review or reproduce the authors' preprocessing steps. However, the code repository does contain preprocessed dataset files, which allowed us to perform all other steps of our reproducibility analysis.

Research Artifacts As mentioned earlier, the authors released a five star repository according to the ML Completeness Checklist. The authors listed the required external libraries, as well as Python- and Lua-specific dependencies. Moreover, the authors included a dockerfile containing the computing environment used for the experiments. Unfortunately, since a self-contained docker container was not included, it is not possible to rebuild the dockerfile, and most dependencies

[13] https://crow.ece.uw.edu/tial/projects/simplification/

[14] https://archive.org/web/

have been deprecated for years. These dependencies include Ubuntu 14.04 with an End of Life (EOL) of 2019, Python 2.7 with EOL of 2020, Torch7 with last active development in 2017, and OpenNMT made obsolete in 2018 due to lack of support for Torch7, among others. Ultimately, we switched to another docker image based on Nvidia's CUDA 10.1 images that comes with Torch7 installed. This introduced further complications as recently released GPUs (*e.g.,* , those in the RTX 3000 series) require CUDA 11 or higher. We avoided this problem for now, but fixing this problem (which is beyond the scope of our present work) requires porting Torch7 and rebuilding it using the appropriate CUDA toolkit, which could be extremely challenging.[15]

Aside from the initial hurdle to get the repository to a running state, we did not face any major issues in using the software artifacts. [86] provided the training code, evaluation code, and pre-trained models.[16] The README file contains instructions and required commands to produce the reported results. There were a few minor discrepancies between the provided instructions and real-world use, but we managed to resolve these issues. We note that the repository does not contain all configuration files used for each model variant. Hence, we use the information provided in the paper to recreate those.

In reviewing the source code, we found three issues affecting *NTS-w2v* variants. We contacted the authors regarding these issues, and they graciously confirmed the first two. At the time of writing this dissertation, we have not heard back regarding the third reported issue. We investigate the impact of the first two issues on the results. These issues are described below.

- **Issue 1: Data Contamination**: The *NTS-w2v* models use a multi-step process to concatenate the pre-trained Google News word2vec embeddings and another embedding trained by the authors using the skip-gram technique. We found that during the skip-gram training process, this embedding utilized all datasets (including the development and test set), introducing data contamination that may call into question those models' results.

[15] Issue is reported here: `https://github.com/nagadomi/distro/issues/11`.

[16] `https://github.com/senisioi/NeuralTextSimplification`

The models affected by this issue are expected to have an advantage over other models. However, the validation and test sets are many times smaller than the training set, so performance gains may be negligible to non-existent.

- **Issue 2: Mismatched Embedding**: This issue occurs during the concatenation process itself. This process uses two dictionaries, one for the encoder and one for the decoder, to generate the embedding matrix. However, we found that these embeddings were mismatched: the encoder used the decoder's dictionary, and the decoder used the encoder's dictionary. We expect fixing this issue will improve the performance of affected models.

- **Issue 3: Zero Embedding Weight**: Lastly, we found that the final embedding matrix is missing the concatenation step, which results in zero vectors for all the words. Using a zero embedding weight nullifies the embedding pre-training altogether.

Automatic Evaluation We follow the exact training setup provided by Nisioi et al. [86], training models for 15 epochs with early stopping applied. Unlike the original paper, we did not tune the model using SARI or BLEU, and used the validation perplexity (lower is better) for model selection and early stopping. The text simplification is performed using beam search. Beam search generates the first k hypotheses at each step sorted by log-likelihood of the target sentence given the input sentence. While the authors experimented with using beam sizes of 5 and 12 and various hypotheses, we limit the scope of our experiments to 5 beams and 1 hypothesis. The hardware used for the experiments in the original paper is not explicitly specified. In our case, we use an *RTX 2080 ti* GPU to train the models. Training took approximately 3 hours. The model had 84 million parameters, of which 50 million belong to the embedding layer. With a maximum sequence length of 80 and a batch size of 1, this model used roughly 3G fpo (3×10^9) in a forward pass.

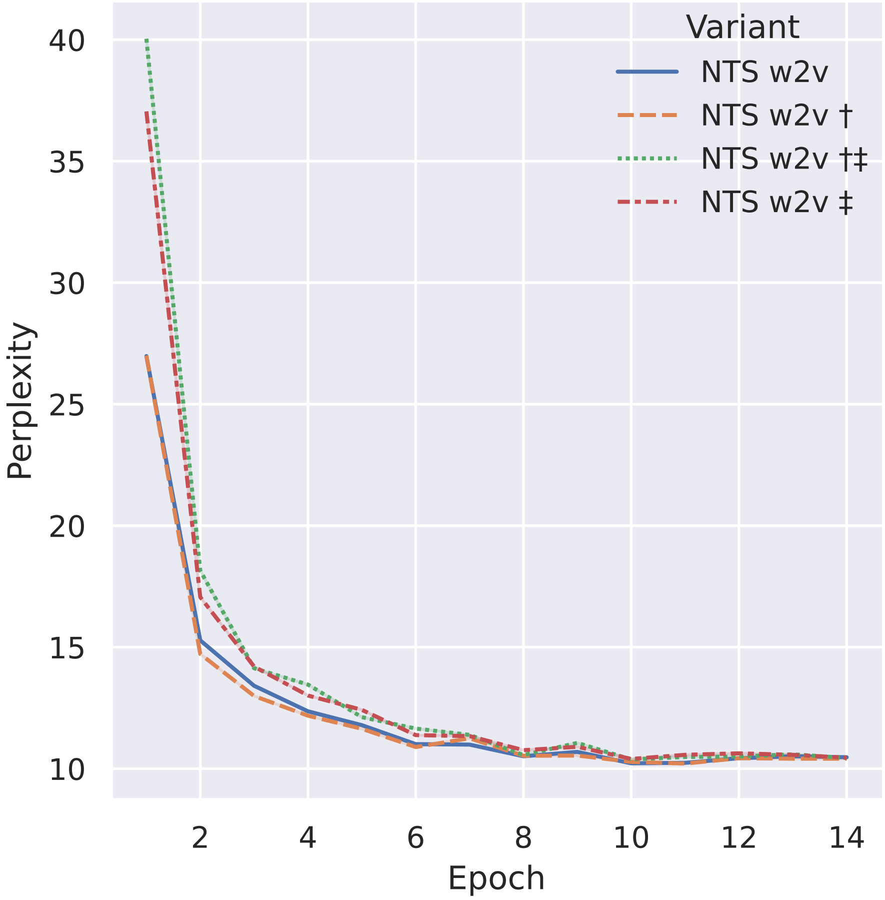


Figure 3.3: Validation perplexity of *NTS-w2v* variants during training (lower is better). † indicates contaminated conditions, and ‡ indicates mismatched conditions.

Table 3.4 and 3.5 contains the results of the original work [86] referred to as $t1$, reproducibility studies of Cooper et al. [128] ($t2$) and Belz et al. [10] ($t3$), and the results calculated by this work ($t4$) with their associated conditions. To ease the analysis, the results of every variant, measure, and output are grouped together. With the two evaluation scripts for calculating BLEU, we have added six values for BLEU and three for SARI. To be more specific, we have added automatic evaluation results for the output generated by Nisioi et al. [86] ($o1$), our own output generated by running the trained model provided by Nisioi et al. [86] ($o4$), and our own output generated by running our own version of the model ($o5$). We note that the model that we trained uses a source with all the fixes applied; however, to the best of our knowledge, all the other *NTS-w2v* variants are trained with the mentioned issues. We present the precision results of the QRA framework in Table 3.6.

The *NTS* variant has CV* values of 1.92 and 1.94 for SARI and BLEU, respectively. With 3.28 and 2.85 for SARI and BLEU, CV* values for *NTS-w2v* are slightly worse. However, the BLEU score of the *NTS-w2v* variant reported by Cooper et al. [128] seems to be an outlier. By excluding their score (80.75), CV* reduces to 1.22. There are two other interesting observations gleaned from Table 3.4 and 3.5. First, our reported results for $o1$ exactly match the reported results by the original paper; this suggests that we successfully recreated the environment they used for their evaluation. Second, the difference between the reported BLEU for $o1$ using the sacreBLEU evaluation script [10] and that found by our study implies there are still several unaccounted conditions. We believe the version of sacreBLEU and the process of running this evaluation script are possible causes for this variation.

Table 3.7 shows results from the paired bootstrap resampling statistical significance test, with an objective of determining whether the performance of *NTS-w2v* in terms of BLEU score is better than the *NTS* variant. With $p = 0.0079$, the difference is indeed statistically significant. Since the output of *NTS-w2v* is generated using a model affected by the zero weight embedding issue (Issue 3 described in Subsection 3.5.2), these two variants are essentially the same.

Finally, we investigate the issues reported for the *NTS-w2v* variant. We exclude Issue 3.5.2, as it simply converts *NTS-w2v* to *NTS* with zero embedding weight. We introduce three new variants:

- *NTS-w2v*†: *NTS-w2v* only affected by data contamination.
- *NTS-w2v*‡: *NTS-w2v* only affected by mismatched embeddings.

Object	Measurand	Output	Trained by	Comp. by	Eval. Script by	Performed by	Measured Value
NTS	BLEU	o1	t1	t1	t1	t1	84.51
		o1	t1	t1	t1	t2	84.50
		o1	t1	t1	≈t1	t3	85.60
		o1	t1	t1	sb	t3	84.20
		o1	t1	t1	t1	t4	84.51
		o1	t1	t1	sb2.1	t4	84.60
		o2	t2	t2	t1	t2	87.46
		o3	t1	t3	≈t1	t3	86.61
		o3	t1	t3	sb	t3	86.20
		o4	t1	t4	t1	t4	86.53
		o4	t1	t4	sb2.1	t4	86.60
		o5	t4	t4	t1	t4	88.81
		o5	t4	t4	sb2.1	t4	88.80
	SARI	o1	t1	t1	t1	t1	30.65
		o1	t1	t1	t1	t2	30.65
		o1	t1	t1	t1	t3	30.65
		o1	t1	t1	t1	t4	30.65
		o2	t2	t2	t1	t2	29.13
		o3	t1	t3	t1	t3	29.96
		o4	t1	t4	t1	t4	29.96
		o5	t4	t4	t1	t4	30.23

Table 3.4: Detailed results, all utilizing the source code released by Nisioi et al. [86]. Outputs $o1$ to $o5$ are generated based on their training conditions: $t1$=Nisioi et al. [86], $t2$=Cooper et al. [128], $t3$=Belz et al. [10], and $t4$= this work. sacreBLEU versions are sb=unknown version, and $sb2.1$=version 2.1.

- *NTS-w2v*†‡: *NTS-w2v* affected by data contamination and mismatched embeddings.

The results are shown in Table 3.8. Overall, the results are extremely close. We found that the variant with the data contaminated outperformed others while *NTS-w2v*, the variant without any issues performed worse than the rest. We expected to observe a noticeable performance difference for the models affected by the mismatched embedding issue, but the performance gap was ultimately marginal and inconsistent. We report the validation performance during

Object	Measurand	Output	Trained by	Comp. by	Eval. Script by	Performed by	Measured Value
NTS-w2v	BLEU	o1	t1	t1	t1	t1	87.50
		o1	t1	t1	≈t1	t3	89.36
		o1	t1	t1	sb	t3	88.10
		o1	t1	t1	t1	t4	87.50
		o1	t1	t1	sb2.1	t4	87.90
		o2	t2	t2	t1	t2	80.75
		o3	t1	t3	≈t1	t3	89.64
		o3	t1	t3	sb	t3	88.80
		o4	t1	t4	t1	t4	89.40
		o4	t1	t4	sb2.1	t4	89.40
		o5	t4	t4	t1	t4	87.04
		o5	t4	t4	sb2.1	t4	87.10
	SARI	o1	t1	t1	t1	t1	31.11
		o1	t1	t1	t1	t3	31.11
		o1	t1	t1	t1	t4	31.11
		o2	t2	t2	t1	t2	30.28
		o3	t1	t3	t1	t3	29.12
		o4	t1	t4	t1	t4	29.12
		o5	t4	t4	t1	t4	29.70

Table 3.5: Detailed results, all utilizing the source code released by Nisioi et al. [86]. Outputs $o1$ to $o5$ are generated based on their training conditions: $t1$=Nisioi et al. [86], $t2$=Cooper et al. [128], $t3$=Belz et al. [10], and $t4$= this work. sacreBLEU versions are sb=unknown version, and $sb2.1$=version 2.1.

Object	Measurand	Sample Size	Mean	Unbiased STDEV	STDEV 95% CI	CV*
NTS	SARI	8	30.23	0.56	[0.23, 0.89]	1.92
NTS	BLEU	13	86.07	1.64	[0.94, 2.34]	1.94
NTS-w2v	SARI	7	30.22	0.96	[0.34, 1.58]	3.28
NTS-w2v	BLEU	12	87.71	2.45	[1.35, 3.54]	2.85

Table 3.6: CV* and component measures (mean, standard deviation, standard deviation confidence intervals) for measured quantity values obtained in multiple measurements of the two *NTS* systems.

System	BLEU (μ ± 95% CI)
Baseline: NTS-w2v	87.9 (87.9 ± 2.0)
NTS	84.6 (84.6 ± 2.9)

Table 3.7: Statistical significance analysis performed on Nisioi et al. [86]'s released output. With $p = 0.0079$, the difference in reported results between the two variants is statistically significant.

Object	Measurand	Eval. Script by	Measured Value
NTS-w2v	BLEU	t1	87.04
NTS-w2v	BLEU	sb2.1	87.10
NTS-w2v	SARI	t1	29.70
NTS-w2v †	BLEU	t1	89.43
NTS-w2v †	BLEU	sb2.1	89.40
NTS-w2v †	SARI	t1	29.80
NTS-w2v †‡	BLEU	t1	89.12
NTS-w2v †‡	BLEU	sb2.1	89.10
NTS-w2v †‡	SARI	t1	29.58
NTS-w2v ‡	BLEU	t1	88.01
NTS-w2v ‡	BLEU	sb2.1	88.00
NTS-w2v ‡	SARI	t1	29.18

Table 3.8: Results of the experiments tracking performance impacts for identified issues, computed for this paper using our version of the model, our output, and the evaluation script provided by Nisioi et al. [86] and sacreBLEU. † indicates contaminated conditions, and ‡ indicates mismatched conditions.

training to analyze whether there are any differences between these four variants. As shown in Figure 3.3, the models with mismatched embeddings had a worst start, by a perplexity gap of almost 15; however, as training progressed, they closed the gap and ended with perplexity differences of less than 1.

Besides the mentioned analysis, we found it hard to provide distinct and unique observations from the results. This is likely due to the fact that the results are not conclusive and the variance is high. We do not believe this is a flaw in our experimental design but rather a good representation of the complexities of comparing different models across varying conditions. The number of experiments conducted in this study is more than 60, a number that exceeds the number of experiments conducted in most other studies by a large margin.

One of the concerning issues we encountered is the issue of deprecation. While this is not a new problem, and it is as old as the software itself, it is becoming more and more prevalent. This is due to extreme reliance on empirical results and the complexity of publications that utilize neural networks. Often source codes use several external libraries and dependencies, any of which may become deprecated at any time. Increased availability of source code and the abundance of tools are signs of a healthy research community. Seeing new tools and libraries developed and improved daily is encouraging. At the same time, we believe researchers should practice caution when introducing new tools and libraries into their experiments, as doing so may shorten the usability of their source code.

3.5.3 Discussion

Taking all our experiments into account, we cannot claim that the performance difference between different variants comes from the design decisions made during their development. Perhaps our most surprising finding is that the *NTS-w2v* variants affected by mismatched embeddings performed on par with the other variants once training was complete. This extreme level of resilience is, in fact, quite alarming. Silent bugs like these can easily go unnoticed, and the results of the experiments can be misleading. Nearly all publications utilizing neural networks report top-performing empirical results; yet, aside from manual code review and deep analysis of the final results, there are no other clear signs or warnings that may suggest a bug is impacting the model.

Due to the age of this repository, getting the project to a running state consumed the most time. We suspect that the situation will deteriorate as most dependencies are no longer being actively maintained. Researchers should be hesitant with introducing new dependencies into their projects. Additionally, we believe it would be fruitful to redirect the time and effort used for identifying and reporting dependencies toward exporting self-contained environments. This is an inadequacy that we found in nearly all of the checklists; in the case of this project, even though we knew all the requirements, we spent hours debugging different errors.

Many factors can affect the results of an experiment. Some of these factors are under the experimenter's control, and some are not. Scientific experiments are developed as a counterpart to abstraction of real-world problems. Hence, while we use such experiments as "benchmarks," focusing on factors and variables that solely affect experimental results undermines the real-world goal. Datasets are created with this in mind, consisting of training, validation, and test sets of which the latter, in particular, is created to represent unseen real-world data. Research on improving the generalization of machine learning algorithms is another good example of leveraging scientific experiments to understand real-world challenges. In other words, over-emphasis on miniscule performance gains that could be attributed to variance in evaluation only serves to undermine the real-world applicability of the research, and the field as a whole.

3.6 Conclusions

While we observed an upward trend in releasing source code within the NLP community and NeurIPS, submitting the code alone does not seem to be adequate. Even though the community is heading towards the right direction, our results suggest that often times, the released source code does not meet a minimum requirement for reproducibility, defined as achieving the results using the provided source code. This requirement is by no means comprehensive, and it will evolve as the state of reproducibility improves. However, it could be the first step to assess the reproducibility of a scientific publication. Furthermore, it aids in debugging and understanding why a work is not reproducible. After all, determining *whether* a work is reproducible or not is not as useful as understanding *why* that is the case. When it comes to reproducibility standards, we believe it is time for them to evolve to address the concerns we raise in regard to the quality of released source codes. We need to shift the focus from source code to software artifacts. These artifacts should include a self-contained runtime environment containing scripts for achieving every single result reported in the paper. Another missing step in current reproducibility assessments is third-party evaluation at the time of submission. Incorporating this into the submission pipeline would reduce the creation of additional workload for the reviewers of the publication. Students and practitioners could be encouraged to partake in the reproducibility evaluation process.

We explored various challenges faced when attempting to reproduce automatic evaluations of scientific publications. These challenges vary in complexity from basic issues like syntax errors and incomplete instructions, to more intricate problems including unavailable datasets, data contamination, and logical errors in the source code. Despite the progress towards increased availability of source code, community organizers and conferences should take steps to elevate the minimum standards for releasing reproducible scientific artifacts in their respective fields. Any change should be gradual, otherwise it may lead to frustration and resistance from the community. It is understandable that researchers may not be able to share their data or research artifacts due to protections for human subjects, General Data Protection Regulation (GDPR), or other confidentiality and privacy issues. These restrictions are out of the control of the researchers and should be considered when evaluating the reproducibility of the results. However, researchers should be encouraged to share as much as possible.

Our recommendations are different from the current focus on reproducibility challenges more common in the ML and NLP communities, in which researchers are encouraged to not use research artifacts and instead, reproduce results from scratch. The problem with this approach is that if the results do not match (which happens often), it is unclear whether this is because of missing features, bugs in the new implementation, or the irreproducibility of the original results. In other words, this is jumping two steps ahead. Finding conditions that are required to be controlled should be considered an important area of research. This process is also more aligned with principles of software engineering; a concept that is often ignored in the field of machine learning despite the increased reliance on empirical results.

Reproducibility is a desired attribute for deep learning models, but it comes with a cost. There may exist cases in which the required conditions for reproducing the results are not practical. Defining what is practical, of course, depends on the problem at hand. For example, reproducing a named entity classifier that requires copying networks' weights and using the same hardware may not be considered practical. This is another instance of bias-variance tradeoff. Bias-variance, a property of statistical and machine learning models, suggests that the variance of the parameter estimated across samples can be reduced by increasing the bias. A dilemma exists when trying to minimize these two sources of error simultaneously. We have a dilemma when it comes to assessing the reproducibility of results. Many attempts have focused on controlling all the variables. Yet, while they have their use cases, their complexity makes them less viable. Perhaps a better alternative is to reduce the emphasis on the top-performing results and utilize techniques that attempt to aggregate and report the results of a set of experiments.

We conclude by underlining that conducting reproducibility evaluations may result in many failed attempts. While some of these failures may actually be due to bugs or other issues, many of them may be due to the lack of self-containment or the lack of proper documentation. These should be seen as another data point in the evaluation process. Ultimately, such reproducibility evaluations are always impacted by the subjective nature of the evaluation process. However, the additional scrutiny recommended as a result of this case study would help the field in the long run. We believe that the recommendations we have provided in this work will help the field to improve the reproducibility of the research results.

With chapters focused on uncertainty in evaluation and reproducibility of automatic evaluation finished, we now turn our attention to reproducibility of human evaluation.

Chapter 4

Human Evaluation Reproducibility[1]

4.1 Introduction

Human evaluations play a critical role in assessing the effectiveness of text generation systems in NLP. Despite progress in automatic metrics, their limitations, such as poor correlation with human judgments, persist [129, 130, 131, 132, 133, 83, 134]. Addressing challenges that affect the reproducibility of human evaluation experiments is thus essential [135].

Using human raters to evaluate algorithms introduces unique challenges. Unlike automated evaluations, human assessments are not as cost-effective and require the recruitment of skilled raters, often limiting the number of evaluated samples due to financial constraints. This frequently necessitates the use of crowd-sourcing platforms such as Amazon Mechanical Turk (AMT)[2] [136, 137, 138].

Efforts to improve the reliability of human evaluations involve measuring inter-rater agreement, estimating the statistical power for sufficient sample sizes, and applying statistical tests to assess outcome significance [139, 140, 141, 142, 134]. However, these strategies focus primarily on

[1] Parts of this chapter were previously published in Arvan et al. [4, 5]

[2] https://www.mturk.com

metrics and overlook the evaluation processes themselves, leading to concerns about systematic approaches to human evaluations [143]. To address these concerns, it is crucial to document and critically assess procedures within human evaluations. This ensures enhanced transparency and rigorous verification, adhering to the principles of Open Science, which are vital for improving reproducibility and validating high-quality research outcomes.

While numerous studies focus on the reproducibility of automated metrics [31, 29, 2, 1], human evaluations have received less attention due to their complexities [144]. The ReproHum Project addresses this gap by developing a methodological framework aimed specifically at enhancing the reproducibility of human assessments in NLP. Drawing from other meta-analytical efforts [145, 146, 147], this initiative seeks to improve evaluation rigor, transparency, and reliability. The insights gained will enhance not only reproducibility checks but also the refinement of human evaluation processes, thus boosting their reliability and academic credibility. Ultimately, the ReproHum project plans to conduct a large-scale, multi-lab reproducibility study on NLP studies involving human evaluations.

4.2 Contents of this Chapter

This chapter details the results and analyses of two human evaluation reproductions that we conducted, in addition to discussing two other paired reproductions executed by colleagues. The section begins with an overview of the ReproHum project, outlining the methodologies employed in these human evaluation reproductions. Further, it deepens into the specific experiments selected for reproduction, including discussions of the original studies, the setups for human evaluations, and additional insights obtained from the original authors.

Subsequent sections describe any deviations from the original experiments and the adaptations made to ensure the fidelity of the reproductions. The chapter concludes by presenting the results of these reproductions, featuring a quantified reproducibility assessment, and discussing the broader implications of these findings. The contents of this chapter are edited from our publications [4, 5] and two paired reproductions conducted by our colleagues [148, 149].

4.3 Background

In the initial phase of the ReproHum Project [144], a pool of 177 papers published in the ACL or the Transactions of the Association for Computational Linguistics (TACL) between 2018 and 2022 was screened. Criteria for this selection included papers that (a) incorporated human evaluation components and (b) were accessible publicly. From this collection, a multi-tiered review process singled out 20 experiments from 15 different papers for reproduction, prioritizing factors such as the responsiveness of the original authors and the availability of essential experimental details. Each selected experiment was then categorized based on the number of evaluators (as few or many), cognitive complexity[3] (low, medium, high), and the training and expertise of evaluators (none, either, or both).

The ReproHum Project subsequently organized the reproduction work into distinct **rounds**. The aim of the first round was to replicate the selected experiments faithfully under consistent conditions across two independent laboratories. Detailed guidelines were provided to these labs to ensure an exact reproduction, utilizing the methodologies outlined in the original studies and supplementing these with additional clarifications from the original authors as necessary. Any deviations from the original experimental framework were meticulously documented, along with the rationales for such changes. Following this, the reproducibility outcomes were collated and assessed against the results published originally, evaluating the fidelity of the reproduction. In this chapter, we present the results of four human evaluation reproductions conducted as part of round one of the ReproHum Project.

Upon completing the first round, a subset of experiments demonstrating a high degree

[3] Scores for criteria are detailed in Appendix E of Howcroft et al. [135].

of reproducibility will advance to the second round. Here, the goal shifted to collaborative improvement: two laboratories would modify specific aspects of a given experiment to explore potential enhancements. Both labs would then conduct the revised experiment, evaluate its reproducibility, and benchmark the new results against the original data. This iterative process aimed not only to affirm the replicability of results but also to refine methodologies and ultimately improve the robustness of research findings. At the time of writing this dissertation, the ReproHum Project is in the final stages of the first round, with the second round set to commence shortly.

4.4 Method

4.4.1 Common Approach to Reproductions

As a participant in the ReproHum project, our laboratory was equipped with essential materials and guidelines to facilitate the reproduction of a specified experiment. These materials included:

- A document outlining a standardized approach to experimental reproduction
- The original paper along with the associated data necessary for replication
- Additional supplementary documents

Direct communication with the original authors was intentionally avoided; instead, all interactions were mediated through the ReproHum organizers. This protocol was established to maintain consistency across different reproductions and to eliminate any potential authorial bias that could influence the outcomes.

The guidelines provided delineated the procedures into two main phases—pre-reproduction and during/post-reproduction. Initially, our tasks included a thorough review of the experimental paper and preparations for executing the reproduction. This involved calculating appropriate compensation for crowd workers and adhering to institutional guidelines. For our team at the University of Illinois Chicago, securing Institutional Review Board (IRB) approval was a prerequisite.[4] All subsequent activities conformed strictly to the protocols approved by our IRB.

The second phase centered on the actual execution of the experiment and the analysis of the data obtained. We meticulously filled out the Human Evaluation Data Sheet (HEDS) for each experimental task. This sheet catalogued detailed entries about the tasks, participants, and their responses. This facilitated a systematic analysis where we identified various error types and juxtaposed the newly generated data against the original findings. This side-by-side comparison was crucial in assessing the fidelity of our reproduction.

4.4.2 Quantified Reproducibility Assessment

We followed the standardized procedure for reproducibility assessment as outlined by the ReproHum team. For single numerical result scores, we calculated the CV* (defined in Section 1.3) to quantify the precision of the results. For sets of numerical scores, we calculated Pearson and Spearman correlations between the reproduced and original results. The Pearson correlation measures the linear relationship between two sets of scores, and the Spearman correlation measures the monotonic relationship between two sets of scores. Both correlations range from -1 to 1, where 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation (suggesting that the outcomes are diametrically opposed), and 0 indicates no correlation. Using these metrics, we assessed how closely the reproduced results aligned with the original results.

4.4.3 Experiments Selected for Reproduction

As part of our involvement in the ReproHum project, our laboratory was tasked with replicating two specific human evaluation experiments. These are *Data-to-text Generation with Macro*

[4] IRB Protocol Numbers: STUDY2023-0240 and STUDY2023-1217.

Planning by Puduppully et al. [150] and *Hierarchical Sketch Induction for Paraphrase Generation* by Hosking et al. [151]. With invaluable support from both the original authors and the ReproHum leadership, we successfully obtained the necessary data and tools to conduct these replication studies. In the subsequent sections, we detail both the original studies and our methodologies in replicating these experiments.

4.4.4 Data-to-text Generation with Macro Planning

Paper Summary. In the paper *Data-to-text Generation with Macro Planning*, Puduppully et al. [150] enhance a neural model by introducing a macro planning phase for data-to-text generation tasks. These tasks involve generating descriptive natural language from structured inputs, such as tabular data or knowledge graphs. Traditional techniques in this domain have been overtaken by the effectiveness of modern neural models, though these newer approaches still suffer from significant drawbacks such as inaccuracies, hallucination effects, and suboptimal context selection and document structuring.

To mitigate these problems, Puduppully et al. [150] suggest the integration of macro planning, which deals with high-level organization and presentation of information. They point out limitations within current datasets used for data-to-text generation, noting how most expect outputs to be organized into various paragraphs—this structure is ideal for applying planned paragraphing strategies. The methodological framework proposed by the authors involves a two-step process: generating a macro plan from training data, followed by using this plan to drive the generation process within a text model.

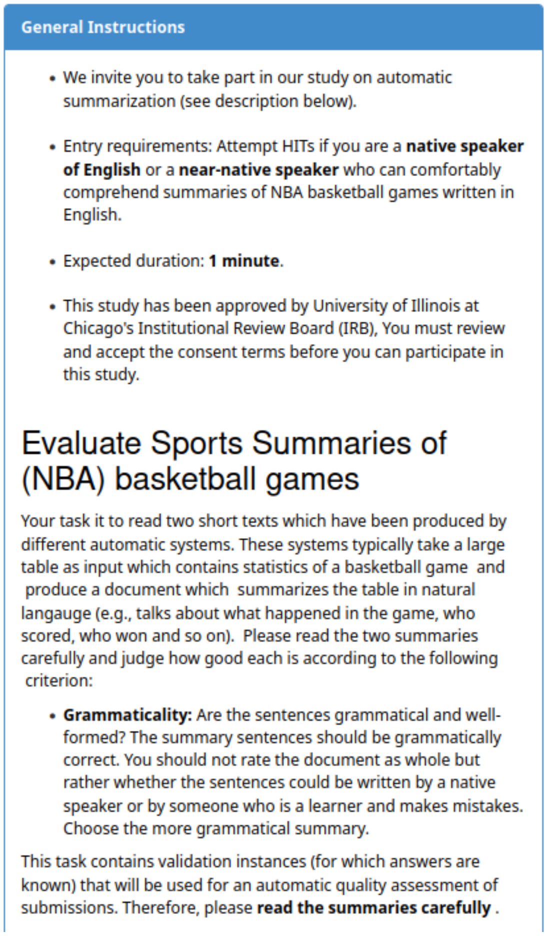
The efficacy of their method is tested using the RotoWire [152] and MLB [153] datasets, both of which encompass detailed game statistics and summaries from basketball and baseball, respectively. Evaluations through both automated metrics and human assessments demonstrated that the text generated by their approach outperformed current leading models in terms of accuracy, coherency, and fluency. The human evaluations compared the gold-standard outputs and four different text generation systems, including the authors' model (Macro), a template-based generator (Templ), the formerly best-performing system ED+CC [152], and the prior state-of-the-art model RBF-2020 [154].

Human Evaluation. For their human evaluation, Puduppully et al. [150] utilized AMT. To ensure the quality of the responses, they set specific participation criteria for the workers. These criteria included a minimum approval rate of 98% from at least 1000 tasks completed previously. Additionally, participation was restricted to crowd workers residing in primarily English-speaking countries, namely the US, UK, Canada, Ireland, Australia, and New Zealand.

The evaluation consisted of two distinct tasks. The first task assessed the factual accuracy of the summaries by examining the number of supporting and contradicting facts related to the games. The second task, which was our primary focus for reproduction, evaluated the quality of the generated texts. This evaluation specifically looked at coherence, grammatical correctness, and conciseness of the text. For this task Puduppully et al. [150] employed a comparative approach, where workers were prompted to assess two randomly selected summaries. Detailed instructions were provided to guide the evaluators, as demonstrated in Figures 4.1 and 4.2, depicting the precise instructions and the input interfaces respectively. Our replication strictly followed the original setup.

To process the results, the authors implemented a Best-Worst Scaling (BWS) method [155, 156]. The performance of each system was quantified by computing scores based on the instances a system was chosen as the best minus instances it was chosen as the worst, normalized by the total appearances of the system. Each of the five competing models (four systems and one gold standard) were paired into ten combinations, and each pair was then evaluated on three key aspects: grammar, coherence, and conciseness.

For a thorough analysis, each pair of summaries was presented to three different workers,



General Instructions

- We invite you to take part in our study on automatic summarization (see description below).
- Entry requirements: Attempt HITs if you are a **native speaker of English** or a **near-native speaker** who can comfortably comprehend summaries of NBA basketball games written in English.
- Expected duration: **1 minute**.
- This study has been approved by University of Illinois at Chicago's Institutional Review Board (IRB), You must review and accept the consent terms before you can participate in this study.

Evaluate Sports Summaries of (NBA) basketball games

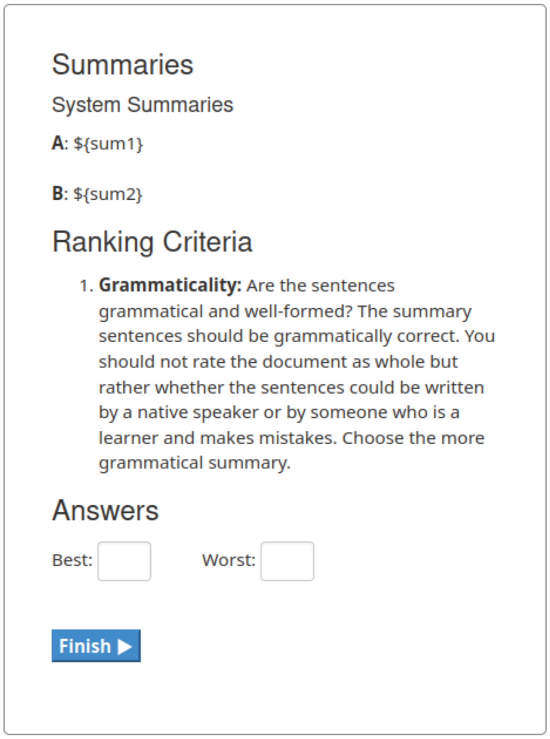
Your task it to read two short texts which have been produced by different automatic systems. These systems typically take a large table as input which contains statistics of a basketball game and produce a document which summarizes the table in natural langauge (e.g., talks about what happened in the game, who scored, who won and so on). Please read the two summaries carefully and judge how good each is according to the following criterion:

- **Grammaticality:** Are the sentences grammatical and well-formed? The summary sentences should be grammatically correct. You should not rate the document as whole but rather whether the sentences could be written by a native speaker or by someone who is a learner and makes mistakes. Choose the more grammatical summary.

This task contains validation instances (for which answers are known) that will be used for an automatic quality assessment of submissions. Therefore, please **read the summaries carefully** .

Figure 4.1: Instructions given to AMT workers for this task.

amassing a total of three unique preference assessments per pair. The evaluation spanned 40 summaries (20 for each dataset used) across ten pairs of systems. Factoring in the three criteria and three evaluations per criterion, the task generated a substantial volume of 3,600 individual preference ratings. In total, 206 crowd workers participated in this nuanced evaluation task, contributing to a robust dataset for analysis.



Summaries

System Summaries

A: ${sum1}

B: ${sum2}

Ranking Criteria

1. **Grammaticality:** Are the sentences grammatical and well-formed? The summary sentences should be grammatically correct. You should not rate the document as whole but rather whether the sentences could be written by a native speaker or by someone who is a learner and makes mistakes. Choose the more grammatical summary.

Answers

Best: Worst:

Figure 4.2: Specific input regions that AMT workers used to rank criteria associated with system summaries.

Additional Details from the Authors. During our effort to replicate the study outlined by Puduppully et al. [150], we received crucial additional insights through a document prepared during interactions between the authors and the ReproHum project team. This document detailed various aspects of the original human evaluation process, including task setup, instructions for crowd workers, and quality control measures implemented.

The original authors provided us access to the forms used in AMT for gathering responses, which were crucial for our accurate replication of the study. They explained that while each task was to be completed by three different crowd workers, the same workers had the option to take on multiple tasks. They also specified exclusion criteria for crowd workers to ensure the response quality was maintained, which we adhered to during our replication.

In our reproduction study, we structured each evaluation criterion into four mini-batches, each consisting of a quarter of the total tasks, to maintain manageability and ensure thorough analysis by the crowd workers. This approach mirrored the organization and the order utilized by the original authors.

Known Deviations from the Original Experiment

We are aware of several deviations in our reproduction from the original experiment, and we detail these below. We do not believe that these deviations had a major impact on our reproduction results.

Scope of Reproduction: The scope of our reproduction was limited to the second human evaluation task reported in their paper, examining the quality of generated text based on coherence, grammaticality, and conciseness. Furthermore, we only reproduced the results on the RotoWire dataset. This was due to our inability to verify the results on the MLB dataset. Unfortunately, this means the sample size of the reproduction is 20, half of the original study.

Attention Checks: A critical component of maintaining data integrity in the original study involved attention checks, with specific exclusions applied to workers who failed these checks during tasks assessing conciseness and coherence. The original authors had exclusion criteria, particularly excluding workers based on their responses involving the template-based system; unfortunately, these detailed criteria were no longer accessible. Consequently, we slightly adjusted our approach based on the guidelines from the ReproHum team.

For conciseness, where the original process annotated specific comparisons, our adaptation involved using an n-gram-based similarity score computed via NLTK.[5] We identified and excluded workers based on their ratings of twelve pairs that exhibited the highest differences between

[5] https://www.nltk.org

the gold standard and system outputs. Workers who rated highly divergent system outputs as superior were excluded from further participation in the study.

In the coherence assessments, the exclusion was more straightforward. We excluded workers who rated the template system's output as superior to the gold standard. Moreover, leveraging the results of the other team assigned to this paper—workers excluded from their study were also excluded from our evaluations, ensuring consistency across parallel studies.

Payment: All participants were compensated for each task completed, irrespective of subsequent exclusion, with a payment rate of $0.22 per task. This rate was an increase from the $0.15 paid in the initial study, adjusted to account for inflation and the local minimum wage requirements.

4.4.5 Hierarchical Sketch Induction for Paraphrase Generation

Paper Summary. Hosking et al. [151] introduced the Hierarchical Refinement Quantized Variational Autoencoder (HRQ-VAE), a generative model that employs a syntactic sketch for paraphrase generation. This approach mirrors human strategies for planning utterances by incorporating a sketching step into the model to enhance paraphrase generation. The performance of HRQ-VAE was evaluated against several baseline models on the Paralex [157], Quora Question Pairs (QQP),[6] and MSCOCO datasets [158].

For baseline comparisons, the authors assessed the Gaussian Variational AutoEncoder (VAE) [159], Latent Bag-of-Words (BoW) [160], and Separator [161], along with other paraphrase generation systems. They employed iBLEU [162], BLEU, Self-BLEU, and P-BLEU metrics to automatically evaluate these models on the indicated datasets. iBLEU, the primary assessment metric, evaluates the quality of paraphrases by checking the faithfulness to the original reference

[6] https://kaggle.com/competitions/quora-question-pairs

paraphrases and the diversity incorporated within the outputs. The evaluation results revealed top performance from the VAE, Latent BoW, Separator, and HRQ-VAE models.

Human Evaluation. The top four models were subsequently assessed through human evaluation conducted on AMT. This evaluation consisted of 180 Human Intelligence Tasks (HITs), each comprising 32 paraphrase pairs. Note that the term *task* is synonymous with HIT in this context. Each task included two attention checks to validate response quality. AMT workers were tasked with selecting the superior paraphrase from an input text and outputs from two competing models, judging based on fluency, meaning, and dissimilarity. Figure 4.3 depicts the user interface from the original study used during this human evaluation. The exact wording of the instructions provided in the user interface is included below:

- Which system output is the most **fluent and grammatical**?
- To what extent is the **meaning** expressed in the original sentence **preserved** in the rewritten version, with no additional information added?
- Does the rewritten version use **different words or phrasing** to the original? You should choose the system that uses the most different words or word order.

The authors provided additional information regarding the human evaluation in the appendix of their paper. Importantly, they reported utilizing AMT 's feature to make HITs available only in specific regions, setting their region availability to the United States and the United Kingdom. Furthermore, they reported that participants were compensated for their time at a rate above the living wage in the regions selected.

Ultimately, in comparing paraphrase pairs the authors evaluated 300 sentences sampled equally from the three datasets, with paraphrases generated by each model resulting in a total of 1800 paraphrases.[7] For a particular pair of two system outputs for a given input sentence, separately for each of the three criteria, a given system received $+1$ or -1 depending on whether it was chosen as the best ($+1$) or worst (-1). The final scores for each model were then calculated by averaging the scores across all of that model's scored samples for a particular criterion. We refer to this score as BWS [155, 156].

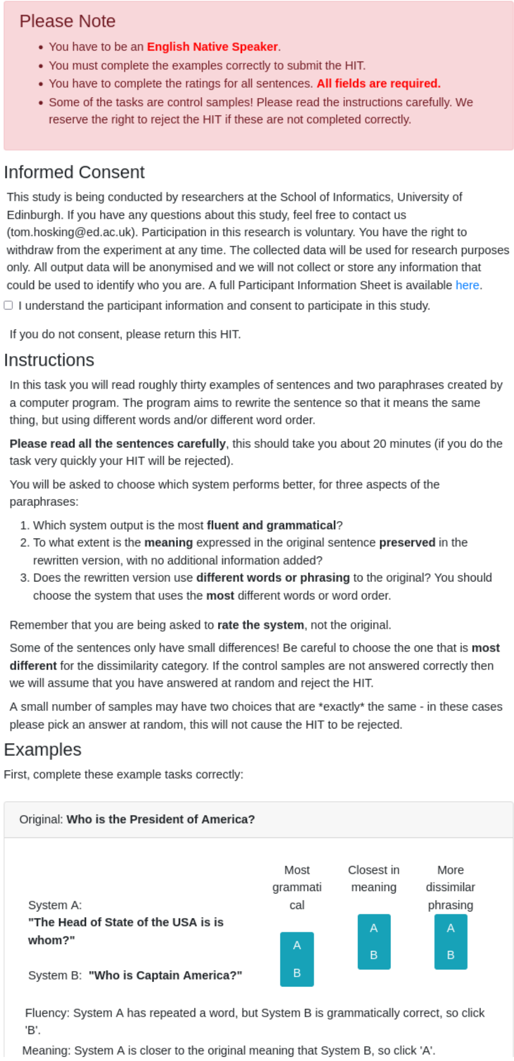
According to the authors, HRQ-VAE was found to be *more fluent* and *more diverse* while maintaining a *similar meaning* to the original sentence. Figure 4 in their paper shows the results of the human evaluation. We identified five unique **claims** based on the human evaluation results in the original paper. Reproducing the human evaluation experiments allowed us to verify these claims:

- **Claim 1:** The VAE baseline is the best at preserving meaning.
- **Claim 2:** The VAE baseline is the worst at introducing variation to the output.
- **Claim 3:** HRQ-VAE better preserves the original intent compared to the other systems.
- **Claim 4:** HRQ-VAE introduces more diversity than VAE.
- **Claim 5:** HRQ-VAE generates much more fluent output than VAE.

Our goal was to repeat the allocated experiment as closely as possible to the original study. We set up the experiment using all information available to us from the original paper [151] and from follow-up communications with the authors by the ReproHum leadership team.

Additional Details from the Authors. The ReproHum team provided us with additional information obtained from them. Specifically, the authors shared the exact outputs that they evaluated and the user interface that they used for the human evaluation. Crucially, the authors

[7] There were four systems; for each comparison, we selected two out of four: $\binom{4}{2} = 6$. With the resulting six unique comparisons for each of the 300 sentences, we have a total of $6 \times 300 = 1800$ comparisons.



Please Note

- You have to be an **English Native Speaker**.
- You must complete the examples correctly to submit the HIT.
- You have to complete the ratings for all sentences. **All fields are required.**
- Some of the tasks are control samples! Please read the instructions carefully. We reserve the right to reject the HIT if these are not completed correctly.

Informed Consent

This study is being conducted by researchers at the School of Informatics, University of Edinburgh. If you have any questions about this study, feel free to contact us (tom.hosking@ed.ac.uk). Participation in this research is voluntary. You have the right to withdraw from the experiment at any time. The collected data will be used for research purposes only. All output data will be anonymised and we will not collect or store any information that could be used to identify who you are. A full Participant Information Sheet is available here.

□ I understand the participant information and consent to participate in this study.

If you do not consent, please return this HIT.

Instructions

In this task you will read roughly thirty examples of sentences and two paraphrases created by a computer program. The program aims to rewrite the sentence so that it means the same thing, but using different words and/or different word order.

Please read all the sentences carefully, this should take you about 20 minutes (if you do the task very quickly your HIT will be rejected).

You will be asked to choose which system performs better, for three aspects of the paraphrases:

1. Which system output is the most **fluent and grammatical**?
2. To what extent is the **meaning** expressed in the original sentence **preserved** in the rewritten version, with no additional information added?
3. Does the rewritten version use **different words or phrasing** to the original? You should choose the system that uses the **most** different words or word order.

Remember that you are being asked to **rate the system**, not the original.

Some of the sentences only have small differences! Be careful to choose the one that is **most different** for the dissimilarity category. If the control samples are not answered correctly then we will assume that you have answered at random and reject the HIT.

A small number of samples may have two choices that are *exactly* the same - in these cases please pick an answer at random, this will not cause the HIT to be rejected.

Examples

First, complete these example tasks correctly:

Original: **Who is the President of America?**

	Most grammatical	Closest in meaning	More dissimilar phrasing
System A: **"The Head of State of the USA is is whom?"**	A B	A B	A B
System B: **"Who is Captain America?"**			

Fluency: System A has repeated a word, but System B is grammatically correct, so click 'B'.

Meaning: System A is closer to the original meaning that System B, so click 'A'.

Figure 4.3: The user interface used for human evaluation in the original study.

noted that they used *attention checks* (control samples with known labels). Each task contained two control samples; in one control sample, the system was a "distractor" and the output was a random sample with a completely different meaning that should clearly never be chosen as best for the *meaning* criterion. The other control sample was when the system's output was the same as the input, which should clearly never be chosen as best for the *dissimilarity* criterion. Note that the second control sample was not relevant to our reproduction, as we were reproducing the results for the *meaning* criterion. In their communication, the authors mentioned that HITs for which either of these attention checks were failed were rejected and resubmitted to AMT. Additionally, they reported compensating participants with $3.50 per HIT with an expected completion time of 20 minutes.

Known Deviations from the Original Experiment

Similar to the reproduction of Puduppully et al.'s experiment [150], we encountered several deviations from the original experiment. We detail these deviations below, along with the rationale for each change.

Scope of Reproduction: We focused on a narrow scope of the original paper: we sought to reproduce the outcomes of the human evaluation experiments for the *meaning* criterion.

Crowdsourcing Platform: Our biggest deviation from the original experiment was in the crowdsourcing platform used. While the original study had utilized AMT, we used Prolific.[8] This decision was made across all experiments in the ReproHum project to ensure consistency, due to limitations in credit usage on AMT and the administrative overhead of managing the funds for different experiments.

Prolific survey design is different from AMT, and we had to adapt the original survey design to the Prolific platform. To be more specific, setting up a survey similar to the structure of HITs was only possible using external survey tools. The ReproHum team shared the code for hosting a server to run the survey. We used a modified version of the code with additional checks to ensure the validity of the responses. Furthermore, we added thread safety to prevent

[8] https://www.prolific.com/

race conditions, where two or more threads try to access or modify the same data at the same time, leading to unpredictable or incorrect results.

Region Control: Our reproduction also deviated slightly in terms of participant region control. While the original authors had limited their HIT availability regions to the United States and the United Kingdom, we followed the region control guidelines of all experiments in the ReproHum project. This meant that participants from Australia and Canada were also included in addition to the United States and the United Kingdom.

Participant Selection: The authors reported filtering participants with approval rates less than 96%, and required that participants had completed at least 5000 HITs. In contrast, we set the approval rate to 99% and the minimum number of HITs completed to 200. This decision was based on the recommendations from Prolific to ensure high-quality participants.[9]

Failed Attention Checks: The original authors reported rejecting HITs for which the attention checks were failed. We did not reject any HITs based on attention checks per recommendations from the ReproHum team; however, we solicited new responses for tasks that failed attention checks.

Participation Limit: The original paper did not report whether a participant could respond to multiple HITs; we assume that no controls were in place for this. In Prolific, participants cannot respond to the same study more than once, even though the input data may be different.

Expected Completion Time: The original authors reported that the expected completion time for a HIT was 20 minutes. Our survey differed from the original study since we only

[9] https://www.prolific.com/resources/find-filter-favourite-how-to-select-participants-for-ai-tasks

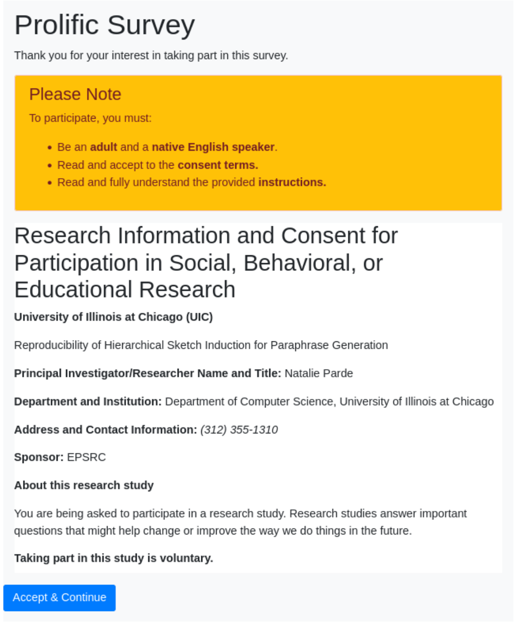
collected responses for the *meaning* criterion. We ran several surveys to estimate the time it would take to complete the task. Ultimately, we set the expected completion time to 8 minutes.

Payment: The original authors reported compensating participants with $3.50 per HIT for 20 minutes of work, resulting in an hourly rate of $10.50. We followed the guidelines of the ReproHum project, setting the wage as the minimum living wage in the United Kingdom (which was higher than our local minimum wage). At the time of data collection, this value was £12 which was equivalent to $15.14 using the exchange rate between UK and US currency at that time. To be more specific, the participants received £1.60 or $2 for 8 minutes of work.

User Interface: Our institutional consent forms were required to be much more detailed than those used in the original study, and this was beyond our control. To ensure that the participants were not overwhelmed, we split the welcome, instructions, and task into three separate pages. We have included images of the user interface used for the reproduction (Figures 4.4, 4.5, and 4.6).

4.5 Results

We present the results of each reproduction in its dedicated subsection. The selected statistical test for each reproduction is Analysis of Variance (ANOVA). ANOVA is a statistical method used to compare the means of three or more groups. We used ANOVA to compare the means of the scores for each system in the human evaluation experiments. We also report the results of the power analysis for each reproduction. The power analysis was conducted using the *pwr* package in R.



Prolific Survey

Thank you for your interest in taking part in this survey.

Please Note

To participate, you must:

- Be an **adult** and a **native English speaker**.
- Read and accept to the **consent terms.**
- Read and fully understand the provided **instructions.**

Research Information and Consent for Participation in Social, Behavioral, or Educational Research

University of Illinois at Chicago (UIC)

Reproducibility of Hierarchical Sketch Induction for Paraphrase Generation

Principal Investigator/Researcher Name and Title: Natalie Parde

Department and Institution: Department of Computer Science, University of Illinois at Chicago

Address and Contact Information: *(312) 355-1310*

Sponsor: EPSRC

About this research study

You are being asked to participate in a research study. Research studies answer important questions that might help change or improve the way we do things in the future.

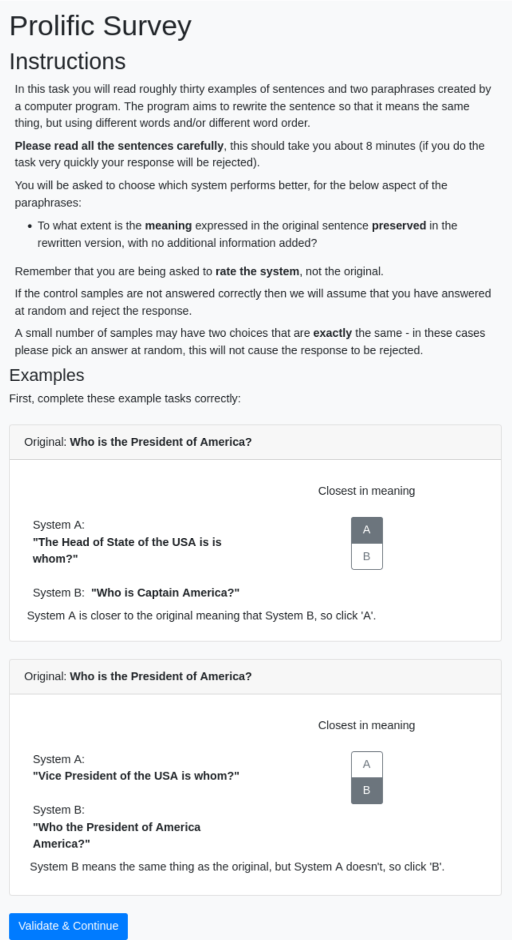
Taking part in this study is voluntary.

Figure 4.4: The user interface used for the reproduction of human evaluation

4.5.1 Data-to-text Generation with Macro Planning

We begin with the outcomes of the power analysis. The sample size per evaluation criterion stands at 20. With this size, to detect a large effect size (Cohen's f=0.4), the power of our study amounts to 0.90. For a medium effect size (Cohen's f=0.25), the power is 0.47. Lastly, for a small effect size (Cohen's f=0.1), the power is 0.10. These results suggest that our study is well-powered to detect large effect sizes, but underpowered for medium and small effect sizes. Note that the original experiment had double the sample size. While that renders the power of detecting medium effect to a reasonable level of 0.8, the power of detecting small effect remains low at 0.17. The required sample size to achieve a power of 0.8 for a small effect size is 240.

Our results for the human evaluation of the second task in *Data-to-text Generation with Macro Planning* by Puduppully et al. [150] are presented in Tables 4.1, 4.2, and 4.3 for coherence, grammar, and conciseness, respectively. The results were computed using 1800



Prolific Survey

Instructions

In this task you will read roughly thirty examples of sentences and two paraphrases created by a computer program. The program aims to rewrite the sentence so that it means the same thing, but using different words and/or different word order.

Please read all the sentences carefully, this should take you about 8 minutes (if you do the task very quickly your response will be rejected).

You will be asked to choose which system performs better, for the below aspect of the paraphrases:

- To what extent is the **meaning** expressed in the original sentence **preserved** in the rewritten version, with no additional information added?

Remember that you are being asked to **rate the system**, not the original.

If the control samples are not answered correctly then we will assume that you have answered at random and reject the response.

A small number of samples may have two choices that are **exactly** the same - in these cases please pick an answer at random, this will not cause the response to be rejected.

Examples

First, complete these example tasks correctly:

Original: **Who is the President of America?**

Closest in meaning

System A: **"The Head of State of the USA is is whom?"**

System B: **"Who is Captain America?"**

[✓ A] B

System A is closer to the original meaning that System B, so click 'A'.

Original: **Who is the President of America?**

Closest in meaning

System A: **"Vice President of the USA is whom?"**

System B: **"Who the President of America America?"**

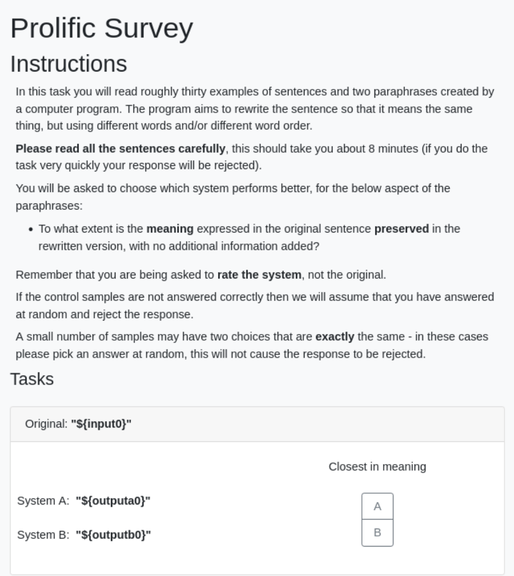
A [✓ B]

System B means the same thing as the original, but System A doesn't, so click 'B'.

Figure 4.5: The user interface used for the reproduction of human evaluation

responses collected through twelve mini-batches (four for each of the three evaluation criteria). Each batch took approximately a day to finish collecting all responses. Overall, 262 crowd workers participated in this task.

The original study reported a Krippendorff's $\alpha = 0.47$; the details of how this was calculated were not provided. We calculated Krippendorff's α for our results for each evaluation criterion. The results were as follows: $\alpha = -0.01$ for coherence, $\alpha = 0.01$ for grammar, and $\alpha = -0.04$ for conciseness. This suggests that the crowd workers' responses were not in agreement with



Prolific Survey

Instructions

In this task you will read roughly thirty examples of sentences and two paraphrases created by a computer program. The program aims to rewrite the sentence so that it means the same thing, but using different words and/or different word order.

Please read all the sentences carefully, this should take you about 8 minutes (if you do the task very quickly your response will be rejected).

You will be asked to choose which system performs better, for the below aspect of the paraphrases:

- To what extent is the **meaning** expressed in the original sentence **preserved** in the rewritten version, with no additional information added?

Remember that you are being asked to **rate the system**, not the original.

If the control samples are not answered correctly then we will assume that you have answered at random and reject the response.

A small number of samples may have two choices that are **exactly** the same - in these cases please pick an answer at random, this will not cause the response to be rejected.

Tasks

Original: **"${input0}"**

Closest in meaning

System A: **"${outputa0}"**

System B: **"${outputb0}"**

Figure 4.6: The user interface used for the reproduction of human evaluation

System	**Orig**	**Ours**	**CV***	r	ρ
Gold	46.25*	12.50	26.01		
Templ	-52.92*	-20.00*	51.65		
ED+CC	-8.33	-7.50	0.90	0.93	0.90
RBF-2020	4.58	9.17	4.28		
Macro	10.42	5.83	4.23		

Table 4.1: Comparison of ROTOWIRE performance metrics for coherence. * indicates a statistically significant difference ($p < 0.05$) between Macro and the other systems. Pearson's correlation is represented by r and Spearman's correlation is represented by ρ. CV* is computed using n=2. Note that the **Original** column numbers are from Table 5 of the original paper, while the **Ours** column numbers are from our reproduction.

each other.

We can observe from the results that the magnitude of difference reported between conditions in the original study's results is much higher than ours. For example, when evaluating grammaticality, the original study reports a BWS score of -61.67 for the template system (the lowest score reported among all conditions), while ours is -23.33 (the lowest score reported

System	**Orig**	**Ours**	**CV***	r	ρ
Gold	38.33	14.17	19.08		
Templ	-61.67*	-23.33*	66.48		
ED+CC	5.00	-8.33	13.52	0.91	0.97
RBF-2020	13.33	9.17	3.73		
Macro	5.00	8.33	3.11		

Table 4.2: Comparison of ROTOWIRE performance metrics for grammar. * indicates a statistically significant difference ($p < 0.05$) between Macro and the other systems. Pearson's correlation is represented by r and Spearman's correlation is represented by ρ. CV* is computed using n=2. Note that the **Original** column numbers are from Table 5 of the original paper, while the **Ours** column numbers are from our reproduction.

System	**Orig**	**Ours**	**CV***	r	ρ
Gold	30.83	5.83	21.06		
Templ	-36.67*	-5.83	39.04		
ED+CC	-4.58	-5.00	0.44	0.87	1.00
RBF-2020	3.75	0.83	2.85		
Macro	6.67	4.17	2.36		

Table 4.3: Comparison of ROTOWIRE performance metrics for conciseness. * indicates a statistically significant difference ($p < 0.05$) between Macro and the other systems. Pearson's correlation is represented by r and Spearman's correlation is represented by ρ. CV* is computed using n=2. Note that the **Original** column numbers are from Table 5 of the original paper, while the **Ours** column numbers are from our reproduction.

among all conditions in our reproduction). Similarly, for coherence, our BWS score of 12.50 is much smaller than the reported BWS=46.25. We utilized the same statistical significance test as the original study (a one-way ANOVA with post-hoc Tukey HSD tests). The results of this test suggest that only two conditions (the Template system's scores for grammar and coherence) yield results with statistically significant differences from the Macro system. This is a different finding from the original study, which reported statistically significant different results for four measures: Templ for grammar, coherence, and conciseness, and Gold for coherence.

In our analyses of the observed errors, we found a high level of similarity between the original experiment and our reproduction. We used Pearson's r and Spearman's ρ to measure the correlation between the two experiments. The Pearson correlations for coherence, grammar, and conciseness were 0.93, 0.91, and 0.87, respectively. Similarly, the Spearman correlations for coherence, grammar, and conciseness were 0.90, 0.97, and 1.00, respectively.

Next, we discuss the results of a parallel reproduction performed by Miltenburg et al. [148], henceforth referred to as Lab 1. Our replication efforts are denoted as Lab 2. Given that Miltenburg et al. [148] have made the raw data publicly available, it facilitates a comprehensive joint analysis of both sets of responses. The combined study is referred to as Labs 1 and 2.

Figure 4.7 presents the BWS scores for various systems evaluated in Lab 1, while Figure 4.8 illustrates similar scores for the systems assessed in Lab 2. Moreover, Figure 4.9 depicts the BWS scores for the integrated data from both labs. Surprisingly, although both labs followed the identical procedure and utilized the same input data, their results are dramatically different. Additionally, when combining the results from both labs, we observe that the differences between the systems effectively neutralize each other, indicating little to no disparity between the systems evaluated.

Given the discrepancies between the results of Labs 1 and 2, low inter-rater agreement, and low power of the experiment, drawing any meaningful conclusions from the combined data is challenging. We further discuss the implications of these results in the next section.

4.5.2 Hierarchical Sketch Induction for Paraphrase Generation

Similar to the previous reproduction, we start with the power analysis of the study. There are four systems, and our sample size is 300. The experiment's power for detecting a small effect size (Cohen's f=0.1) is 0.84. This value is above the desired threshold of 0.80, indicating that our study is well-powered to detect small effect sizes.

For statistical analysis, we employed ANOVA to determine significant differences among the

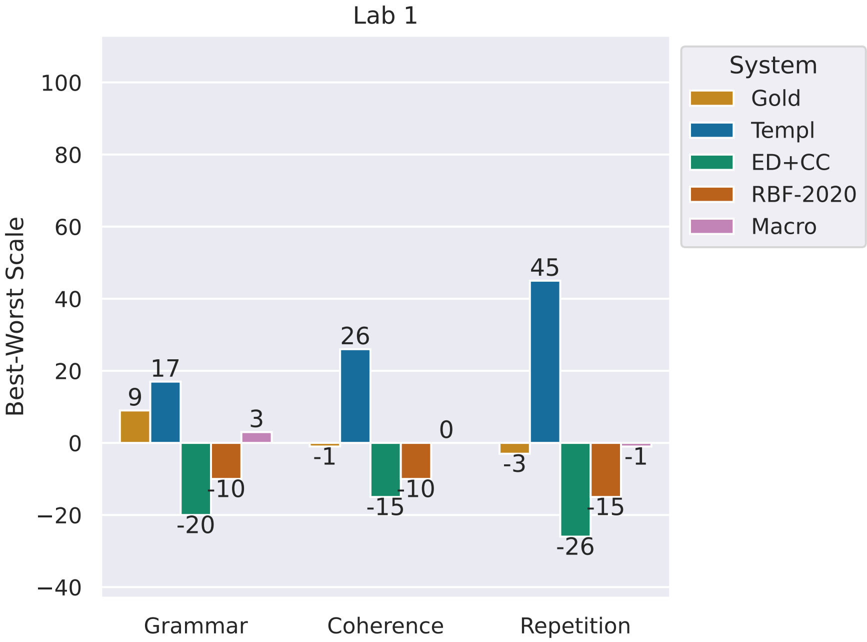


Figure 4.7: Bar plot of the BWS scores for the different systems in Lab 1 (Miltenburg et al. [148]).

System	**Orig**	**Ours**	**CV***	r	ρ
VAE	36	37.04	0.76	0.99	1
Latent BoW	-16	-14.52	1.74		
Separator	-24	-29.78	7.88		
HRQ-VAE	4	7.26	3.08		

Table 4.4: Overview of the results of the human evaluation alongside precision metrics to reflect the degree of reproducibility. Pearson's correlation is represented by r and Spearman's correlation is represented by ρ. CV* is computed using n=2. *Orig* refers to the original results reported by Hosking et al. [151].

means of multiple independent groups. In conducting the ANOVA test, we observed an F value of 79.93 with a corresponding p=3.97e-47. Subsequently, we used Tukey's HSD test to identify significant differences between individual groups, revealing significant distinctions among all

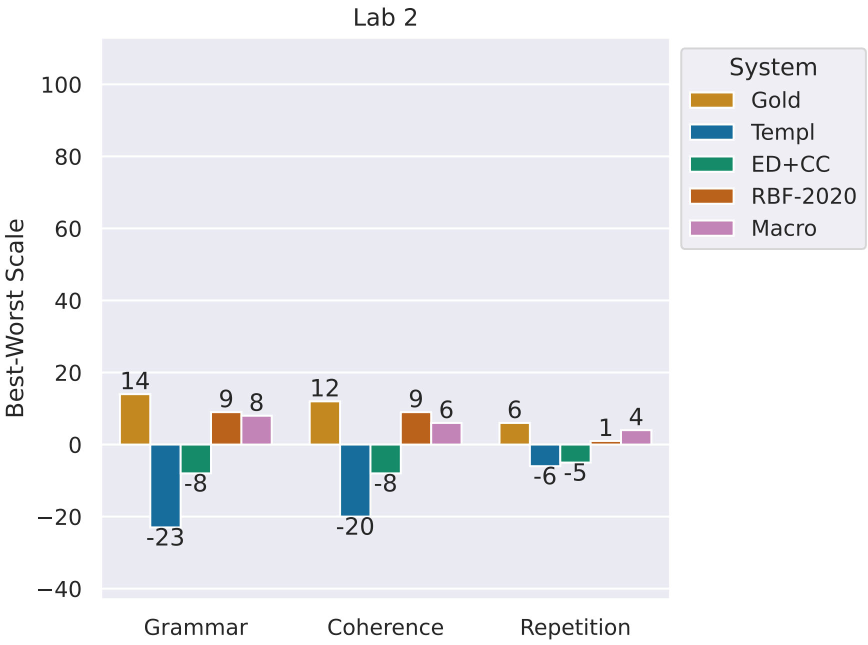


Figure 4.8: Bar plot of the BWS scores for the different systems in Lab 2 (ours).

System	**Orig**	**Lab 3**	**CV***	r	ρ
VAE	36	23	43.93	0.99	1
Latent BoW	-16	-8.67	59.24		
Separator	-24	-17.89	29.08		
HRQ-VAE	4	3.56	11.60		

Table 4.5: Overview of the results of the human evaluation alongside precision metrics to reflect the degree of reproducibility.

groups.

Table 4.4 shows the results of the human evaluation for the selected criterion, comparing the outcomes from the original and reproduced experiments. Overall, we observe that our results are very close to the scores originally reported [151]. This is reflected in low CV* values for all the systems. Pearson correlation and p-value are r=0.99 and p=0.01, respectively.

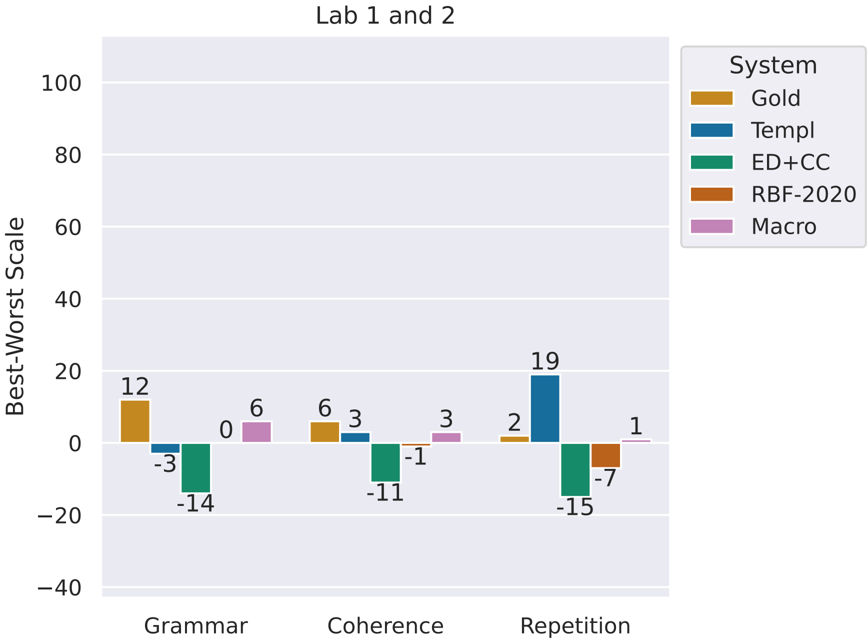


Figure 4.9: Bar plot of the BWS scores for the different systems in combined responses from Lab 1 and 2.

Similarly, Spearman correlation and p-value are ρ=1.00 and p=0.00. Both Pearson and Spearman correlations are very high, indicating a strong relationship between the original and reproduced scores. Figure 4.10 presents this same information in the format used by the original paper, showing BWS outcomes for the four systems compared in the original paper and in our reproduction. We used Krippendorff's alpha to evaluate the agreement among the categorical responses collected, resulting in a value of α=0.51. This metric was not included in the original study, preventing a direct comparison of our findings.

Next, we discuss the results of a parallel reproduction performed by Watson et al. [149], henceforth referred to as Lab 3. Table 4.5 presents the results of the human evaluation for the selected criterion, comparing the outcomes from the original and reproduced experiments. Unlike the previous reproduction, both reproductions are highly correlated to the original study.

Claim	Verification
The VAE baseline is the best at preserving meaning.	Verified
The VAE baseline is the worst at introducing variation to the output.	Out of Scope
HRQ-VAE better preserves the original intent compared to the other systems.	Verified
HRQ-VAE introduces more diversity than VAE.	Out of Scope
HRQ-VAE generates much more fluent output than VAE.	Out of Scope

Table 4.6: Claims and verifications.

Another interesting observation is that the results from Lab 3 have a higher CV* value than ours. We discuss potential reasons for these discrepancies in the next section.

Overall, given reproduced results' similarity to and correlation with the originally reported results, we could easily confirm two out of five of the original claims based on the human evaluation results. The other three claims were out of scope for our reproduction, as they pertained to criteria other than *meaning*. We summarize the claims and our verification in Table 4.6.

4.6 Discussion

It is important to note that the experiments discussed in this chapter were selected after several rounds of rigorous filtering. Moreover, the authors of the respective papers demonstrated exceptional cooperation by supplying additional data and details that are frequently omitted in scientific publications within this field. Such a commitment to transparency and collaboration is highly commendable and facilitates a deeper level of critique and understanding. With this in mind, we can discuss potential reasons for the discrepancies observed in the results of the reproduced experiments.

To discuss the implications of our findings, we first reiterate the contributions of the work

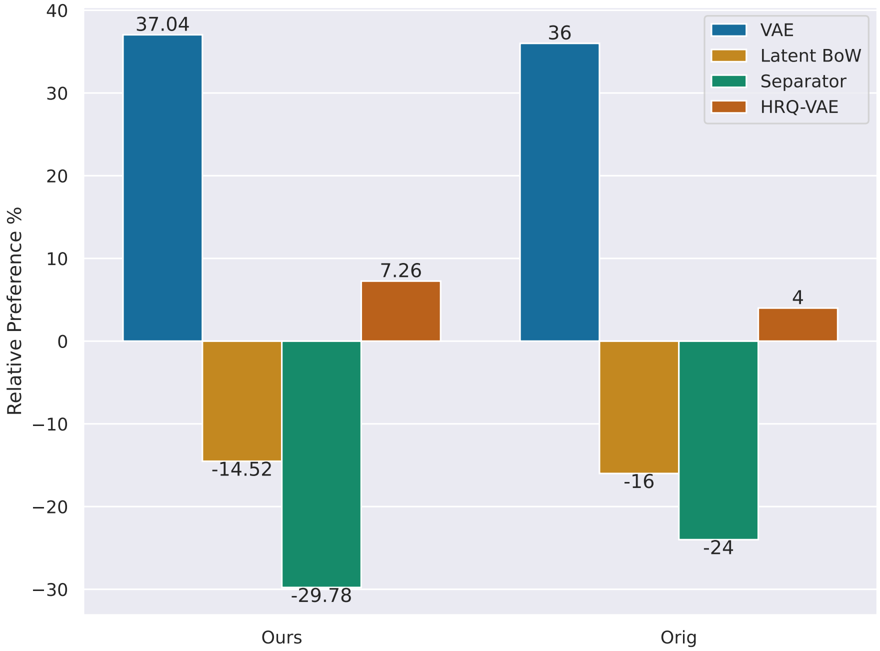


Figure 4.10: Results of human evaluation, comparing the original and reproduced systems.

of Puduppully et al. [150] and the scope of our reproduction. Puduppully et al. [150] presented a novel technique with the goal of improving the quality of data-to-text generation. They used a combination of automatic and human evaluation methods to show that their approach was superior to existing state-of-the-art models on two datasets, RotoWire and MLB. The scope of our reproduction was limited to the second human evaluation task reported in their paper, examining the quality of generated text based on coherence, grammaticality, and conciseness. Furthermore, we only reproduced the results on the RotoWire dataset. To provide a better perspective, the MLB dataset is larger (nearly ten times as many tokens) than the RotoWire dataset. Hence, we cannot form conclusive judgments based on a full reproduction of this experiment; rather, we focus on a subset of it that was reproduced in our study.

The variations observed in the reproduction of the experiment by Puduppully et al. [150]

highlight potential flaws in the experimental design. A power analysis indicates that the experiment is severely underpowered. Specifically, to detect a small effect size, the sample size would need to increase twelvefold, from 20 to 240, to achieve an acceptable level of statistical power. This high variability could likely be attributed to non-representative samples and insufficient sample sizes.

Unfortunately, reproducing the first experiment with the necessary sample size of 240 is highly impractical. Such replication would require complete access to the outputs of the trained models. In the absence of this data, one would need to recreate the entire research from scratch. This approach is fraught with potential complications as it could introduce additional, unintended changes and variations in the results. Consequently, achieving a faithful replication under these circumstances is challenging. Some of these challenges were discussed in the previous chapter (Chapter 3.1).

Additionally, even within the collected responses, the inter-annotator agreement was notably low across all evaluation criteria. This low agreement suggests several possible issues: the tasks may not have been clearly defined, the instructions provided might have lacked clarity, or the complexity of the task may have exceeded the crowd workers' domain knowledge, leading to varied interpretations and responses. Moreover, not every participant encountered attention checks, which could have further contributed to inconsistent responses. These factors collectively suggest that the experiment's design and execution could benefit from major revisions to enhance reliability and clarity.

Aside from the missing attention checks, the minimum requirements for participation in the study were set at a 98% approval rate across at least 1000 previously completed tasks. The impact of pre-screening participants based on these criteria is unclear and could be a future area of investigation. While these requirements may help ensure the quality of the data collected, they could also introduce biases. For one, the number of completed tasks is not a reliable indicator of a worker's experience or expertise. Additionally, over time, this number inflates. A similar concern was raised by González Corbelle et al. [163]. Considering that data collection is essential to machine learning and NLP research, it is important to ensure the quality of the data collected. Selecting based on the top k% of workers could be a simple workaround to eliminate this reliance on the number of completed tasks.

The findings from the second experiment strongly support the reproducibility of the original study. This suggests that this study would be an excellent candidate for the subsequent phase of the ReproHum project. In this next phase, minor modifications could be implemented to further examine the robustness of the original findings. One potential alteration could involve replacing the current attention check, which is irrelevant to the *meaning* criterion, with one that is more pertinent, thus enhancing the experimental design.

Results from the parallel reproduction conducted by Watson et al. [149] align with our findings, further affirming the reproducibility of the original study. Nevertheless, we observed higher CV* values in their results, indicating a higher level of variability. The only documented difference between the data collection processes of the two studies was the prescreening filters used for the participants. We utilized the Prolific recommended guidelines for selecting participants, setting the approval rate to 99% and the minimum number of accepted tasks to 200 while the other study did not. This analysis is post-hoc and should be interpreted with caution. However, it does suggest that the number of accepted tasks could be a potential source of variability in the data collected. Prior to reproductions, researchers should analyze potential impact of each condition. If the outcome of applying a certain condition varies over time, adjustments should be made to ensure this factor does not affect the results.

4.7 Conclusions

In this chapter, we have presented the outcomes of our replication of two experiments from the ReproHum project, along with the results of a parallel reproduction. Our focus lay on the human evaluations associated with two papers: *Data-to-text Generation with Macro Planning* by Puduppully et al. [150] and *Hierarchical Sketch Induction for Paraphrase Generation* by Hosking et al. [151].

Our power analysis of the first experiment disclosed that the study was substantially underpowered, necessitating a twelvefold increase in sample size to detect a small effect size. A key conclusion from this chapter is the realization that human evaluations can yield non-reproducible results if not meticulously designed. Although crowd-sourcing platforms provide a cost-effective and expedient means for data collection, the integrity of the collected data is not inherently assured. Realistically, it is advisable to eschew human evaluation in scenarios where there is insufficient funding to gather an adequate number of responses to discern the desired effect size rather than risk obtaining unreliable results. Furthermore, power analysis would ground an empirical study in its potential limitations, thereby enhancing the credibility of its findings. This is by no means a new insight, yet its importance is ignored in many studies.

This chapter concludes the primary discussions and analyses of the dissertation. In the following chapter, we summarize the key contributions and conclusions of the dissertation, providing essential insights and recommendations for the academic community and other researchers.

Chapter 5

Conclusions and Final Remarks

5.1 Summary of Contributions

This dissertation investigates the application of Open Science principles within the domains of ML and NLP. It initially addresses the challenges of uncertainty in evaluations from a metrological standpoint, particularly through the use of NHST. Despite its underutilization in comparison to other fields, NHST plays a critical role in identifying false discoveries and determining the required sample sizes for different effect sizes. This dissertation stresses the necessity of accurately reporting uncertainty to enhance the reproducibility of results and cautions against under-specification that may hinder future replication efforts.

We identified positive trends in the availability of research artifacts, particularly at conferences that emphasize the importance of sharing. Subsequently, this work underscores the significance of releasing self-contained research artifacts that include both the environment and the source code. Such a practice would obviate the need for manual setup and configuration, which can introduce complications. Furthermore, it shifts the effort required to setup the project from the late stage to the very beginning.

Several case studies were conducted, focusing on the reproducibility assessment of various publications. The outcomes of these studies were mixed. In some instances, the authors were responsive and willing to address the issues identified. In other cases, however, the authors were unresponsive or unable to provide the necessary information. This situation reiterates the importance of releasing self-contained research artifacts. During the reproduction of a neural text simplification pipeline, three bugs were identified and subsequently fixed; surprisingly, the impact of these bugs was negligible.

The penultimate chapter evaluates the reproducibility of human evaluation conducted as part of the ReproHum initiative. This initiative represents the largest meta-study on the reproducibility of human evaluations in NLP and related areas. It aims to identify prevalent issues and establish standards for future human evaluation studies. We reported one successful and one failed reproduction attempts. Further investigation revealed low statistical power and low quality of the responses as potential culprits for the failure.

Ultimately, this dissertation focused on reporting and evaluation practices in ML and NLP. Given the exponential growth of these fields, even modest improvements in research reproducibility could have substantial impacts. It is hoped that the findings presented here will encourage researchers to adopt more rigorous, open, and accessible practices, thereby fostering more reliable and reproducible research landscapes.

5.2 Final Conclusions

This dissertation examined challenges in reproducing both automatic and human evaluations in ML and NLP research. Although our findings are constrained by the limited number of studies reviewed, the breadth of identified issues raises important concerns. Despite potential pathways for enhancement, a comprehensive and immediate resolution seems unattainable without a drastic shift in research practices and publication norms. Moreover, the current academic setting discourages thorough and rigorous evaluation, favoring novelty and marginal performance improvements instead. Nevertheless, this work demonstrates that focusing on reproducibility and rigorous evaluation, although challenging, remains achievable. If the problems observed in this study reflect a broader trend, enhancing the evaluation and reproducibility of machine learning models will become increasingly crucial. Broader implications of the findings and potential future directions are as follows.

Perspective of Negative Results and Failed Reproductions: A fundamental part of the problem resides in the peer-review process. Although most conferences offer specific reviewing guidelines, more can be done to ensure that reviewers recognize and value studies addressing reproducibility issues of earlier works. More importantly, it is vital both for the scientific community and the general public to recognize that scientific publications do not represent the absolute truth. They should instead be viewed as reflections of the current understanding of specific issues, inclusive of the methodologies applied. Through critical reviews and replication efforts, these understandings can be continuously refined and improved. Additionally, a failed reproduction does not necessarily imply that the original study was incorrect. By reporting negative results and failed reproduction attempts, researchers can contribute to the collective knowledge base, fostering a culture of transparency and growth. Arguably, reproduction attempts are one of the best ways to assess the specificity and adequacy of scientific reports. Moreover, they can assist in identifying weak aspects of the experiment design.

Reproducibility as a Secondary Priority: In the realm of academic research, the rapid pace of development coupled with the intense pressure to publish has inadvertently relegated the importance of reproducibility. Researchers often engage in the iterative modification of complex models, striving to achieve results deemed publishable, a practice fundamentally flawed as it bypasses systematic evaluation crucial for verifying model legitimacy over time. As a result, distinguishing genuine improvements from anomalies becomes increasingly challenging, and even significant errors may remain undetected—a situation that is alarming within the scholarly community. Furthermore, the tendency to provide post-hoc justifications for positive outcomes exacerbates the issue, undermining the integrity of scientific reporting. It is imperative for the academic community to recalibrate its focus, prioritizing rigorous verification processes that enhance reproducibility and uphold the highest standards of scientific inquiry.

Assessing the prevalence of bugs in scientific publications is challenging, yet given the ubiquitous nature of bugs in software development and the reliance on empirical evidence, it is plausible to presume a similar situation in scientific literature. Such concerns are magnified in the realm of machine learning, characterized by the complexity of models and the often impenetrable nature of the results. In the interest of scientific integrity, researchers need to prioritize reproducibility from the inception of their projects. Utilizing containers to create isolated and controllable experimental environments is an effective strategy. This method allows for the packaging of code, data, and the necessary computational environment, ensuring that results can be faithfully reproduced by others. Moreover, researchers must diligently manage the details of their experiments; for instance, inadequately powered experiments can lead to misleading findings and irreproducible results, undermining the value of the research.

Evaluation and Benchmarks: While benchmarks are commonly used to estimate the real-world performance of a model on specific tasks, their role has become disproportionately emphasized. If the ultimate objective is to develop models suited for application in real-world scenarios, it becomes critical to evaluate them on real-world data. Although benchmarks provide valuable starting points, they should not represent the final goal. Deploying machine learning pipelines involves a series of complex steps including data collection, preprocessing, model training, and evaluation. To move beyond benchmark-focused strategies, utilizing online evaluation emerges as a more effective method. This approach allows for the continuous assessment of a model in a real-world context, offering direct insights into its performance. Online evaluations, which may be conducted on unlabeled data, enable the definition and application of task-specific metrics. Additionally, practices such as error analysis and A/B testing enrich understanding by revealing how models perform under varying conditions. This comprehensive evaluation framework is essential for developing robust models capable of functioning effectively in practical applications. In sensitive fields such as healthcare, where model performance directly impacts patient outcomes, the importance of rigorous evaluation is even more pronounced. Machine learning holds the potential to significantly improve human lives; however, rushing its deployment without first building trust could hinder advancement.

Bibliography

[1] Mohammad Arvan, Luís Pina, and Natalie Parde. “Reproducibility of Exploring Neural Text Simplification Models: A Review”. In: *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*. Waterville, Maine, USA and virtual meeting: Association for Computational Linguistics, July 2022, pp. 62–70. URL: https://aclanthology.org/2022.inlg-genchal.10.

[2] Mohammad Arvan, Luís Pina, and Natalie Parde. “Reproducibility in Computational Linguistics: Is Source Code Enough?” In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Association for Computational Linguistics, 2022, pp. 2350–2361. URL: https://aclanthology.org/2022.emnlp-main.150.

[3] Mohammad Arvan, A. Seza Doğruöz, and Natalie Parde. “Investigating Reproducibility at Interspeech Conferences: A Longitudinal and Comparative Perspective”. In: *Proc. INTERSPEECH 2023*. 2023, pp. 3929–3933. DOI: 10.21437/Interspeech.2023-2252.

[4] Mohammad Arvan and Natalie Parde. “Human Evaluation Reproduction Report for Data-to-text Generation with Macro Planning”. In: *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*. Ed. by Anya Belz, Maja Popović, Ehud Reiter, Craig Thomson, and João Sedoc. Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria, Sept. 2023, pp. 89–96. URL: https://aclanthology.org/2023.humeval-1.8.

[5] Mohammad Arvan and Natalie Parde. “ReproHum #0712-01: Human Evaluation Reproduction Report for “Hierarchical Sketch Induction for Paraphrase Generation””. In: *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*. Sept. 2023.

[6] UNESCO. *UNESCO Recommendation on Open Science*. Nov. 2021. DOI: 10.5281/zenodo.5834767. URL: https://doi.org/10.5281/zenodo.5834767.

[7] Joint Committee for Guides in Metrology. "Evaluation of measurement data—Guide to the expression of uncertainty in measurement". In: *Int. Organ. Stand. Geneva ISBN* 50 (2008), p. 134.

[8] JCGM. "JCGM 200: 2012 International Vocabulary of Metrology: Basic and General Concepts and Associated Terms (VIM)". In: *The Joint Committee for Guides in Metrology and The Bureau International des Poids et Mesures: Paris, France* (2012).

[9] Anne Plant and Robert Hanisch. "Reproducibility in Science: A Metrology Perspective". In: *Harvard Data Science Review* 2.4 (Dec. 2020). https://hdsr.mitpress.mit.edu/pub/0r4v4k4z.

[10] Anya Belz, Maja Popovic, and Simon Mille. "Quantified Reproducibility Assessment of NLP Results". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 16–28. URL: https://aclanthology.org/2022.acl-long.2.

[11] Anya Belz. "A Metrological Perspective on Reproducibility in NLP". In: *Comput. Linguistics* 48.4 (2022), pp. 1125–1135. DOI: 10.1162/coli_a_00448. URL: https://doi.org/10.1162/coli%5C_a%5C_00448.

[12] Philip Wadler and Sophia Drossopoulou, eds. *Proceedings of the ACM on Programming Languages*. OOPSLA. Association for Computing Machinery, 2021.

[13] Babak Falsafi, Michael Ferdman, Shan Lu, and Thomas F. Wenisch, eds. *ASPLOS '22: 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Lausanne, Switzerland, 28 February 2022 - 4 March 2022*. ACM, 2022.

[14] Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. "A Systematic Review of Reproducibility Research in Natural Language Processing". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*. Ed. by Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty. Association for Computational Linguistics, 2021, pp. 381–393. ISBN: 978-1-954085-02-2. DOI: 10.18653/v1/2021.eacl-main.29. URL: https://doi.org/10.18653/v1/2021.eacl-main.29.

[15] Nicolas P. Rougier, Konrad Hinsen, Frédéric Alexandre, Thomas Arildsen, Lorena A. Barba, Fabien C. Y. Benureau, C. Titus Brown, Pierre de Buyl, Ozan Caglayan, Andrew P. Davison, Marc-André Delsuc, Georgios Detorakis, Alexandra K. Diem, Damien Drix, Pierre Enel, Benoît Girard, Olivia Guest, Matt G. Hall, Rafael Neto Henriques, Xavier Hinaut, Kamil S. Jaron, Mehdi Khamassi, Almar Klein, Tiina Manninen, Pietro Marchesi, Dan McGlinn, Christoph Metzner, Owen L. Petchey, Hans Ekkehard Plesser, Timothée Poisot, Karthik Ram, Yoav Ram, Etienne B. Roesch, Cyrille Rossant, Vahid Rostami, Aaron Shifman, Joseph Stachelek, Marcel Stimberg, Frank Stollmeier, Federico Vaggi, Guillaume Viejo, Julien Vitay, Anya E. Vostinar, Roman Yurchak, and Tiziano Zito.

"Sustainable computational science: the ReScience initiative". In: *PeerJ Comput. Sci.* 3 (2017), e142. DOI: 10.7717/peerj-cs.142. URL: https://doi.org/10.7717/peerj-cs.142.

[16] ACM. *ACM Artifact Review and Badging*. 2022. URL: https://web.archive.org/web/20220624192023/https://www.acm.org/publications/policies/artifact-review-and-badging-current (visited on 06/24/2022).

[17] Kirstie Whitaker. "Showing your working: a how to guide to reproducible research". In: *"slideshare"* (Sept. 2017). DOI: 10.6084/m9.figshare.5443201.

[18] Patrick D Schloss. "Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research". In: *MBio* 9.3 (2018), e00525–18.

[19] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. "PaLM: Scaling Language Modeling with Pathways". In: *J. Mach. Learn. Res.* 24 (2023), 240:1–240:113. URL: http://jmlr.org/papers/v24/22-1144.html.

[20] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015. URL: http://arxiv.org/abs/1409.0473.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is All you Need". In: *Annual Conference on Neural Information Processing Systems (NeurIPS 2017)*. Ed. by Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett. 2017, pp. 5998–6008. URL: http://papers.nips.cc/paper/7181-attention-is-all-you-need.

[22] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting". In: *J. Mach. Learn. Res.* 15.1 (2014), pp. 1929–1958. URL: http://dl.acm.org/citation.cfm?id=2670313.

[23] Nature. *Nature's Reporting standards and availability of data, materials, code and protocols*. 2022. URL: https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards#availability-of-computer-code (visited on 06/24/2022).

[24] AAAI. *AAAI Reproducibility Checklist*. 2022. URL: https://web.archive.org/web/20220624192004/https://aaai.org/Conferences/AAAI-21/reproducibility-checklist/ (visited on 06/24/2022).

[25] ACL. *ACL Responsible NLP Research*. 2022. URL: https://web.archive.org/web/20220622114506/https://aclrollingreview.org/responsibleNLPresearch/ (visited on 06/24/2022).

[26] Daniel Deutsch, Yash Kumar Lal, Annie Louis, Pete Walsh, Jesse Dodge, and Niranjan Balasubramanian. *2022 North American Chapter of the Association for Computational Linguistics Reproducibility Track*. 2022. URL: https://web.archive.org/web/20220531224424/https://2022.naacl.org/blog/reproducibility-track (visited on 05/31/2022).

[27] Joelle Pineau. *ML Reproducibility Checklist*. 2019. URL: https://web.archive.org/web/20220624192746/https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf (visited on 06/24/2022).

[28] Robert Stojnic. *ML Code Completeness Checklist*. 2022. URL: https://web.archive.org/web/20220624192742/https://GitHub.com/paperswithcode/releasing-research-code (visited on 06/24/2022).

[29] Edward Raff. "A Step Toward Quantifying Independently Reproducible Machine Learning Research". In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Ed. by Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett. 2019, pp. 5486–5496. URL: https://proceedings.neurips.cc/paper/2019/hash/c429429bf1f2af051f2021dc92a8ebea-Abstract.html.

[30] Martijn Wieling, Josine Rawee, and Gertjan van Noord. "Reproducibility in Computational Linguistics: Are We Willing to Share?" In: *Comput. Linguistics* 44.4 (2018). DOI: 10.1162/coli_a_00330. URL: https://doi.org/10.1162/coli%5C_a%5C_00330.

[31] Babatunde Kazeem Olorisade, Pearl Brereton, and Peter Andras. "Reproducibility of studies on text mining for citation screening in systematic reviews: Evaluation and checklist". In: *J. Biomed. Informatics* 73 (2017), pp. 1–13. DOI: 10.1016/j.jbi.2017.07.010. URL: https://doi.org/10.1016/j.jbi.2017.07.010.

[32] Odd Erik Gundersen and Sigbjørn Kjensmo. "State of the Art: Reproducibility in Artificial Intelligence". In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. Ed. by Sheila A. McIlraith and Kilian Q. Weinberger. AAAI Press, 2018, pp. 1644–1651. URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17248.

[33] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Hugo Larochelle. "Improving Reproducibility in Machine Learning Research(A Report from the NeurIPS 2019 Reproducibility Program)". In: *J. Mach. Learn. Res.* 22 (2021), 164:1–164:20. URL: http://jmlr.org/papers/v22/20-303.html.

[34] Koustuv Sinha, Jesse Dodge, Sasha Luccioni, Jessica Forde, Sharath Chandra Raparthy, François Mercier, Joelle Pineau, and Robert Stojnic. *ML Reproducibility Challenge 2021*. 2021. URL: https://web.archive.org/web/20220624192701/https://paperswithcode.com/rc2021 (visited on 06/24/2022).

[35] Anya Belz, Anastasia Shimorina, Shubham Agarwal, and Ehud Reiter. "The ReproGen Shared Task on Reproducibility of Human Evaluations in NLG: Overview and Results". In: *Proceedings of the 14th International Conference on Natural Language Generation, INLG 2021, Aberdeen, Scotland, UK, 20-24 September, 2021*. Ed. by Anya Belz, Angela Fan, Ehud Reiter, and Yaji Sripada. Association for Computational Linguistics, 2021, pp. 249–258. ISBN: 978-1-954085-51-0. URL: https://aclanthology.org/2021.inlg-1.24.

[36] Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. "Deep Reinforcement Learning That Matters". In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. Ed. by Sheila A. McIlraith and Kilian Q. Weinberger. AAAI Press, 2018, pp. 3207–3214. URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16669.

[37] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. "SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceed

ings.neurips.cc/paper_files/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf.

[38] Aarohi Srivastava, Abhinav Rastogi, and BIG-bench authors. "Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models". In: *Transactions on Machine Learning Research* (2023). ISSN: 2835-8856. URL: https://openreview.net/forum?id=uyTL5Bvosj.

[39] Noah Weber, Leena Shekhar, and Niranjan Balasubramanian. "The Fine Line between Linguistic Generalization and Failure in Seq2Seq-Attention Models". In: *Proceedings of the Workshop on Generalization in the Age of Deep Learning*. Ed. by Yonatan Bisk, Omer Levy, and Mark Yatskar. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 24–27. DOI: 10.18653/v1/W18-1004. URL: https://aclanthology.org/W18-1004.

[40] Gábor Melis, Chris Dyer, and Phil Blunsom. "On the State of the Art of Evaluation in Neural Language Models". In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL: https://openreview.net/forum?id=ByJHuTgA-.

[41] Xiang Zhou, Yixin Nie, Hao Tan, and Mohit Bansal. "The Curse of Performance Instability in Analysis Datasets: Consequences, Source, and Suggestions". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Association for Computational Linguistics, 2020, pp. 8215–8228. ISBN: 978-1-952148-60-6. DOI: 10.18653/v1/2020.emnlp-main.659. URL: https://doi.org/10.18653/v1/2020.emnlp-main.659.

[42] Hung Viet Pham, Shangshu Qian, Jiannan Wang, Thibaud Lutellier, Jonathan Rosenthal, Lin Tan, Yaoliang Yu, and Nachiappan Nagappan. "Problems and Opportunities in Training Deep Learning Software Systems: An Analysis of Variance". In: *35th IEEE/ACM International Conference on Automated Software Engineering, ASE 2020, Melbourne, Australia, September 21-25, 2020*. IEEE, 2020, pp. 771–783. DOI: 10.1145/3324884.3416545. URL: https://doi.org/10.1145/3324884.3416545.

[43] Stephanie C. Y. Chan, Samuel Fishman, Anoop Korattikara, John F. Canny, and Sergio Guadarrama. "Measuring the Reliability of Reinforcement Learning Algorithms". In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL: https://openreview.net/forum?id=SJlpYJBKvH.

[44] Scott M. Jordan, Yash Chandak, Daniel Cohen, Mengxue Zhang, and Philip S. Thomas. "Evaluating the Performance of Reinforcement Learning Algorithms". In: *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 4962–4973. URL: http://proceedings.mlr.press/v119/jordan20a.html.

[45] Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C. Courville, and Marc G. Bellemare. "Deep Reinforcement Learning at the Edge of the Statistical Precipice". In: *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*. Ed. by Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan. 2021, pp. 29304–29320. URL: https://proceedings.neurips.cc/paper/2021/hash/f514cec81cb148559cf475e7426eed5e-Abstract.html.

[46] Maja Popović, Mohammad Arvan, Natalie Parde, and Anya Belz. "Exploring Variation of Results from Different Experimental Conditions". In: *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada, July 2023.

[47] Julian Michael, Ari Holtzman, Alicia Parrish, Aaron Mueller, Alex Wang, Angelica Chen, Divyam Madaan, Nikita Nangia, Richard Yuanzhe Pang, Jason Phang, and Samuel R. Bowman. "What Do NLP Researchers Believe? Results of the NLP Community Metasurvey". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 16334–16368. DOI: 10.18653/v1/2023.acl-long.903. URL: https://aclanthology.org/2023.acl-long.903.

[48] Gerd Gigerenzer, Zeno Swijtink, Theodore Porter, Lorraine Daston, John Beatty, and Lorenz Kruger. "The inference experts". In: *The Empire of Chance: How Probability Changed Science and Everyday Life*. Ideas in Context. Cambridge University Press, 1989, pp. 70–122.

[49] Russell T. Warne. "Null Hypothesis Statistical Significance Testing and z-Tests". In: *Statistics for the Social Sciences: A General Linear Model Approach*. Cambridge University Press, 2020, pp. 152–182.

[50] David S Moore and George P McCabe. *Introduction to the practice of statistics*. Array. WH Freeman/Times Books/Henry Holt & Co, 1989.

[51] Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. "The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 1383–1392. DOI: 10.18653/v1/P18-1128. URL: https://aclanthology.org/P18-1128.

[52] Frederik Michel Dekking. *A Modern Introduction to Probability and Statistics: Understanding why and how*. Springer Science & Business Media, 2005.

[53] Katherine S Button, John PA Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson, and Marcus R Munafò. "Power failure: why small sample size undermines the reliability of neuroscience". In: *Nature reviews neuroscience* 14.5 (2013), pp. 365–376.

[54] Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. "With Little Power Comes Great Responsibility". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 9263–9274. DOI: 10.18653/v1/2020.emnlp-main.745. URL: https://aclanthology.org/2020.emnlp-main.745.

[55] Jacob Cohen. "The earth is round ($p<.05$)." In: *American psychologist* 49.12 (1994), p. 997.

[56] Gail M Sullivan and Richard Feinn. "Using effect size—or why the P value is not enough". In: *Journal of graduate medical education* 4.3 (2012), pp. 279–282.

[57] Haotian Zhu, Denise Mak, Jesse Gioannini, and Fei Xia. "NLPStatTest: A Toolkit for Comparing NLP System Performance". In: *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*. Ed. by Derek Wong and Douwe Kiela. Suzhou, China: Association for Computational Linguistics, Dec. 2020, pp. 40–46. URL: https://aclanthology.org/2020.aacl-demo.7.

[58] Joseph Berkson. "Tests of significance considered as evidence". In: *Journal of the American Statistical Association* 37.219 (1942), pp. 325–335.

[59] Denton E Morrison and Ramon E Henkel. *The significance test controversy: A reader*. Transaction Publishers, 2006.

[60] John PA Ioannidis. "Why most published research findings are false". In: *PLoS medicine* 2.8 (2005), e124.

[61] Wolfgang Forstmeier, Eric-Jan Wagenmakers, and Timothy H Parker. "Detecting and avoiding likely false-positive findings–a practical guide". In: *Biological Reviews* 92.4 (2017), pp. 1941–1968.

[62] Anne M Scheel. "Why most psychological research findings are not even wrong". In: *Infant and Child Development* 31.1 (2022), e2295.

[63] Edwin D. Simpson. "Statistical Significance Testing for Natural Language Processing". In: *Computational Linguistics* 46.4 (Feb. 2021), pp. 905–908. ISSN: 0891-2017. DOI: 10.1162/coli_r_00388. eprint: https://direct.mit.edu/coli/article-pdf/46/4/905/1888268/coli_r_00388.pdf. URL: https://doi.org/10.1162/coli%5C_r%5C_00388.

[64] Daniel J Simons. "The value of direct replication". In: *Perspectives on psychological science* 9.1 (2014), pp. 76–80.

[65] Blakeley B McShane, David Gal, Andrew Gelman, Christian Robert, and Jennifer L Tackett. "Abandon statistical significance". In: *The American Statistician* 73.sup1 (2019), pp. 235–245.

[66] Ronald L Wasserstein and Nicole A Lazar. *The ASA statement on p-values: context, process, and purpose*. 2016.

[67] Daniel J Benjamin, James O Berger, Magnus Johannesson, Brian A Nosek, E-J Wagenmakers, Richard Berk, Kenneth A Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, et al. "Redefine statistical significance". In: *Nature human behaviour* 2.1 (2018), pp. 6–10.

[68] Geoff Cumming. "The new statistics: Why and how". In: *Psychological science* 25.1 (2014), pp. 7–29.

[69] Ramal Moonesinghe, Muin J Khoury, and A Cecile J W Janssens. "Most published research findings are false—but a little replication goes a long way". In: *PLoS medicine* 4.2 (2007), e28.

[70] John PA Ioannidis. "How to make more published research true". In: *Revista Cubana de Información en Ciencias de la Salud (ACIMED)* 26.2 (2015), pp. 187–200.

[71] Y Andre Wang and Mijke Rhemtulla. "Power analysis for parameter estimation in structural equation modeling: A discussion and tutorial". In: *Advances in Methods and Practices in Psychological Science* 4.1 (2021), p. 2515245920918253.

[72] Kou Murayama, Reinhard Pekrun, and Klaus Fiedler. "Research practices that can prevent an inflation of false-positive rates". In: *Personality and Social Psychology Review* 18.2 (2014), pp. 107–118.

[73] Etienne P LeBel, Randy J McCarthy, Brian D Earp, Malte Elson, and Wolf Vanpaemel. "A unified framework to quantify the credibility of scientific findings". In: *Advances in Methods and Practices in Psychological Science* 1.3 (2018), pp. 389–402.

[74] Marcus R Munafò, Brian A Nosek, Dorothy VM Bishop, Katherine S Button, Christopher D Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J Ware, and John Ioannidis. "A manifesto for reproducible science". In: *Nature human behaviour* 1.1 (2017), pp. 1–9.

[75] Nils Reimers and Iryna Gurevych. "Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*. Ed. by Martha Palmer, Rebecca Hwa, and Sebastian Riedel. Association for Computational Linguistics, 2017, pp. 338–348. URL: https://www.aclweb.org/anthology/D17-1035/.

[76] Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah A. Smith. "Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping". In: *CoRR* abs/2002.06305 (2020). arXiv: 2002.06305. URL: https://arxiv.org/abs/2002.06305.

[77] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. "Underspecification presents challenges for credibility in modern machine learning". In: *The Journal of Machine Learning Research* 23.1 (2022), pp. 10237–10297.

[78] Siva Sivaganesan. "An Introduction to the Bootstrap (Bradley Efron and Robert J. Tibshirani)". In: *SIAM Rev.* 36.4 (1994), pp. 677–678. DOI: 10.1137/1036171. URL: https://doi.org/10.1137/1036171.

[79] Philipp Koehn. "Statistical Significance Tests for Machine Translation Evaluation". In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain*. ACL, 2004, pp. 388–395. URL: https://aclanthology.org/W04-3250/.

[80] Eustasio Del Barrio, Juan A Cuesta-Albertos, and Carlos Matrán. "An optimal transportation approach for assessing almost stochastic order". In: *The Mathematics of the Uncertain*. Springer, 2018, pp. 33–44.

[81] Rotem Dror, Segev Shlomov, and Roi Reichart. "Deep Dominance - How to Properly Compare Deep Neural Models". In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. Ed. by Anna Korhonen, David R. Traum, and Lluís Màrquez. Association for Computational Linguistics, 2019, pp. 2773–2785. ISBN: 978-1-950737-48-2. DOI: 10.18653/v1/p19-1266. URL: https://doi.org/10.18653/v1/p19-1266.

[82] Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. "An Empirical Investigation of Statistical Significance in NLP". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Ed. by Jun'ichi Tsujii, James Henderson, and Marius Pașca. Jeju Island, Korea: Association for Computational Linguistics, July 2012, pp. 995–1005. URL: https://aclanthology.org/D12-1091.

[83] Matt Post. "A Call for Clarity in Reporting BLEU Scores". In: *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*. Ed. by Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana L. Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor. Association for Computational Linguistics, 2018, pp. 186–191. ISBN: 978-1-948087-81-0. DOI: 10.18653/v1/w18-6319. URL: https://doi.org/10.18653/v1/w18-6319.

[84] Student. "The probable error of a mean". In: *Biometrika* (1908), pp. 1–25.

[85] Sho Takase and Shun Kiyono. "Rethinking Perturbations in Encoder-Decoders for Fast Training". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*. Ed. by Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou. Association for Computational Linguistics, 2021, pp. 5767–5780. ISBN: 978-1-954085-46-6. DOI: 10.18653/v1/2021.naacl-main.460. URL: https://doi.org/10.18653/v1/2021.naacl-main.460.

[86] Sergiu Nisioi, Sanja Stajner, Simone Paolo Ponzetto, and Liviu P. Dinu. "Exploring Neural Text Simplification Models". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*. Ed. by Regina Barzilay and Min-Yen Kan. Association for Computational Linguistics, 2017, pp. 85–91. ISBN: 978-1-945626-76-0. DOI: 10.18653/v1/P17-2014. URL: https://doi.org/10.18653/v1/P17-2014.

[87] Cédric Colas, Olivier Sigaud, and Pierre-Yves Oudeyer. "How Many Random Seeds? Statistical Power Analysis in Deep Reinforcement Learning Experiments". In: *CoRR* abs/1806.08295 (2018). arXiv: 1806.08295. URL: http://arxiv.org/abs/1806.08295.

[88] Cédric Colas, Olivier Sigaud, and Pierre-Yves Oudeyer. "A Hitchhiker's Guide to Statistical Comparisons of Reinforcement Learning Algorithms". In: *Reproducibility in Machine Learning, ICLR 2019 Workshop, New Orleans, Louisiana, United States, May 6, 2019*. OpenReview.net, 2019. URL: https://openreview.net/forum?id=ryx0N3IaIV.

[89] Yoav Benjamini and Daniel Yekutieli. "The control of the false discovery rate in multiple testing under dependency". In: *Annals of statistics* (2001), pp. 1165–1188.

[90] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Routledge, 2013.

[91] William Coster and David Kauchak. "Simple English Wikipedia: A New Text Simplification Task". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, June 2011, pp. 665–669. URL: https://aclanthology.org/P11-2117.

[92] Stephane Champely, Claus Ekstrom, Peter Dalgaard, Jeffrey Gill, Stephan Weibelzahl, Aditya Anandkumar, Clay Ford, Robert Volcic, Helios De Rosario, and Maintainer Helios De Rosario. "Package 'pwr'". In: *R package version* 1.2 (2018).

[93] Skipper Seabold and Josef Perktold. "statsmodels: Econometric and statistical modeling with python". In: *9th Python in Science Conference*. 2010.

[94] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[95] Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. "deep-significance-Easy and Meaningful Statistical Significance Testing in the Age of Neural Networks". In: *arXiv preprint arXiv:2204.06815* (2022).

[96] LREC Organizers. *LREC Call for Papers*. 2022. URL: https://web.archive.org/web/20220624201614/https://lrec2022.lrec-conf.org/en/about/calls-papers/1st-call-papers/ (visited on 06/24/2022).

[97] COLING Organizers. *COLING Call for Papers*. 2022. URL: https://web.archive.org/web/20220624201518/http://coling2022.org/Cpapers (visited on 06/24/2022).

[98] EMNLP Organizers. *EMNLP Call for Papers*. 2022. URL: https://web.archive.org/web/20220128225038/https://2020.emnlp.org/call-for-papers (visited on 06/24/2022).

[99] ACL-IJCNLP Organizers. *ACL-IJCNLP Call for Papers*. 2022. URL: https://web.archive.org/web/20211123051202/https://2021.aclweb.org/calls/papers/#reproducibility-criteria (visited on 06/24/2022).

[100] NAACL Organizers. *NAACL Call for Papers*. 2022. URL: https://web.archive.org/web/20220624202112/https://www.aclweb.org/portal/content/naacl-hlt-2021-first-call-papers (visited on 06/24/2022).

[101] Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, eds. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021. URL: https://aclanthology.org/2021.emnlp-main.0.

[102] Alexander Jones, William Yang Wang, and Kyle Mahowald. "A Massively Multilingual Analysis of Cross-linguality in Shared Embedding Space". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Association for Computational Linguistics, 2021, pp. 5833–5847. DOI: 10.18653/v1/2021.emnlp-main.471. URL: https://doi.org/10.18653/v1/2021.emnlp-main.471.

[103] Dian Yu and Kenji Sagae. “Automatically Exposing Problems with Neural Dialog Models”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Association for Computational Linguistics, 2021, pp. 456–470. DOI: 10.18653/v1/2021.emnlp-main.37. URL: https://doi.org/10.18653/v1/2021.emnlp-main.37.

[104] Jingfeng Yang, Federico Fancellu, Bonnie Webber, and Diyi Yang. “Frustratingly Simple but Surprisingly Strong: Using Language-Independent Features for Zero-shot Cross-lingual Semantic Parsing”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Association for Computational Linguistics, 2021, pp. 5848–5856. DOI: 10.18653/v1/2021.emnlp-main.472. URL: https://doi.org/10.18653/v1/2021.emnlp-main.472.

[105] Lu Zhang, Jiandong Ding, Yi Xu, Yingyao Liu, and Shuigeng Zhou. “Weakly-supervised Text Classification Based on Keyword Graph”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Association for Computational Linguistics, 2021, pp. 2803–2813. DOI: 10.18653/v1/2021.emnlp-main.222. URL: https://doi.org/10.18653/v1/2021.emnlp-main.222.

[106] Xiang Deng, Yu Su, Alyssa Lees, You Wu, Cong Yu, and Huan Sun. “ReasonBERT: Pre-trained to Reason with Distant Supervision”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Association for Computational Linguistics, 2021, pp. 6112–6127. DOI: 10.18653/v1/2021.emnlp-main.494. URL: https://doi.org/10.18653/v1/2021.emnlp-main.494.

[107] Sangwoo Cho, Franck Dernoncourt, Tim Ganter, Trung Bui, Nedim Lipka, Walter Chang, Hailin Jin, Jonathan Brandt, Hassan Foroosh, and Fei Liu. “StreamHover: Livestream Transcript Summarization and Annotation”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Association for Computational Linguistics, 2021, pp. 6457–6474. DOI: 10.18653/v1/2021.emnlp-main.520. URL: https://doi.org/10.18653/v1/2021.emnlp-main.520.

[108] Autumn Toney and Aylin Caliskan. “ValNorm Quantifies Semantics to Reveal Consistent Valence Biases Across Languages and Over Centuries”. In: *Proceedings of the 2021*

Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Association for Computational Linguistics, 2021, pp. 7203–7218. DOI: 10.18653/v1/2021.emnlp-main.574. URL: https://doi.org/10.18653/v1/2021.emnlp-main.574.

[109] Sarah Wiegreffe, Ana Marasovic, and Noah A. Smith. “Measuring Association Between Labels and Free-Text Rationales”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Association for Computational Linguistics, 2021, pp. 10266–10284. DOI: 10.18653/v1/2021.emnlp-main.804. URL: https://doi.org/10.18653/v1/2021.emnlp-main.804.

[110] Chin-Yew Lin. “ROUGE: A Package for Automatic Evaluation of Summaries”. In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.

[111] Daniel Deutsch and Dan Roth. “SacreROUGE: An Open-Source Library for Using and Developing Summarization Evaluation Metrics”. In: *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 120–125. URL: https://www.aclweb.org/anthology/2020.nlposs-1.17.

[112] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”. In: *J. Mach. Learn. Res.* 21 (2020), 140:1–140:67. URL: http://jmlr.org/papers/v21/20-074.html.

[113] NeurIPS. *NeurIPS 2022 Code and Data Submission Guidelines*. 2022. URL: https://web.archive.org/web/20220624192902/https://neurips.cc/Conferences/2022/PaperInformation/CodeSubmissionPolicy (visited on 06/24/2022).

[114] Ranjit Jhala and Isil Dillig, eds. *PLDI '22: 43rd ACM SIGPLAN International Conference on Programming Language Design and Implementation, San Diego, CA, USA, June 13 - 17, 2022*. ACM, 2022.

[115] Karim Ali and Jan Vitek, eds. *36th European Conference on Object-Oriented Programming, ECOOP 2022, June 6-10, 2022, Berlin, Germany*. Vol. 222. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022.

[116] Sepp Hochreiter and Jürgen Schmidhuber. “LSTM can Solve Hard Long Time Lag Problems”. In: *Advances in Neural Information Processing Systems 9, NIPS, Denver, CO, USA, December 2-5, 1996*. Ed. by Michael Mozer, Michael I. Jordan, and Thomas Petsche. MIT Press, 1996, pp. 473–479. URL: http://papers.nips.cc/paper/1215-lstm-can-solve-hard-long-time-lag-problems.

[117] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735. URL: https://doi.org/10.1162/neco.1997.9.8.1735.

[118] Thang Luong, Hieu Pham, and Christopher D. Manning. "Effective Approaches to Attention-based Neural Machine Translation". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*. Ed. by Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton. The Association for Computational Linguistics, 2015, pp. 1412–1421. ISBN: 978-1-941643-32-7. DOI: 10.18653/v1/d15-1166. URL: https://doi.org/10.18653/v1/d15-1166.

[119] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient Estimation of Word Representations in Vector Space". In: *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2013. URL: http://arxiv.org/abs/1301.3781.

[120] Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. "Distributed Representations of Words and Phrases and their Compositionality". In: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*. Ed. by Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger. 2013, pp. 3111–3119. URL: https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.

[121] William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. "Aligning Sentences from Standard Wikipedia to Simple Wikipedia". In: *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*. Ed. by Rada Mihalcea, Joyce Yue Chai, and Anoop Sarkar. The Association for Computational Linguistics, 2015, pp. 211–217. ISBN: 978-1-941643-49-5. DOI: 10.3115/v1/n15-1022. URL: https://doi.org/10.3115/v1/n15-1022.

[122] Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. "Optimizing Statistical Machine Translation for Text Simplification". In: *Trans. Assoc. Comput. Linguistics* 4 (2016), pp. 401–415. URL: https://transacl.org/ojs/index.php/tacl/article/view/741.

[123] Jenny Rose Finkel, Trond Grenager, and Christopher D. Manning. "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling". In: *ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA*. Ed. by Kevin Knight, Hwee

Tou Ng, and Kemal Oflazer. The Association for Computer Linguistics, 2005, pp. 363–370. DOI: 10.3115/1219840.1219885. URL: https://aclanthology.org/P05-1045/.

[124] Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. "Show Your Work: Improved Reporting of Experimental Results". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*. Ed. by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. Association for Computational Linguistics, 2019, pp. 2185–2194. ISBN: 978-1-950737-90-1. DOI: 10.18653/v1/D19-1224. URL: https://doi.org/10.18653/v1/D19-1224.

[125] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. "Green AI". In: *Commun. ACM* 63.12 (2020), pp. 54–63. DOI: 10.1145/3381831. URL: https://doi.org/10.1145/3381831.

[126] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*. ACL, 2002, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: https://aclanthology.org/P02-1040/.

[127] Wiebke Wagner. "Steven Bird, Ewan Klein and Edward Loper: Natural Language Processing with Python, Analyzing Text with the Natural Language Toolkit - O'Reilly Media, Beijing, 2009, ISBN 978-0-596-51649-9". In: *Lang. Resour. Evaluation* 44.4 (2010), pp. 421–424. DOI: 10.1007/s10579-010-9124-x. URL: https://doi.org/10.1007/s10579-010-9124-x.

[128] Michael Cooper and Matthew Shardlow. "CombiNMT: An Exploration into Neural Text Simplification Models". In: *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis. European Language Resources Association, 2020, pp. 5588–5594. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.686/.

[129] Robert J. Gaizauskas. "Karen Sparck Jones and Julia Galliers, *Evaluating Natural Language Processing Systems: An Analysis and Review*. Berlin: Springer-Verlag, 1996. ISBN 3 540 61309 9, Price DM54.00 (paperback), 228 pages". In: *Nat. Lang. Eng.* 4.2 (1998), pp. 175–190. URL: http://journals.cambridge.org/action/displayAbstract?aid=48405.

[130] Anja Belz and Ehud Reiter. "Comparing Automatic and Human Evaluation of NLG Systems". In: *EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 3-7, 2006, Trento, Italy*. Ed. by Diana McCarthy and Shuly Wintner. The Association for Computer Linguistics, 2006. URL: https://aclanthology.org/E06-1040/.

[131] Ehud Reiter and Anja Belz. "An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems". In: *Comput. Linguistics* 35.4 (2009), pp. 529–558. DOI: 10.1162/coli.2009.35.4.35405. URL: https://doi.org/10.1162/coli.2009.35.4.35405.

[132] Natalie Schluter. "The limits of automatic summarisation according to ROUGE". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*. Ed. by Mirella Lapata, Phil Blunsom, and Alexander Koller. Association for Computational Linguistics, 2017, pp. 41–45. DOI: 10.18653/v1/e17-2007. URL: https://doi.org/10.18653/v1/e17-2007.

[133] Jekaterina Novikova, Ondrej Dusek, Amanda Cercas Curry, and Verena Rieser. "Why We Need New Evaluation Metrics for NLG". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*. Ed. by Martha Palmer, Rebecca Hwa, and Sebastian Riedel. Association for Computational Linguistics, 2017, pp. 2241–2252. DOI: 10.18653/v1/d17-1238. URL: https://doi.org/10.18653/v1/d17-1238.

[134] Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. "Best practices for the human evaluation of automatically generated text". In: *Proceedings of the 12th International Conference on Natural Language Generation, INLG 2019, Tokyo, Japan, October 29 - November 1, 2019*. Ed. by Kees van Deemter, Chenghua Lin, and Hiroya Takamura. Association for Computational Linguistics, 2019, pp. 355–368. DOI: 10.18653/v1/W19-8643. URL: https://aclanthology.org/W19-8643/.

[135] David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. "Twenty Years of Confusion in Human Evaluation: NLG Needs Evaluation Sheets and Standardised Definitions". In: *Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020, Dublin, Ireland, December 15-18, 2020*. Ed. by Brian Davis, Yvette Graham, John D. Kelleher, and Yaji Sripada. Association for Computational Linguistics, 2020, pp. 169–182. URL: https://aclanthology.org/2020.inlg-1.23/.

[136] Joseph K. Goodman, Cynthia E. Cryder, and Amar Cheema. "Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples". In: *Journal of Behavioral Decision Making* 26.3 (2013), pp. 213–224. DOI: https://doi.org/10.10

02/bdm.1753. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bdm.1753. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bdm.1753.

[137] Haotian Zhou and Ayelet Fishbach. "The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions." In: *Journal of personality and social psychology* 111.4 (2016), p. 493.

[138] Kimberly A Arditte, Demet Çek, Ashley M Shaw, and Kiara R Timpano. "The importance of assessing clinical phenomena in Mechanical Turk research." In: *Psychological assessment* 28.6 (2016), p. 684.

[139] Janyce Wiebe, Rebecca F. Bruce, and Thomas P. O'Hara. "Development and Use of a Gold-Standard Data Set for Subjectivity Classifications". In: *27th Annual Meeting of the Association for Computational Linguistics, University of Maryland, College Park, Maryland, USA, 20-26 June 1999*. Ed. by Robert Dale and Kenneth Ward Church. ACL, 1999, pp. 246–253. ISBN: 1-55860-609-2. DOI: 10.3115/1034678.1034721. URL: https://aclanthology.org/P99-1032/.

[140] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. "Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks". In: *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*. ACL, 2008, pp. 254–263. URL: https://aclanthology.org/D08-1027/.

[141] James Pustejovsky and Amber Stubbs. *Natural Language Annotation for Machine Learning - a Guide to Corpus-Building for Applications*. O'Reilly, 2012. ISBN: 978-1-449-30666-3. URL: http://www.oreilly.de/catalog/9781449306663/index.html.

[142] Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. "The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*. Ed. by Iryna Gurevych and Yusuke Miyao. Association for Computational Linguistics, 2018, pp. 1383–1392. ISBN: 978-1-948087-32-2. DOI: 10.18653/v1/P18-1128. URL: https://aclanthology.org/P18-1128/.

[143] Anastasia Shimorina and Anya Belz. "The Human Evaluation Datasheet 1.0: A Template for Recording Details of Human Evaluation Experiments in NLP". In: *CoRR* abs/2103.09710 (2021). arXiv: 2103.09710. URL: https://arxiv.org/abs/2103.09710.

[144] Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M. Alonso-Moral, Mohammad Arvan, Jackie Cheung, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, Tanvi Dinkar, Ondřej Dušek, Steffen Eger, Qixiang Fang, Albert Gatt, Dimitra Gkatzia, Javier González-Corbelle, Dirk Hovy, Manuela Hürlimann, Takumi Ito, John D. Kelleher,

Filip Klubicka, Huiyuan Lai, Chris van der Lee, Emiel van Miltenburg, Yiru Li, Saad Mahamood, Margot Mieskes, Malvina Nissim, Natalie Parde, Ondřej Plátek, Verena Rieser, Pablo Mosteiro Romero, Joel Tetreault, Antonio Toral, Xiaojun Wan, Leo Wanner, Lewis Watson, and Diyi Yang. "Missing Information, Unresponsive Authors, Experimental Flaws: The Impossibility of Assessing the Reproducibility of Previous Human Evaluations in NLP". In: *The Fourth Workshop on Insights from Negative Results in NLP*. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 1–10. URL: https://aclanthology.org/2023.insights-1.1.

[145] Open Science Collaboration. "Estimating the reproducibility of psychological science". In: *Science* 349.6251 (2015), aac4716.

[146] Timothy M Errington, Alexandria Denis, Nicole Perfito, Elizabeth Iorns, and Brian A Nosek. "Challenges for assessing replicability in preclinical cancer biology". In: *elife* 10 (2021), e67995.

[147] Timothy M Errington, Maya Mathur, Courtney K Soderberg, Alexandria Denis, Nicole Perfito, Elizabeth Iorns, and Brian A Nosek. "Investigating the replicability of preclinical cancer biology". In: *Elife* 10 (2021), e71601.

[148] Emiel van Miltenburg, Anouck Braggaar, Nadine Braun, Debby Damen, Martijn Goudbeek, Chris van der Lee, Frédéric Tomas, and Emiel Krahmer. "How reproducible is best-worst scaling for human evaluation? A reproduction of 'Data-to-text Generation with Macro Planning"'. In: *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*. Ed. by Anya Belz, Maja Popović, Ehud Reiter, Craig Thomson, and João Sedoc. Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria, Sept. 2023, pp. 75–88. URL: https://aclanthology.org/2023.humeval-1.7.

[149] Lewis N. Watson and Dimitra Gkatzia. "ReproHum #0712-01: Reproducing Human Evaluation of Meaning Preservation in Paraphrase Generation". In: *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*. Ed. by Simone Balloccu, Anya Belz, Rudali Huidrom, Ehud Reiter, Joao Sedoc, and Craig Thomson. Torino, Italia: ELRA and ICCL, May 2024, pp. 221–228. URL: https://aclanthology.org/2024.humeval-1.19.

[150] Ratish Puduppully and Mirella Lapata. "Data-to-text Generation with Macro Planning". In: *Trans. Assoc. Comput. Linguistics* 9 (2021), pp. 510–527. DOI: 10.1162/tacl_a_00381. URL: https://doi.org/10.1162/tacl%5C_a%5C_00381.

[151] Tom Hosking, Hao Tang, and Mirella Lapata. "Hierarchical Sketch Induction for Paraphrase Generation". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Association for Computational Linguistics, 2022, pp. 2489–2501. DOI: 10.18653/V1/2022.ACL-LONG.178. URL: https://doi.org/10.18653/v1/2022.acl-long.178.

[152] Sam Wiseman, Stuart M. Shieber, and Alexander M. Rush. "Challenges in Data-to-Document Generation". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*. Ed. by Martha Palmer, Rebecca Hwa, and Sebastian Riedel. Association for Computational Linguistics, 2017, pp. 2253–2263. DOI: 10.18653/v1/d17-1239. URL: https://doi.org/10.18653/v1/d17-1239.

[153] Ratish Puduppully, Li Dong, and Mirella Lapata. "Data-to-text Generation with Entity Modeling". In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. Ed. by Anna Korhonen, David R. Traum, and Lluís Màrquez. Association for Computational Linguistics, 2019, pp. 2023–2035. DOI: 10.18653/v1/p19-1195. URL: https://doi.org/10.18653/v1/p19-1195.

[154] Clément Rebuffel, Laure Soulier, Geoffrey Scoutheeten, and Patrick Gallinari. "A Hierarchical Model for Data-to-Text Generation". In: *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part I*. Ed. by Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins. Vol. 12035. Lecture Notes in Computer Science. Springer, 2020, pp. 65–80. ISBN: 978-3-030-45438-8. DOI: 10.1007/978-3-030-45439-5_5. URL: https://doi.org/10.1007/978-3-030-45439-5%5C_5.

[155] Jordan J Louviere and George G Woodworth. *Best-worst scaling: A model for the largest difference judgments*. Tech. rep. Working paper, 1991.

[156] Jordan J Louviere, Terry N Flynn, and Anthony Alfred John Marley. *Best-worst scaling: Theory, methods and applications*. Cambridge University Press, 2015.

[157] Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. "Paraphrase-Driven Learning for Open Question Answering". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*. The Association for Computer Linguistics, 2013, pp. 1608–1618. URL: https://aclanthology.org/P13-1158/.

[158] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. "Microsoft COCO: Common Objects in Context". In: *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*. Ed. by David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars. Vol. 8693. Lecture Notes in Computer Science. Springer, 2014, pp. 740–755. DOI: 10.1007/978-3-319-10602-1_48. URL: https://doi.org/10.1007/978-3-319-10602-1%5C_48.

[159] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. "Generating Sentences from a Continuous Space". In: *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*. Ed. by Yoav Goldberg and Stefan Riezler. ACL, 2016, pp. 10–21. DOI: 10.18653/V1/K16-1002. URL: https://doi.org/10.18653/v1/k16-1002.

[160] Yao Fu, Yansong Feng, and John P. Cunningham. "Paraphrase Generation with Latent Bag of Words". In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Ed. by Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett. 2019, pp. 13623–13634. URL: https://proceedings.neurips.cc/paper/2019/hash/5e2b66750529d8ae895ad2591118466f-Abstract.html.

[161] Tom Hosking and Mirella Lapata. "Factorising Meaning and Form for Intent-Preserving Paraphrasing". 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 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*. Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Association for Computational Linguistics, 2021, pp. 1405–1418. DOI: 10.18653/V1/2021.ACL-LONG.112. URL: https://doi.org/10.18653/v1/2021.acl-long.112.

[162] Hong Sun and Ming Zhou. "Joint Learning of a Dual SMT System for Paraphrase Generation". In: *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 2: Short Papers*. The Association for Computer Linguistics, 2012, pp. 38–42. URL: https://aclanthology.org/P12-2008/.

[163] Javier González Corbelle, Jose Alonso, and Alberto Bugarín-Diz. "Some lessons learned reproducing human evaluation of a data-to-text system". In: *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*. Ed. by Anya Belz, Maja Popović, Ehud Reiter, Craig Thomson, and João Sedoc. Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria, Sept. 2023, pp. 49–68. URL: https://aclanthology.org/2023.humeval-1.5.

[164] Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, eds. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. Association for Computational Linguistics, 2021. URL: https://aclanthology.org/volumes/2021.emnlp-main/.

[165] Sheila A. McIlraith and Kilian Q. Weinberger, eds. *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances*

in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018. AAAI Press, 2018. URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/schedConf/presentations.

Appendix

Copyright Permissions for Previously Published Material

In the course of this dissertation, I have utilized materials previously published by the Association for Computational Linguistics (ACL) and the International Speech Communication Association (ISCA). The use of these materials is governed by the respective organizations' Creative Commons licenses and publication policies, which outline the permissions required for teaching, research, and other uses.

ACL Materials

For materials published by the ACL prior to 2016, the content is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 International License. This license permits the use and reproduction of the materials for non-commercial purposes, including teaching and research, without the need for explicit permission from the copyright holder.

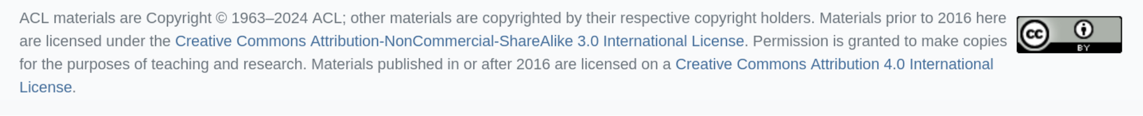
For materials published in or after 2016, the content is licensed under the Creative Commons Attribution 4.0 International License. This license allows for broader use, including commercial purposes, provided appropriate credit is given to the original authors and sources. Under this license, I am permitted to use these materials in my dissertation without obtaining further permission from the copyright holder.

INTERSPEECH Materials

For any article published in the INTERSPEECH proceedings, ISCA grants each author permission to use their article in their dissertation or in institutional repositories (both paper and electronic versions), provided that the article is correctly referenced, including page numbers and/or paper numbers. All authors of a paper have the same permission for reprinting under the same conditions. This permission allows me to include these materials in my dissertation without requiring further approval from the copyright holder.

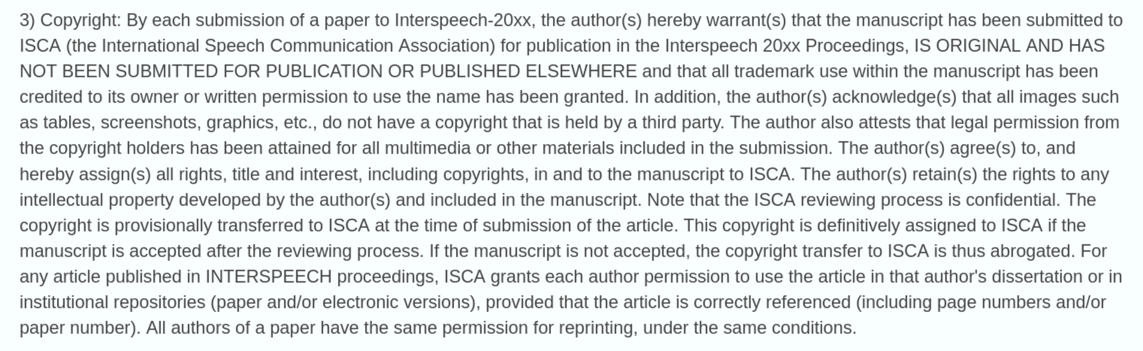
Documentation of Permissions

To verify the applicability of these licenses and permissions, I have included screenshots of the ACL's and ISCA's online statements regarding their copyright policies in the following pages of this appendix. These screenshots serve as proof that no additional permissions were necessary for the use of the previously published materials included in this dissertation.



ACL materials are Copyright © 1963–2024 ACL; other materials are copyrighted by their respective copyright holders. Materials prior to 2016 here are licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 International License. Permission is granted to make copies for the purposes of teaching and research. Materials published in or after 2016 are licensed on a Creative Commons Attribution 4.0 International License.

Figure 5.1: ACL's copyright policy



3) Copyright: By each submission of a paper to Interspeech-20xx, the author(s) hereby warrant(s) that the manuscript has been submitted to ISCA (the International Speech Communication Association) for publication in the Interspeech 20xx Proceedings, IS ORIGINAL AND HAS NOT BEEN SUBMITTED FOR PUBLICATION OR PUBLISHED ELSEWHERE and that all trademark use within the manuscript has been credited to its owner or written permission to use the name has been granted. In addition, the author(s) acknowledge(s) that all images such as tables, screenshots, graphics, etc., do not have a copyright that is held by a third party. The author also attests that legal permission from the copyright holders has been attained for all multimedia or other materials included in the submission. The author(s) agree(s) to, and hereby assign(s) all rights, title and interest, including copyrights, in and to the manuscript to ISCA. The author(s) retain(s) the rights to any intellectual property developed by the author(s) and included in the manuscript. Note that the ISCA reviewing process is confidential. The copyright is provisionally transferred to ISCA at the time of submission of the article. This copyright is definitively assigned to ISCA if the manuscript is accepted after the reviewing process. If the manuscript is not accepted, the copyright transfer to ISCA is thus abrogated. For any article published in INTERSPEECH proceedings, ISCA grants each author permission to use the article in that author's dissertation or in institutional repositories (paper and/or electronic versions), provided that the article is correctly referenced (including page numbers and/or paper number). All authors of a paper have the same permission for reprinting, under the same conditions.

Figure 5.2: ISCA's copyright policy

VITA

Mohammad Arvan

EDUCATION

Ph.D., Computer Science, University of Illinois at Chicago, Chicago, Illinois, 2024.
B.Sc., Software Engineering, Qazvin Islamic Azad University, Qazvin, Iran, 2016.

EXPERIENCE

Research Assistant, UIC Natural Language Processing Laboratory, Department of Computer Science, University of Illinois at Chicago, 2018 - 2024.
Research Assistant, Mechatronic Research Laboratory, Qazvin Islamic Azad University, 2013 - 2015

HONORS

Finalist in the 2024 Three Minute Thesis (3MT) Competition at the University of Illinois at Chicago.
Recipient of the 2020 Provost's Graduate Research Award, University of Illinois at Chicago ($5000).
Innovative User Interface Award, RoboCup World Championship, Rescue Robot League, Hefei, China, 2015.
Ranked 1st, RoboCup World Championship, Rescue Robot League, Hefei, China, 2015.