Researching Responsible and Trustworthy Natural Language Processing

Replication, Reproducibility, and Generalisation

Emily Allaway

20 October 2025

School of Informatics
University of Edinburgh
eallaway@ed.ac.uk

Overview

Replication and Reproducibility

Generalisation

Overview

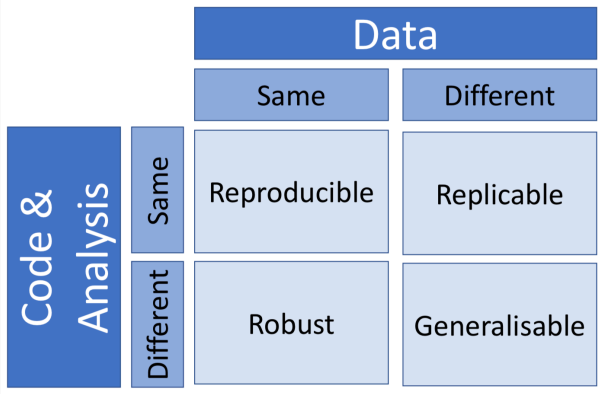
Replication and Reproducibility

Generalisation

What do we expect of scientific results?

Statistically significant and relevant
Hypothesis driven
Accessibility-wise understandable
Falsifiable
Clearly scoped
Good sense of real performance (median, mean)
Reproducible
Grounded in existing literature
Clarity
Lack of quotations in science
Defining terms
Methodologically sound - awareness of potential issues
Overly anthropomorphic

Terms



From: Pineau et al. (2021) and https://github.com/WhitakerLab/ReproducibleResearch

Issue in NLP

Table 1
Distribution of data and code availability in both 2011 and 2016.

	2011: data		2016: data		2011: code		2016: code	
Data / code available	**116**	**75.8%**	**196**	**86.3%**	**48**	**33.1%**	**131**	**59.3%**
- working link in paper	98	64.1%	179	78.9%	27	18.6%	80	36.2%
- link sent	11	7.2%	15	6.6%	17	11.7%	50	22.6%
- repaired link sent	7	4.6%	2	0.9%	4	2.8%	1	0.5%
Data / code unavailable	**37**	**24.2%**	**31**	**13.7%**	**97**	**66.9%**	**90**	**40.7%**
- sharing impossible	19	12.4%	14	6.2%	46	31.7%	42	19.0%
- no reply	17	11.1%	12	5.3%	43	29.7%	32	14.5%
- good intentions	0	0.0%	2	0.9%	5	3.4%	12	5.4%
- link down	1	0.7%	3	1.3%	3	2.0%	4	1.8%

From Wieling et al. (2018)

Efforts to improve reproducibility

The ARR Responsible NLP Research checklist, based on:

✓ **For all reported experimental results**

- ☐ Description of computing infrastructure
- ☐ Average runtime for each approach
- ☐ Details of train/validation/test splits
- ☐ Corresponding validation performance for each reported test result
- ☐ A link to implemented code

✓ **For experiments with hyperparameter search**

- ☐ Bounds for each hyperparameter
- ☐ Hyperparameter configurations for best-performing models
- ☐ Number of hyperparameter search trials
- ☐ The method of choosing hyperparameter values (e.g., uniform sampling, manual tuning, etc.) and the criterion used to select among them (e.g., accuracy)
- ☐ Expected validation performance, as introduced in §3.1, or another measure of the mean and variance as a function of the number of hyperparameter trials.

Dodge et al. (2019)

3. ☐ **Safe use** of data is ensured. *(Check all that apply)*

3.1. ☐ The data does not include any protected information (e.g. sexual orientation or political views under GDPR), or a specified exception applies. *See Section ____*

3.2. ☐ The paper is accompanied by a data statement describing the basic demographic and geographic characteristics of the population that is the source of the language data, and the population that it is intended to represent. *See ____*

3.3. ☐ If applicable: the paper describes whether any characteristics of the human subjects were self-reported (preferably) or inferred (in what way), justifying the methodology and choice of description categories. *See Section ____*

3.4. ☐ The paper discusses the harms that may ensue from the limitations of the data collection methodology, especially concerning marginalized/vulnerable populations, and specifies the scope within which the data can be used safely. *See Section ____*

3.5. ☐ If any personal data is used: the paper specifies the standards applied for its storage and processing, and any anonymization efforts. *See Section ____*

3.6. ☐ If the individual speakers remain identifiable via search: the paper discusses possible harms from misuse of this data, and their mitigation. *See Section ____*

Rogers et al. (2021)

Limitations & Risks

Limitations (week 3), Risks (now)

- Examples
  - potential malicious or unintended harmful effects & uses
  - environmental impact
  - fairness
  - privacy
  - security
- Consider particularly
  - Dual use
  - Variety of stakeholders impacted
  - Relevant mitigation strategies

From A2: https://aclrollingreview.org/responsibleNLPresearch/

Exercise - why are items on the checklist

For each item on the checklist, discuss

- Why is the information useful?
- What area of reproducible research does it contribute to?

B - Scientific Artifacts

B1. Cite creators? credit and accountability

B2. Licenses and terms of use? legal protections in data creation to be aware of

B3. Intended use?
clear whether it can be used for commerical purposes, especially if sensitive data

B4. Uniquely identifying information or offensive content?
protections for original data creators or legal protections for protecting

B5. Documentation of artifacts?
datasets/benchmarks - scoping the results and sample population, also for replication

B6. Statistics of artifacts?
grounding findings in numbers, clear about what you specifically did (replication and good practice)

C - Compute Experiments

C1. Parameters and compute information?

context for readers and clear about scale needed to replicate

C2. Hyperparameters? know if people optimise for the test set (for example the random seed), extensive search is computationally expensive maybe, know what you can expect from the results, transparency, need for reproduction

C3. Descriptive statistics of results?

know what you can expect from the results, what is being described (is it mean, median) and not just cherry picking

C4. Packages and settings?

different implementations can lead to different results, replication

D - Human Annotations & Participants

D1. Instructions to participants?

reliability and generalisability, how to reproduce across populations, ethical check

D2. Recruitment information?

recruitment influences the results because may limit the demographic applicability,
compensation - make sure to pay well (ethics), whether pay is tied to performance

D3. Consent?

reliability and generalisability,
how to reproduce across populations - going through a consenting process impacts the population,
make sure to get assent even if can't consent,
ethical check,

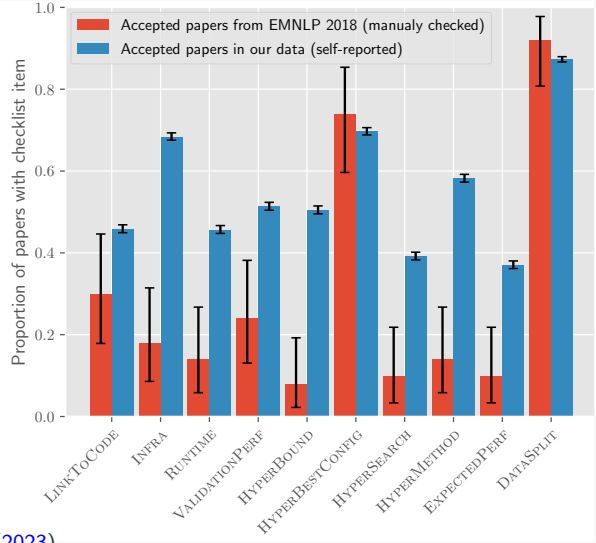
D4. Ethics approval?

reliability and generalisability, ethics check clearly stated,
whether research will have the intended impacts

D5. Annotator population characteristics?

reliability and generalisability, how to reproduce across populations, ethical check,
should be collecting the minimal necessary data and the amount should be justified

Is the checklist effective?



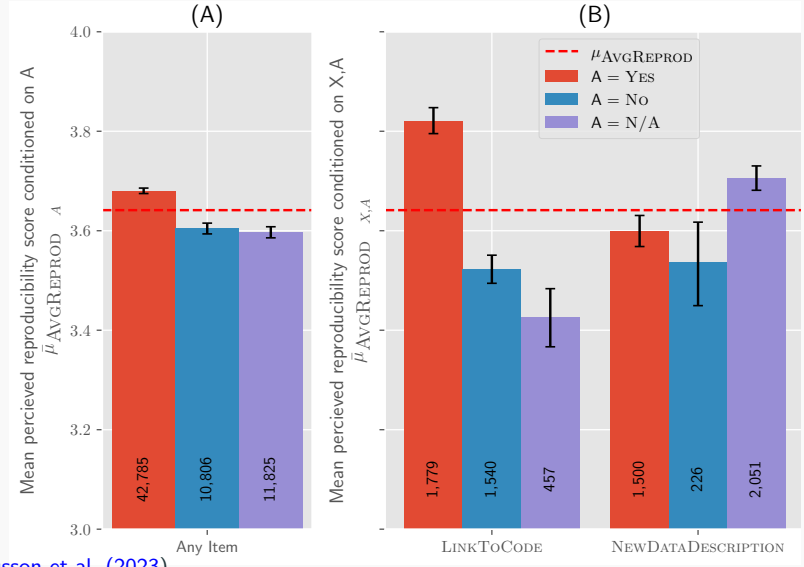
From Magnusson et al. (2023)

Potentially bad faith responses

Response	Conference	Submissions	ACCEPT
YES	EMNLP 2020	134 (4.5%)	−9.9%
	EMNLP 2021	238 (7.3%)	−6.7%
	NAACL 2021	79 (6.4%)	−3.3%
	ACL 2021	213 (7.3%)	−8.2%
NO	EMNLP 2020	1 (0.0%)	−39.7%
	EMNLP 2021	0 (0.0%)	-
	NAACL 2021	0 (0.0%)	-
	ACL 2021	0 (0.0%)	-

From: Magnusson et al. (2023)

Perceived reproducibility from the checklist



From: Magnusson et al. (2023)

Work still difficult to reproduce

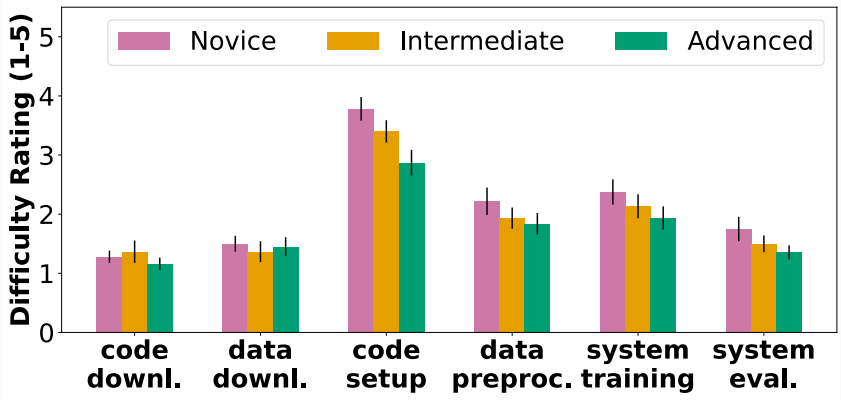


Figure 2: Mean reproducibility difficulty rating (1-5 5 being most difficult) for each step of experiments

From Storks et al. (2023)

Code is a major blockers to reproduction

Reproducibility Blocker	Frequency
Insufficient Code Dependency Specification	38
Difficult-to-Access External Resources	27
Unclear Code Usage Documentation	17
Pre-Existing Bugs in Code	16
Difficult-to-Read Code	11
Other	30

From Storks et al. (2023)

Challenges doing reproducible research

- *Takes time!*

- Held to higher standards
- Openness to mistakes
- Publication bias towards novel findings
- IP/confidentiality issues

From Data et al. (2017)

Exercise - recommendations

Magnusson et al. (2023) and Storks et al. (2023) both make recommendations for the checklist:

1. Checklist responses made public
2. Extra time allowed for submitting the checklist & accompanying items

Suggested ACLRC Addition	Frequency
Standards for Documentation Clarity	22
Full Specification of Code Dependencies	18
Demonstration of Code Usage	9
Provision of Support for Issues	8
Standards for Code Clarity	5

Should these be implemented?

Further reading

- Reproducibility, correctness, and buildability: The three principles for ethical public dissemination of computer science and engineering research (Rozier and Rozier, 2014)
- Three Dimensions of Reproducibility in Natural Language Processing (Cohen et al., 2018)
- A practical taxonomy of reproducibility for machine learning research (Tatman et al., 2018)