NLP Reproducibility For All: Understanding Experiences of Beginners

Shane Storks, Keunwoo Peter Yu, Ziqiao Ma, & Joyce Chai
Computer Science and Engineering Division, University of Michigan

Introduction & Data Collection

As NLP has recently seen an unprecedented level of excitement, and more people are eager to enter the field, it is unclear whether current research reproducibility efforts are sufficient for this group of **beginners** to apply the latest developments, and what key factors impact their experience doing so.

We run a user study with 93 beginners from an introductory NLP course, where students each reproduced results from 1 of 3 recent reproducible ACL conference papers. This included several steps:

1. **Pre-survey on student skill level**: collected data on students' programming background and understanding of coursework, which was used to divide them into 3 skill levels: *novice*, *intermediate*, and *advanced*.
2. **Paper result reproduction**: students reproduced results, tracking their time spent on setting up and running the code associated with their assigned paper.
3. **Post-survey on student experience**: students shared their reproduced results, and answered questions about their assigned paper and experience reproducing its results.

Expert reproduction time by paper:

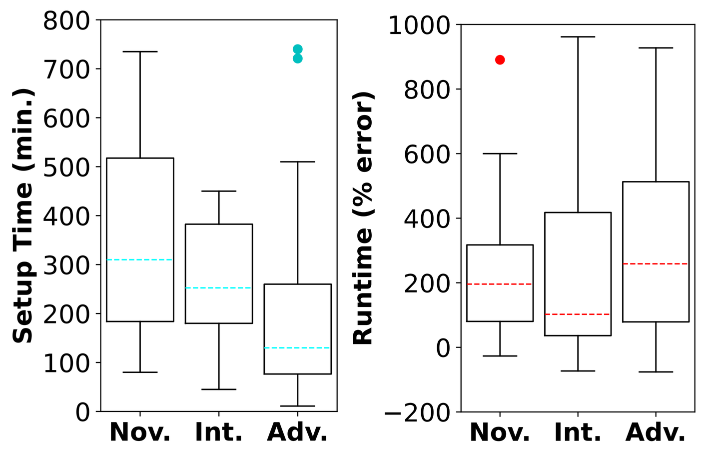
Paper	Reference	Setup	Runtime
A	[1]	2 hrs.	0.5 hr.
B	[2]	2 hrs.	3 hrs.
C	[3]	2 hrs.	2 hrs.

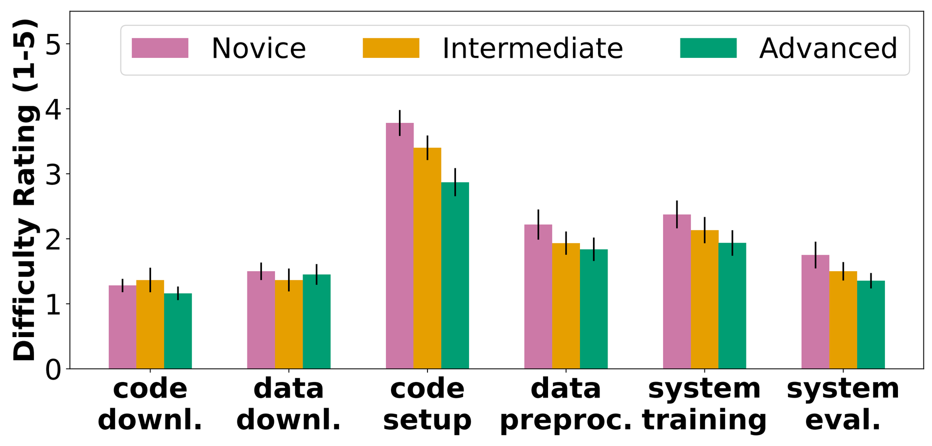
Paper assignments by skill level:

Paper	Nov.	Int.	Adv.	Total
A	12	11	11	34
B	10	10	10	30
C	10	9	10	29

Role of Skill Level

First, we examine the impact of student **skill level** on their *experience*, i.e., their reported time spent and difficulty in reproducing experimental results.





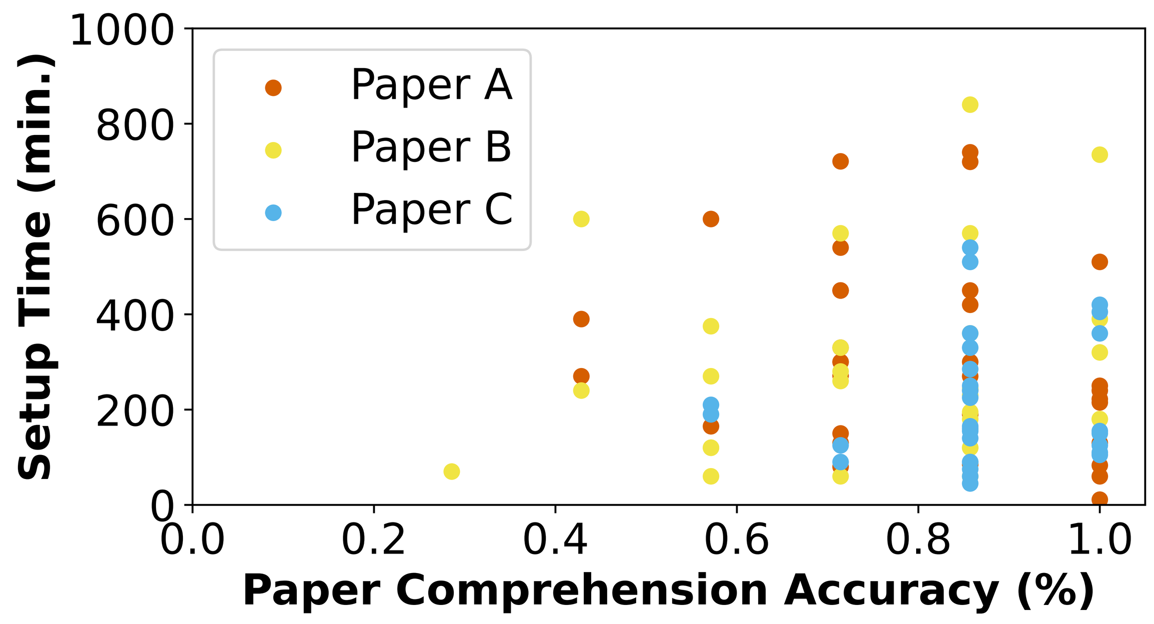
We find significant Spearman correlations between skill level factors and experience factors, but they **only explain up to ρ^2=18.5% of variance**.

Skill Level Factor	ρ (time)	ρ (diff.)
Python Experience (Years)	-0.291	-0.230
PyTorch Experience (Years)	-0.251	-0.259
LSTM Understanding (1-5)	**-0.430**	**-0.396**
Transformer Understanding (1-5)	-0.317	-0.338

Role of Paper Comprehension

We characterized students' **comprehension** of the work by measuring their accuracy on standard multiple-choice questions about their assigned paper's *motivation*, *problem definition*, *approaches*, *implementation*, *results*, and *conclusion*.

We find **no correlation between paper comprehension and code setup time or difficulty**. Beyond writing a strong, well-understood paper, **effectively open-sourcing code is a separate and important issue for reproducibility**.



Role of Reproducibility Efforts

We examine the relationship between **reproducibility efforts** made for each paper and students' experience. Students identified which items of the ACL Reproducibility Checklist (ACLRC, inspired by [4]) were most important in reproducing the results of their assigned paper. We ran a multiple linear regression for how well their choices predicted students' setup time and runtime, and an ordinal logistic regression for how they predicted reported setup difficulty.

Paper	Top ACLRC Item, Setup Time	β	R^2
A	10. Best Hyperparameters	**4.24**	0.53
B	1. Model Description	**8.47**	0.15
C	14. Dataset Partition Info	**4.08**	0.62
All	1. Model Description	1.89	0.40

Paper	Top ACLRC Item, Runtime	β	R^2
A	9. Hyperparameter Bounds	**46.43**	0.17
B	11. Model Selection Strategy	**-13.20**	0.66
C	6. Val. Set Metrics	-3.26	-0.04
All	9. Hyperparameter Bounds	6.61	0.07

Paper	Top ACLRC Item, Setup Difficulty	β
A	10. Best Hyperparameters	**1.82**
B	11. Model Selection Strategy	**4.26**
C	5. Model Complexity Info	**-4.40**
All	15. Data Preprocessing Info	**0.65**

We found these reproducibility efforts correlated more strongly with setup time, runtime, and setup difficulty, explaining up to R^2=66% of these experience factors. Lastly, we surveyed students on what helped and blocked them in reproducing results, and their suggested additions to the ACLRC:

Reproducibility Helper	Frequency
Clear Code Usage Documentation	56
Example Scripts and Commands	27
Easy-to-Read Code	15
Easy-to-Access External Resources	13
Sufficient Code Dependency Specification	12
Other	11

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

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
Other	23
Already Included	23

Student comments commonly identified **code usage documentation**, **code clarity and functionality**, **availability of external resources**, and specification of **code dependencies** in their feedback, suggesting these aspects are most important for beginners to reproduce NLP results. As such, we recommend that researchers in NLP (and perhaps neighboring disciplines) take extra care toward these efforts when releasing experiment code and data.

Acknowledgements & References

We thank the authors of [1, 2, 3] for making our study possible by sharing reproducible NLP research. We also thank Advanced Research Computing (ARC) at University of Michigan for providing computational resources and services.

[1] Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. Temporal reasoning on implicit events from distant supervision. In *NAACL: HLT 2021*, pages 1361–1371, Online, 2021. Association for Computational Linguistics.

[2] Lucia Donatelli, Theresa Schmidt, Debanjali Biswas, Arne Köhn, Fangzhou Zhai, and Alexander Koller. Aligning actions across recipe graphs. In *EMNLP 2021*, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics.

[3] Vivek Gupta, Maitrey Mehta, Pegah Nokhiz, and Vivek Srikumar. INFOTABS: Inference on tables as semi-structured data. In *ACL 2020*, Online, 2020. Association for Computational Linguistics.

[4] Joelle Pineau. The machine learning reproducibility checklist. https://www.cs.mcgill.ca/ jpineau/ReproducibilityChecklist.pdf, 2020.

Links

Pre-Print

Code