Evolving Stances on Reproducibility: A Longitudinal Study of NLP and ML Researchers' Views and Experience of Reproducibility

Craig Thomson[1,2], Ehud Reiter[2], João Sedoc[3], Anya Belz[1]
[1]Dublin City University / ADAPT, Ireland
[2]University of Aberdeen, UK
[3]New York University, USA
Correspondence: craig.thomson@dcu.ie, anya.belz@dcu.ie

Abstract

Like other fields of science, NLP/ML has seen growing interest in, and work on, reproducibility and methods for improving it over the past 10 years. Identical experiments producing different results can be due to variation between samples of evaluation items or evaluators, but it can also be due to poor experimental practice. Both can be mitigated by bringing multiple comparable studies together in systematic reviews that can draw conclusions beyond the level of the individual studies, but such systematic reviews barely exist in NLP/ML. The alternative is to focus on improving experimental practice and study-level reproducibility, and the first step in this direction is awareness of the importance of reproducibility and knowledge of how to improve it. In the work reported in this paper, we aim to assess (i) what NLP/ML practitioners' current views and experience of reproducibility are, and (ii) to what extent they have changed over the past two years, a period of rapidly growing interest in reproducibility. We report, for the first time, results from two identical surveys, the first carried out in 2022 and the second in 2024, each time surveying 149 NLP and ML researchers. We report the results from the 2024 survey to address *i* above. We then compare the results of the two surveys in order to address *ii* above. We find that views and experience overall are moving towards better practice and appreciation of reproducibility.

1 Introduction

Poor reproducibility of experimental results is a point of concern not just in NLP/ML (Pedersen, 2008; Wieling et al., 2018; Belz et al., 2021), but also in science more generally (Baker, 2016). In NLP/ML, reproducibility concerns have led to a flurry of activity aimed at diagnosing the problem (Belz et al., 2020; Arvan et al., 2022; Belz et al., 2022; Belz, 2022; Belz et al., 2023; Storks et al., 2023; Kapoor and Narayanan, 2023; Xue et al., 2023; Thomson et al., 2024; Semmelrock et al., 2025), encouraging researchers to share full details and resources of their work (Pineau et al., 2021; Shimorina and Belz, 2022; Belz and Thomson, 2024b; Ruan et al., 2024), and understanding how differences between evaluation methods impact reproducibility (van Miltenburg et al., 2024; Belz et al., 2025b; Popp et al., 2025). Some conferences have introduced checklists and/or guidelines for reproducibility, e.g. ACL (an NLP venue) and AAAI (a more general ML venue) use reproducibility checklists.[1,2] ICLR encourages authors to include a reproducibility statement, wheras ICML encourages authors to "submit code to foster reproducibility." See Appendix B for more details.

Some work has looked at whether previous interventions have had any effect on poor levels of reproducibility. Raff and Farris (2023) describe existing interventions as a 'siren song,' pointing out that whilst many conferences have appointed reproducibility chairs, they rely mainly on self-reporting of code inclusion, clearly not the same thing as reproducibility. Magnusson et al. (2023) perform an analysis of the *CL reproducibility checklist data, albeit without answering the question of whether it is effective at improving reproducibility. For interventions like checklists to be effective, they need to result in changed understanding and practices relating to reproducibility, beyond a small increase in reviewer scores for reproducibility. What we currently do not know is whether the checked boxes and reviewer scores actually lead to improved reproducibility and understanding of related issues.

The work presented here, in tandem with other research on reproducibility of NLP evaluations Belz and Thomson (2023, 2024a); Belz et al. (2025b) is intended to address this question. We conducted two identical surveys of researchers'

[1]https://aclrollingreview.org/responsibleNLPresearch
[2]https://aaai.org/conference/aaai/aaai-25/aaai-25-reproducibility-checklist

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 25738–25760
November 4-9, 2025

views and experience of reproducibility two years apart, in 2022 and 2024, to determine (i) what current stances on reproducibility are, and (ii) how they have changed over the past two years, a period of increasing research focus on reproduciblity. We start below with a summary of related research (Section 2). We describe our study design (Section 3), and present the results of the 2024 survey (Section 4) and an analysis of the differences between the 2024 survey and an identical survey conducted in 2022 (Sections 5 and 6). We finish with some discussion and conclusions in Section 7.

2 Related Research

Sporadic surveys have investigated reproducibility for science in general (Baker, 2016), and NLP (Mieskes et al., 2019) in particular. Baker (2016) asked whether respondents thought there was a reproducibility crisis, as well as how severe they felt it was, both generally and within their field. They also asked about attempts to repeat experiments and reproduce results, respondents' experience publishing such work, and contributing factors and barriers that lead to poor reproduction practices. Baker (2016) surveyed scientists in all disciplines, finding that 70% of researchers had attempted to reproduce the work of others, and failed. Biologists made up 45% of responses, with only around 1.5% working broadly in computer science, including one participant in AI (none in NLP). It is therefore not representative of NLP and the issues it faces.

Mieskes et al. (2019) asked about respondents' experiences when trying to repeat their own work, and that of others. Questions on where respondents accessed code, data, and parameters were also included, as was a question on the outcome of attempts at contacting original authors. This survey focuses on whether similar results were obtained, where researchers were able to access resources (the authors website, GitHub, etc., and what respondents' experiences were when contacting authors. Respondents thought the issues were important, although the majority obtained different results when reproducing the work of others.

The survey conducted by van Miltenburg et al. (2023) looked at the barriers encountered when attempting to perform an error analysis of previously published results. Whilst not directly investigating reproducibility, it similarly found that whilst researchers felt error analysis to be important, there are barriers such as time, money, publication page limits, and peer-reviewer interest.

3 Overview of Study Design

We designed a survey of NLP/ML researchers' views and experience of reproducibility and conducted it once in 2022, near the start of the ReproHum Project,[3] and again in 2024, after the project concluded. For the 2024 survey, we collected responses from the same number of respondents with the same characteristics (see below) as in 2022.

Following survey question design principles (described in Appendix C), we created our survey of 25 multiple-choice/select-one questions, and 6 multiple-choice/select-all-that-apply questions, to take about 10 minutes on average. The survey flow structure is shown in Appendix Figure 9.

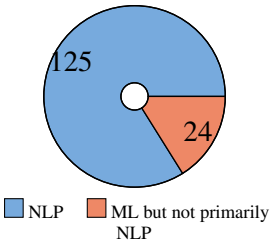
Respondents were first shown a landing page and introduction (Appendix, Figure 10). The survey itself was structured into seven sections (as reflected both in the flow chart and in Ssections 4.1–4.4 which present the 2024 results):

1. Participant characteristics (Section 4.1): Questions 1–7 (Q1–Q7) collect demographic information, e.g. field of work, career stage and sector.
2. Recreating research (Section 4.2):
   (a) Other people's research: Q8–Q10 are about respondents' experience recreating other people's research.
   (b) Recreating own research: Q11–Q13 are about respondents' experience recreating their own research.
3. Recreating human evaluations (Section 4.3): Q11–Q17 are about recreating human evaluations specifically.
4. Barriers and self-reflection (Section 4.4):
   (a) Q19–Q21 cover aspects that might stop researchers from addressing reproducibility in their work.
   (b) Q22–Q27 are about things researchers do in their work that makes recreation/reproduction easier.
   (c) Q28–Q31 address some remaining views and practice.

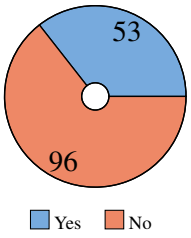
We ensured that the proportions of high-level participant characteristics in the 2024 survey were the same as in the 2022 instance: 125 NLP researchers (45 students, 80 non-students); and 24 ML researchers (8 students, 16 non-students).

We first present results from the 2024 survey (Section 4), then analyse the change from 2022 to

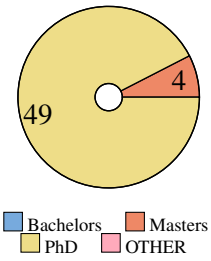
[3] https://reprohum.github.io



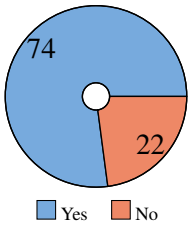
(a) Question 01: How would you describe your broad field of research?



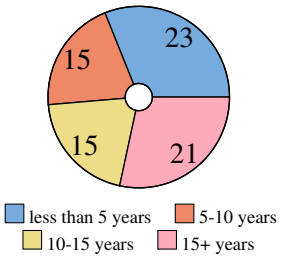
(b) Question 02: Are you currently a student?



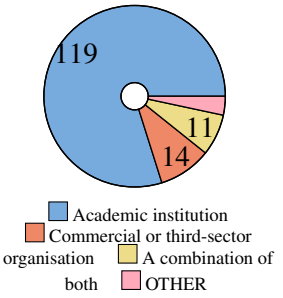
(c) Question 03: Towards which degree are you currently studying?



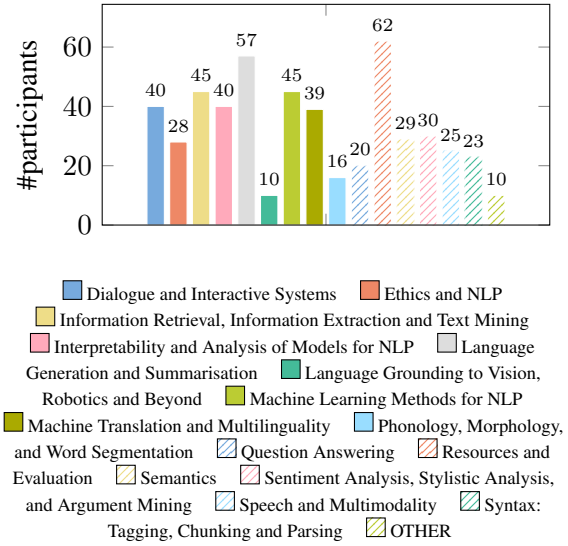
(d) Question 04: Do you have a PhD?



(e) Question 05: How many years ago did you obtain your PhD?



(f) Question 06: Which of the following best describes your current primary affiliation(s)?



(g) Question 07: Which of the following NLP subfields have you worked in in the past 4 years (select all that apply)?

Figure 1: 2024 Results: Answers to Q1–Q7 about demographics.

2024 in terms of (i) percentage increase/decrease in response values (Section 5), and (ii) significant changes in responses when recoded for three specific aspects (Section 6).

3.1 Statistical methods

Our repeated cross-sectional longitudinal study of reproducibility in NLP/ML utilises straightforward statistical methods. Participants are deidentified and thus we cannot study individual-level longitudinal change. Instead, we make a weaker independence assumption. It is unlikely that the overlap between survey participants is large.

For ordinal survey items, we use the Mann–Whitney U test to assess statistical significance. While a t-test could also be applied, prior work suggests it would yield similar conclusions in practice (de Winter and Dodou, 2010). For categorical data, we apply Chi-squared tests, justified by our sufficiently large sample size (Agresti, 2018).

As shown in the survey flow chart (Figure 9), the survey includes conditional branching, where some follow-up questions are shown only to certain respondents. However, the branching proportions remain stable across the two time points, allowing us to simplify the analysis by disregarding the branching structure while normalising for sample size. While a bootstrap-based conditional analysis could be applied for greater rigour, the consistently high rate of *Yes* responses suggests that such complexity would not alter our conclusions. For multi-item scales, we follow standard aggregation practices: items are typically summed, unless evidence suggests they reflect non-ordinal artifacts, in which case we treat a single positive response as sufficient evidence of the construct.

4 2024 Survey Results

This section reports the results of the 2024 survey, reflecting current views in NLP/ML. We present results for multiple-choice/select-one questions as pie charts unless the possible answers form an ordinal range in which case they are presented in hor-

izontal bar charts. If the responses are in the form (Yes/No/Unsure) they are presented in vertical bar charts. Results from multiple-choice/select-all-that-apply questions are presented as vertical bar charts. All figures include the verbatim question(s) in the caption, and the set(s) of possible answers in the chart legend. To make charts accessible, we use the light colour scheme by Paul Tol.[4]

4.1 Participant characteristics (2024)

In this section, we briefly discuss the results from the participant characteristics questions Q1–Q7 (Section 4.1) at the start of the survey. Briefly, about 5/6 respondents worked in NLP/ML, with 1/6 working in ML (but not primarily on NLP). Of the 149 respondents, 53 were students, all except 4 of whom were PhD students. Of the 96 non-students, 74 had a PhD, obtained in roughly equal numbers 0–5, 5–10, 10–15, and 15+ years ago, with a few more from the first and last time spans.

Most respondents (80%) were from academic institutions, 9% from industry or third sector, 9% had both affiliation types. Free text responses for the *Other* option were either public research institutions (3), or no affiliation (2). Respondents tended to work in different subfields of NLP/ML, with a high number selecting resources and evaluation. This may not be representative of the field as a whole; this may reflect a degree of self-selection of participants with experience in this subfield.

4.2 Recreating research (2024)

In this section, we look at responses to Q8–Q13 (Figures 2 and 3) which ask participants about their general experience recreating their own, and other researchers' experiments.

From Q8 (Figure 2a) we can see that just under 5/6 of respondents had attempted a **recreation of others' research** in the past 5 years. Of those, 42% obtained broadly the same results *always* or *frequently* (Q10, Figure 2c). From Q9 (Figure 2b) we see that difficulties obtaining and running code, protocols, models and other resources were encountered by almost all researchers who had tried a recreation; compute, communication with original authors and results analysis posed problems for about half, with financial cost not much of an issue. The issues identified under *Other* were differences between code and paper, as well as changes in crowdsourcing site participant pools.

From Q11 (Figure 3a) we can see that over half of respondents had attempted **recreating their own work** (fewer than for other people's work), and from Q13 (3c) that in all but 6 cases they obtained broadly the same results *always* or *frequently* (a far larger proportion than for others' research).

Fewer issues were encountered than when recreating the work of others (Q12, Figure 3b). With (obviously) no barrier in contacting the author of the research, respondents found it much easier to obtain code, models, etc. However, difficulty running code, protocols or models remained the biggest issue. Seven of the eight *Other* responses were to do with the passage of time, resulting in changes in APIs and infrastructure, as well as staff turnover. Some respondents also mentioned difficulties remembering how their code worked.

4.3 Recreating human evaluations (2024)

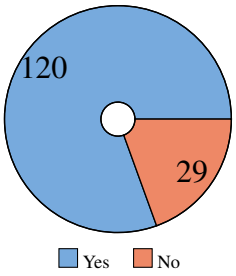
Participants were first asked if they had performed a human evaluation in the past 5 years (Q14). From Figure 4a we can see that 72% had, and of those, two thirds had attempted to recreate one (Q15, Figure 4b). Of the third who had *not*, a lack of time and funding, as well as difficulty recruiting participants were among the top reasons both for not performing a human evaluation (Q16, Figure 4d), and not recreating one (Q17, Figure 4e).

Comparing Q16 and Q17, the top two ranked issues are the same: time and money. Difficulty recruiting participants was also ranked highly. The biggest differences were that no respondents who had recreated human evaluations (Q17) felt that human evaluations were unnecessary (joint top in Q16), and, proportionately, more than three times as many respondents to Q17 indicated difficulty in publishing recreations of human evaluations than those who indicated it was a barrier to performing them in the first place (respondents to Q16).[5]

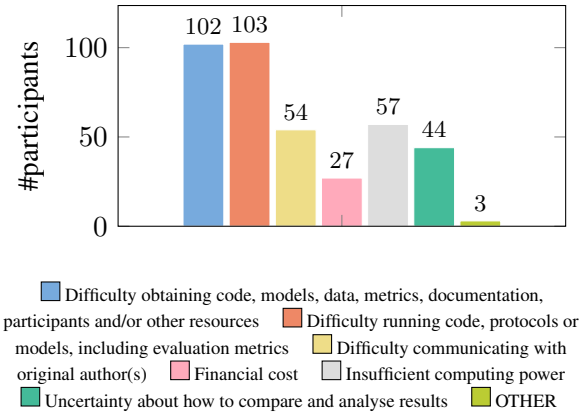
Respondents who had recreated a human evaluation reported difficulty obtaining design details as the most common issue (over two thirds), followed by some experimental conditions being impossible to recreate (about three fifths). For Q18 (Figure 4c), about half of respondents reported difficulty recruiting participants, and a similar number that tools and other resources were no longer available. Responses under *Other* were difficulty assessing degree of reproducibility, and poor memory

[4] https://cran.r-project.org/web/packages/khroma/vignettes/tol.html. Original URL unavailable.

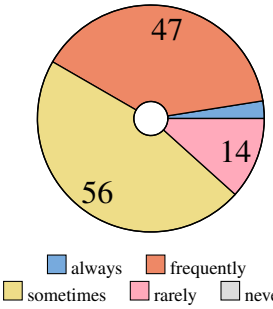
[5] Whilst the raw numbers indicate 5–6 times as many responses, we need to account for Q16 only being shown to 41 respondents, whilst Q17 was shown to 72.



(a) Question 08: Have you in the past 5 years tried to recreate any aspect of other people's research?

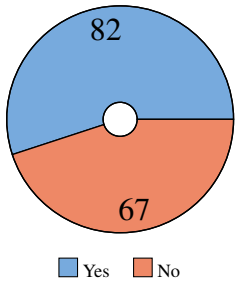


(b) Question 09: Which of the following issues did you encounter when trying to recreate other people's research (select all that apply)?

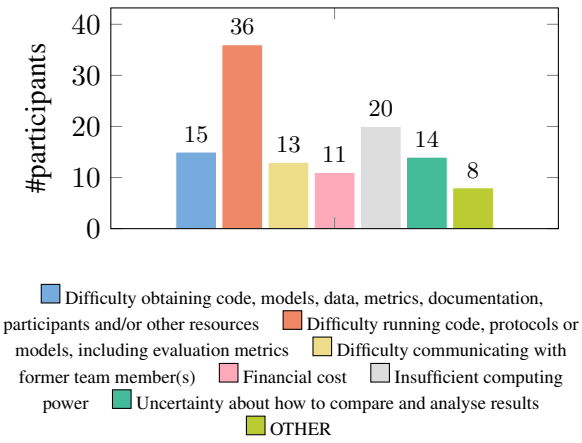


(c) Question 10: Thinking about results when you recreated other people's research, were your results broadly the same as those reported in the original work?

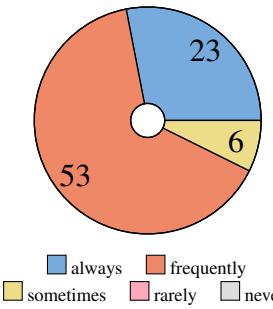
Figure 2: 2024 Results: Answers to Q8–Q10 about recreating the work of others.



(a) Question 11: Have you in the past 5 years tried to recreate any aspect of your own research?



(b) Question 12: Which of the following issues did you encounter when trying to recreating your own research (select all that apply)?



(c) Question 13: Thinking about results when you recreated your own research, were your results broadly the same as your earlier results?

Figure 3: 2024 Results: Answers to Q11–Q13 about recreating own work.

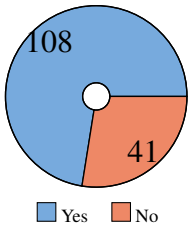
(suggesting a lack of documentation).

4.4 Barriers and self-reflection (2024)

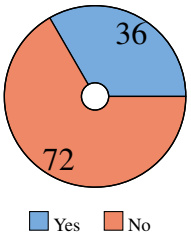
In Q19–21 (Figure 5), participants were asked about **barriers in performing reproduction studies** that participants faced. From Q19, 64% of participants agreed (either somewhat or strongly) they would submit more reproduction studies if some conference acceptance rate was reserved for them. 60% of participants agreed they would submit more such applications if funding was reserved for reproduction studies (Q20). 62% agreed that they would perform more reproduction work if a toolkit was available (Q21).

Next, we asked about **participants' views of their own efforts toward recreatable research** (Figure 6). 57% agreed *strongly* or *somewhat* that others could easily reproduce their work (Q22). From Q23, we can see that 81% of participants believe they always/frequently take steps to ensure their published work can be reproduced. For reporting code and data, 77% always/frequently make model code available (Q25), and 76% the data (Q26). Lastly, 66% of participants reported always/frequently performing a human evaluation alongside metric scores (Q27). Looking at Q25 in detail, we see that more junior researchers are the more likely they are to share their code: 86% of PhD students, 87% of researchers with <5 years experience, 80% of those with 5-10 years, 66% of those with 10-15 years, and 71% of those with 15+ years of experience *always* or *frequently* share their code.

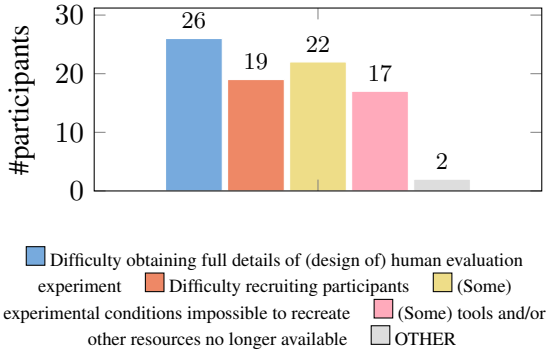
Finally, we asked about **views on reproducibility more generally** (Figure 7), and level of engage-



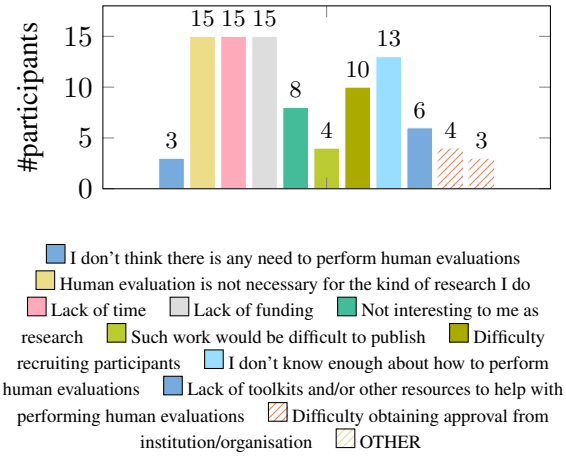
(a) Question 14: Have you in the past 5 years performed a human evaluation of an NLP system?



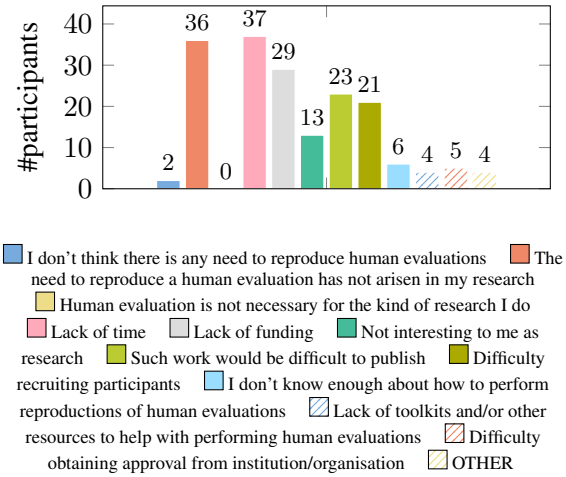
(b) Question 15: Have you in the past 5 years tried to recreate your own or someone else's human evaluation of an NLP system?



(c) Question 18: Which of the following issues did you encounter when trying to recreate your own or someone else's human evaluation in the past 5 years over and above the issues mentioned in previous questions (select all that apply)?



(d) Question 16: Which of the following are among the reasons why you have not carried out human evaluations in the past 5 years (select all that apply)?



(e) Question 17: Which of the following are among the reasons why you have not attempted to recreate a human evaluation in the past 5 years (select all that apply)?

Figure 4: 2024 Results: Answers to Q14–Q18 about human evaluation.

	PhD Student	<5 years	5–10 years	10–15 years	15+ years
always	28	14	5	7	3
frequently	14	6	7	3	12
sometimes	6	3	3	4	4
rarely	1	0	0	1	1
never	0	0	0	0	1

Table 1: Response frequencies by career stage for Q25; always/frequently make model code available.

ment on reproducibility issues. Nearly all participants (97%) agreed that work in their field should be easy to recreate (Q28), with only 9% happy with how their field currently addresses reproducibility (Q29). As many as 31% of participants were aware of recreation attempts of their own work (Q30); 56% indicated that they comment on reproducibility issues as reviewers.

5 Change Analysis 2022 vs. 2024

In this section, we look at the proportional differences (up or down) between the 2022 and 2024 response counts. We can't just describe the change in terms of how many more/fewer people selected different responses, because in some cases, a question was answered by a different number of people in 2022 and 2024 (due to the conditional logic of the question flow, see Figure 9, Appendix). Instead, we discuss change in terms of the difference $\Delta_{A_{i,j}}^{Y_1,Y_2}$ in the proportion of people who selected a given answer $A_{i,j}$ to a question in two different years Y_1 and Y_2:

$$\Delta_{A_{i,j}}^{Y_1,Y_2} = (A_{i,j}^{Y_2}/R_i^{Y_2} - A_{i,j}^{Y_1}/R_i^{Y_1}) \quad (1)$$

where $A_{i,j}^{Y_k}$ is the answer count for the jth response to the ith question in year Y_k, and $R_i^{Y_k}$ is the number of respondents who answered the ith question

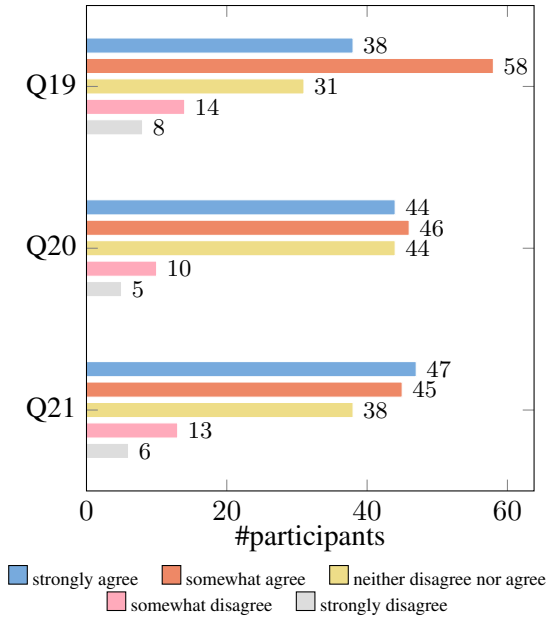


Figure 5: 2024 Results: Answers to Q19–Q21 about things organisations could do to enable reproduction. Participants were asked indicate the degree to which they agreed with the below statements. **Q19**: "If the conferences I usually submit to reserved some of their acceptance rate for reproduction papers, I would submit (more) such papers." **Q20**: "If the funding bodies I usually apply to reserved some of their funding for reproduction projects, I would submit (more) such applications." **Q21**: "If there was an easy to use toolkit available for this purpose, I would carry out (more) reproduction work."

in year Y_k. In our comparison, $Y_1 = 2022$ and $Y_2 = 2024$. We discuss some of the bigger differences in the remainder of this section, grouped into subsections as in Section 4. See also Figures 11–39 in the Appendix.

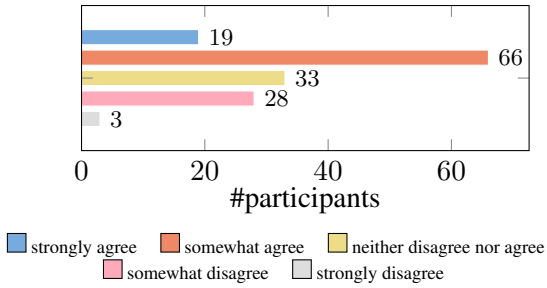
5.1 Participant Demographics (2022 → 2024)

Q3 (Figure 11) saw a significant increase in number of PhD vs. other students. Other changes were minimal; Q1–2 had no changes by design.

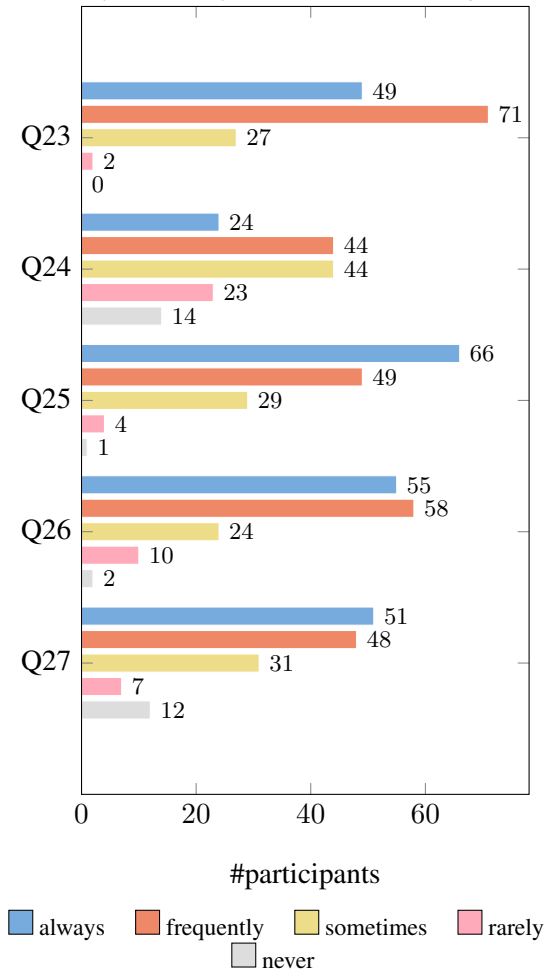
5.2 Recreating Research (2022 → 2024)

When recreating the research of others, interestingly, whilst there was a 7.59%pt drop in difficulty communicating with authors, there was a a 7.41%pt increase for difficulty obtaining resources (Q9, Figure 17).

When recreating own research, there was a 5.2%pt drop in difficulties communicating with former team members (Q12, Figure 20), accompanied by a 5.39%pt drop in issues obtaining resources.



(a) Question 22: Please indicate the degree to which you agree with the following statements: [Other researchers could easily recreate my work, without contacting me]



(b) 2024 Results: Answers to Q23–Q27 about efforts researchers make to enable reproduction. Participants were asked to indicate the frequency with which they do the following, **Q23**: "I take steps to ensure that my published work can be easily recreated", **Q24**: "When reporting work with metric scores, I also carry out human evaluations", **Q25**: "When reporting work with models, I make the code available", **Q26**: "When reporting work with data, I make the data available", **Q27**: "When reporting human evaluations, I make full details available, including evaluation interface and evaluator instructions"

Figure 6: 2024 Results: Answers to Q22–Q27 about efforts researchers make to enable reproduction.

There was also a 7.28%pt increase in researchers reporting insufficient compute, not unexpected given

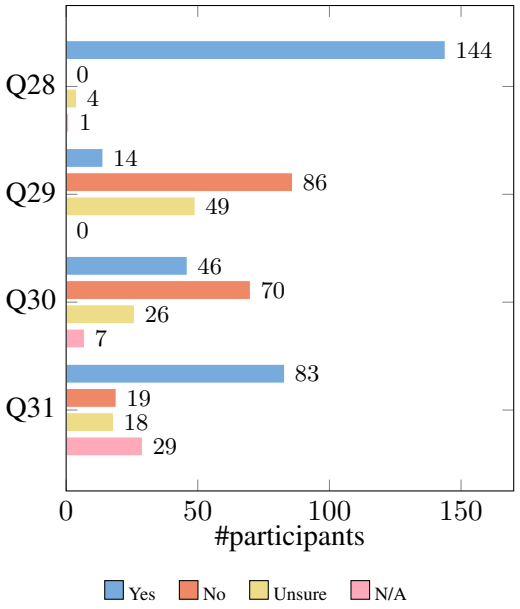


Figure 7: 2024 Results: Answers to Q28–Q31 about researcher's opinions on reproducibility. Participants were asked to indicate whether or not the below statements applied in your case. **Q28**: "I think it's important that work in my field should be easy to recreate." **Q29**: "I'm happy with how my field currently addresses reproducibility." **Q30**: "I'm aware of recreation attempts of my own work." **Q31**: "As a reviewer, I have commented on reproducibility issues."

the ever increasing size of many NLP models. Finally, there was an 8.79%pt drop in results always being broadly the same (Q13, Figure 21).

5.3 Recreating human evaluations (2022 → 2024)

It was encouraging to see increases in respondents who had performed a human evaluation (+3.36%pt for Q14, Figure 22), and the number who had recreated a human evaluation (+6.15%pt for Q15, Figure 23). Looking at issues raised in Q16 (performing human evaluations, Figure 24) and Q17 (recreating human evaluations, Figure 25) together, there was a drop in both of about 15%pt for those reporting a lack of toolkits and/or other resources being an issue. For Q18 (Figure 26), issues encountered when recreating human evaluations, there was a 21.83%pt increase in experiment conditions being impossible to recreate, as well as a 7.94%pt increase in resources no longer being available.

5.4 Barriers and self-reflection (2022 → 2024)

Looking at things that would encourage researchers to perform more reproducibility work, there was an 7.38%pt increase in those agreeing they would submit more reproductions if conferences reserved acceptance rate for them (Q19, Figure 27). When asked whether they would perform more reproduction studies if a toolkit was available, there was an 6.04%pt decrease in those who agreed. There was a 6.04%pt drop in respondents always making data available (Q26, Figure 34) and a 7.38%pt drop in participants being happy with how their field addresses reproducibility (Q29, Figure 37). Question 30 (Figure 38) showed a 14.77%pt increase in those who were unaware of recreation attempts of their work. Finally, for Q31 (Figure 39), addressing whether respondents comment on reproducibility during peer review, we see about a 10%pt drop.

6 High-level recoded analysis

Whilst there were no significant differences in individual questions (other than Q3 in the participant characteristics section), we were also interested in whether there was a higher-level improvement in reproducibility work and awareness of reproducibility issues. We therefore recoded responses, to address three higher-level questions:

> **HQ1** (reproduction activity): Have researchers increased in engaging in reproduction efforts?
>
> **HQ2** (enabling reproducibility practices): Has there been an increase in researchers actively taking steps to make their work reproducible?
>
> **HQ3** (awareness): Is there an increase in researchers valuing reproducibility and understanding its relevance to the field?

For recoding responses, we used the following questions for each construct: HQ1–*Are researchers doing, or attempting, recreation/reproduction studies?*; HQ2–*Are researchers doing things to make recreation/reproduction of research easier?*; HQ3–*Are researchers aware of the importance of issues related to recreatability/reproducibility?* In order to prevent bias, the recoding was performed by two non-author annotators from the same group.

For each construct, we mapped each question/response pair, excluding demographic questions (Q1–7), to one of the below options:

> **Supports Yes**: The Question-response pair supports an answer of "Yes" to the construct question.
>
> **Supports No**: The Question-response pair supports an answer of "No" to the construct question.

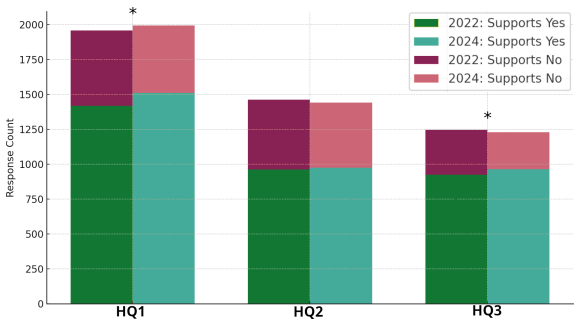


Figure 8: Change in Construct Support Between 2022 and 2024 (* = statistically significant change at $\alpha = 0.05$).

> **Supports Neither**: The Question-response pair does not address the construct question.

The two non-author annotators first annotated three questions together before discussing the annotation protocol so that they could agree upon an interpretation of the instructions. They each then annotated roughly half of question-response pairs.

Figure 8 shows the recoded results for each high-level construct: for HQ1 (reproduction activity) and HQ3 (awareness) there is a statistically significant increase for *Yes* in 2024. However, whilst there was an increase in HQ2 (enabling reproducibility practices), it was not statistically significant. The *Supports Neither* mapped responses were excluded from our analysis, as they indicate question-response pairs that are not relevant to the construct.

7 Discussion and Conclusion

Among the results that stand out in the 2024 survey are the large proportion of respondents that have tried to recreate other people's work (120/149), yet only a small proportion (3/120) was always able to obtain the same result. The most common response to this question of whether they had obtained the same results (Q10) was *Sometimes* (56/149).

108 of 149 participants reported that they have performed human evaluations, with 65 of the 108 (60%) stating that if they report metric scores they also carry out human evaluations.[6] This does not align with findings by Belz et al. (2025a), who performed a systematic review of 120 papers from ACL (2022–2024), finding that of the 115 papers that reported an evaluation, only 21 included a human evaluation.

It is encouraging that more researchers are engaging with reproducibility work (HQ1, from the high-level analysis in Section 6), and that they are more aware of reproducibility issues (HQ3). Only 4 of the 41 (10%) respondents who had not performed a human evaluation in the previous 5 years reported difficulty publishing as a reason. Of the 72 who had performed a human evaluation, but not reproduction, 23 (32%) reported it as a reason (Fig. 4). This indicates a perceived lack of interest in reproducibility work at conferences, from organisers and/or reviewers. That 64% of all respondents agreed they would submit more reproduction papers if conferences set aside acceptance rate for them also supports this.

Unsurprisingly, lack of time and funding were the two top issues preventing people from doing reproduction work, with 60% of respondents saying they would perform more reproduction work if funding bodies earmarked money for it. Just behind time and money for reasons respondents did not perform human evaluations, was that they did not know how (13 of 41). Six participants also suggested a toolkit would be useful.

That significant numbers of respondents are not doing more things to make reproducibility easier for others (HQ2, from the high-level analysis) is perhaps less because they are unwilling, and more that individual researchers are limited in what they can do. Giving reproduction work the status it deserves as a cornerstone of science is something that funders and conference organisers can address, and our survey indicates that researchers would do more work on reproduction if this was done. Given the lack of confidence in carrying out reproduction studies, developing guidance, methodologies and frameworks for performing evaluations would enable NLP researchers to learn how to perform repeatable experiments, and reduce their workload by not having to reinvent the wheel.

Overall, we have seen that virtually all 2024 respondents agree that NLP/ML work *should* be easy to recreate (up from 2022). At the same time, falling numbers are happy with how the field addresses reproducibility, and fewer researchers are sharing data, and are confident in their own work being reproducible. These results may in fact reflect growing understanding not only of the importance of reproducibility, but also of the difficulties in doing it properly. Our recoded results show that researchers *are* now (i) engaging in reproduction efforts more, and (ii) valuing reproducibility and understanding its relevance to the field more. Taken together, these surely are encouraging signs.

[6] 3 participants had not performed a human evaluation in the past 5 years, but do so when they report metrics.

Limitations

Our survey may have sampling bias and not reflect the NLP/ML communities as a whole. This may be inevitable for a survey such as this; researchers who are interested in issues related to reproducibility might be more inclined to respond to the survey. However, given that our work focuses on longitudinal change, the influence of this sampling bias on our conclusions is somewhat mitigated. Our survey was conducted in English, which may have limited geographical diversity. Whilst respondents did work on a wide variety of tasks (see Figure 1g), it is unclear whether this distribution is representative of the NLP community; the number of submitted or accepted papers per track is not publicly available for ACL 2024.

Acknowledgements

The ReproHum project was funded by EPSRC grant EP/V05645X/1. Craig Thomson is currently funded by the ADAPT SFI Centre for Digital Media Technology. João Sedoc thanks NYU Stern for support. We would like to thank Michela Lorandi and Massimiliano Pronesti for performing the recoding annotations. Finally, we would like to thank all participants who responded to our surveys.

References

Alan Agresti. 2018. *Statistical methods for the social sciences*. Pearson.

Mohammad Arvan, Luís Pina, and Natalie Parde. 2022. Reproducibility in computational linguistics: Is source code enough? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2350–2361, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Monya Baker. 2016. Is there a reproducibility crisis? *Nature*, 533:452–454.

Anya Belz. 2022. A metrological perspective on reproducibility in NLP*. *Computational Linguistics*, 48(4):1125–1135.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021. A systematic review of reproducibility research in natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 381–393.

Anya Belz, Simon Mille, and David M. Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 183–194, Dublin, Ireland. Association for Computational Linguistics.

Anya Belz, Simon Mille, and Craig Thomson. 2025a. Standard quality criteria derived from current NLP evaluations for guiding evaluation design and grounding comparability and AI compliance assessments. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26685–26715, Vienna, Austria. Association for Computational Linguistics.

Anya Belz, Maja Popovic, and Simon Mille. 2022. Quantified reproducibility assessment of NLP results. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16–28, Dublin, Ireland. Association for Computational Linguistics.

Anya Belz and Craig Thomson. 2023. The 2023 ReproNLP shared task on reproducibility of evaluations in NLP: Overview and results. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 35–48, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Anya Belz and Craig Thomson. 2024a. The 2024 ReproNLP shared task on reproducibility of evaluations in NLP: Overview and results. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 91–105, Torino, Italia. ELRA and ICCL.

Anya Belz and Craig Thomson. 2024b. Heds 3.0: The human evaluation data sheet version 3.0. *Preprint*, arXiv:2412.07940.

Anya Belz, Craig Thomson, Javier González Corbelle, and Malo Ruelle. 2025b. The 2025 ReproNLP shared task on reproducibility of evaluations in NLP: Overview and results. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 1002–1016, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M. Alonso-Moral, Mohammad Arvan, Anouck Braggaar, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, Tanvi Dinkar, Ondřej Dušek, Steffen Eger, Qixiang Fang, Mingqi Gao, Albert Gatt, Dimitra Gkatzia, Javier González-Corbelle, Dirk Hovy, Manuela Hürlimann, Takumi Ito, John D. Kelleher, Filip Klubicka, Emiel Krahmer, Huiyuan Lai, Chris van der Lee, Yiru Li, Saad Mahamood, Margot Mieskes, Emiel van Miltenburg, Pablo Mosteiro, Malvina Nissim, Natalie Parde, Ondřej Plátek, Verena Rieser, Jie Ruan, Joel Tetreault, Antonio Toral, Xiaojun Wan, Leo Wanner, Lewis Watson, and Diyi Yang. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous

human evaluations in NLP. In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Herbert H. Clark and Michael F. Schober. 1992. *Asking questions and influencing answers.*, pages 15–48. Questions about questions: Inquiries into the cognitive bases of surveys. Russell Sage Foundation, New York, NY, US.

Joost FC de Winter and Dimitra Dodou. 2010. Five-point likert items: t test versus mann-whitney-wilcoxon (addendum added october 2012). *Practical Assessment, Research, and Evaluation*, 15(1).

Floyd J. Fowler and Carol Cosenza. 2008. *Writing effective questions.*, pages 136–160. International Handbook of Survey Methodology. Lawrence Erlbaum Associates, New York, NY, US.

Sayash Kapoor and Arvind Narayanan. 2023. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*. Publisher: Elsevier.

Ian Magnusson, Noah A. Smith, and Jesse Dodge. 2023. Reproducibility in NLP: What have we learned from the checklist? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12789–12811, Toronto, Canada. Association for Computational Linguistics.

Margot Mieskes, Karën Fort, Aurélie Névéol, Cyril Grouin, and Kevin Cohen. 2019. Community perspective on replicability in natural language processing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 768–775, Varna, Bulgaria. INCOMA Ltd.

Ted Pedersen. 2008. Empiricism is not a matter of faith. *Computational Linguistics*, 34(3):465–470.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. 2021. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *Journal of Machine Learning Research*, 22(20-303):1–20.

Birgit Popp, Sarah Keck, Androniki Mertsiotaki, Emily Kratsch, and Alexander Daum. 2025. Which method(s) to pick when evaluating large language models with humans? - a comparison of 6 methods. CC BY-NC-ND 4.0.

Edward Raff and Andrew L. Farris. 2023. A siren song of open source reproducibility, examples from machine learning. In *Proceedings of the 2023 ACM Conference on Reproducibility and Replicability*, ACM REP '23, page 115–120, New York, NY, USA. Association for Computing Machinery.

Jie Ruan, Wenqing Wang, and Xiaojun Wan. 2024. Defining and detecting vulnerability in human evaluation guidelines: A preliminary study towards reliable NLG evaluation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7965–7989, Mexico City, Mexico. Association for Computational Linguistics.

Harald Semmelrock, Tony Ross-Hellauer, Simone Kopeinik, Dieter Theiler, Armin Haberl, Stefan Thalmann, and Dominik Kowald. 2025. Reproducibility in machine-learning-based research: Overview, barriers, and drivers. *AI Magazine*, 46(2):e70002.

Anastasia Shimorina and Anya Belz. 2022. The human evaluation datasheet: A template for recording details of human evaluation experiments in NLP. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 54–75, Dublin, Ireland. Association for Computational Linguistics.

Shane Storks, Keunwoo Yu, Ziqiao Ma, and Joyce Chai. 2023. NLP reproducibility for all: Understanding experiences of beginners. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10199–10219, Toronto, Canada. Association for Computational Linguistics.

Craig Thomson, Ehud Reiter, and Anya Belz. 2024. Common flaws in running human evaluation experiments in NLP. *Computational Linguistics*, 50(2):795–805.

Emiel van Miltenburg, Anouck Braggaar, Nadine Braun, Martijn Goudbeek, Emiel Krahmer, Chris van der Lee, Steffen Pauws, and Frédéric Tomas. 2024. ReproHum: #0033-03: How reproducible are fluency ratings of generated text? a reproduction of August et al. 2022. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 132–144, Torino, Italia. ELRA and ICCL.

Emiel van Miltenburg, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Stephanie Schoch, Craig Thomson, and Luou Wen. 2023. Barriers and enabling factors for error analysis in NLG research. *Northern European Journal of Language Technology*, 9.

Martijn Wieling, Josine Rawee, and Gertjan van Noord. 2018. Squib: Reproducibility in computational linguistics: Are we willing to share? *Computational Linguistics*, 44(4):641–649.

Yan Xue, Xuefei Cao, Xingli Yang, Yu Wang, Ruibo Wang, and Jihong Li. 2023. We need to talk about reproducibility in NLP model comparison. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9424–9434, Singapore. Association for Computational Linguistics.

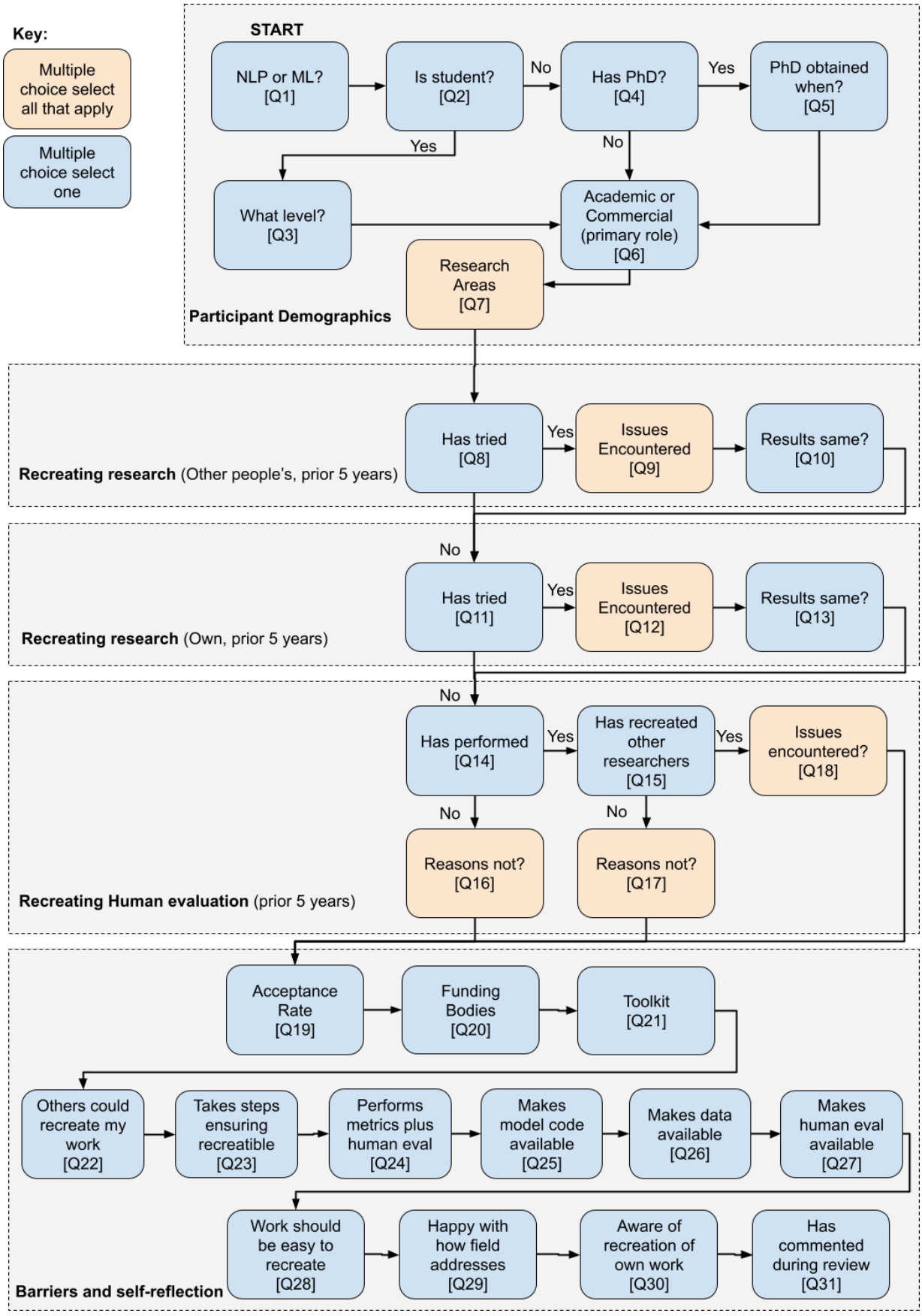


Figure 9: Flow chart of the survey design and question control.

A Reproducibility & Ethics

ChatGPT (4o-mini-high) was used as an aid in writing python code that generated tikz pictures.[7] The code was manually checked to ensure correct working.

All participants who volunteered for the survey completed our institution's standard consent form, that includes questions on data use. The Google Form instructions they were shown are included as Figure 10.

B Conference Checklists & Guidelines

The checklists and guidelines mentioned in Section 1 can be found here:

ACL: https://aclrollingreview.org/static/responsibleNLPresearch.pdf

AAAI: https://aaai.org/conference/aaai/aaai-25/aaai-25-reproducibility-checklist

ICLR: https://iclr.cc/Conferences/2022/AuthorGuide

ICML: https://icml.cc/Conferences/2025/AuthorInstructions

C Survey Design Principles

In this section we briefly summarise the design principles we applied in creating our surveys (Clark and Schober, 1992). They address issues fundamental to the process of asking questions, and are important for even the most basic of surveys.

Clark and Schober (1992) note that whilst survey design can appear to be a simple task, creating a shared meaning of questions between researcher and respondent is difficult. This shared meaning is described as a construct by Fowler and Cosenza (2008), by which they mean the abstract concept that can be measured using the answers provided to well designed questions. They describe four key stages a respondent will go through when answering a question:

1. Understand intended meaning of the question.
2. Have or be able to retrieve the information required to answer it.
3. Translate the information into the question response format, for example, with close-ended response options.
4. Be willing and able to provide the answer.

The vast majority of issues that prevent respondents from following this process are related to (i) ambiguity, and (ii) lack of familiarity. Questions must be clear, using unambiguous words that are either familiar to respondents, or explained. Only one question should be asked at a time and any close-ended questions should have exhaustive and mutually exclusive response options. If a timeframe or selection of events might be ambiguous, it should be clearly specified. Finally, things should not be assumed of the respondent, any presuppositions that do not apply to an individual will change how they comprehend the question.

D 2022 Survey Results

Figures 40, 41, 42, 43, 44, 45, and 46 show the results of the 2022 survey in the same format and figure grouping as the 2024 results in Section 4

[7] https://openai.com/index/chatgpt

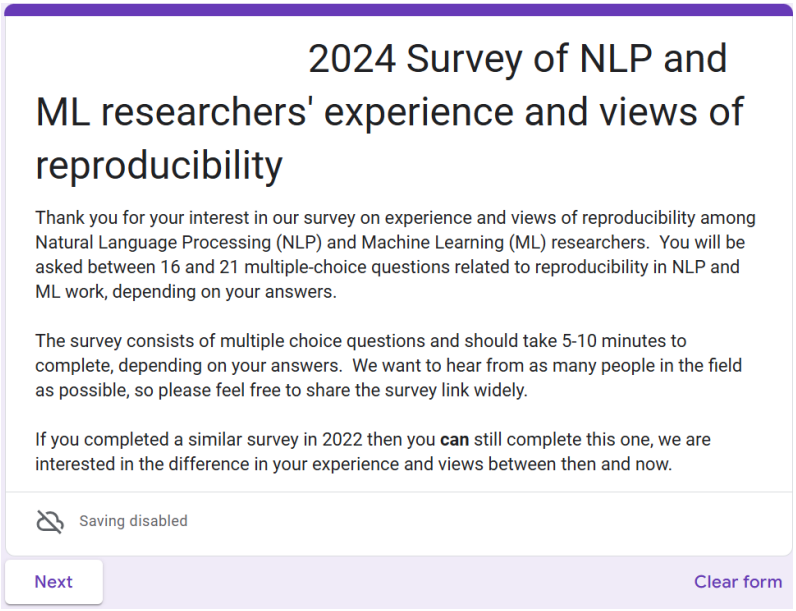


Figure 10: Screenshot of the 2024 survey Google Form showing the introduction. The 2022 version was the same, except that it omitted the final sentence, and the year in the title was 2022.

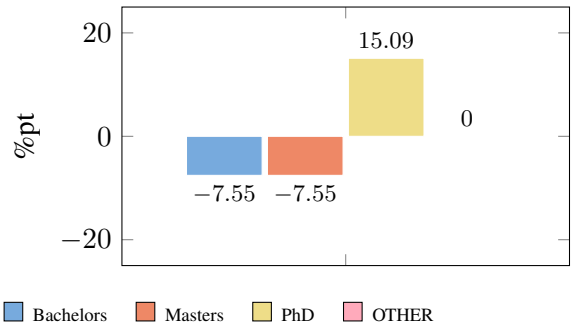


Figure 11: Question 03: Towards which degree are you currently studying?

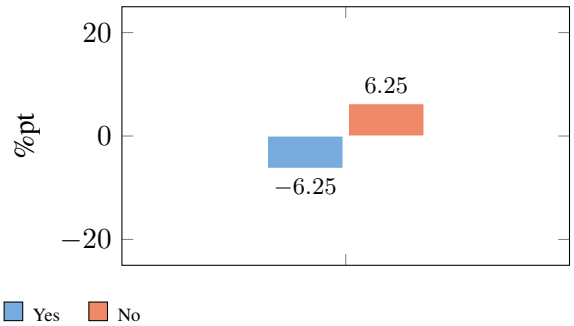


Figure 12: Question 04: Do you have a PhD?

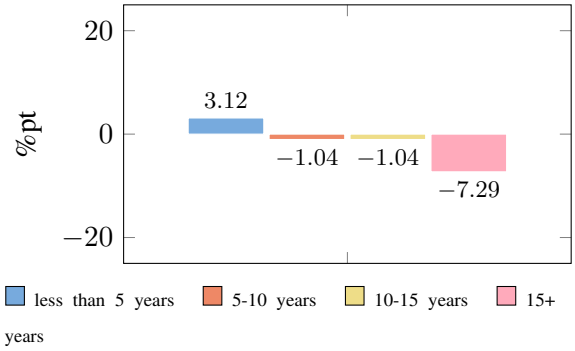


Figure 13: Question 05: How many years ago did you obtain your PhD?

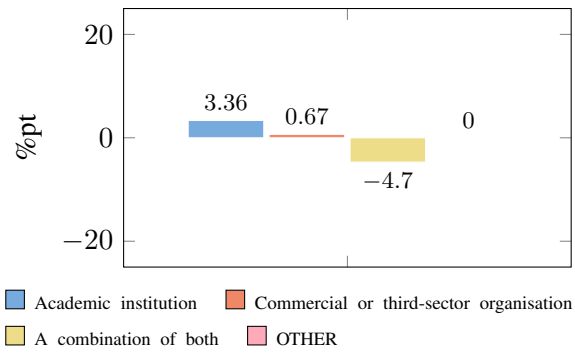


Figure 14: Question 06: Which of the following best describes your current primary affiliation(s)?

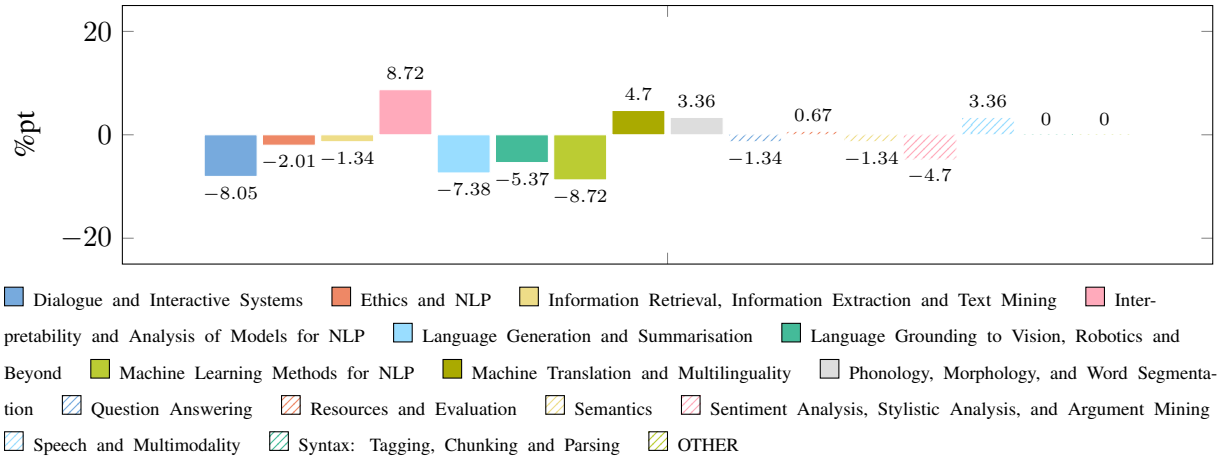


Figure 15: Question 07: Which of the following NLP subfields have you worked in in the past 4 years (select all that apply)?

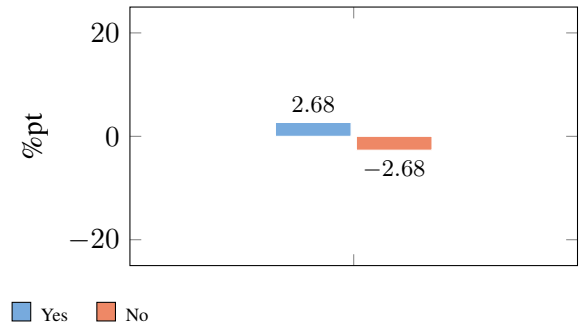


Figure 16: Question 08: Have you in the past 5 years tried to recreate any aspect of other people's research?

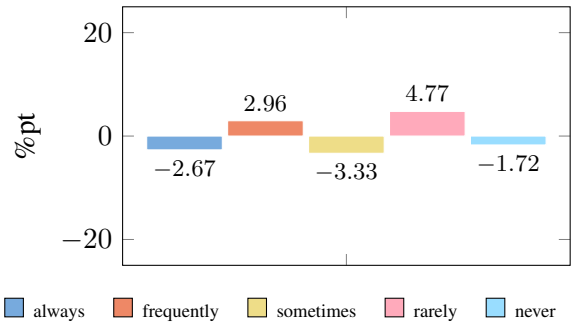


Figure 18: Question 10: Thinking about results when you recreated other people's research, were your results broadly the same as those reported in the original work?

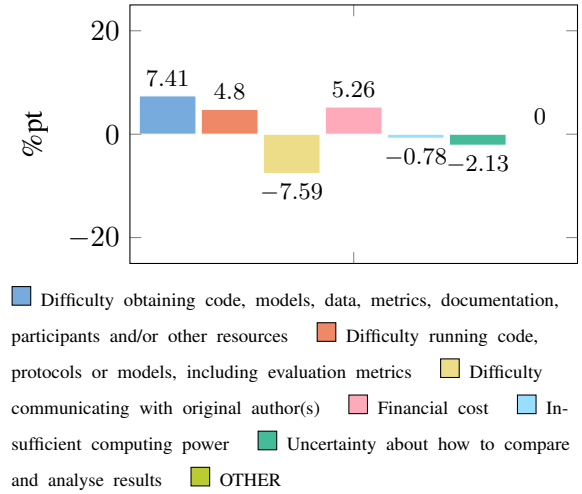


Figure 17: Question 09: Which of the following issues did you encounter when trying to recreate other people's research (select all that apply)?

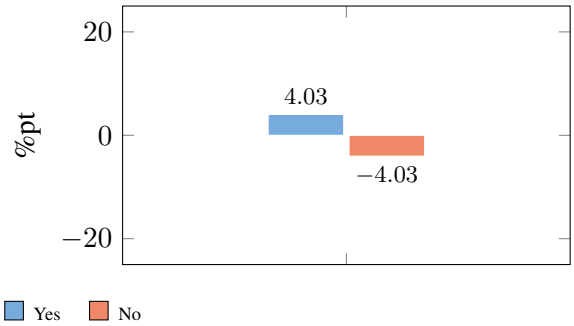


Figure 19: Question 11: Have you in the past 5 years tried to recreate any aspect of your own research?

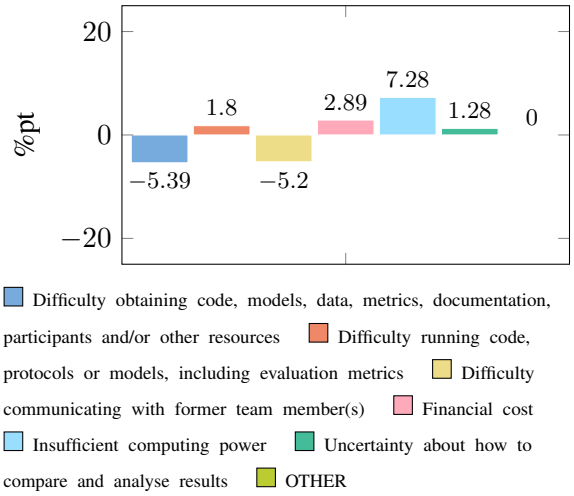


Figure 20: Question 12: Which of the following issues did you encounter when trying to recreating your own research (select all that apply)?

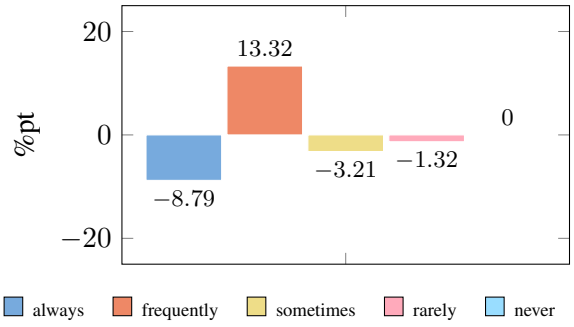


Figure 21: Question 13: Thinking about results when you recreated your own research, were your results broadly the same as your earlier results?

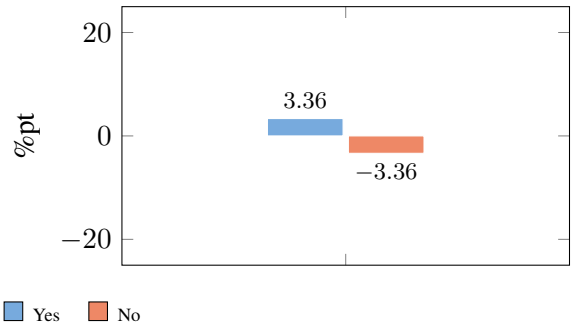


Figure 22: Question 14: Have you in the past 5 years performed a human evaluation of an NLP system?

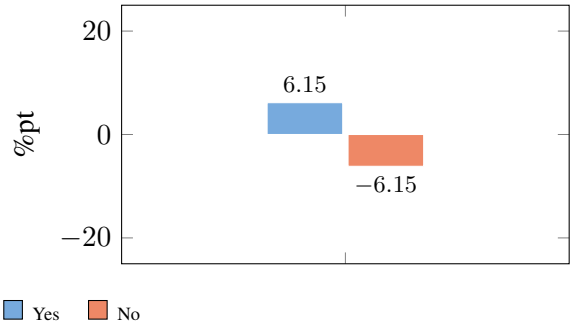


Figure 23: Question 15: Have you in the past 5 years tried to recreate your own or someone else's human evaluation of an NLP system?

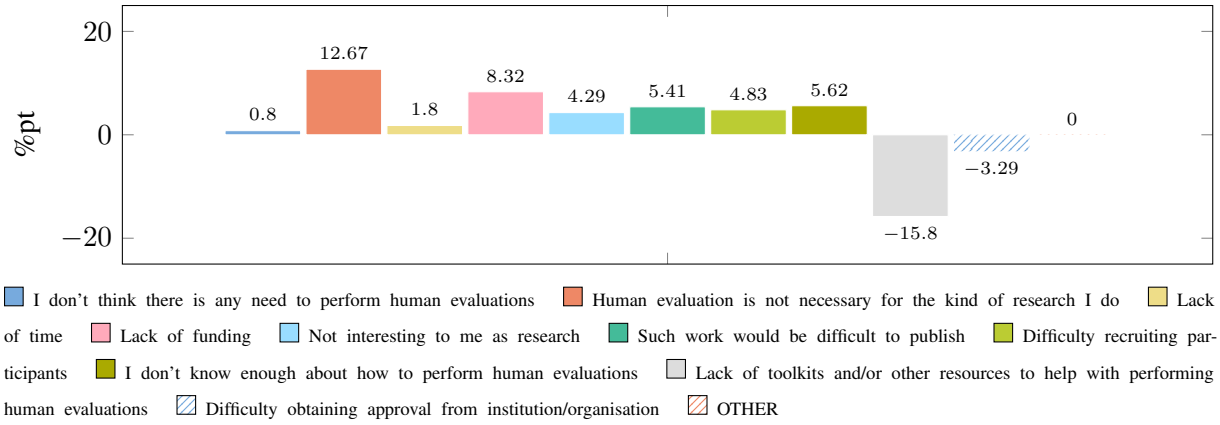


Figure 24: Question 16: Which of the following are among the reasons why you have not carried out human evaluations in the past 5 years (select all that apply)?

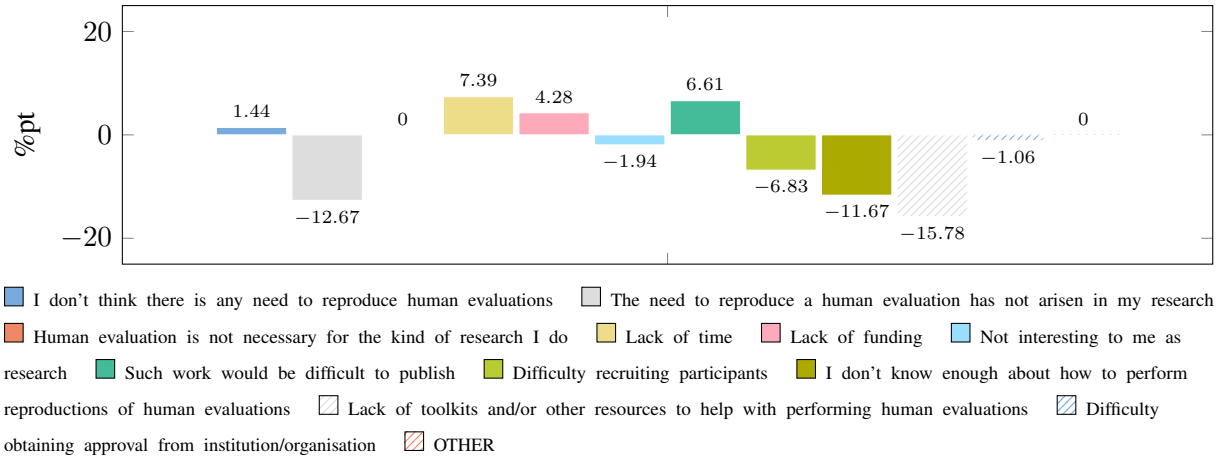


Figure 25: Question 17: Which of the following are among the reasons why you have not attempted to recreate a human evaluation in the past 5 years (select all that apply)?

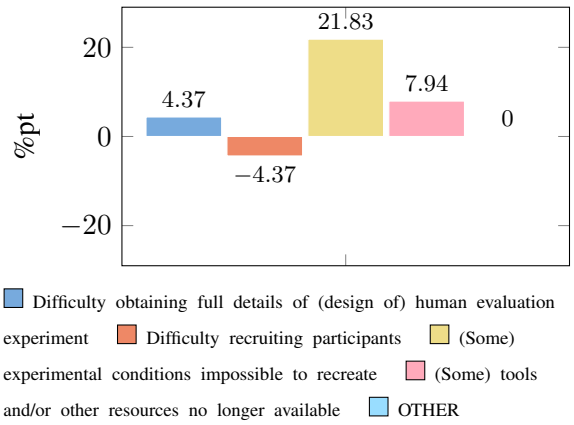


Figure 26: Question 18: Which of the following issues did you encounter when trying to recreate your own or someone else's human evaluation in the past 5 years over and above the issues mentioned in previous questions (select all that apply)?

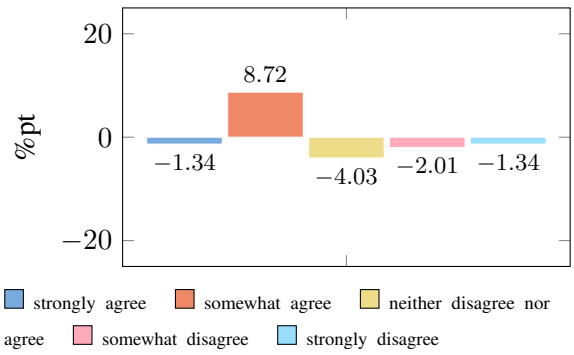


Figure 27: Question 19: Please indicate the degree to which you agree with the following statements: [If the conferences I usually submit to reserved some of their acceptance rate for reproduction papers, I would submit (more) such papers]

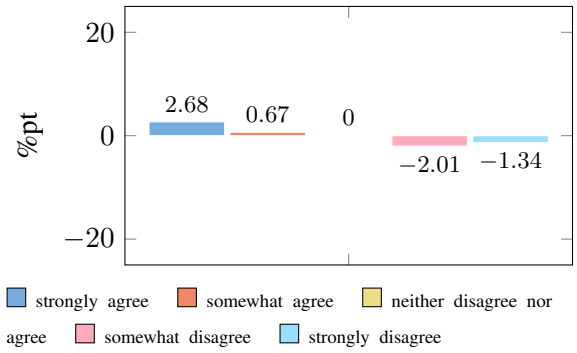


Figure 28: Question 20: Please indicate the degree to which you agree with the following statements: [If the funding bodies I usually apply to reserved some of their funding for reproduction projects, I would submit (more) such applications]

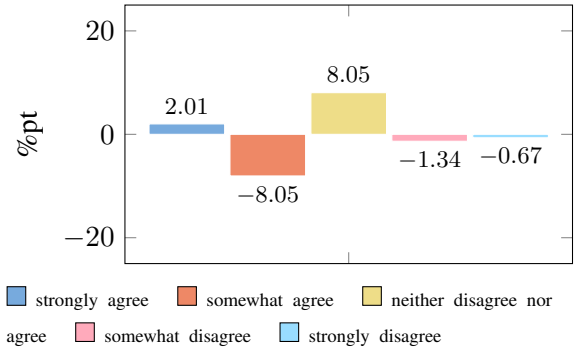


Figure 29: Question 21: Please indicate the degree to which you agree with the following statements: [If there was an easy to use toolkit available for this purpose, I would carry out (more) reproduction work]

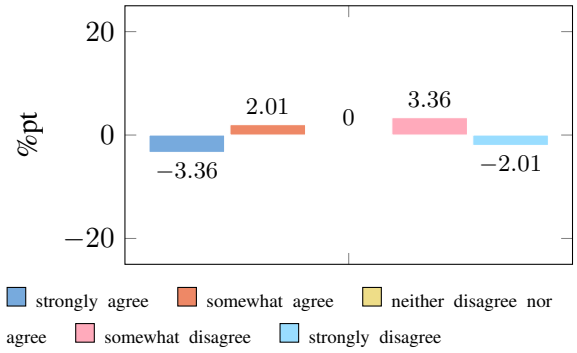


Figure 30: Question 22: Please indicate the degree to which you agree with the following statements: [Other researchers could easily recreate my work, without contacting me]

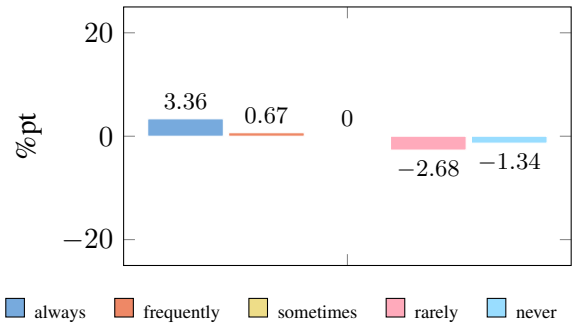


Figure 31: Question 23: Please indicate the frequency with which you do the following : [I take steps to ensure that my published work can be easily recreated]

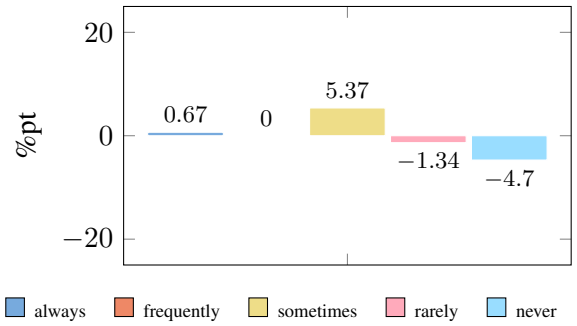


Figure 32: Question 24: Please indicate the frequency with which you do the following : [When reporting work with metric scores, I also carry out human evaluations]

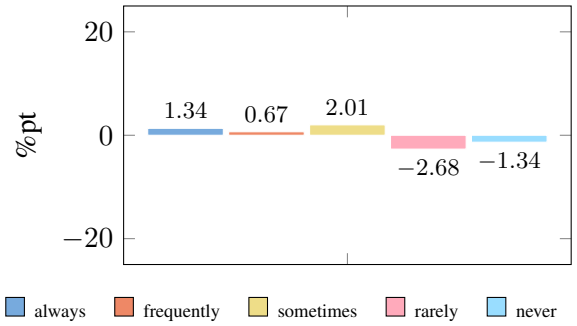


Figure 33: Question 25: Please indicate the frequency with which you do the following : [When reporting work with models, I make the code available]

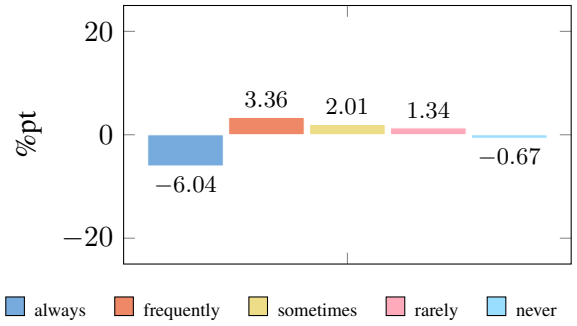


Figure 34: Question 26: Please indicate the frequency with which you do the following : [When reporting work with data, I make the data available]

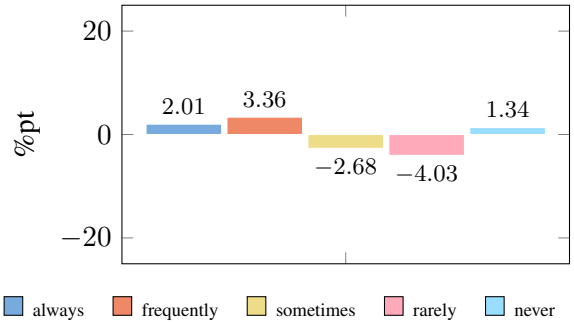


Figure 35: Question 27: Please indicate the frequency with which you do the following : [When reporting human evaluations, I make full details available, including evaluation interface and evaluator instructions]

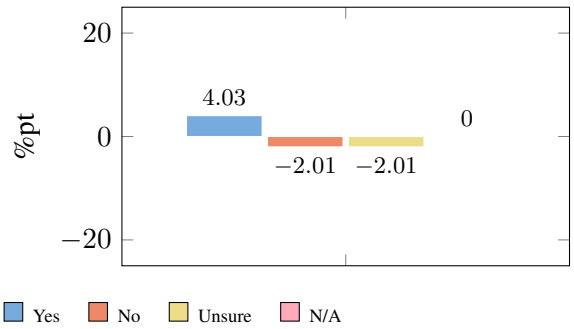


Figure 36: Question 28: Please indicate whether or not the following statements apply in your case [I think it's important that work in my field should be easy to recreate]

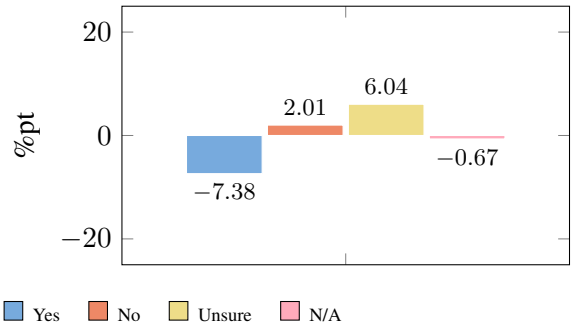


Figure 37: Question 29: Please indicate whether or not the following statements apply in your case [I'm happy with how my field currently addresses reproducibility]

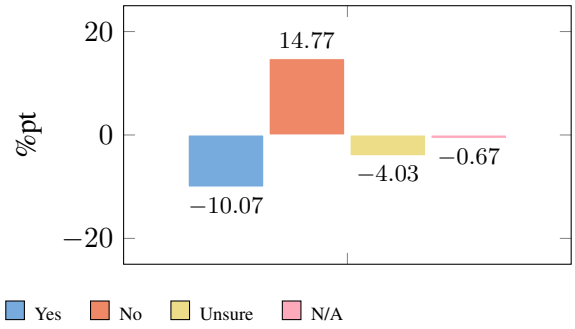


Figure 38: Question 30: Please indicate whether or not the following statements apply in your case [I'm aware of recreation attempts of my own work]

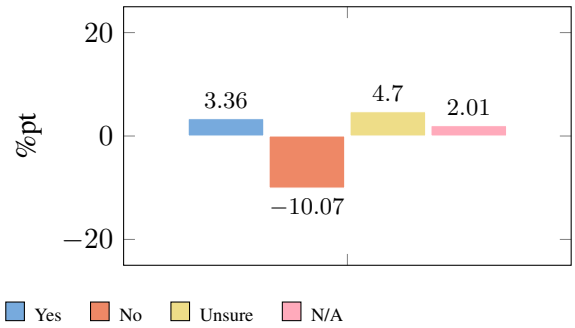
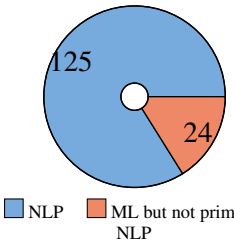
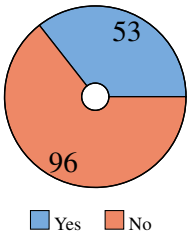


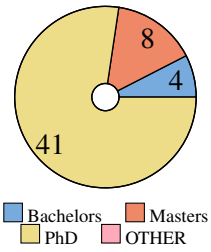
Figure 39: Question 31: Please indicate whether or not the following statements apply in your case [As a reviewer, I have commented on reproducibility issues]



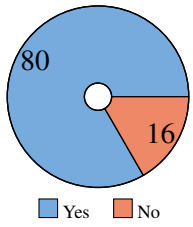
(a) Question 01: How would you describe your broad field of research?



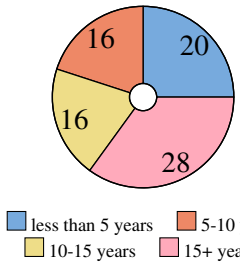
(b) Question 02: Are you currently a student?



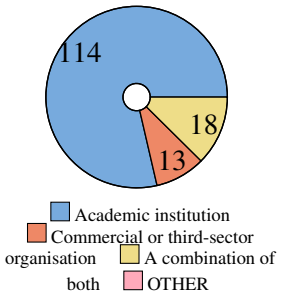
(c) Question 03: Towards which degree are you currently studying?



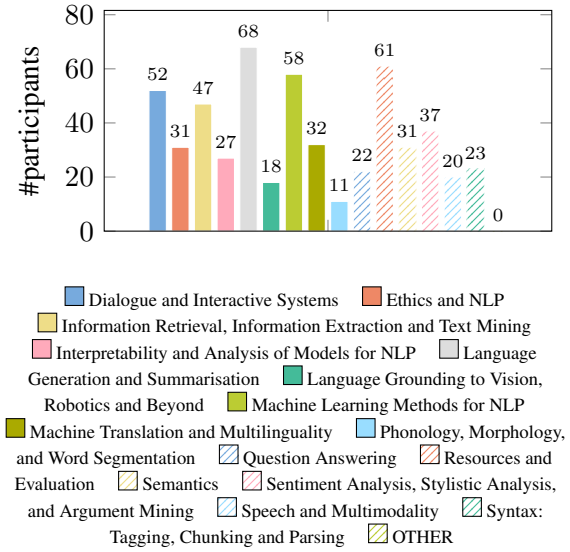
(d) Question 04: Do you have a PhD?



(e) Question 05: How many years ago did you obtain your PhD?

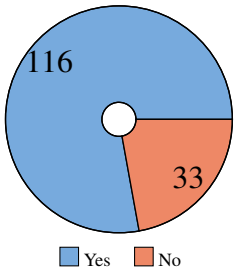


(f) Question 06: Which of the following best describes your current primary affiliation(s)?

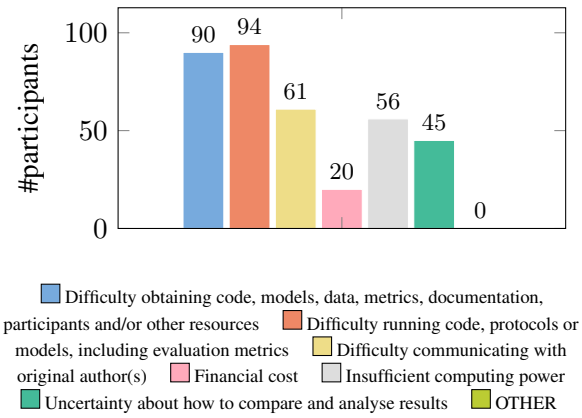


(g) Question 07: Which of the following NLP subfields have you worked in in the past 4 years (select all that apply)?

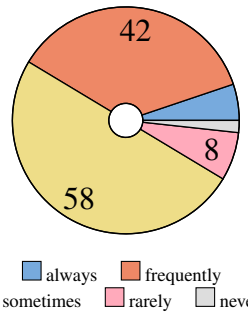
Figure 40: 2022 Results: Answers to Q1–Q7 about demographics.



(a) Question 08: Have you in the past 5 years tried to recreate any aspect of other people's research?

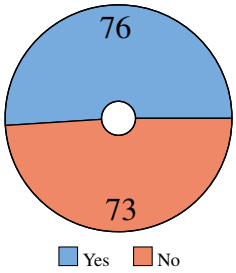


(b) Question 09: Which of the following issues did you encounter when trying to recreate other people's research (select all that apply)?

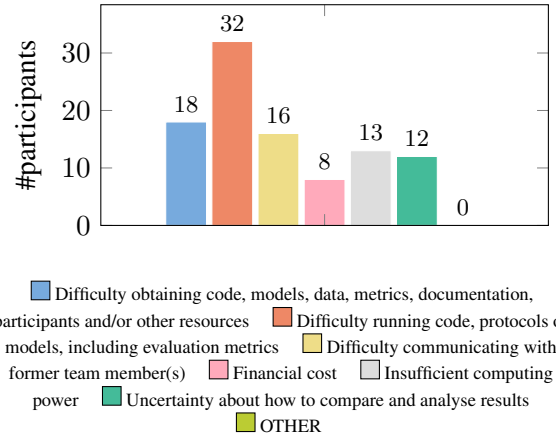


(c) Question 10: Thinking about results when you recreated other people's research, were your results broadly the same as those reported in the original work?

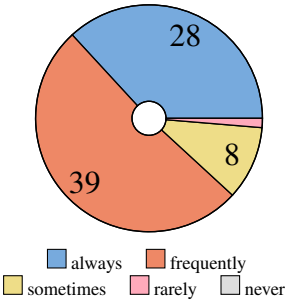
Figure 41: 2022 Results: Answers to Q8–Q10 about recreating the work of others.



(a) Question 11: Have you in the past 5 years tried to recreate any aspect of your own research?

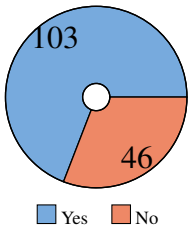


(b) Question 12: Which of the following issues did you encounter when trying to recreating your own research (select all that apply)?

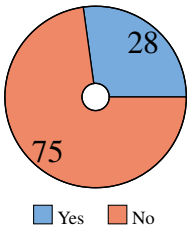


(c) Question 13: Thinking about results when you recreated your own research, were your results broadly the same as your earlier results?

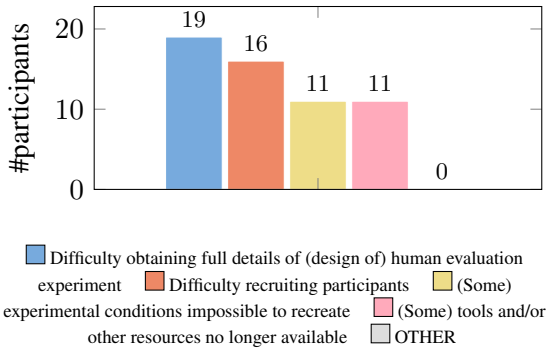
Figure 42: 2022 Results: Answers to Q11–Q13 about recreating own work.



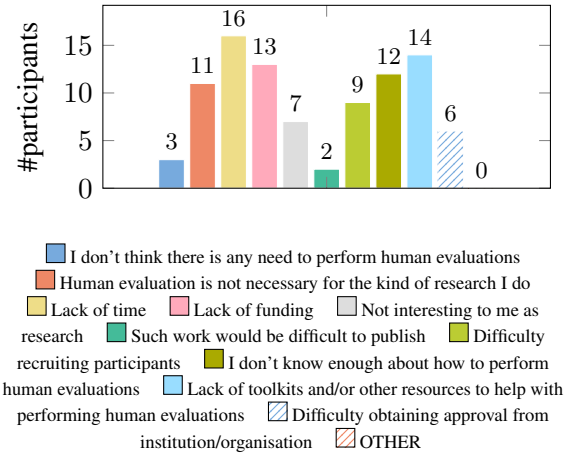
(a) Question 14: Have you in the past 5 years performed a human evaluation of an NLP system?



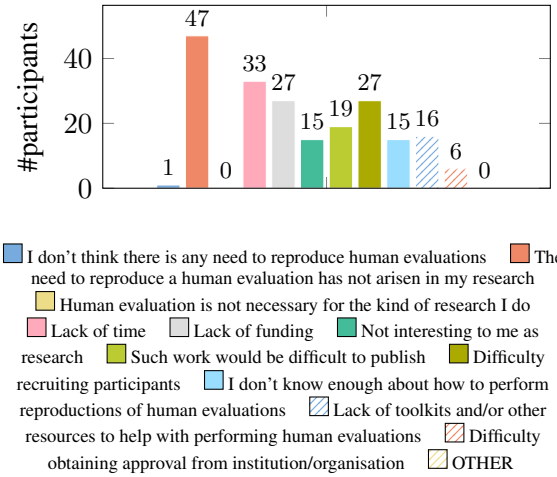
(b) Question 15: Have you in the past 5 years tried to recreate your own or someone else's human evaluation of an NLP system?



(c) Question 18: Which of the following issues did you encounter when trying to recreate your own or someone else's human evaluation in the past 5 years over and above the issues mentioned in previous questions (select all that apply)?



(d) Question 16: Which of the following are among the reasons why you have not carried out human evaluations in the past 5 years (select all that apply)?



(e) Question 17: Which of the following are among the reasons why you have not attempted to recreate a human evaluation in the past 5 years (select all that apply)?

Figure 43: 2022 Results: Answers to Q14–Q18 about human evaluation.

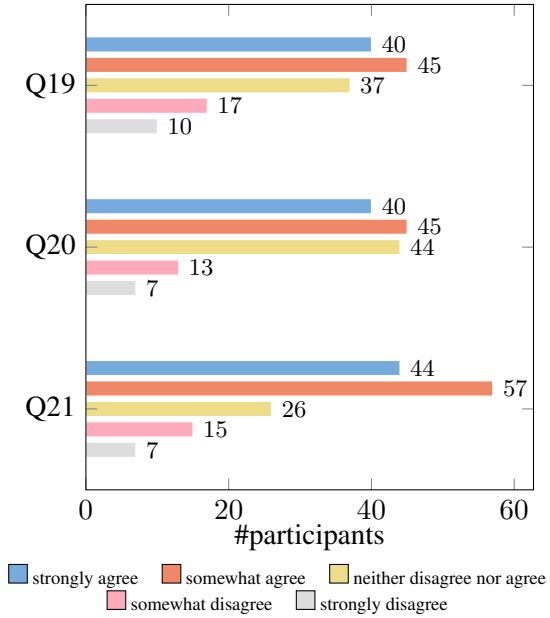
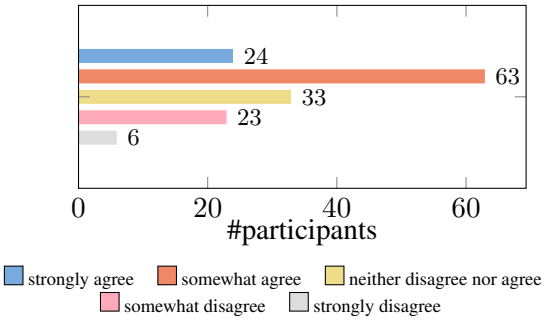
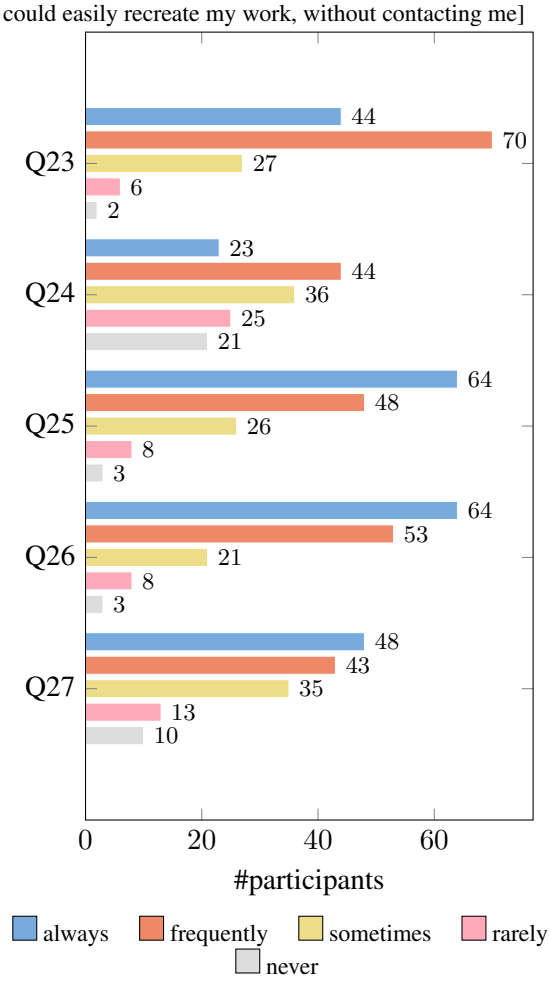


Figure 44: 2022 Results: Answers to Q19–Q21 about things organisations could do to enable reproduction. Q19: Please indicate the degree to which you agree with the following statements: [If the conferences I usually submit to reserved some of their acceptance rate for reproduction papers, I would submit (more) such papers] / Q20: Please indicate the degree to which you agree with the following statements: [If the funding bodies I usually apply to reserved some of their funding for reproduction projects, I would submit (more) such applications] / Q21: Please indicate the degree to which you agree with the following statements: [If there was an easy to use toolkit available for this purpose, I would carry out (more) reproduction work]



(a) Question 22: Please indicate the degree to which you agree with the following statements: [Other researchers could easily recreate my work, without contacting me]



(b) 2022 Results: Answers to Q23–Q27 about efforts researchers make to enable reproduction. Q23: Please indicate the frequency with which you do the following : [I take steps to ensure that my published work can be easily recreated] / Q24: Please indicate the frequency with which you do the following : [When reporting work with metric scores, I also carry out human evaluations] / Q25: Please indicate the frequency with which you do the following : [When reporting work with models, I make the code available] / Q26: Please indicate the frequency with which you do the following : [When reporting work with data, I make the data available] / Q27: Please indicate the frequency with which you do the following : [When reporting human evaluations, I make full details available, including evaluation interface and evaluator instructions]

Figure 45: 2022 Results: Answers to Q22–Q27 about efforts researchers make to enable reproduction.

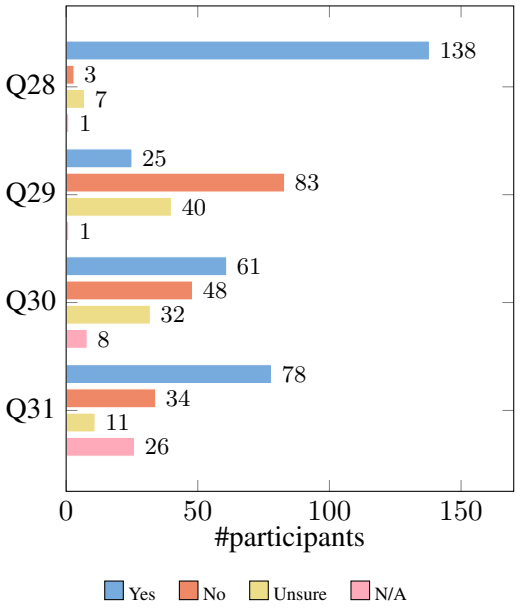


Figure 46: 2022 Results: Answers to Q28–Q31 about researcher‘s opinions on reproducibility. Q28: Please indicate whether or not the following statements apply in your case [I think it’s important that work in my field should be easy to recreate] / Q29: Please indicate whether or not the following statements apply in your case [I’m happy with how my field currently addresses reproducibility] / Q30: Please indicate whether or not the following statements apply in your case [I’m aware of recreation attempts of my own work] / Q31: Please indicate whether or not the following statements apply in your case [As a reviewer, I have commented on reproducibility issues]