Code Contribution and Authorship

Eva Maxfield Brown Nicholas Weber

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur eget porta erat. Morbi consectetur est vel gravida pretium. Suspendisse ut dui eu ante cursus gravida non sed sem. Nullam sapien tellus, commodo id velit id, eleifend volutpat quam. Phasellus mauris velit, dapibus finibus elementum vel, pulvinar non tellus. Nunc pellentesque pretium diam, quis maximus dolor faucibus id. Nunc convallis sodales ante, ut ullamcorper est egestas vitae. Nam sit amet enim ultrices, ultrices elit pulvinar, volutpat risus.

1 Random Citations

- may be interesting / useful: https://doi.org/10.1145/1772690.1772766

2 Introduction

- Contemporary scientific research fundamentally depends on specialized software tools and computational methods (Edwards et al. 2013; Mayernik et al. 2017; Howison et al. 2015).
  - define scientific software (analysis scripts, research tools, computational infrastructure) (Hasselbring et al. 2024)
  - software enables reproducible research and large-scale experiments (Krafczyk et al. 2019; Trisovic et al. 2021)
  - code serves as a detailed log of research methodology (Ram 2013)
  - due to all of the above, code is increasingly being shared alongside research articles (Cao et al. 2023; Trujillo, Hébert-Dufresne, and Bagrow 2022)
- The development and maintenance of scientific software requires substantial contribution, yet faces persistent challenges in receiving academic recognition (Muna et al. 2016).
  - software contributions often receive only acknowledgments rather than authorship (Philippe et al. 2019)

  - lack of formal credit affects career advancement in academia (Carver et al. 2022; Biagioli and Galison 2014)
  - other general discussion of software citations and credit systems (Merow et al. 2023; Westner et al. 2024; Katz et al. 2020)

- Recent initiatives to expand academic credit systems, while promising, have not fully addressed the challenges of recognizing software contributions.
  - describe the Contributor Roles Taxonomy (CRediT) (Brand et al. 2015)
  - previous research using CRediT (and prior systems) to understand research labor distribution (Larivière, Pontille, and Sugimoto 2020; Larivière et al. 2016; Sauermann and Haeussler 2017; K. Li, Zhang, and Larivière 2023; Lu et al. 2019)
  - CRediT research is still centered on traditional author lists (historic and systematic bias, self-reporting without verification, etc.) (Haeussler and Sauermann 2013; Gøtzsche et al. 2007; Ni et al. 2021)

- Our novel predictive model addresses these challenges by enabling systematic matching between scientific article authors and source code developer accounts.
  - we use predictive modeling due to the lack of standardized identifiers (i.e. ORCID) for developers (Haak et al. 2012)
  - further, lack consistency in naming and email overlap [GET CITATION FOR SORTING HAT FROM GOGGINS??]
  - semantic models handle subtle variations in identity information (general entity matching has moved to transformers and semantic embeddings) (Y. Li et al. 2020; Brunner and Stockinger 2020)

- By applying our model across a corpus of 138596 paired research articles and repositories, we provide unique insight into the dynamics of code contribution within research teams, the impact of code contribution on research outcomes, and an understanding of the authors who are and who aren't code contributors.
  - move from self-reporting to verifiable source code repository commit histories
  - provide preliminary quantitative evidence of exclusion of code contributors from academic authorship
  - model article level impact metrics as a function of software development dynamics to show the benefit code contributors have on research
  - find that first authors are more likely to be code contributors than not
  - find that code-contributing authors have reduced individual level impact metrics compared to their non-coding counterparts

- These findings not only illuminate the relationship between code contribution and scientific impact but also provide an empirical foundation for reforming academic credit systems to better recognize software development contributions in research.

3 Background

- The relationship between scientific software development and academic credit systems represents a complex intersection of traditional academic practices and modern research requirements.
  - academic credit traditionally focuses on analytical, theoretical, and experimental contributions (Larivière et al. 2016; X. Liu, Zhang, and Li 2023)
  - software development historically viewed as technical rather than scholarly work
  - growing recognition that research software development requires deep domain expertise (Heroux 2022; Carver et al. 2022)
  - increased emphasis on large scale (big data) projects has resulted in larger need for software development (Jin et al. 2015; Hampton et al. 2013; Fan, Han, and Liu 2014)
  - understanding this relationship requires examining both team-level dynamics and individual contributions
- (H1) Modern research increasingly depends on collaborative software development, yet we lack systematic evidence of how code contribution patterns affect research outcomes.
  - existing research focuses primarily on general team size and diversity (Franceschet and Costantini 2010; Larivière et al. 2014; AlShebli et al. 2024; Yang, Ding, and Liu 2024; L. Liu et al. 2021; Naik et al. 2023)
  - software engineering literature shows correlation between team size and code quality (many eyes make all bugs shallow) (Wyss, De Carli, and Davidson 2023; Meirelles et al. 2010)
  - limited understanding of how code contribution is associated to research impact
  - need to understand relationship between code contributors and citation metrics to understand the value of these technical, potentially uncredited, contributions
  - we believe that more code contributors may signal a more technical research project and that technical complexity may be rewarded with more citations
- (H2) Despite formal taxonomies like CRediT attempting to standardize contribution recognition, the criteria for granting authorship to technical contributors remain inconsistent and poorly understood across research communities.
  - existing contribution frameworks provide definitions for software development roles (K. Li, Zhang, and Larivière 2023; Ding et al. 2021)
  - however, these frameworks may not capture the full spectrum of technical contributions
  - repository histories allow us to examine how sustained technical engagement relates to authorship status (Ram 2013)
  - we believe that longer project involvement increases likelihood of authorship recognition

  - specifically, we hypothesize that projects with longer durations will show higher proportions of author-developers compared to non-author developers

- (H3 and H4) Academic authorship conventions signal both intellectual contribution and project responsibilities, yet their relationship to software development remains poorly understood.

  - first authors traditionally responsible for primary intellectual and experimental contributions (Larivière et al. 2016; Larivière, Pontille, and Sugimoto 2020; Júnior et al. 2016)
  - corresponding authors serve as primary points of contact and often maintain research artifacts
  - varying expectations across academic disciplines regarding technical contributions (E. Smith 2023)
  - limited research examining how these authorship roles relate to direct code contributions
  - potential insights into how software development responsibilities are distributed within research teams
  - we believe that first authors and corresponding authors will have higher proportions of code contribution than not.
  - conversely, middle and last authors and non-corresponding authors will have lower proportions of code contribution than not.

- (H5) Academic career advancement has historically depended on traditional impact metrics, creating potential tension for researchers who dedicate significant time to software development.

  - time invested in code development may reduce traditional scholarly output as code development is a time consuming activity (Springmeyer, Blattner, and Max 1992; Goodman et al. 2014)
  - potential career implications for researchers who prioritize coding (Hannay et al. 2009; Heroux 2022; A. M. Smith, Norman, and Cruz 2019; Cosden, McHenry, and Katz 2022)
  - need to understand relationship between code contributions and academic impact
  - we believe that code contributing researchers will have lower individual level impact metrics than non-coding researchers, likely due to a combination of lack of recognition and authorship credit as well as reduced time for traditional scholarly output

- Understanding these relationships is crucial for developing equitable academic credit systems that recognize the full spectrum of research contributions.

  - findings will inform policy making around academic credit
  - importance of large-scale quantitative evidence for understanding current credit systems
  - implications for academic hiring and promotion decisions

- potential to develop new impact metrics that capture software contributions

4 Data and Methods

4.1 Linking Scientific Articles and Source Code Repositories

- Modern scientific research increasingly requires the public sharing of research code, creating unique opportunities to study the relationship between academic authorship and software development.
  - many journals and platforms now require or recommend code and data sharing (Stodden, Guo, and Ma 2013; Sharma et al. 2024)
  - this requirement creates traceable links between publications and code
  - these links enable systematic study of both article-repository and author-developer relationships (Hata et al. 2021; Kelley and Garijo 2021; Stankovski and Garijo 2024; Milewicz, Pinto, and Rodeghero 2019)
- Our data collection process leverages multiple complementary sources of linked scientific articles and code repositories to ensure comprehensive coverage.
  - PLOS: Traditional research articles with code requirements
  - JOSS and SoftwareX: Specialized software-focused publications
  - Papers with Code / ArXiv: Capturing pre-print landscape
  - to reduce the complexity of dataset processing and enrichment, we filter out any article-source-code-repository pairs which store code somewhere other than GitHub, note: we do this for simplicity of processing but recognize that there is work elsewhere to understand code outside of GitHub (Trujillo, Hébert-Dufresne, and Bagrow 2022)
- Through integration of multiple data sources, we extract detailed information about both the academic and software development aspects of each project.
  - specifically we utilize the Semantic Scholar API for article DOI resolution to ensure that we find the latest version for each article.
  - this is particularly important for working with preprints as they may have been published in a journal since their inclusion in the Papers with Code dataset
  - we then utilize the OpenAlex API to gather publication metadata (i.e. open access status, domain, publication date), author details (i.e. name, author position, corresponding author status), and article- and individual-level metrics (i.e. citation count, FWCI, h-index).
  - the GitHub API provides similar information for source code repositories, including repository metadata (i.e. name, description, languages, creation date), contributor details (i.e. username, name, email), and repository-level metrics (i.e. star count, fork count, issue count).

- while the majority of our data is sourced from Papers with Code, our additional collection from PLOS, JOSS, and SoftwareX as well as the enrichment from GitHub and OpenAlex together form one of the largest collections of linked, metadata enriched, datasets of paired scientific articles and associated source code repositories.
  - in total, we collect and enrich data for 163292 article-repository pairs

4.2 A Predictive Model for Matching Article Authors and Source Code Contributors

4.2.1 Annotated Dataset Creation

- The development of an accurate author-developer matching model requires high-quality labeled training data that captures the complexity of real-world identity matching.
  - entity matching between authors and developers is non-trivial
  - multiple forms of name variation and incomplete information
  - add figure showing example matches/non-matches
- We developed an annotation process to create a robust training dataset while maximizing efficiency and accuracy.
  - focus on JOSS articles to increase positive match density
  - we create author-developer pairs for annotation by creating all possible combinations of authors and developers within a single JOSS article-repository pair
  - we take a random sample of 3000 pairs from the full set and have two independent annotators label each
  - after all 3000 pairs are annotated, we resolve any disagreements between the two annotators
- The resulting annotated dataset provides a comprehensive foundation for training our predictive model while highlighting common patterns in author-developer identity matching.
  - after resolution of all annotated pairs, our annotated dataset contains 451 (15.0%) positive and 2548 (85.0%) negative author-developer-account pairs
  - there are 2027 unique authors and 2733 unique developer accounts within this annotated set
  - however, not all developer accounts contain complete information, in our set 2191 (80.2%) have associated names and 839 (30.7%) have associated emails

4.2.2 Training and Evaluation

- To optimize our predictive model for author-contributor matching, we evaluate a variety of Transformer-based base models and input features.

- specifically, we fine-tune from three different base transformer models:
    - deberta-v3-base (He, Gao, and Chen 2021; He et al. 2021)
    - bert-base-multilingual-cased (Devlin et al. 2018)
    - distilbert-base-uncased (Sanh et al. 2019)
- these three models are all variations or built-upon BERT, and while significant time has passed since BERT was first introduced, BERT-based models remain a strong base for many NLP tasks across a number of domains while being relatively "small" compared to the much larger decoder transformer relatives (GPT, Llama, etc.) (Tran et al. 2024; Yu et al. 2024; Jeong and Kim 2022)

- We employed a systematic evaluation to identify optimal combination of base models and input features.
    - first, to ensure that there was no data leakage, we split our dataset into training and test sets
    - specifically, we created two random sets of 10% of all unique authors and 10% of all unique developers, any pairs containing either the author or developer were placed into the test set
    - in doing so, we ensured that the model was never trained on any author or developer information later used for evaluation
    - due to the fact that each author and developer-account can be included in multiple annotated pairs, our final training set contains 2442 (81.4%) and our test set contains 557 (18.6%) author-developer-account pairs
    - we fine-tuned each of our three base models using all combinations of available developer-account features, from including only the developer account username to including the developer's username, name, and email.
    - to avoid overfitting and ensure generalizability, we fine-tuned each of the base models for only a single training epoch.
    - model evaluation was performed using standard classification metrics, including accuracy, precision, recall, and F1 score

- After extensive model comparison we find that fine-tuning from Microsoft's deberta-v3-base and including the developer's username and name achieves the best performance for author-developer matching.
    - our best model achieves a binary F1 score of 0.944, with an accuracy of 0.984, precision of 0.938, and recall of 0.95 (see Figure 1 for a confusion matrix of model predictions on the test set).
    - analysis of feature importance
        - note that the addition of developer's name has a "larger effect" on model performance but that could simply be because of how many more developers have a name available than an email
        - also note that there is a model that performs just as well as this one using bert-multilingual and includes the developers email however we choose to use the

deberta and name only version for its simplicity as well as the fact that deberta is a much more recently developed and released model which was pre-trained on a much larger dataset.

    * considering that in most cases, deberta out-performs bert-multilingual, we believe that while the overall evaluation metrics between the top two performing models are the same, the deberta based model will generalize to other unseen data better than the bert-multilingual model

  – all model and feature set combination results are available in Table 10

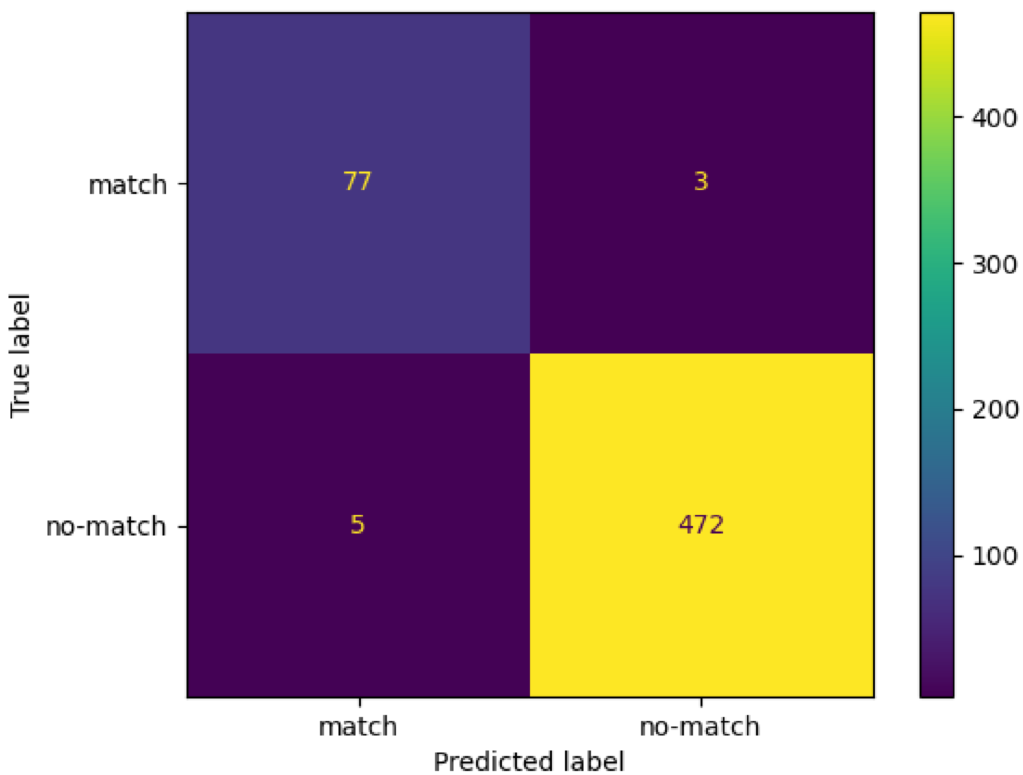


Figure 1: Confusion Matrix Produced From Evaluation of Best Performing Model (deberta-v3 with developer username, developer name, and author name).

- To enable future research, we have made our trained model and supporting application library publicly available.
  – Python library implementation: sci-soft-models
  – HuggingFace model deployment: dev-author-em-clf

4.3 Linking Authors and GitHub Developer Accounts

- Our trained entity-matching model enables comprehensive identification of author-developer relationships while accounting for the complex realities of academic software development practices.
  - in practice, to complete our dataset, we apply our trained model to all possible author and developer-account combinations within each article-repository pair
  - The presence of multiple developer accounts per individual reflects common practices in academic software development that must be accommodated in our analysis.
  - developers often maintain separate accounts for different projects or institutions
  - account transitions are common as researchers move between roles
- Further, while our model performs well overall, we note that there are some limitations to our approach.
  - in most cases predictions are trivial due to minor differences in text (spelling of author name to username)
  - however we do observe a few cases in which our model may not perform as well
  - namely, shorter names, articles and repositories which have contributors with the same last name (i.e. siblings or other relationship), and "organization" accounts (i.e. research lab GitHub accounts used for management, administration, and documentation or a project)
  - TODO: should we take a sample and estimate how widespread these problems are?
  - we include appropriate filtering during analysis to ensure that we do not include author-developer pairs which are unlikely to be the same individual
- Our final dataset provides unprecedented scale and scope for analyzing the relationship between academic authorship and software development contributions.
  - Specifically, our dataset contains 138596 article-repository pairs, 295806 distinct authors, and 152170 distinct developer accounts.
  - a detailed breakdown of these counts by data source, domain, document type, and open access status is available in Table 1

5 Preliminary Analysis Code Contributor Authorship and Development Dynamics of Research Teams

- To enrich our pre-existing dataset, we apply our trained predictive model across pairs of authors and developer accounts.
  - again, these pairs are all combinations of author and developer account within an individual paper
  - specifics, how many unique author-developer account pairs are we able to find

Table 1: Counts of Article-Repository Pairs, Authors, and Developers by Data Sources, Domains, Document Types, and Access Status.

Category	Subset	Article-Repository Pairs	Authors	Developers
By Domain	Physical Sciences	116600	240545	130592
	Social Sciences	8838	29269	14043
	Life Sciences	7729	31649	12150
	Health Sciences	5172	25979	7248
By Document Type	preprint	72177	170301	87311
	research article	63528	173183	78935
	software article	2891	9294	12868
By Access Status	Open	132856	286874	147831
	Closed	5740	23668	9352
By Data Source	pwc	129615	262889	134926
	plos	6090	30233	8784
	joss	2336	7105	11362
	softwarex	555	2244	1628
Total		**138596**	**295806**	**152170**

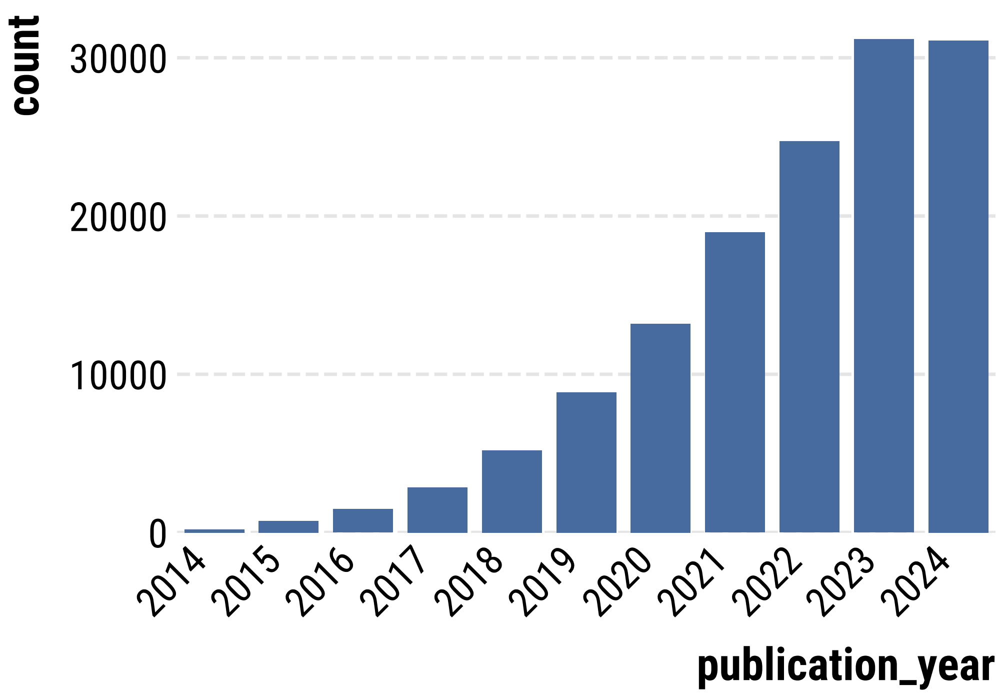


Figure 2: Number of articles by publication year. Only publication years with 100 or more articles are included.

  - we find 108754 unique author-developer account pairs
  - we next use this enriched dataset to understand software development dynamics within research teams, and characterize the authors who are and who aren't code contributors.

5.1 Software Development Dynamics Within Research Teams

- We begin by measuring the distributions of different coding and non-coding contributors across all of the article-code-repository pairs within our dataset.
  - individuals in our dataset can fall into three categories:
    * Code-Contributing Authors (CC-A): authors for which our model predicted a match with at least one developer account which contributed code to the associated repository
    * Non-Code-Contributing Authors (NCC-A): authors for which our model did not predict any matches with developer accouts which contributed code to the associated repository
    * Code-Contributing Non-Authors (CC-NA): developer accounts which contributed code to the associated repository but were not predicted to be a match with any author
  - within our dataset, we find that papers on average have 4.9 ± 1.9 total authors, 3.9 ± 2.0 non-code-contributing authors, 1.0 ± 0.7 code-contributing authors, and 0.4 ± 1.7 code-contributing non-authors (see Figure 3 for a visualization of these distributions).
  - Table 2 provides a detailed breakdown of these distributions by domain, article type, and open access status.
  - Our finding of, on average, only a single code-contributing author on a paper, is similar to previous work in understanding distributions of labor in knowledge production from Larivière, Pontille, and Sugimoto (2020) which found that the CRediT tasks of "Data Curation", "Formal Analysis", "Visualization", and "Software" were all predominantly performed by a first author.
  - However, we do also see that code-contributing non-authors are present in the data, albeit, on average, with less than one code-contributing non-author per article.

- Next we investigate how these distributions have changed over time and how they change by the total number of authors on a paper.
  - Figure 4 shows thats both, the mean number of code-contributing authors and code-contributing non-authors have remained relatively stable over time.
  - Figure 5 similarly shows that the mean number of code-contributing authors and code-contributing non-authors remains relatively stable across different team sizes.

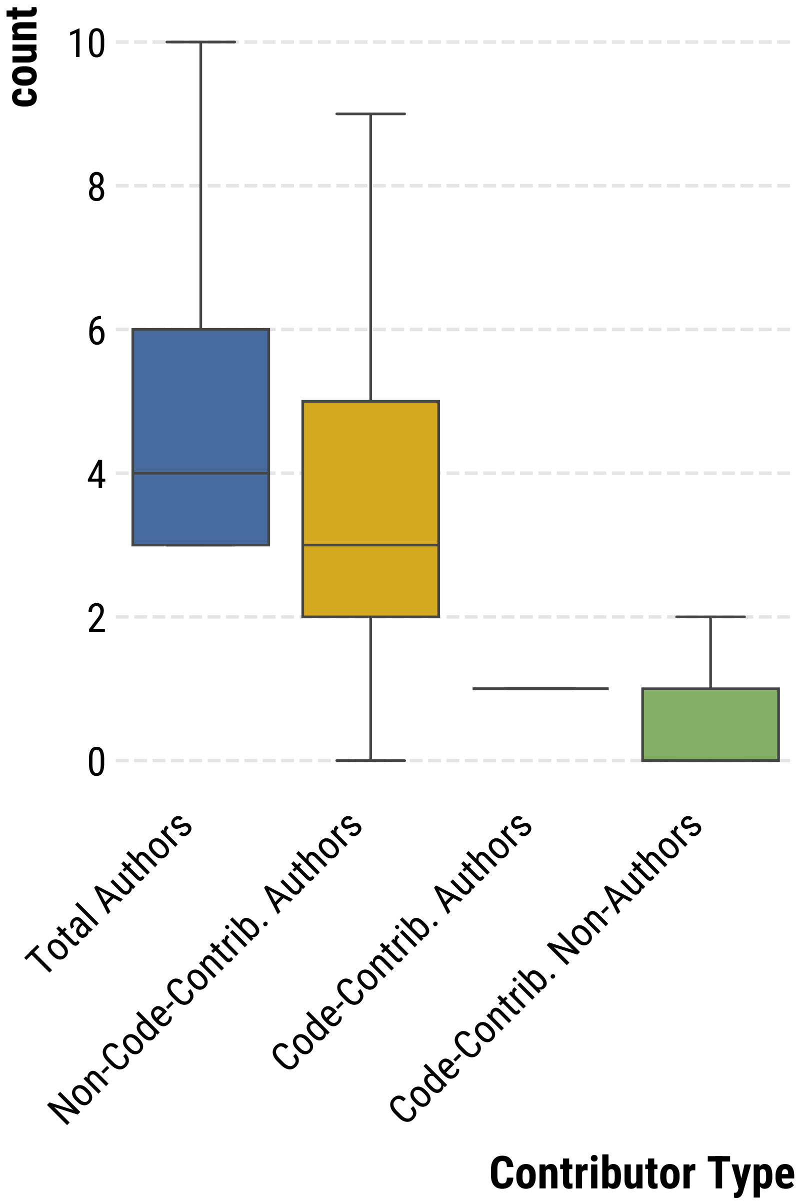


Figure 3: Distribution of the number of Total Authors, Non-Code-Contributing Authors, Code-Contributing Authors, and Code-Contributing Non-Authors across all article-repository pairs. Only includes article-repository pairs with a most recent commit no later than 90 days after publication.

Table 2: Mean and Standard Deviation of Non-Code-Contributing Authors (NCC-A), Code-Contributing Authors (CC-A), and Code-Contributing Non-Authors (CC-NA) Research Team Members by Domain, Article Type, and Open Access Status. Only includes research teams from article-repository pairs with a most recent commit no later than 90 days after publication and excludes research teams which are in the top 3% of total author sizes.

Control	Subset	Total Authors	NCC-A	CC-A	CC-NA
OA Status	Closed	5.1 ± 1.9	4.0 ± 1.9	1.1 ± 0.7	0.5 ± 2.1
	Open	4.9 ± 1.9	3.9 ± 2.0	1.0 ± 0.7	0.4 ± 1.7
Domain	Health Sciences	6.1 ± 2.5	5.1 ± 2.5	1.0 ± 0.6	0.4 ± 1.2
	Life Sciences	5.2 ± 2.1	4.2 ± 2.2	1.0 ± 0.7	0.4 ± 1.2
	Physical Sciences	4.8 ± 1.8	3.8 ± 1.9	1.0 ± 0.7	0.5 ± 1.8
	Social Sciences	4.5 ± 1.7	3.5 ± 1.8	1.1 ± 0.7	0.3 ± 1.1
Article Type	preprint	4.8 ± 1.8	3.8 ± 1.9	1.1 ± 0.7	0.5 ± 2.2
	research article	4.9 ± 1.9	3.9 ± 2.0	1.0 ± 0.7	0.4 ± 1.6
	software article	4.7 ± 1.9	3.2 ± 1.9	1.5 ± 1.4	0.9 ± 1.1

  - these descriptive statistics suggest that code-contributing non-authors are an inconsistent feature of research teams, and that at the very least, their exclusion from authorship is both, not a recent phenomenon and does not appear to be getting worse over time.

5.1.1 Modeling Citations

- We model an article's total citations by the coding contributorship of the research team and controlled by a number of different factors.
- Each control variable is modeled separately and the results are presented in the tables below:
  - Controlling for article open access status: Table 3
  - Controlling for article domain: Table 4
  - Controlling for article type: Table 5

5.1.1.1 Open Access Status

- open access articles
  - n_author_devs is not significant
  - gain 1.1621828452888712 more citations per non-author code contributor
- closed access articles

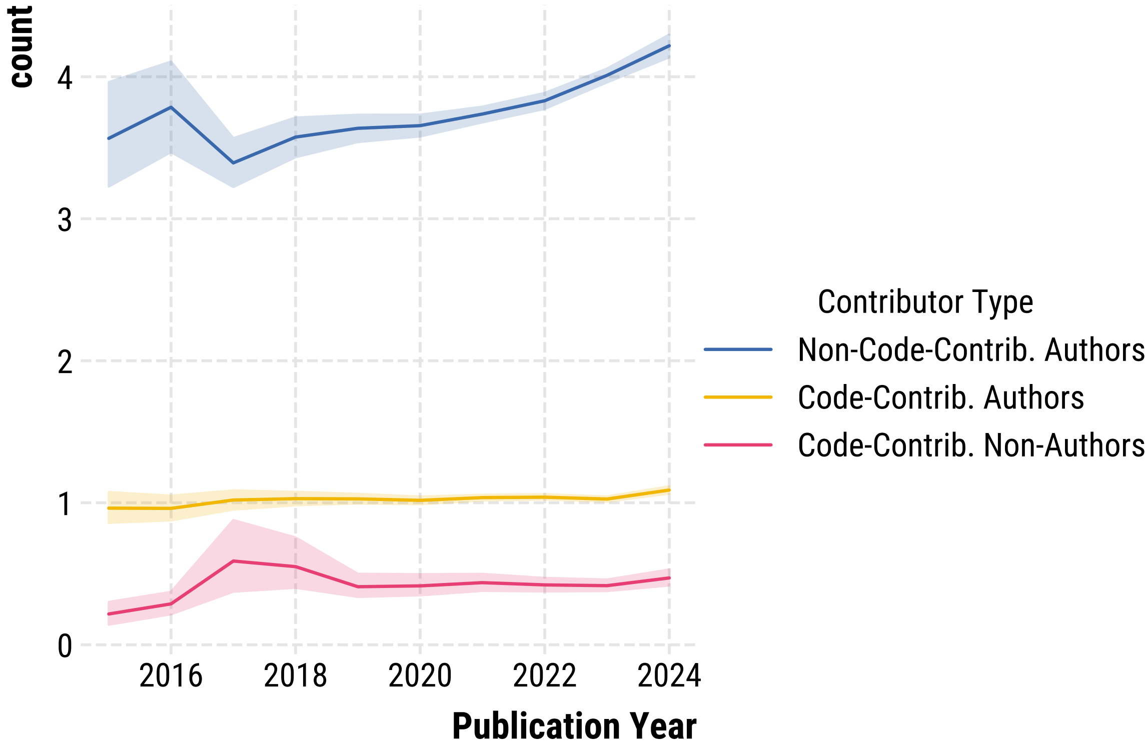


Figure 4: Mean number of Non-Code-Contributing Authors, Code-Contributing Authors, and Code-Contributing Non-Authors over time. Only includes article-repository pairs with a most recent commit no later than 90 days after publication and excludes research-teams which are in the top 3% of total author sizes for publication years with 50 or more articles.

Table 3: Article Citations by Code Contributorship of Research Team Controlled by Open Access Status

Variable	coef	P>|z|	[0.025	0.975]
const ***	**0.61**	**0.00**	**0.49**	**0.73**
n authors ***	**0.07**	**0.00**	**0.06**	**0.08**
n author devs	0.05	0.24	-0.04	0.14
n non author devs	0.00	0.96	-0.03	0.03
years since publication ***	**0.38**	**0.00**	**0.37**	**0.39**
is open access ***	**0.42**	**0.00**	**0.30**	**0.54**
n author devs * is open access	-0.01	0.82	-0.10	0.08
n non author devs * is open access	-0.00	0.94	-0.03	0.03

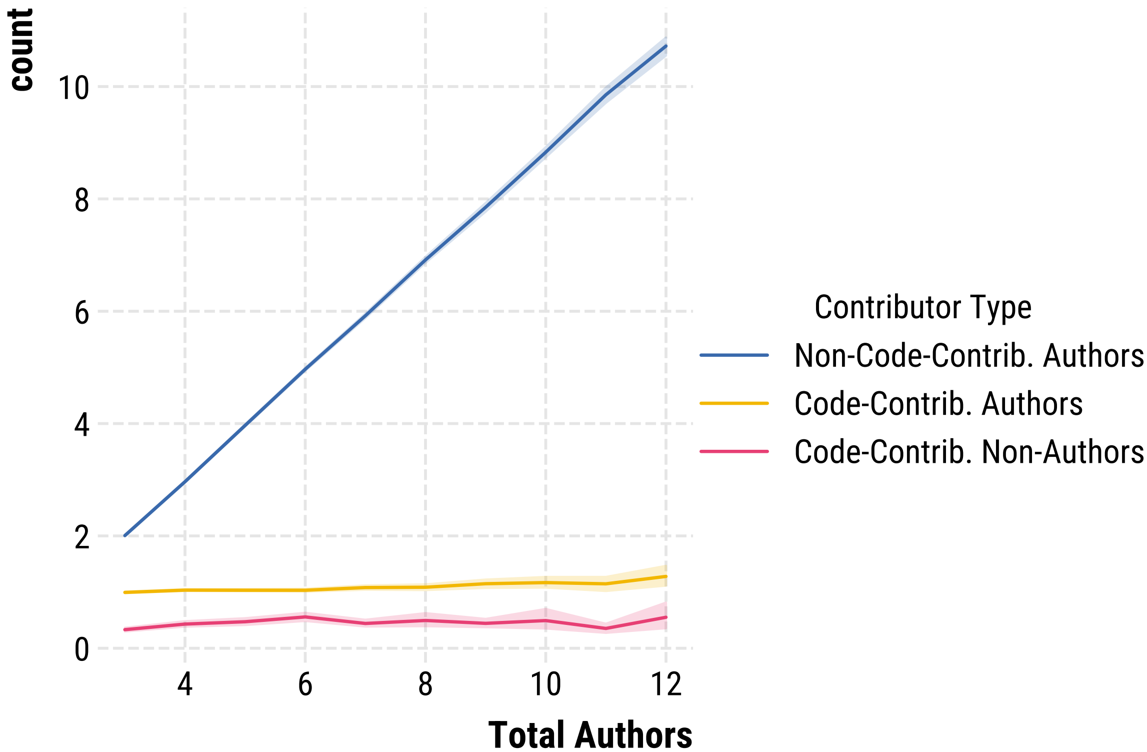


Figure 5: Mean number of Non-Code-Contributing Authors, Code-Contributing Authors, and Code-Contributing Non-Authors by Total Authors. Only includes article-repository pairs with a most recent commit no later than 90 days after publication and excludes research-teams which are in the top 3% of total author sizes.

Table 4: Article Citations by Code Contributorship of Research Team Controlled by Domain

Variable	coef	P>|z|	[0.025	0.975]
const ***	**0.88**	**0.00**	**0.75**	**1.01**
n authors ***	**0.07**	**0.00**	**0.06**	**0.08**
n author devs	0.03	0.60	-0.08	0.13
n non author devs	0.01	0.61	-0.04	0.07
years since publication ***	**0.40**	**0.00**	**0.39**	**0.41**
domain Life Sciences ***	**-0.21**	**0.01**	**-0.36**	**-0.06**
domain Physical Sciences ***	**0.14**	**0.03**	**0.01**	**0.26**
domain Social Sciences ***	**-0.18**	**0.03**	**-0.34**	**-0.02**
n author devs * domain Life Sciences	0.07	0.29	-0.06	0.19
n author devs * domain Physical Sciences	0.00	0.93	-0.10	0.11
n author devs * domain Social Sciences	0.10	0.14	-0.03	0.23
n non author devs * domain Life Sciences	-0.04	0.23	-0.11	0.03
n non author devs * domain Physical Sciences	-0.02	0.57	-0.07	0.04
n non author devs * domain Social Sciences	-0.04	0.31	-0.11	0.03

  - n_author_devs is not significant
  - gain 1.1681251177146603 more citations per non-author code contributor

5.1.1.2 Domain

- health sciences
  - gain 1.0989992577120393 more citations per author code contributor compared to no author code contributor health science papers
  - gain 1.0733665310933769 more citations per non-author code contributor compared to no non-author code contributor health science papers
- life sciences
  - n_author_devs is not significant
  - n_non_author_devs is not significant
- physical sciences
  - n_author_devs is not significant
  - gain 1.0836121025480543 more citations per non-author code contributor compared to no non-author code contributor physical science papers
- social sciences
  - n_author_devs is not significant
  - n_non_author_devs is not significant

Table 5: Article Citations by Code Contributorship of Research Team Controlled by Article Type

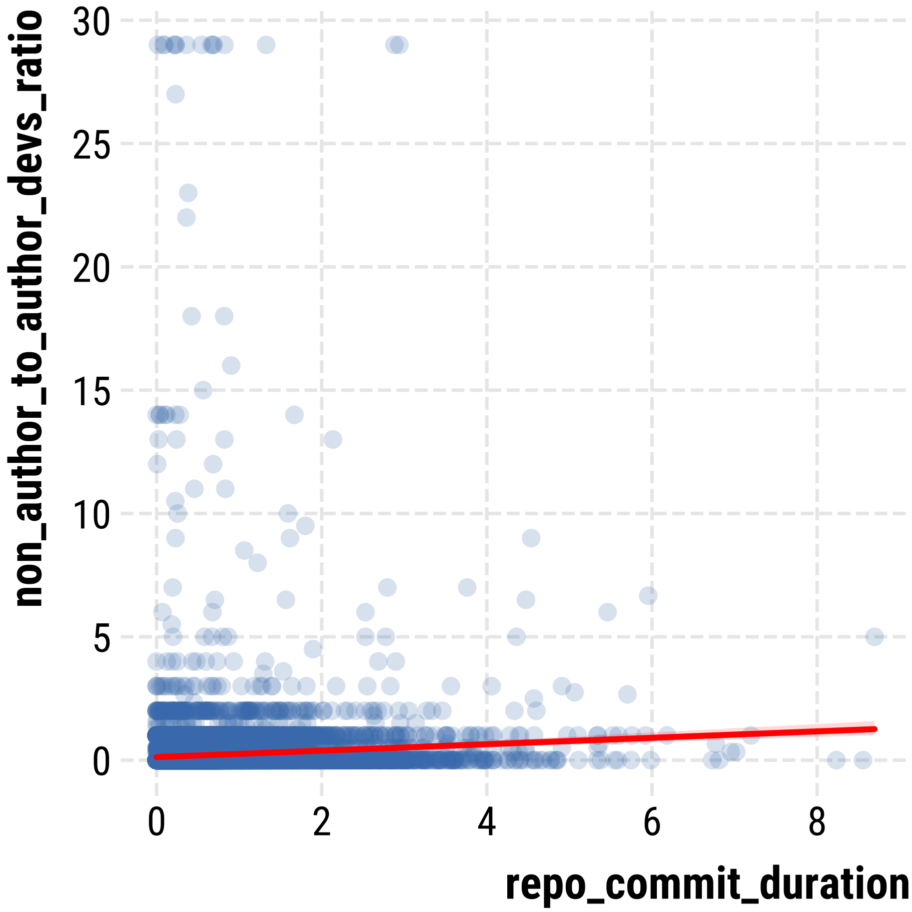
Variable	coef	P>|z|	[0.025	0.975]
const ***	**0.53**	**0.00**	**0.45**	**0.61**
n authors ***	**0.07**	**0.00**	**0.06**	**0.08**
n author devs	-0.03	0.20	-0.09	0.02
n non author devs ***	**-0.03**	**0.00**	**-0.05**	**-0.01**
years since publication ***	**0.40**	**0.00**	**0.39**	**0.41**
article type research article ***	**0.47**	**0.00**	**0.40**	**0.55**
article type software article ***	**-0.47**	**0.00**	**-0.73**	**-0.22**
n author devs * article type research article ***	**0.10**	**0.00**	**0.05**	**0.16**
n author devs * article type software article	-0.06	0.37	-0.19	0.07
n non author devs * article type research article ***	**0.04**	**0.00**	**0.02**	**0.06**
n non author devs * article type software article	0.09	0.24	-0.06	0.24

5.1.1.3 Article Type

- preprint
  - n_author_devs is not significant
  - gain 1.132242296815202 more citations per non-author code contributor compared to no non-author code contributor preprints
- research article
  - gain 1.0934086297167769 more citations per author code contributor compared to no author code contributor research articles
  - gain 1.0381081810730037 more citations per non-author code contributor compared to no non-author code contributor research articles
- software article
  - n_author_devs is not significant
  - n_non_author_devs is not significant

5.1.2 Team Composition and Project Duration

- We next investigate the relationship between the duration of a project and the coding contributorship of the research team.




```
PearsonRResult(statistic=0.07782139893046706, pvalue=5.319679494173922e-23)
```

5.2 Characteristics of Scientific Code Contributors

- Next we investigate the differences between coding and non-coding article authors.
  - specifics, author position in authorship list is a commonly used tool in scientometrics
  - similarly, metrics of "scientific impact" such as h-index, i10 index, and two-year mean citedness are also available to us.
  - plot / table of the distributions between coding and non-coding authors
  - ANOVA / Chi2 tests to see if these differences are significant
  - results in summary
- Just as before, we next investigate if these results are affected by article type and research domain.

- subplot + stats tests for differences by each article type
- subplot + stats tests for differences by each domain
- results in summary

5.2.1 Author Positions of Code Contributing Authors

```
Overall
is_code_contributor   False  True
position
first                 16428  35893
last                  45829   4569
middle               131433  12920
Chi2: 86552.98658551606, p: 0.0, n: 247072
position: first, p: 0.0, statistic: 0.6860151755509261, n: 52321
position: middle, p: 0.0, statistic: 0.0895028160135224, n: 144353
position: last, p: 0.0, statistic: 0.09065835945870868, n: 50398
```

5.2.1.1 Domain

```
Physical Sciences
is_code_contributor   False  True
position
first                 13068  28919
last                  36977   3433
middle               101231  10507
Chi2: 68220.51167362447, p: 0.0, n: 194135
position: first, p: 0.0, statistic: 0.6887608069164265, n: 41987
position: middle, p: 0.0, statistic: 0.09403246881096852, n: 111738
position: last, p: 0.0, statistic: 0.08495421925266024, n: 40410
Life Sciences
is_code_contributor  False  True
position
first                 1309   2586
last                  3293    491
middle               11023    852
Chi2: 6554.038088460125, p: 0.0, n: 19554
position: first, p: 1.3228240814067576e-94, statistic: 0.6639281129653402, n: 3895
position: middle, p: 0.0, statistic: 0.07174736842105263, n: 11875
position: last, p: 0.0, statistic: 0.1297568710359408, n: 3784
Social Sciences
is_code_contributor  False  True
```

```
position
first                1225   2813
last                 3444    411
middle               8228   1021
Chi2: 5715.82059242655, p: 0.0, n: 17142
position: first, p: 2.2055993260095607e-141, statistic: 0.6966319960376424, n: 4038
position: last, p: 0.0, statistic: 0.1066147859922179, n: 3855
position: middle, p: 0.0, statistic: 0.11039031246621256, n: 9249
Health Sciences
is_code_contributor  False  True
position
first                 826   1575
last                 2115    234
middle              10951    540
Chi2: 5998.504729922407, p: 0.0, n: 16241
position: first, p: 1.8308248282746384e-53, statistic: 0.6559766763848397, n: 2401
position: middle, p: 0.0, statistic: 0.04699329910364633, n: 11491
position: last, p: 0.0, statistic: 0.09961685823754789, n: 2349
```

5.2.1.2 Article Type

```
research article
is_code_contributor  False  True
position
first               10141  21940
last                28279   2909
middle              82646   7345
Chi2: 54917.35337329746, p: 0.0, n: 153260
position: first, p: 0.0, statistic: 0.6838938935818709, n: 32081
position: middle, p: 0.0, statistic: 0.08161927303841496, n: 89991
position: last, p: 0.0, statistic: 0.09327305373861741, n: 31188
preprint
is_code_contributor  False  True
position
first                6102  13421
last                17105   1493
middle              47791   5134
Chi2: 31620.70287233051, p: 0.0, n: 91046
position: first, p: 0.0, statistic: 0.6874455770117297, n: 19523
position: last, p: 0.0, statistic: 0.08027744918808474, n: 18598
position: middle, p: 0.0, statistic: 0.09700519603212092, n: 52925
```

```
software article
is_code_contributor  False  True
position
first                  185   532
last                   445   167
middle                 996   441
Chi2: 436.6562463251282, p: 1.5181016986238546e-95, n: 2766
position: first, p: 2.8268609500046396e-39, statistic: 0.7419804741980475, n: 717
position: last, p: 4.070740833733961e-30, statistic: 0.272875816993464, n: 612
position: middle, p: 4.079797536800063e-49, statistic: 0.3068893528183716, n: 1437
```

5.2.1.3 Open Access Status

```
Open Access
is_code_contributor   False  True
position
first                 15323  33256
last                  42466   4290
middle               121386  12002
Chi2: 79588.66090114709, p: 0.0, n: 228723
position: first, p: 0.0, statistic: 0.684575639679697, n: 48579
position: middle, p: 0.0, statistic: 0.08997810897532012, n: 133388
position: last, p: 0.0, statistic: 0.09175293010522714, n: 46756
Closed Access
is_code_contributor  False  True
position
first                 1105   2637
last                  3363    279
middle               10047    918
Chi2: 6990.309755750654, p: 0.0, n: 18349
position: first, p: 2.5647584282503314e-142, statistic: 0.7047033671833244, n: 3742
position: last, p: 0.0, statistic: 0.07660626029654036, n: 3642
position: middle, p: 0.0, statistic: 0.08372093023255814, n: 10965
```

5.2.2 Corresponding Status of Code Contributing Authors

```
Overall
is_code_contributor   False  True
is_corresponding
```

```
Corresponding         20112   8701
Not Corresponding    173578  44681
Chi2: 1421.1208069996246, p: 5.40576546314e-311, n: 247072
is_corresponding: Not Corresponding, p: 0.0, statistic: 0.2047154985590514, n: 218259
is_corresponding: Corresponding, p: 0.0, statistic: 0.301981744351508, n: 28813
```

5.2.2.1 Domain

```
Physical Sciences
is_code_contributor   False  True
is_corresponding
Corresponding          7239   5248
Not Corresponding    144037  37611
Chi2: 3086.5504905422213, p: 0.0, n: 194135
is_corresponding: Not Corresponding, p: 0.0, statistic: 0.2070543028274465, n: 181648
is_corresponding: Corresponding, p: 3.0892901003685015e-71, statistic: 0.42027708817169857, n
Life Sciences
is_code_contributor  False  True
is_corresponding
Corresponding         6247   1772
Not Corresponding     9378   2157
Chi2: 33.80559720544868, p: 6.090338640251236e-09, n: 19554
is_corresponding: Not Corresponding, p: 0.0, statistic: 0.1869960988296489, n: 11535
is_corresponding: Corresponding, p: 0.0, statistic: 0.2209751839381469, n: 8019
Social Sciences
is_code_contributor  False  True
is_corresponding
Corresponding         1655    803
Not Corresponding    11242   3442
Chi2: 95.74853878212862, p: 1.304427435637673e-22, n: 17142
is_corresponding: Corresponding, p: 2.2084447510143967e-67, statistic: 0.32668836452400324, n
is_corresponding: Not Corresponding, p: 0.0, statistic: 0.23440479433396894, n: 14684
Health Sciences
is_code_contributor  False  True
is_corresponding
Corresponding         4971    878
Not Corresponding     8921   1471
Coding by is_corresponding not significant
```

5.2.2.2 Article Type

```
research article
is_code_contributor   False  True
is_corresponding
Corresponding         19623   7716
Not Corresponding    101443  24478
Chi2: 1044.0017267689143, p: 4.898689779565618e-229, n: 153260
is_corresponding: Not Corresponding, p: 0.0, statistic: 0.1943917217938231, n: 125921
is_corresponding: Corresponding, p: 0.0, statistic: 0.2822341709645561, n: 27339
preprint
is_code_contributor  False  True
is_corresponding
Corresponding          264    772
Not Corresponding    70734  19276
Chi2: 1678.8685858544789, p: 0.0, n: 91046
is_corresponding: Not Corresponding, p: 0.0, statistic: 0.2141539828907899, n: 90010
is_corresponding: Corresponding, p: 2.737621889235552e-58, statistic: 0.7451737451737451, n:
software article
is_code_contributor  False  True
is_corresponding
Corresponding          225    213
Not Corresponding     1401    927
Chi2: 11.450236853781284, p: 0.0007148488745836787, n: 2766
is_corresponding: Corresponding, p: 0.5992164561363355, statistic: 0.4863013698630137, n: 438
is_corresponding: Not Corresponding, p: 1.5581829692815375e-22, statistic: 0.398195876288659
```

5.2.2.3 Open Access Status

```
Open Access
is_code_contributor   False  True
is_corresponding
Corresponding         19897   8448
Not Corresponding    159278  41100
Chi2: 1263.1447877833145, p: 1.1546882926708494e-276, n: 228723
is_corresponding: Not Corresponding, p: 0.0, statistic: 0.20511233768178144, n: 200378
is_corresponding: Corresponding, p: 0.0, statistic: 0.29804198271300053, n: 28345
Closed Access
is_code_contributor  False  True
is_corresponding
Corresponding          215    253
Not Corresponding    14300   3581
Chi2: 317.52534860502976, p: 5.011183853047046e-71, n: 18349
```

Table 6: Counts of Researcher Coding Status Used in H5

Control	Subset	Any Coding	Majority Coding	Always Coding	Total
Freq. Author Pos.	first	1689	4952	3322	11444
	last	2413	568	191	10188
	middle	12184	3327	701	31911
Freq. Domain	Health Sciences	345	202	86	1501
	Life Sciences	369	241	129	1436
	Physical Sciences	15289	8179	3821	49311
	Social Sciences	283	225	178	1295
Freq. Article Type	preprint	9572	4963	2214	28948
	research article	6669	3765	1871	24217
	software article	45	119	129	378

```
is_corresponding: Not Corresponding, p: 0.0, statistic: 0.20026844136233993, n: 17881
is_corresponding: Corresponding, p: 0.08709725881534564, statistic: 0.5405982905982906, n: 4
```

5.2.3 Modeling H-Index

- We model an authors total citations by their coding status and controlled by a number of different factors.
- Each control variable is modeled separately and the results are presented in the tables below:
  - Controlling for an authors most frequent author position: Table 7
  - Controlling for an authors most frequent domain: Table 8
  - Controlling for an authors most frequent article type: Table 9

5.2.3.1 Author Position

- first authors
  - any coding has a positive association with citations (1.2636444922077779)
  - majority coding has a negative association with citations (0.8860339595928756)
  - always coding has a negative association with citations (0.764143255648199)
- middle authors
  - any coding has a negative association with citations (0.7482635675785653)
  - majority coding has a negative association with citations (0.9039330328858641)
  - always coding is not significant

Table 7: Researcher H-Index by Coding Status Controlled by Most Freq. Author Position

Variable	coef	P>|z|	[0.025	0.975]
const ***	**2.38**	**0.00**	**2.31**	**2.44**
works count ***	**0.00**	**0.00**	**0.00**	**0.00**
any coding ***	**0.14**	**0.00**	**0.05**	**0.22**
majority coding	-0.07	0.08	-0.14	0.01
always coding ***	**-0.24**	**0.00**	**-0.33**	**-0.16**
common author position last ***	**1.05**	**0.00**	**0.98**	**1.11**
common author position middle ***	**0.74**	**0.00**	**0.68**	**0.81**
any coding * common author position last ***	**-0.28**	**0.00**	**-0.37**	**-0.19**
any coding * common author position middle ***	**-0.45**	**0.00**	**-0.53**	**-0.36**
majority coding * common author position last ***	**-0.35**	**0.00**	**-0.45**	**-0.26**
majority coding * common author position middle ***	**-0.61**	**0.00**	**-0.69**	**-0.52**
always coding * common author position last ***	**-0.37**	**0.00**	**-0.51**	**-0.22**
always coding * common author position middle ***	**-0.51**	**0.00**	**-0.64**	**-0.38**

- last authors
  - any coding has a negative association with citations (0.8737159116880344)
  - majority coding is not significant
  - always coding is not significant

In general, any coding has a positive association while majority and always coding have negative associations with citations. “The more you code the less you are cited” – granted that we don’t have a lot of data for always coding authors (which itself backs up qual lit).

in general, coding is associated with about a ~10 - 30% decrease in citations for a number of conditions when compared to non-coding first authors.

5.2.3.2 Domain

- health sciences
  - any coding is not significant
  - majority coding has a negative association with citations (0.7527666447061963)
  - always coding has a negative association with citations (0.6120140740013499)
- life sciences
  - any coding is not significant
  - majority coding is not significant
  - always coding has a positive association with citations (1.5983949987546404)

Table 8: Researcher H-Index by Coding Status Controlled by Most Freq. Domain

Variable	coef	P>|z|	[0.025	0.975]
const ***	**3.32**	**0.00**	**3.29**	**3.36**
works count ***	**0.00**	**0.00**	**0.00**	**0.00**
any coding ***	**-0.39**	**0.00**	**-0.48**	**-0.31**
majority coding ***	**-1.45**	**0.00**	**-1.65**	**-1.25**
always coding ***	**-1.22**	**0.00**	**-1.58**	**-0.86**
common domain Life Sciences ***	**0.10**	**0.00**	**0.05**	**0.15**
common domain Physical Sciences ***	**-0.15**	**0.00**	**-0.18**	**-0.11**
common domain Social Sciences ***	**-0.16**	**0.00**	**-0.22**	**-0.10**
any coding * common domain Life Sciences	0.11	0.07	-0.01	0.22
any coding * common domain Physical Sciences	0.08	0.09	-0.01	0.17
any coding * common domain Social Sciences	0.01	0.94	-0.14	0.15
majority coding * common domain Life Sciences ***	**0.81**	**0.00**	**0.58**	**1.04**
majority coding * common domain Physical Sciences ***	**0.70**	**0.00**	**0.50**	**0.90**
majority coding * common domain Social Sciences ***	**0.72**	**0.00**	**0.46**	**0.98**
always coding * common domain Life Sciences	0.33	0.12	-0.08	0.74
always coding * common domain Physical Sciences	0.25	0.18	-0.12	0.62
always coding * common domain Social Sciences	0.30	0.17	-0.13	0.73

- physical sciences
  - any coding is not significant
  - majority coding is not significant
  - always coding is not significant
- social sciences
  - any coding has a negative associate with citations (0.8033217181536265)
  - majority coding is not significant
  - always coding has a positive association with citations (1.587245303225596)

We couldn't significant results for a majority of the groupings so I don't think that we can say anything too strongly, however, health sciences doesn't seem to favor coding (which I think is inline most with "RSE" dynamics). Social sciences is split, and social science is broad so this may be a "qual" vs "quant" split, if you include mixed methods researchers in there its hard to parse.

Physical sciences being entirely non-significant is interesting because a majority of our data comes from Physical sciences. this could indicate that coding is just a part of the culture and doesn't have a significant impact on citations (which is inline with CS being the bulk of our physical sciences data).

Table 9: Researcher H-Index by Coding Status Controlled by Most Freq. Article Type

Variable	coef	P>|z|	[0.025	0.975]
const ***	**3.10**	**0.00**	**3.09**	**3.11**
works count ***	**0.00**	**0.00**	**0.00**	**0.00**
any coding ***	**-0.30**	**0.00**	**-0.32**	**-0.27**
majority coding ***	**-0.77**	**0.00**	**-0.81**	**-0.73**
always coding ***	**-0.99**	**0.00**	**-1.07**	**-0.92**
common article type research article ***	**0.18**	**0.00**	**0.17**	**0.20**
common article type software article ***	**0.22**	**0.00**	**0.12**	**0.33**
any coding * common article type research article	-0.02	0.15	-0.05	0.01
any coding * common article type software article	0.19	0.06	-0.01	0.39
majority coding * common article type research article	0.01	0.80	-0.05	0.06
majority coding * common article type software article ***	**0.36**	**0.00**	**0.19**	**0.54**
always coding * common article type research article	0.00	0.93	-0.10	0.10
always coding * common article type software article ***	**0.37**	**0.00**	**0.15**	**0.58**

5.2.3.3 Article Type

- preprint
  - any coding is not significant
  - majority coding has a negative association with citations (0.6736800392488677)
  - always coding has a negative association with citations (0.6306526773980542)
- research article
  - any coding is not significant
  - majority coding has a positive association with citations (1.0565406146754943)
  - always coding has a positive association with citations (1.0843708965667604)
- software article
  - any coding has a positive association with citations (1.7471746543074462)
  - majority coding is not significant
  - always coding has a positive association with citations (1.4681454416819895)

Preprints (from arXiv) have a negative association with coding generally. Research articles have a very slim positive association with coding. Software articles have a strongly positive association with coding (unsuprising).

Table 10: Comparison of Models for Author-Developer-Account Matching

Optional Feats.	**Model**	**Accuracy**	**Precision**	**Recall**	**F1**
name	deberta	0.984	0.938	0.950	0.944
name, email	bert-multilingual	0.984	0.938	0.950	0.944
name, email	deberta	0.982	0.907	0.975	0.940
name	bert-multilingual	0.982	0.938	0.938	0.938
name	distilbert	0.978	0.936	0.912	0.924
name, email	distilbert	0.978	0.936	0.912	0.924
email	deberta	0.957	0.859	0.838	0.848
email	bert-multilingual	0.950	0.894	0.738	0.808
n/a	deberta	0.946	0.847	0.762	0.803
n/a	bert-multilingual	0.941	0.862	0.700	0.772
n/a	distilbert	0.856	0.000	0.000	0.000
email	distilbert	0.856	0.000	0.000	0.000

6 Appendix

6.1 Full Comparison of Models and Optional Features for Author-Developer-Account Matching

References

AlShebli, Bedoor, Shahan Ali Memon, James A Evans, and Talal Rahwan. 2024. "China and the US Produce More Impactful AI Research When Collaborating Together." *Scientific Reports* 14 (1): 28576.

Biagioli, Mario, and Peter Galison. 2014. *Scientific Authorship: Credit and Intellectual Property in Science*. Routledge.

Brand, Amy, Liz Allen, Micah Altman, Marjorie Hlava, and Jo Scott. 2015. "Beyond Authorship: Attribution, Contribution, Collaboration, and Credit." *Learned Publishing* 28 (2).

Brunner, Ursin, and Kurt Stockinger. 2020. "Entity Matching with Transformer Architectures - a Step Forward in Data Integration." In *International Conference on Extending Database Technology*.

Cao, Hancheng, Jesse Dodge, Kyle Lo, Daniel A. McFarland, and Lucy Lu Wang. 2023. "The Rise of Open Science: Tracking the Evolution and Perceived Value of Data and Methods Link-Sharing Practices." *ArXiv* abs/2310.03193.

Carver, Jeffrey C., Nic Weber, Karthik Ram, Sandra Gesing, and Daniel S. Katz. 2022. "A Survey of the State of the Practice for Research Software in the United States." *PeerJ Computer Science* 8.

Cosden, Ian A., Kenton McHenry, and Daniel S. Katz. 2022. "Research Software Engineers: Career Entry Points and Training Gaps." *Computing in Science & Engineering* 24 (6): 14–21. https://doi.org/10.1109/MCSE.2023.3258630.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." *CoRR* abs/1810.04805. http://arxiv.org/abs/1810.04805.

Ding, Jingda, Chao Liu, Qiao Zheng, and Wei Cai. 2021. "A New Method of Co-Author Credit Allocation Based on Contributor Roles Taxonomy: Proof of Concept and Evaluation Using Papers Published in PLOS ONE." *Scientometrics* 126: 7561–81. https://api.semanticscholar.org/CorpusID:237206568.

Edwards, Paul N, Steven J Jackson, Melissa K Chalmers, Geoffrey C Bowker, Christine L Borgman, David Ribes, Matt Burton, and Scout Calvert. 2013. "Knowledge Infrastructures: Intellectual Frameworks and Research Challenges."

Fan, Jianqing, Fang Han, and Han Liu. 2014. "Challenges of Big Data Analysis." *National Science Review* 1 (2): 293–314. https://doi.org/10.1093/nsr/nwt032.

Franceschet, Massimo, and Antonio Costantini. 2010. "The Effect of Scholar Collaboration on Impact and Quality of Academic Papers." *J. Informetrics* 4: 540–53. https://api.semanticscholar.org/CorpusID:17315541.

Goodman, Alyssa, Alberto Pepe, Alexander W Blocker, Christine L Borgman, Kyle Cranmer, Merce Crosas, Rosanne Di Stefano, et al. 2014. "Ten Simple Rules for the Care and Feeding of Scientific Data." *PLoS Computational Biology* 10 (4): e1003542.

Gøtzsche, Peter C, Asbjørn Hróbjartsson, Helle Krogh Johansen, Mette T Haahr, Douglas G Altman, and An-Wen Chan. 2007. "Ghost Authorship in Industry-Initiated Randomised Trials." *PLoS Medicine* 4 (1): e19.

Haak, Laurel L, Martin Fenner, Laura Paglione, Ed Pentz, and Howard Ratner. 2012. "ORCID: A System to Uniquely Identify Researchers." *Learned Publishing* 25 (4): 259–64.

Haeussler, Carolin, and Henry Sauermann. 2013. "Credit Where Credit Is Due? The Impact of Project Contributions and Social Factors on Authorship and Inventorship." *Research Policy* 42 (3): 688–703. https://doi.org/10.1016/j.respol.2012.09.009.

Hampton, Stephanie E, Carly A Strasser, Joshua J Tewksbury, Wendy K Gram, Amber E Budden, Archer L Batcheller, Clifford S Duke, and John H Porter. 2013. "Big Data and the Future of Ecology." *Frontiers in Ecology and the Environment* 11 (3): 156–62.

Hannay, Jo Erskine, Carolyn MacLeod, Janice Singer, Hans Petter Langtangen, Dietmar Pfahl, and Greg Wilson. 2009. "How Do Scientists Develop and Use Scientific Software?" In *2009 ICSE Workshop on Software Engineering for Computational Science and Engineering*, 1–8. https://doi.org/10.1109/SECSE.2009.5069155.

Hasselbring, Wilhelm, Stephan Druskat, Jan Bernoth, Philine Betker, Michael Felderer, Stephan Ferenz, Anna-Lena Lamprecht, Jan Linxweiler, and Bernhard Rumpe. 2024. "Toward Research Software Categories." https://arxiv.org/abs/2404.14364.

Hata, Hideaki, Jin L. C. Guo, Raula Gaikovina Kula, and Christoph Treude. 2021. "Science-Software Linkage: The Challenges of Traceability Between Scientific Knowledge and Software Artifacts." *ArXiv* abs/2104.05891.

He, Pengcheng, Jianfeng Gao, and Weizhu Chen. 2021. "DeBERTaV3: Improving De-

BERTa Using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing." https://arxiv.org/abs/2111.09543.

He, Pengcheng, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. "DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION." In *International Conference on Learning Representations*. https://openreview.net/forum?id=XPZIaotutsD.

Heroux, Michael A. 2022. "Research Software Science: Expanding the Impact of Research Software Engineering." *Computing in Science & Engineering* 24 (6): 22–27. https://doi.org/10.1109/MCSE.2023.3260475.

Howison, James, Ewa Deelman, Michael J. McLennan, Rafael Ferreira da Silva, and James D. Herbsleb. 2015. "Understanding the Scientific Software Ecosystem and Its Impact: Current and Future Measures." *Research Evaluation* 24 (4): 454–70. https://doi.org/10.1093/reseval/rvv014.

Jeong, Yuna, and Eunhui Kim. 2022. "Scideberta: Learning Deberta for Science Technology Documents and Fine-Tuning Information Extraction Tasks." *IEEE Access* 10: 60805–13.

Jin, Xiaolong, Benjamin W Wah, Xueqi Cheng, and Yuanzhuo Wang. 2015. "Significance and Challenges of Big Data Research." *Big Data Research* 2 (2): 59–64.

Júnior, Edilson Anselmo Corrêa, Filipi Nascimento Silva, Luciano da Fontoura Costa, and Diego Raphael Amancio. 2016. "Patterns of Authors Contribution in Scientific Manuscripts." *J. Informetrics* 11: 498–510. https://api.semanticscholar.org/CorpusID:9967627.

Katz, Daniel S., Neil P. Chue Hong, Tim Clark, August Muench, Shelley Stall, Daina R. Bouquin, Matthew Cannon, et al. 2020. "Recognizing the Value of Software: A Software Citation Guide." *F1000Research* 9.

Kelley, Aidan, and Daniel Garijo. 2021. "A Framework for Creating Knowledge Graphs of Scientific Software Metadata." *Quantitative Science Studies* 2: 1423–46.

Krafczyk, Matthew, August Shi, Adhithya Bhaskar, Darko Marinov, and Victoria Stodden. 2019. "Scientific Tests and Continuous Integration Strategies to Enhance Reproducibility in the Scientific Software Context." In *Proceedings of the 2nd International Workshop on Practical Reproducible Evaluation of Computer Systems*, 23–28. P-RECS '19. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3322790.3330595.

Larivière, Vincent, Nadine Desrochers, Benoît Macaluso, Philippe Mongeon, Adèle Paul-Hus, and Cassidy R Sugimoto. 2016. "Contributorship and Division of Labor in Knowledge Production." *Social Studies of Science* 46 (3): 417–35. https://doi.org/10.1177/0306312716650046.

Larivière, Vincent, Yves Gingras, Cassidy R. Sugimoto, and Andrew Tsou. 2014. "Team Size Matters: Collaboration and Scientific Impact Since 1900." *Journal of the Association for Information Science and Technology* 66. https://api.semanticscholar.org/CorpusID:13505127.

Larivière, Vincent, David Pontille, and Cassidy R. Sugimoto. 2020. "Investigating the Division of Scientific Labor Using the Contributor Roles Taxonomy (CRediT)." *Quantitative Science Studies*, 1–18.

Li, Kai, Chenwei Zhang, and Vincent Larivière. 2023. "Are Research Contributions Assigned Differently Under the Two Contributorship Classification Systems in PLoS ONE?" *ArXiv* abs/2310.11687. https://api.semanticscholar.org/CorpusID:264289300.

Li, Yuliang, Jinfeng Li, Yoshihiko Suhara, AnHai Doan, and Wang Chiew Tan. 2020. "Deep Entity Matching with Pre-Trained Language Models." *Proceedings of the VLDB Endowment* 14: 50–60.

Liu, Linlin, Jianfei Yu, Junming Huang, Feng Xia, and Tao Jia. 2021. "The Dominance of Big Teams in China's Scientific Output." *Quantitative Science Studies* 2 (1): 350–62.

Liu, Xin, Chengjing Zhang, and Jiang Li. 2023. "Conceptual and Technical Work: Who Will Disrupt Science?" *Journal of Informetrics* 17 (3): 101432. https://doi.org/https://doi.org/10.1016/j.joi.2023.101432.

Lu, Chao, Yingyi Zhang, Yong-Yeol Ahn, Ying Ding, Chenwei Zhang, and Dandan Ma. 2019. "Co-contributorship Network and Division of Labor in Individual Scientific Collaborations." *Journal of the Association for Information Science and Technology* 71: 1162–78. https://api.semanticscholar.org/CorpusID:208267519.

Mayernik, Matthew S, David L Hart, Keith E Maull, and Nicholas M Weber. 2017. "Assessing and Tracing the Outcomes and Impact of Research Infrastructures." *Journal of the Association for Information Science and Technology* 68 (6): 1341–59.

Meirelles, Paulo, Carlos Santos Jr, João Miranda, Fabio Kon, Antonio Terceiro, and Christina Chavez. 2010. "A Study of the Relationships Between Source Code Metrics and Attractiveness in Free Software Projects." In *2010 Brazilian Symposium on Software Engineering*, 11–20. IEEE.

Merow, Cory, Brad L. Boyle, Brian J. Enquist, Xiao Feng, Jamie M. Kass, Brian Salvin Maitner, Brian McGill, et al. 2023. "Better Incentives Are Needed to Reward Academic Software Development." *Nature Ecology & Evolution* 7: 626–27.

Milewicz, Reed, Gustavo Pinto, and Paige Rodeghero. 2019. "Characterizing the Roles of Contributors in Open-Source Scientific Software Projects." In *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*, 421–32. https://doi.org/10.1109/MSR.2019.00069.

Muna, Demitri, Michael Alexander, Alice Allen, Richard Ashley, Daniel Asmus, Ruyman Azzollini, Michele Bannister, et al. 2016. "The Astropy Problem." https://arxiv.org/abs/1610.03159.

Naik, Cian, Cassidy R Sugimoto, Vincent Larivière, Chenlei Leng, and Weisi Guo. 2023. "Impact of Geographic Diversity on Citation of Collaborative Research." *Quantitative Science Studies* 4 (2): 442–65.

Ni, Chaoqun, Elise Smith, Haimiao Yuan, Vincent Larivière, and Cassidy R. Sugimoto. 2021. "The Gendered Nature of Authorship." *Science Advances* 7 (36): eabe4639. https://doi.org/10.1126/sciadv.abe4639.

Philippe, Olivier, Martin Hammitzsch, Stephan Janosch, Anelda van der Walt, Ben van Werkhoven, Simon Hettrick, Daniel S. Katz, et al. 2019. "Softwaresaved/International-Survey: Public Release for 2018 Results." Zenodo. https://doi.org/10.5281/zenodo.2585783.

Ram, Karthik. 2013. "Git Can Facilitate Greater Reproducibility and Increased Transparency in Science." *Source Code for Biology and Medicine* 8: 1–8.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter." *ArXiv* abs/1910.01108.

Sauermann, Henry, and Carolin Haeussler. 2017. "Authorship and Contribution Disclosures." *Science Advances* 3 (11): e1700404. https://doi.org/10.1126/sciadv.1700404.

Sharma, Nitesh Kumar, Ram Ayyala, Dhrithi Deshpande, Yesha Patel, Viorel Munteanu, Dumitru Ciorba, Viorel Bostan, et al. 2024. "Analytical Code Sharing Practices in Biomedical Research." *PeerJ Computer Science* 10: e2066.

Smith, Arfon M, Dara Norman, and Kelle Cruz. 2019. "Elevating the Role of Software as a Product of the Research Enterprise." *Bulletin of the American Astronomical Society* 51 (7): 52.

Smith, Elise. 2023. "'Technical' Contributors and Authorship Distribution in Health Science." *Science and Engineering Ethics* 29 (4): 22.

Springmeyer, Rebecca R, Meera M Blattner, and Nelson L Max. 1992. "A Characterization of the Scientific Data Analysis Process." In *IEEE Visualization*, 92:235–42.

Stankovski, Aleksandar, and Daniel Garijo. 2024. "RepoFromPaper: An Approach to Extract Software Code Implementations from Scientific Publications." In *NSLP*.

Stodden, Victoria, Peixuan Guo, and Zhaokun Ma. 2013. "Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals." *PloS One* 8 (6): e67111.

Tran, Hanh Thi Hong, Nishan Chatterjee, Senja Pollak, and Antoine Doucet. 2024. "DeBERTa Beats Behemoths: A Comparative Analysis of Fine-Tuning, Prompting, and PEFT Approaches on LegalLensNER." In *Proceedings of the Natural Legal Language Processing Workshop 2024*, 371–80.

Trisovic, Ana, Matthew K. Lau, Thomas Pasquier, and Mercè Crosas. 2021. "A Large-Scale Study on Research Code Quality and Execution." *Scientific Data* 9.

Trujillo, Milo Z, Laurent Hébert-Dufresne, and James Bagrow. 2022. "The Penumbra of Open Source: Projects Outside of Centralized Platforms Are Longer Maintained, More Academic and More Collaborative." *EPJ Data Science* 11 (1): 31.

Westner, Britta U., Daniel R. McCloy, Eric Larson, Alexandre Gramfort, Daniel S. Katz, Arfon M. Smith, and invited co-signees. 2024. "Cycling on the Freeway: The Perilous State of Open Source Neuroscience Software." *ArXiv* abs/2403.19394.

Wyss, Elizabeth, Lorenzo De Carli, and Drew Davidson. 2023. "(Nothing but) Many Eyes Make All Bugs Shallow." In *Proceedings of the 2023 Workshop on Software Supply Chain Offensive Research and Ecosystem Defenses*, 53–63. SCORED '23. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3605770.3625216.

Yang, Alex J., Ying Ding, and Meijun Liu. 2024. "Female-Led Teams Produce More Innovative Ideas yet Receive Less Scientific Impact." *Quantitative Science Studies* 5 (4): 861–81. https://doi.org/10.1162/qss_a_00335.

Yu, Liqiang, Bo Liu, Qunwei Lin, Xinyu Zhao, and Chang Che. 2024. "Semantic Similarity Matching for Patent Documents Using Ensemble BERT-Related Model and Novel Text Processing Method." *arXiv Preprint arXiv:2401.06782*.