A Major Obstacle for NLP Research: Let's Talk about Time Allocation!

Katharina Kann[♠] and **Shiran Dudy**[♠] and **Arya D. McCarthy**[♣]
[♠]University of Colorado Boulder
firstname.lastname@colorado.edu
[♣]Johns Hopkins University
lastname@jhu.edu

Abstract

The field of natural language processing (NLP) has grown over the last few years: conferences have become larger, we have published an incredible amount of papers, and state-of-the-art research has been implemented in a large variety of customer-facing products. However, this paper argues that we have been less successful than we *should* have been and reflects on where and how the field fails to tap its full potential. Specifically, we demonstrate that, in recent years, **subpar time allocation has been a major obstacle for NLP research**. We outline multiple concrete problems together with their negative consequences and, importantly, suggest remedies to improve the status quo. We hope that this paper will be a starting point for discussions around which common practices are – or are *not* – beneficial for NLP research.

1 Introduction

Why did I get nothing done today? is a question many people ask themselves frequently throughout their professional careers. Psychologists agree on good time management skills being of utmost importance for a healthy and productive lifestyle (Lakei, 1973; Claessens et al., 2007; Major et al., 2002, *inter alia*). However, many academics and industry researchers lack time management skills, working long days and getting not enough done – not even the interesting experiment they had wanted to start over a year ago.

In this position paper, we argue that natural language processing (NLP) as a field has a similar problem: we do not allocate our time well. Instead, we spend it on things that seem more urgent than they are, are easy but unimportant, or result in the largest short-term gains. This paper identifies the largest traps the authors believe the NLP community falls into. We then provide, for each of the four identified problems (P1–P4), suggested remedies. While we know that – just as for individuals – change takes time, we hope that this paper, in combination with the EMNLP 2022 special theme *Open questions, major obstacles, and unresolved issues in NLP*, will ignite critical discussions.

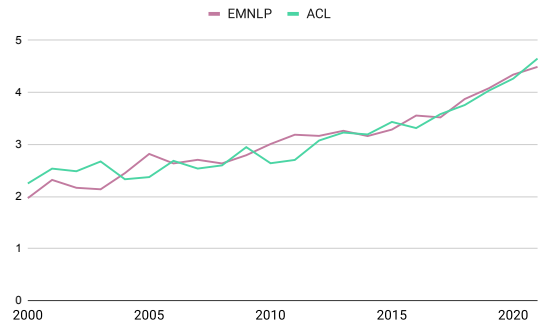

Figure 1: Avg. # of authors per paper; 2000–2021.

Related Work Over the last couple of years, multiple papers have provided critical reflections on the state of affairs in NLP research: Bender and Koller (2020) criticizes the hype around language models and argues, similarly to Bisk et al. (2020), that true understanding is impossible when language is detached from the physical world. In contrast, Bowman (2022) talks about the risks associated with *underclaiming*. Turning to evaluation, Bowman and Dahl (2021) provides a critical view on benchmarking, and Rodriguez et al. (2021) proposes ways to improve leaderboards in order to truly track progress. Other position papers discuss the importance of data curation (Rogers, 2021) and the need for focusing on the user for natural language generation (Dudy et al., 2021; Flek, 2020). Bianchi and Hovy (2021) identifies general concerning trends in NLP research. Parcalabescu et al. (2021) discusses our use of the term *multimodality* and proposes to use *task-specific* definitions of multimodality in the machine learning era. Church (2020) discusses downward trends in reviewing quality and whether these can be mitigated. We add to those meta-level papers by discussing subpar use of time as a major problem.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 8959–8969

2 What Is Going Wrong?

2.1 P1: Too Many Papers per Author

The Situation Publications in NLP are cheap compared to many other fields: there is no need to set up complicated real-world experiments (as, e.g., in physics), existing data can be used for many studies, and lately even much of the code we use is readily available. Thus, the time from idea to final paper can be extremely short. Some researchers also split one substantial paper's work into 2–5 less dense and structurally similar papers.

Consequently, NLP researchers publish a lot: Rei[1] finds that the 14 most productive first authors in NLP published 9 (1 researcher), 6 (2 researchers), and 5 (11 researchers) papers in 2021. And this number only counts the *most prestigious conferences in NLP*: Google Scholar shows that, across all venues, the first 3 authors published 16, 7, and 7 papers.

While some enjoy writing, many – especially junior – NLP researchers feel external pressure to publish in large volumes; quantity often overshadows quality of publications for hiring decisions, and PhD applicants struggle to find advisors if they do not have multiple relevant publications.

Negative Consequences A straightforward consequence of the pressure to publish is that *much of an NLP researcher's time goes into writing*: conservatively assuming one week of full-time writing per paper, the authors with the most papers respectively spend 16, 7, and 7 weeks per year just writing; this is nearly $\frac{1}{3}$ of the most productive author's year.

The second negative consequence is the *time needed to review this many papers*: reviewing one substantial paper would be quicker than reviewing 5 separate ones, especially if reviewers are not shared. This lowers review quality, frustrates authors, and causes errors to be missed. The latter then misinforms other researchers, also wasting their time.

Third, the ongoing race for publications makes it *difficult for researchers to stop and reflect on if what they are currently working on is worthwhile*. It also leads to *mixed feelings regarding the start of ambitious, high-risk/high-reward research*: many researchers are scared away by the prospect of potentially not obtaining their expected outcomes and being unable to publish. Thus, the need to constantly produce large quantities of output not only reduces the quality of individual papers, but also hinders meaningful progress of the field by encouraging the pursuit of superficial research questions.

Finally, *thorough scholarship is extremely difficult* in this environment. This leads to all sorts of shortcomings in NLP publications – missing references, mathematical errors, and even nonsensical experimental designs – which are then overlooked by overworked reviewers (Church, 2020).

Suggested Remedies To change the state of the field, we can either change our expectations or the available opportunities. For the former, it is crucial that quality is valued more than quantity for hiring. To start, we recommend having reviews be publicly available (as done, e.g., by the Conference on Neural Information Processing Systems[2]), to help people from adjacent fields understand the value of a candidate's publication. Another option is to standardize requesting reviews from experts (in addition to letters of recommendation). To reduce the opportunities for submitting large amounts of less impactful papers, we could set an upper limit for the number of (first-author) papers one can submit. This could be a hard limit or a soft limit with a penalty for too many low-quality submissions, such as blocking papers with low average scores from resubmission for a fixed period of time.[3]

2.2 P2: Too Many Authors per Paper

The Situation The second problem we highlight is the inverse of the first: too many authors per paper,[4] given the strategies we employ to manage collaborations. As shown in Figure 1, author lists are, on average, becoming longer and longer: in 2000, the average number of authors on ACL and EMNLP papers was 2.25 and, respectively, 1.97, but that number had increased to 4.65 and, respectively, 4.49 in 2021. Large collaborations can greatly advance science and, if done well, are beneficial to all participating researchers. However, they also pose an unintended challenge: many times, each author's expected, as well as actual, contribution becomes unclear. The former is often a consequence of a lack of communication or team management skills. The latter is the result of NLP not having a standardized way to communicate each researcher's contribution to collaborative projects.

In a traditional two-author setting with a student

[1] https://www.marekrei.com/blog/ml-and-nlp-publications-in-2021/

[2] https://neurips.cc

[3] This is current practice in TACL but not at conferences.

[4] Examples with >20 authors are Nan et al. (2021), Srivastava et al. (2022), and Gehrmann et al. (2021, 2022).

and their advisor, it is generally understood that the student does most of the hands-on work and their advisor guides the research and writing process. However, with more authors, the situation becomes less clear to both authors and readers.

Negative Consequences When expected contributions are unclear to the authors themselves, it is easy to have *too many cooks spoil the broth*: e.g., one author could write one section while one of their colleagues rewrites another section in a way that makes combining them non-trivial and time-consuming. Additionally, being vague about each author's contributions can lead to *friction around authorship*, which take time as well as mental energy and a toll on the relationship between two people; also, authorship discussions tend to disadvantage members of underrepresented groups (Ni et al., 2021).

Worse, however, is a situation in which it is the reader to whom it is not obvious what each authors' contribution has been. Some researchers giving authorship to people whose contribution was minimal *devalues the time and work of middle authors who actually do contribute a lot*.

Another problem with too many authors is that *miscommunication easily wastes time and resources*. For instance, it is easy to be inconsistent if experiments are run by multiple researchers, who might not use the same codebase.

Suggested Remedies In order to avoid situations where the contributions of individual authors are unclear to the reader – and, thus, accurate assignment of credit is impossible – we propose a straightforward solution that can completely eliminate this negative consequence of large collaborations: publishing a contribution statement (Brand et al., 2015) for each paper. This is common in other fields but very rare in NLP (a notable exception is, e.g., Srivastava et al. (2022)). Making a contribution statement mandatory for NLP publications would be easy but extremely effective.

For group management, setting expectations together and communicating the expected roles of all involved parties, including the possible authorship order can save time and energy toll.[5] We suggest that doing this right at the beginning of each collaborative project should become common practice in NLP ("#EMNLP2022Rule"). However, it has been shown that many principal investigators (PIs) lack training in lab and personnel management skills (Van Noorden, 2018). Thus, PIs and their research groups would likely benefit from explicit training. One possible way to achieve this could be to extend existing mentoring initiatives at NLP conferences to focus more on leadership skills. Another suggestion mentioned by Van Noorden (2018) – which we recommend for NLP – is that PIs should ask for feedback from their groups more regularly.

[5] Moster et al. (2021) offers insights on managing collaborations adjusted to remote work conditions.

2.3 P3: Gatekeeping

The Situation We do like unconventional topics (e.g., the connection between synesthesia and character-level NLP models (Kann and Monsalve-Mercado, 2021)), and statements like "This work is too interdisciplinary to get accepted" or "This work would be better for a workshop on a specific topic" are hardly ever true. However, reviewers in NLP like papers that resemble those they themselves have previously published. They only accept non-mainstream submissions if they are written in a very specific style: authors need to know how to pitch a topic to the NLP community.

For readers new to publishing in NLP, here are the basic guidelines we have found for getting a paper accepted – many of which are nonsensical: 1) Your submitted paper should always have the exact maximum number of pages – not a line more or less. 2) The first section should be called *Introduction*. 3) The last section should be called *Conclusion* – not *Discussion* or similar. 4) You should have a figure that is (somewhat) related to your paper's content on the top right corner of the first page. 5) You should have equations in your paper – complicated equations will increase your chances of acceptance (Lipton and Steinhardt, 2019). 6) Do not explicitly write out popular model equations, e.g., for the LSTM (Hochreiter and Schmidhuber, 1997). 7) The *Related Work* section should come immediately before the *Conclusion*, to make your novelty seem larger. 8) Do not present only a dataset—provide empirical results, even if they are unimportant.

Negative Consequences This gatekeeping especially affects people whose research mentors are not able to teach them the style of the NLP community: 1) people from universities with little experience in NLP research, 2) researchers from countries not traditionally part of the international NLP community, and 3) people from adjacent fields, such as psychology, social science, or even linguistics.

Thus, *gatekeeping reinforces existing social in-*

equalities and harms our research progress, as we get exposed to groundbreaking ideas later than necessary – or never. It is also a huge waste of our time: for instance, there is no reason why content presented in 7.56 pages should be less impactful than content presented in 8 pages. However, we, as a community, make it an issue and *cause researchers to waste hours trimming or extending papers*. Similarly, we force people to *waste their time thinking about which equations they can put into a paper that does not, in fact, benefit from them*.

Suggested Remedies We argue that resolving the problem of gatekeeping is crucial in order to allow our field to grow in a healthy way. We make two suggestions: 1) We need to explicitly educate reviewers to not take superficial properties of papers into account. This could be implemented, e.g., in the form of mandatory training videos for all ACL reviewers. However, this is a type of implicit bias (Greenwald and Banaji, 1995) and we encourage more discussion on possible solutions. 2) While we are waiting for this to be effective, we need to level the playing field by making unofficial rules and tricks widely known. The easiest way would be to publish explanations for first-time submitters together with calls for papers. Mentoring programs are great alternatives: while they are timewise costly for individuals, they will, in the long run, save time for the field as a whole.

2.4 P4: Missing the Point

The Situation NLP aims to build technology that improves the lives of its end users. However, NLP research is often purely technically driven, and actual human needs are investigated little or not at all (Flek, 2020; Dudy et al., 2021); this is especially prevalent when building tools for communities speaking low-resource languages (Caselli et al., 2021). This can – and does – result in researchers focusing on irrelevant problems. A similar problem is what we call *legacy research questions*: research questions that are motivated by problems or tools that are no longer relevant. Examples pointed out by Bowman (2022) are papers motivated by the brittleness of question answering (QA) systems whose performance has long been surpassed by the state of the art or an analysis and drawing of conclusions based on outdated systems like BERT (Devlin et al., 2019).[6]

To quantify this problem, we performed a case study by randomly sampling and examining 30 papers from human-oriented tracks at EMNLP 2021.[7] Only 3 papers engaged with users through evaluation and only 2 papers grounded their research questions in user needs; details can be found in Appendix A.

Last, looking at recent top-tier conferences in the field of NLP, a substantial amount of papers focus on what we call *quick research questions*, i.e., projects which maximize short-term gains for the researcher(s): Baden et al. (2022) identify that the majority of NLP research for text analysis is devoted to "easy problems", instead of aiming to "measure much more demanding constructs."

Negative Consequences Work that is missing the point does not move the field in a meaningful direction. It *wastes the researcher's time* by detracting from topics that truly benefit the community, the public, or the researcher themselves. Next, they *waste the reviewers' time* as well as *the general reader's time* by failing to provide insights. They also needlessly use computing resources, thus contributing to the climate crisis (Strubell et al., 2019). Ignoring user needs further dangerously bears the risk of causing real harm to stakeholders (Raji et al., 2022). Designing technology without the participation of potential users has in the past led to spectacular product failures (Johnson, 2021; Simon, 2020).

Finally, work on superficial research questions can be fast and result in a large amount of research output. In our current system that values quantity over quality for hiring, researchers working on superficial questions tend to have more successful careers. This, in turn, *encourages new researchers to also waste their time* by doing something similar.

Suggested Remedies It is important for NLP researchers to engage more with the intended users of the technology we build. This could be encouraged during the review process, e.g., with targeted questions. Legacy research questions will need to be detected during reviewing as well – raising awareness of this phenomenon will likely reduce impacted submissions and acceptance of papers focused on legacy research questions alike. Regarding quick research questions, one of the remedies suggested for P1 could be a possible solution here as well: moving towards valuing quality over quantity.

[6] It is, of course, possible to perform interesting studies involving older models. However, this requires well motivated research questions.

[7] The tracks we consider are: *Machine translation and Multilinguality*, *Dialogue and Interactive Systems*, *Question Answering*, and *Summarization*.

3 Conclusion

In this paper, we outlined how several problematic practices in NLP research lead to a waste of the most important resource we have – our time – and, thus, constitute major obstacles for NLP research. We suggested multiple possible solutions to existing problems. We hope to foster much-needed discussion around how we, as a community, envision moving forward in the face of these concerns.

Limitations

As we focus on time allocation, this is not an exhaustive list of problems we see in our research community. However, other concerns are beyond the scope of this work. Similarly, not all mentioned problems apply to all groups – it is, for instance, totally possible that individual groups excel at managing large collaborations.

We further do not claim that our suggested remedies are perfect solutions. They come with their own sets of challenges and should be implemented with care: for instance, contribution statements could unintentionally minimize contributions that do not make it into the final paper. Additionally, we do not claim to have listed all possible remedies for the identified problems. By contrast, we explicitly encourage other researchers to start discussing ways to improve the status quo.

Acknowledgments

We would like to thank the anonymous reviewers for their thought-provoking comments as well as the members of University of Colorado Boulder's NALA Group for their helpful feedback. This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL 2019805. The opinions expressed are those of the authors and do not represent views of the NSF. ADM is supported by an Amazon Fellowship and a Frederick Jelinek Fellowship.

References

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Aspect-controllable opinion summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6578–6593.

Christian Baden, Christian Pipal, Martijn Schoonvelde, and Mariken A. C. G van der Velden. 2022. Three gaps in computational text analysis methods for social sciences: A research agenda. *Communication Methods and Measures*, 16(1):1–18.

Cristian-Paul Bara, CH-Wang Sky, and Joyce Chai. 2021. Mindcraft: Theory of mind modeling for situated dialogue in collaborative tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1112–1125.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Federico Bianchi and Dirk Hovy. 2021. On the gap between adoption and understanding in NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3895–3901, Online. Association for Computational Linguistics.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Samuel Bowman. 2022. The dangers of underclaiming: Reasons for caution when reporting how NLP systems fail. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7484–7499, Dublin, Ireland. Association for Computational Linguistics.

Samuel R. Bowman and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.

Amy Brand, Liz Allen, Micah Altman, Marjorie Hlava, and Jo Scott. 2015. Beyond authorship: attribution, contribution, collaboration, and credit. *Learned Publishing*, 28(2):151–155.

Tommaso Caselli, Roberto Cibin, Costanza Conforti, Enrique Encinas, and Maurizio Teli. 2021. Guiding principles for participatory design-inspired natural language processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 27–35, Online. Association for Computational Linguistics.

Guanhua Chen, Shuming Ma, Yun Chen, Li Dong, Dongdong Zhang, Jia Pan, Wenping Wang, and Furu Wei. 2021. Zero-shot cross-lingual transfer of neural machine translation with multilingual pretrained encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 15–26.

Kenneth Ward Church. 2020. Emerging trends: Reviewing the reviewers (again). *Natural Language Engineering*, 26(2):245–257.

Brigitte JC Claessens, Wendelien Van Eerde, Christel G Rutte, and Robert A Roe. 2007. A review of the time management literature. *Personnel review*.

Christopher Clark, Jordi Salvador, Dustin Schwenk, Derrick Bonafilia, Mark Yatskar, Eric Kolve, Alvaro Herrasti, Jonghyun Choi, Sachin Mehta, Sam Skjonsberg, et al. 2021. Iconary: A pictionary-based game for testing multimodal communication with drawings and text. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1864–1886.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Shiran Dudy, Steven Bedrick, and Bonnie Webber. 2021. Refocusing on relevance: Personalization in NLG. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5190–5202, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Avia Efrat, Uri Shaham, Dan Kilman, and Omer Levy. 2021. Cryptonite: A cryptic crossword benchmark for extreme ambiguity in language. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4186–4192.

Tobias Falke and Patrick Lehnen. 2021. Feedback attribution for counterfactual bandit learning in multi-domain spoken language understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1190–1198.

Lucie Flek. 2020. Returning the n to nlp: Towards contextually personalized classification models. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7828–7838.

Siddhant Garg and Alessandro Moschitti. 2021. Will this question be answered? question filtering via answer model distillation for efficient question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7329–7346.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Sebastian Gehrmann, Abhik Bhattacharjee, Abinaya Mahendiran, Alex Wang, Alexandros Papangelis, Aman Madaan, Angelina McMillan-Major, Anna Shvets, Ashish Upadhyay, Bingsheng Yao, Bryan Wilie, Chandra Bhagavatula, Chaobin You, Craig Thomson, Cristina Garbacea, Dakuo Wang, Daniel Deutsch, Deyi Xiong, Di Jin, Dimitra Gkatzia, Dragomir Radev, Elizabeth Clark, Esin Durmus, Faisal Ladhak, Filip Ginter, Genta Indra Winata, Hendrik Strobelt, Hiroaki Hayashi, Jekaterina Novikova, Jenna Kanerva, Jenny Chim, Jiawei Zhou, Jordan Clive, Joshua Maynez, João Sedoc, Juraj Juraska, Kaustubh Dhole, Khyathi Raghavi Chandu, Leonardo F. R. Ribeiro, Lewis Tunstall, Li Zhang, Mahima Pushkarna, Mathias Creutz, Michael White, Mihir Sanjay Kale, Moussa Kamal Eddine, Nico Daheim, Nishant Subramani, Ondrej Dusek, Paul Pu Liang, Pawan Sasanka Ammanamanchi, Qi Zhu, Ratish Puduppully, Reno Kriz, Rifat Shahriyar, Ronald Cardenas, Saad Mahamood, Salomey Osei, Samuel Cahyawijaya, Sanja Štajner, Sebastien Montella, Shailza, Shailza Jolly, Simon Mille, Tahmid Hasan, Tianhao Shen, Tosin Adewumi, Vikas Raunak, Vipul Raheja, Vitaly Nikolaev, Vivian Tsai, Yacine Jernite, Ying Xu, Yisi Sang, Yixin Liu, and Yufang Hou. 2022. GEMv2: Multilingual NLG benchmarking in a single line of code. *arXiv preprint arXiv:2206.11249*.

Daniela Gerz, Pei-Hao Su, Razvan Kusztos, Avishek Mondal, Michał Lis, Eshan Singhal, Nikola Mrkšić, Tsung-Hsien Wen, and Ivan Vulić. 2021. Multilingual and cross-lingual intent detection from spoken data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7468–7475.

Anthony G Greenwald and Mahzarin R Banaji. 1995. Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, 102(1):4.

Jia-Chen Gu, Zhenhua Ling, Yu Wu, Quan Liu, Zhigang Chen, and Xiaodan Zhu. 2021. Detecting speaker personas from conversational texts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1126–1136.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Canming Huang, Weinan He, and Yongmei Liu. 2021. Improving unsupervised commonsense reasoning using knowledge-enabled natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4875–4885.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Taylor Berg-Kirkpatrick. 2021. Investigating robustness of dialog models to popular figurative language constructs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7476–7485.

Khari Johnson. 2021. The efforts to make text-based ai less racist and terrible. https://tinyurl.com/5x8rah4s. Accessed: 17 June 2021.

Patrick Kahardipraja, Brielen Madureira, and David Schlangen. 2021. Towards incremental transformers: An empirical analysis of transformer models for incremental nlu. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1178–1189.

Ashwin Kalyan, Abhinav Kumar, Arjun Chandrasekaran, Ashish Sabharwal, and Peter Clark. 2021. How much coffee was consumed during emnlp 2019? fermi problems: A new reasoning challenge for ai. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7318–7328.

Katharina Kann and Mauro M. Monsalve-Mercado. 2021. Coloring the black box: What synesthesia tells us about character embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2673–2685, Online. Association for Computational Linguistics.

A Lakei. 1973. How to get control of your time and life. *New York: Nal Penguin Inc*.

Ofer Lavi, Ella Rabinovich, Segev Shlomov, David Boaz, Inbal Ronen, and Ateret Anaby Tavor. 2021. We've had this conversation before: A novel approach to measuring dialog similarity. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1169–1177.

Yunlong Liang, Chulun Zhou, Fandong Meng, Jinan Xu, Yufeng Chen, Jinsong Su, and Jie Zhou. 2021. Towards making the most of dialogue characteristics for neural chat translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 67–79.

Zachary C. Lipton and Jacob Steinhardt. 2019. Troubling trends in machine learning scholarship: Some ml papers suffer from flaws that could mislead the public and stymie future research. *Queue*, 17(1):45–77.

Dan Liu, Mengge Du, Xiaoxi Li, Ya Li, and Enhong Chen. 2021. Cross attention augmented transducer networks for simultaneous translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 39–55.

Wenchang Ma, Ryuichi Takanobu, and Minlie Huang. 2021. Cr-walker: Tree-structured graph reasoning and dialog acts for conversational recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1839–1851.

Andrea Madotto, Zhaojiang Lin, Zhenpeng Zhou, Seungwhan Moon, Paul A Crook, Bing Liu, Zhou Yu, Eunjoon Cho, Pascale Fung, and Zhiguang Wang. 2021. Continual learning in task-oriented dialogue systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7452–7467.

Virginia Smith Major, Katherine J Klein, and Mark G Ehrhart. 2002. Work time, work interference with family, and psychological distress. *Journal of applied psychology*, 87(3):427.

Nikita Moghe, Mark Steedman, and Alexandra Birch. 2021. Cross-lingual intermediate fine-tuning improves dialogue state tracking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1137–1150.

Makayla Moster, Denae Ford, and Paige Rodeghero. 2021. " is my mic on?" preparing se students for collaborative remote work and hybrid team communication. In *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering Education and Training (ICSE-SEET)*, pages 89–94. IEEE.

Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto, Jessica Pan, Faiaz Rahman, Ahmad Zaidi, Mutethia Mutuma, Yasin Tarabar, Ankit Gupta, Tao Yu, Yi Chern Tan, Xi Victoria Lin, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. 2021. DART: Open-domain structured data record to text generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 432–447, Online. Association for Computational Linguistics.

Chaoqun Ni, Elise Smith, Haimiao Yuan, Vincent Larivière, and Cassidy R. Sugimoto. 2021. The gendered nature of authorship. *Science Advances*, 7(36):eabe4639.

Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-m: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 27–38.

Letitia Parcalabescu, Nils Trost, and Anette Frank. 2021. What is multimodality? In *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*, pages 1–10, Groningen, Netherlands (Online). Association for Computational Linguistics.

Dinesh Raghu, Shantanu Agarwal, Sachindra Joshi, et al. 2021. End-to-end learning of flowchart grounded task-oriented dialogs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4348–4366.

Inioluwa Deborah Raji, I Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The fallacy of ai functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 959–972.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4486–4503, Online. Association for Computational Linguistics.

Anna Rogers. 2021. Changing the world by changing the data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2182–2194, Online. Association for Computational Linguistics.

Elizabeth Salesky, David Etter, and Matt Post. 2021. Robust open-vocabulary translation from visual text representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7235–7252.

Simon. 2020. Google duplex: The effects of deception on well-being. `https://tinyurl.com/2yadfuer`. Accessed: 11 June 2020.

Jongyoon Song, Sungwon Kim, and Sungroh Yoon. 2021. Alignart: Non-autoregressive neural machine translation by jointly learning to estimate alignment and translate. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1–14.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy

Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramón Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Debnath Shyamolima, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Timothy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in nlp. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650.

Richard Van Noorden. 2018. Leadership problems in the lab. *Nature*, 557(3).

Ivan Vulić, Pei-Hao Su, Samuel Coope, Daniela Gerz, Paweł Budzianowski, Iñigo Casanueva, Nikola Mrkšić, and Tsung-Hsien Wen. 2021. Convfit: Conversational fine-tuning of pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1151–1168.

Haoran Xu, Benjamin Van Durme, and Kenton Murray. 2021. Bert, mbert, or bibert? a study on contextualized embeddings for neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6663–6675.

Shaolei Zhang and Yang Feng. 2021. Universal simultaneous machine translation with mixture-of-experts wait-k policy. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7306–7317.

Shiyue Zhang and Mohit Bansal. 2021. Finding a balanced degree of automation for summary evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6617–6632.

Jinming Zhao, Philip Arthur, Gholamreza Haffari, Trevor Cohn, and Ehsan Shareghi. 2021a. It is not as good as you think! evaluating simultaneous machine translation on interpretation data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6707–6715.

Yangyang Zhao, Zhenyu Wang, Changxi Zhu, and Shihan Wang. 2021b. Efficient dialogue complementary policy learning via deep q-network policy and episodic memory policy. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4311–4323.

Kunrui Zhu, Yan Gao, Jiaqi Guo, and Jian-Guang Lou. 2021. Translating headers of tabular data: A pilot study of schema translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 56–66.

A Appendix

In Table 1 we provide the analysis conducted on selected EMNLP 2021 papers. *Engaging with Users* indicates that researchers engage with humans, either during the design phase or for evaluation. In our analysis none of the papers engage with users throughout the process, leaving humans only to the evaluation part (3 papers). *User-driven* indicates that the motivation is grounded in user needs (2 papers). The following tracks are considered: session 1: track A: Machine translation and multi-linguality 1, session 3: track B: Dialogue and interactive systems 1, session 4: track B: Dialogue and interactive systems 2, session 5: track A: question answering 1, session 6: track B: summarization, session 7: track A: machine translation and multi-linguality 2, session 7: track B: question answering 2.

	Paper	Engaging with Users	User-driven
1	AligNART (Song et al., 2021)	no	no
2	Zero-Shot Cross-Lingual Transfer (Chen et al., 2021)	no	no
3	ERNIE-M (Ouyang et al., 2021)	no	no
4	Cross attention augmented transducer (Liu et al., 2021)	no	no
5	Translating Headers of Tabular Data (Zhu et al., 2021)	no	no
6	Towards Making the Most (Liang et al., 2021)	no	no
7	MindCraft (Bara et al., 2021)	yes	no
8	Detecting Speaker Personas (Gu et al., 2021)	no	no
9	Cross-lingual Intermediate Fine-tuning (Moghe et al., 2021)	no	no
10	ConvFiT (Vulić et al., 2021)	no	no
11	We've had this conversation before (Lavi et al., 2021)	no	no
12	Towards Incremental Transformers (Kahardipraja et al., 2021)	no	no
13	Feedback Attribution (Falke and Lehnen, 2021)	no	yes
14	CR-Walker (Ma et al., 2021)	no	no
15	Iconary (Clark et al., 2021)	yes	no
16	Improving Unsupervised Commonsense (Huang et al., 2021)	no	no
17	Cryptonite (Efrat et al., 2021)	no	no
18	Efficient Dialogue Complementary Policy Learning (Zhao et al., 2021b)	yes	no
19	End-to-End Learning of Flowchart (Raghu et al., 2021)	no	yes
20	Aspect-Controllable Opinion Summarization (Amplayo et al., 2021)	no	no
21	Finding a Balanced Degree of Automation (Zhang and Bansal, 2021)	no	no
22	BERT, mBERT, or BiBERT (Xu et al., 2021)	no	no
23	It Is Not As Good As You Think (Zhao et al., 2021a)	no	no
24	Robust Open-Vocabulary Translation (Salesky et al., 2021)	no	no
25	Universal Simultaneous Machine Translation (Zhang and Feng, 2021)	no	no
26	How much coffee was consumed (Kalyan et al., 2021)	no	no
27	Will this Question be Answered (Garg and Moschitti, 2021)	no	no
28	Continual Learning (Madotto et al., 2021)	no	no
29	Multilingual and Cross-Lingual Intent (Gerz et al., 2021)	no	no
30	Investigating Robustness of Dialog Models (Jhamtani et al., 2021)	no	no

Table 1: Our analysis of 30 randomly chosen papers from EMNLP 2021.