Mind the Gap between the Application Track and the Real World

Ananya Ganesh **Jie Cao** **E. Margaret Perkoff** **Rosy Southwell**
Martha Palmer **Katharina Kann**

University of Colorado Boulder
ananya.ganesh@colorado.edu

Abstract

Recent advances in NLP have led to a rise in inter-disciplinary and application-oriented research. While this demonstrates the growing real-world impact of the field, research papers frequently feature experiments that do not account for the complexities of realistic data and environments. To explore the extent of this gap, we investigate the relationship between the real-world motivations described in NLP papers and the models and evaluation which comprise the proposed solution. We first survey papers from the *NLP Applications* track from ACL 2020 and EMNLP 2020, asking which papers have differences between their stated motivation and their experimental setting, and if so, mention them. We find that many papers fall short of considering real-world input and output conditions due to adopting simplified modeling or evaluation settings. As a case study, we then empirically show that the performance of an educational dialog understanding system deteriorates when used in a realistic classroom environment.

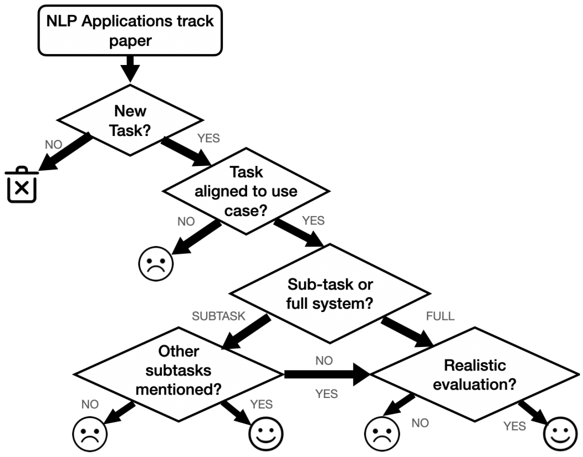

Figure 1: Summary of our survey strategy.

1 Introduction

Modern NLP systems, powered by large language models (LLMs), now have the ability to perform well at foundational natural language understanding and generation tasks (Wang et al., 2018; Brown et al., 2020). Such systems have also increased access and made inter-disciplinary contributions possible across fields such as medicine, law, education, and science. In NLP venues like ACL, the growth in applied and inter-disciplinary work can be witnessed in the NLP Applications track, which received the second-highest number of submissions at EMNLP 2022.

Recently published research from these tracks includes work on complex and important tasks such as synthesizing code for visualization (Chen et al., 2021), classifying operational risk in finance (Zhou et al., 2020), and verifiying scientific claims (Wadden et al., 2020). However, the inherent complexities associated with real-world data distributions and workflows can lead to the actual problem being simplified into an artificial setting that does not realistically reflect the original motivation. For instance, systems may make assumptions about the input available (e.g., require providing pseudocode/docstrings for code generation), or only evaluate on manually curated clean data as opposed to noisier data such as automatic speech recognition (ASR) outputs.

Motivated by this observation and in line with the ACL 2023 theme track, **we set out to investigate the relationship between the motivation described in the introductions and the actual experiments in application-focused NLP papers**. We survey papers from the *NLP applications* tracks of ACL 2020 and EMNLP 2020. Specifically, we ask if there are gaps between motivation and experimentation, in the form of i) sub-tasks that are required for the application, but haven't been mentioned in the paper ii) data distributions that are expected in real-world conditions, but haven't been included in the paper's modeling or evaluation. We find that authors do not always explicitly mention assumptions they make, and often operate in con-

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 1833–1842

Question	Counts
Does the paper comprehensively describe the use case for a reader to understand?	Yes: 15
Is the paper dealing with an entire task or a subtask only?	Entire: 11; Subtask: 4
Does the paper mention the other missing subtasks explicitly?	Yes: 1; No: 3
Is the downstream evaluation realistic?	Yes: 7; No: 7; Unsure: 1

Table 1: Findings from the survey of NLP Application track papers.

strained scenarios highly different from their intended motivation.

To empirically demonstrate the severity of this problem, we then present a case study investigating the performance of an educational dialog system, when the inputs are changed from manually transcribed data to transcripts from a state-of-the-art ASR system. The purpose of the system is to classify utterances made by a student in a classroom into *talkmoves* (Michaels and O'Connor, 2015; O'Connor and Michaels, 2019) that reflect the communication strategies they use, such as *making a claim, relating to another student*. We find that performance drops by 14.6 points (21.2%) when evaluting on Google ASR instead of human transcripts. However, ASR was not identified as a key component of the evaluation pipeline by the original work. We argue that as the field grows and NLP models get better and better at simulated and constrained settings, it is important for us to explicitly consider additional complexities of our systems in practice. We then present suggestions for authors and organizers of conferences, towards this end.

2 Survey

2.1 Method

For the survey of application-oriented research papers, we look at all papers from the *NLP Applications* track of two recent NLP conferences, ACL 2020 and EMNLP 2020, which have a total of 115 papers. These conferences, which were conducted virtually, provide publicly available interfaces,[1] that allow automatically filtering papers by the track they were submitted to.

We then manually filter papers to identify those that propose and work on *new tasks*. We choose these since papers that tackle existing tasks, such as fact checking, might be restricted to existing benchmarks and datasets that are established in a topic (Thorne et al., 2018). In contrast, papers that propose a new task, such as recommending fonts suitable for written text (Shirani et al., 2020), can integrate considerations about the environment where the task will be used, into their problem formulation and evaluation setup. We end up with 12 papers from EMNLP 2020, and 3 papers from ACL 2020 that deal with new tasks.

We then answer four questions about each paper:

1. *Does the paper comprehensively describe the use case for a reader to understand?* This question helps us establish that the motivations of the authors are clear to us before proceeding with the survey. We discard papers if the answer is *no* here.

2. *Is the paper dealing with an entire task or a sub-task only?* An example of the sub-task only would be if the desired application was assisting students with writing by providing feedback, but the actual task worked on was detecting errors in writing, with the task of formulating feedback being a sub-task for future work.

3. *Does the paper mention the other missing sub-tasks explicitly?* We investigate if the authors either mention existing systems that work on the other sub-tasks, or explicitly describe the remaining steps as future work. This is only collected when the answer to Q2 is "sub-task only".

4. *Is the downstream evaluation realistic?* An example of the answer being *No*, is if the expected use-case requires classifying spoken dialog in real-time, but the paper only evaluates on manually transcribed data.

The survey is conducted by three authors of this paper, who have all been working on NLP for 3+ years. In cases where agreement is not perfect, we report the majority answer. While all four questions take either *yes* or *no* for an answer, we optionally collect reasons for answering *no* on Questions 1

[1] https://virtual.2020.emnlp.org/index.html https://virtual.2020.acl.org/index.html

and 4. We only accept *unsure* as an answer when no decision can be made.

2.2 Findings

The results of the survey are presented in Table 1. In response to the second question, we find that 4 out of 15 papers work on sub-tasks of the overall system; however, only one of these papers explicitly mentions the other sub-tasks as components of the pipeline. Overlooked are tasks such as machine translation, performing grammatical error correction, and performing document retrieval prior to classification. In response to the fourth question, we find that 7 out of 15 papers do not include evaluations that are realistic for the setting in which they might be deployed. Some comments provided by the annotators as evidence include "evaluating only on transcribed dialog and not on ASR", "evaluating only on data translated from the original language", "not incorporating retrieval performance into evaluation pipeline" and "not checking the validity of integrated evidence." One of the responses to the last question is *unsure*, provided by two of the annotators, while the third annotator answered *yes*. One annotator's rationale for being unable to decide is that the output space modeled in the paper does not adequately reflect that seen by a user, while the second annotator claims that the task is highly subjective.

We compute inter-rater agreement using Krippendorff's α, used when there are more than two annotators (Artstein and Poesio, 2008). On Questions 2,3 and 4, the α values are 0.39, 0.44, and 0.44. While the relatively low values reflect the subjective nature of assessing application-oriented work qualitatively, our three-way annotation process and majority voting reduces the effect of an overly strict or lenient annotator. Overall, our findings indicate that application-oriented papers display some gaps that need to be addressed before the intended application is viable. While this gap often occurs in the evaluation pipeline, we highlight the importance of adequately describing all components or sub-tasks essential for an application in practice.

3 Case Study

In this section, we present a case study of an application from the domain of education. The task involves classifying student utterances into *talk moves* (Michaels and O'Connor, 2015), which are strategies provided by the Academically Productive Talk framework (Michaels et al., 2008), that students and teachers use for maintaining productive and respective discourse in a classroom. We empirically analyze the impact of evaluating this task only on a constrained, artificial environment, as opposed to a more realistic setting.

3.1 Dataset and Models

Dataset The data consists of conversations among middle school students performing collaborative work in science classrooms, documented in more detail in Southwell et al. (2022). Groups of 2-4 consenting students are seated at each table, and audio is collected through table-top Yeti Blue microphones. In total, 31 five-minute dialogue sessions are chosen for the *talk moves* analysis. Like most papers in our survey, we build a high-quality dataset for our application: samples were filtered and transcribed manually ("human" transcript) by a team of three annotators, resulting in 2003 student utterances. There are five student talk moves under the APT scheme, including *Relating to another student*, *Asking for more info*, *Making a Claim*, *Providing evidence or reasoning*, and *None*. We additionally include the label *Not enough context* when the annotators cannot make a decision. Examples of all labels can be found in Appendix A. Due to label imbalance, we cluster the labels into 3 categories (NONE, LEARNING COMMUNITY (LC) and OTHER) . Our clustering follows the higher-level grouping of talk moves into *Learning Community*, *Content Knowledge*, and *Rigorous Thinking* as defined in (Resnick et al., 2018). The dataset is then divided by session into training/dev/test splits for our model.

Model Following the state-of-the-art model for classifying teacher *talk moves* (Suresh et al., 2022), we build our student *talk moves* model by finetuning the RoBERTa-base (Liu et al., 2019) model for sequence classification. We use the previous $N = 6$ utterances as the context when predicting the *talkmove* label for the current utterance, after experimenting with multiple context windows (N) on our development set. As a baseline, we develop a random classifier using the scikit-learn DummyClassifier (Pedregosa et al., 2011), that ignores input features and uses training label distributions to make a decision. Our models are trained and validated on cleaned human transcriptions. While we do not experiment with *training* on the ASR

	Human		Google$_{\text{filter}}$		Whisper$_{\text{filter}}$	
	train	dev	train	dev	train	dev
Non-Empty	991	371	646	223	869	338
NONE	299	109	153	62	252	96
LC	515	194	361	108	450	176
OTHER	177	73	132	53	167	66

Table 2: Data distribution on our student talkmove datasets, comparing human with two ASR transcripts from Google and Whisper.

transcripts for the current case study, results for this setting can be found in Cao et al. (2023).

3.2 Distribution Shift: Human vs. ASR

However, when deploying our models in the classroom, we do not have access to clean human transcripts, and instead need to work with the outputs of ASR systems. To compare the differences between both, we look at two state-of-the-art ASR systems: Google (Google, 2023) and OpenAI Whisper (Radford et al., 2022).[2] Table 2 shows the distribution shift between human and ASR transcripts. Because of the noisy small-group classroom setting, some student utterances are difficult to recognize, resulting in imperfect ASR transcriptions with incomplete or empty utterances. This causes the input distributions to vary between human and ASR transcripts. Additionally, when the empty utterances are filtered out, the label distribution also shifts across human and different ASRs. To provide as fair a comparison as possible with the original human transcripts, we create two versions of the ASR data. The first version, denoted using the subscript 'filter' is filtered such that empty utterances are removed, which results in its size varying from the human transcripts. The second version, denoted by the subscript 'all', retains all ASR utterances where the corresponding human transcription is not empty, thus resulting in the same number of utterances as the original human transcripts.

3.3 Results

To show the performance gap caused by the above distribution shift, we evaluate our model on both human transcriptions and transcriptions from the two ASR systems. For each ASR transcript, we report both performances on their filtered version (Google$_{\text{filter}}$, Whisper$_{\text{filter}}$) and the all version (Google$_{\text{all}}$, Whisper$_{\text{all}}$). We report macro F1 as well as class-wise F1 for all models, as shown in Table 3. The top rows show performance of the random baseline. Because of the shift in label distributions, as described in Section 3.2, even the input-agnostic random baselines vary for the different versions. Looking at the model performances, we see that overall macro F1 drops by 8.91 points for Whisper$_{\text{all}}$ (a 12% drop) and 14.6 points (a 21% drop) for Google$_{\text{all}}$ when comparing across transcripts that have the same length.

Testing	macro F1	NONE	LC	OTHER
Random Baselines				
Human	0.316	0.393	0.353	0.201
Google$_{\text{filter}}$	0.321	0.379	0.352	0.230
Whisper$_{\text{filter}}$	0.317	0.392	0.357	0.202
Google$_{\text{all}}$	0.306	0.385	0.344	0.190
Whisper$_{\text{all}}$	0.312	0.390	0.354	0.193
Training on Human				
Human	0.689	0.701	0.783	0.581
Google$_{\text{filter}}$	0.591	0.555	0.635	0.581
Whisper$_{\text{filter}}$	0.614	0.625	0.601	0.617
Google$_{\text{all}}$	0.543	0.59	0.572	0.467
Whisper$_{\text{all}}$	0.599	0.641	0.558	0.599

Table 3: Results on student talk move classification.

When considering real-world deployment, the potential for such a dramatic drop in performance should be taken into account by both the designer (including researchers) and the user (such as teachers). However, for similar applications based on classroom discourse analysis, such as classifying teacher talk moves (Suresh et al., 2022), predicting appropriate next teacher talk moves (Ganesh et al., 2021) or measuring teacher uptake of student ideas (Demszky et al., 2021), comparisons to ASR transcriptions to illustrate real-world performance are rarely made, and, in many cases, ASR as a component is never mentioned.

4 Discussion

Through the above survey and case study, we qualitatively and quantitatively examine the gap between task-focused solutions in NLP research, and realistic use cases. We first acknowledge that there has existed a long-standing tradition in NLP to contextualize current research efforts through potential future applications. Looking at task-oriented dialog

[2]We select Google as it has been shown to work as well for children as adults (Rodrigues et al., 2019) and outperform similar services (Filippidou and Moussiades, 2020).

systems for example, early work such as Deutsch (1975) was motivated by the need to design computational assistants to support humans in mechanical tasks, and discussed the construction of essential components such as discourse processors, despite missing key upstream and downstream components such as ASR or dialog generation. Investigating sub-problems and their respective solutions in environments that are distinct from real-word settings has largely been unavoidable and sometimes even desirable. However, we argue that with the growth of the field and with the progress enabled by LLMs and related advances, we now have the opportunity to examine how closely our experimental setups can reflect our long term goals. Additionally, for papers that are explicitly in the *Applications* track, which present new applications intended to satisfy a real-world user need, we believe it is even more important to consider the bigger picture, and accurately describe necessary next steps for making the application a reality.

To bridge this gap, we propose a few initial recommendations: i) we suggest including a question on the Responsible NLP Checklist[3] pertinent to application-oriented papers, asking if the experimental setup has taken into account the real-world conditions of the application, ii) we recommend that authors describe any potential gaps between their motivation and proposed solution, and if so, state what is lost in the gap (such as ASR), and iii) we call for work to investigate ways to explicitly account for the gap, such as simulating noisy input data in cases where accessing the true distributions is not possible. We invite discussion from the research community on other ways forward.

5 Related Work

Our paper adds to a body of work on meta-analysis of NLP papers and the state of NLP research, particularly from the recently introduced theme tracks at *ACL conferences (Bianchi and Hovy, 2021; Bowman, 2022; Kann et al., 2022). Similarly to us in that the authors examine evaluation practices, Bowman and Dahl (2021) points out problems with benchmarking, while Rodriguez et al. (2021) proposes ways to improve leaderboards in order to truly track progress. Other papers that critically examine evaluation and leaderboards include Ribeiro et al. (2020); Dodge et al. (2019) and Ethayarajh and Jurafsky (2020). In contrast, we focus on discrepancies between proposed experimental settings and the stated motivation of research endeavours.

In addition, Bowman (2022) discusses that, similar to problematic hype, underclaiming when talking about NLP models comes with risks, and Bianchi and Hovy (2021) highlights multiple concerning trends in NLP research. More broadly, Lipton and Steinhardt (2019) discuss concerns with ML scholarship, and Church (2020) draws attention to downward trends in reviewing quality and how these can potentially be mitigated.

6 Conclusions

We investigate the "gap" between the motivations of application-focused NLP papers and their actual experimental setting. Through a survey of NLP Applications papers from two NLP conferences, we find that i) necessary components for the application get overlooked when papers focus on subtasks and ii) realistic input sources such as ASR are not being considered in downstream evaluations. We further highlight the severity of the latter issue through a case study on a dialog understanding system intended for classrooms, showing the drop in performance when ASR input, expected in the real-world, is used. While we outline potential strategies to address this issue, we hope our work will spur further discussion about future steps.

Limitations

One of the limitations of our survey is that it covers a limited sample space of 15 papers from EMNLP 2020 and ACL 2020. While a larger sample would be helpful in gathering more evidence, access to specific tracks is limited at NLP conferences, unless hosted online via a virtual or hybrid system. With respect to our case study, we evaluate on the ASR utterances, but with labels corresponding to the original manual transcriptions. For a perfect comparison, the ASR utterances would need to be re-annotated as the talk move could change based on the severity of transcription errors.

Acknowledgments

We thank the anonymous reviewers for their thoughtful feedback and suggestions. 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.

[3] https://aclrollingreview.org/responsibleNLPresearch/

References

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

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.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jie Cao, Ananya Ganesh, Jon Cai, Rosy Southwell, Margaret Perkoff, Michael Regan, Katharina Kann, James Martin, Martha Palmer, and Sidney D'Mello. 2023. A comparative analysis of automatic speech recognition errors in small group classroom discourse. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '23. Association for Computing Machinery.

Xinyun Chen, Linyuan Gong, Alvin Cheung, and Dawn Song. 2021. PlotCoder: Hierarchical decoding for synthesizing visualization code in programmatic context. 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 2169–2181, Online. Association for Computational Linguistics.

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

Dorottya Demszky, Jing Liu, Zid Mancenido, Julie Cohen, Heather Hill, Dan Jurafsky, and Tatsunori Hashimoto. 2021. Measuring conversational uptake: A case study on student-teacher interactions. 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 1638–1653, Online. Association for Computational Linguistics.

Barbara G. Deutsch. 1975. Establishing context in task-oriented dialogs. *American Journal of Computational Linguistics*, pages 4–18. Microfiche 35.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China. Association for Computational Linguistics.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online. Association for Computational Linguistics.

Foteini Filippidou and Lefteris Moussiades. 2020. A benchmarking of ibm, google and wit automatic speech recognition systems. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 73–82. Springer.

Ananya Ganesh, Martha Palmer, and Katharina Kann. 2021. What would a teacher do? Predicting future talk moves. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4739–4751, Online. Association for Computational Linguistics.

Google. 2023. Google speech-to-text. `https://cloud.google.com/speech-to-text/`. [Online; accessed 20-Jan-2022].

Katharina Kann, Shiran Dudy, and Arya D. McCarthy. 2022. A major obstacle for nlp research: Let's talk about time allocation!

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Sarah Michaels and Catherine O'Connor. 2015. Conceptualizing talk moves as tools: Professional development approaches for academically productive discussion. *Socializing intelligence through talk and dialogue*, 347:362.

Sarah Michaels, Catherine O'Connor, and Lauren B Resnick. 2008. Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. *Studies in philosophy and education*, 27(4):283–297.

Catherine O'Connor and Sarah Michaels. 2019. Supporting teachers in taking up productive talk moves: The long road to professional learning at scale. *International Journal of Educational Research*, 97:166–175.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Lauren B Resnick, Christa SC Asterhan, and Sherice N Clarke. 2018. Accountable talk: Instructional dialogue that builds the mind. *Geneva, Switzerland: The International Academy of Education (IAE) and the International Bureau of Education (IBE) of the United Nations Educational, Scientific and Cultural Organization (UNESCO)*.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Ana Rodrigues, Rita Santos, Jorge Abreu, Pedro Beça, Pedro Almeida, and Sílvia Fernandes. 2019. Analyzing the performance of asr systems: The effects of noise, distance to the device, age and gender. In *Proceedings of the XX International Conference on Human Computer Interaction*, pages 1–8.

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.

Amirreza Shirani, Franck Dernoncourt, Jose Echevarria, Paul Asente, Nedim Lipka, and Thamar Solorio. 2020. Let me choose: From verbal context to font selection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8607–8613, Online. Association for Computational Linguistics.

R. Southwell, S. Pugh, E.M. Perkoff, C. Clevenger, J. Bush, and S. D'Mello. 2022. Challenges and feasibility of automatic speech recognition for modeling student collaborative discourse in classrooms. In *Proceedings of the 15th International Conference on Educational Data Mining*. International Educational Data Mining Society.

Abhijit Suresh, Jennifer Jacobs, Margaret Perkoff, James H. Martin, and Tamara Sumner. 2022. Fine-tuning transformers with additional context to classify discursive moves in mathematics classrooms. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 71–81, Seattle, Washington. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Fan Zhou, Shengming Zhang, and Yi Yang. 2020. Interpretable operational risk classification with semi-supervised variational autoencoder. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 846–852, Online. Association for Computational Linguistics.

A Talk Move and Label Clustering

Table 4 shows the original student *talk moves* in our dataset. We merged the two labels related to learning community as a single label LC, and then

Label	TalkMove	Counts	Example
NONE	None	299	‘OK”,‘Alright”,‘Let’s do the next step.”
LC	Relating to another student	512	‘My bad”, ‘Press the button”,‘You need to code that”
	Asking for more info	3	‘I don’t understand number four.”
OTHER	Making a claim	41	‘We should place the wire on P2.”,‘We could do a winky face next.”
	Providing evidence or reasoning	1	‘Because that’s how loud our class usually is.”
	Not Enough Context	139	‘Here”,‘Do you mean [inaudible]”

Table 4: Student Talk Moves included in our talkmove dataset.

merged two rare labels “Making a claim”, and “Providing evidence and reasoning‘ with “Not Enough Context”, and form a new label OTHER

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Left blank.

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

2

☒ B1. Did you cite the creators of artifacts you used?
Left blank.

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Left blank.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Left blank.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Left blank.

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Left blank.

☒ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C ☑ Did you run computational experiments?

3

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
3.1

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Not applicable. Left blank.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Left blank.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Left blank.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Left blank.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Left blank.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Left blank.