Multimodal large language models for inclusive collaboration learning tasks

Armanda Lewis
New York University
726 Broadway
New York, NY 10003
al861@nyu.edu

Abstract

This PhD project leverages advancements in multimodal large language models to build an inclusive collaboration feedback loop, in order to facilitate the automated detection, modeling, and feedback for participants developing general collaboration skills. This topic is important given the role of collaboration as an essential 21st century skill, the potential to ground large language models within learning theory and real-world practice, and the expressive potential of transformer models to support equity and inclusion. We address some concerns of integrating advances in natural language processing into downstream tasks such as the learning analytics feedback loop.

1 Introduction

Collaboration, a coordinated process involving two or more individuals participating in a task in an interdependent way, is an important topic of study given its importance as a major 21st century skill (Lai et al., 2017; Council, 2011; Rios et al., 2020). Though collaboration as a general term is viewed as a learnable competency, notable distinctions emerge when examining how collaboration surfaces within relevant research. One semantic distinction is that the term collaboration is not explicitly defined, or is used interchangeably with concepts such as group collaboration, teamwork, collective problem solving, cooperation, and more (OECD, 2015). These inconsistencies in meaning make it challenging to connect various research agendas that purport the advantages of collaboration. Another distinction to note is modality-related. Some research does not make any modality distinctions when reporting the impacts of results, though much has viewed collaboration via online/computer-mediated interactions, both synchronous and asynchronous, while other research has examined co-located collaborative acts that happen face-to-face. Despite semantic, modality, and other distinctions, various fields have advanced what we know about collaboration, specifically collaboration as a language-mediated process.

Scholars within the fields of NLP, cognitive science, and educational research have focused separately on verbal and written aspects of collaborative exchanges - speech, text-based outputs, and audio such as non-linguistic pauses - to better understand aspects of collaboration. Recent NLP research, for example, has explored neural models equipped with dynamic knowledge graph embeddings, the use of large language models to model real world speech, and the development of collaboration datasets (Ekstedt and Skantze, 2020; He et al., 2017; Lee et al., 2022), while cognitive science has explored general modeling approaches for collaborative behavior and large language models as knowledge sources for intelligent agents (Goldstone and Gureckis, 2009; Huang et al., 2022; Wray et al., 2021). Learning analytics, a subset of educational research that extracts diverse datastreams from the learning process to improve learning, has developed automated multimodal approaches to detect, model and provide feedback about collaborative learning exchanges (Dowell et al., 2019; Pugh et al., 2022; Worsley and Ochoa, 2020). Though these studies differ in their disciplinary perspectives, they view language as essential to individuals' application of collaborative behavior and researchers' understanding of said behavior.

2 Purpose of Research Project

Because language is grounded in experience (Bisk et al., 2020), and collaboration is mediated through language, collaboration is an appropriate skill to be learned, practiced, and analyzed through language-mediated experiences and techniques. This dissertation project, situated at the intersection of NLP, cognitive science, and learning analytics, focuses on how we may support people in their development of complex, dynamic collaborative language

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop, pages 202 - 210

skills. The project extends prior research, but also introduces unexplored areas such as multimodal language modeling and inclusive collaboration. Therefore, the aim is to contribute to several open research questions related to how we may foster collaborative language, a proxy for overall collaboration skills, in people as an explicit act of learning. This project examines these critical gaps in current research to explore the ultimate question of: How can we use multimodal large language models to detect, model, and provide feedback on inclusive collaboration behavior? Sub-questions include:

- How may a multimodal framework offer improved collaborative language detection over and above unimodal language modeling?;
- What are possibilities for detecting and modeling inclusive collaboration language among a group of diverse participants?, and
- How may we leverage multimodal large language modeling in the service of learning to collaborate through automated and feedback mechanisms?

This study explores the potentials of adopting multimodal NLP techniques within a learning analytics lens. Multimodal NLP is an emerging area within NLP that stems from the development of the large language model, a massive-parameter pretrained model. Large language models are an active area of development within NLP, and one set of researchers have demonstrated impressive semantic and generative capabilities (Kaplan et al., 2020; Tay et al., 2021), while others pose ethical, environmental, and interpretability concerns about unbounded scaling of model size (Bender et al., 2021; Strubell et al., 2020; Weidinger et al., 2021). We focus on the potential of multimodal NLP, large language models that integrate multimodal (acoustic, image, tactile, and/or video) data beyond text-based language, and explore potentials of multimodal NLP for automated, fine-grained detection of collaborative processes that will support learners within and across experiences, an important downstream application of the technology (Bommasani et al., 2021; Brown et al., 2020; Islam and Iqbal, 2021; Rahman et al., 2020). We also contribute to current critiques of performance-first modeling that may overlook important opportunities to create real world NLP models that reduce bias. This project operationalizes an inclusive collaboration index with the goal of general equity and inclusion over identity-specific bias mitigation.

3 Integrating Inclusion into Downstream NLP Collaboration Tasks

Within learning analytics (Holstein and Doroudi, 2021), NLP (Blodgett et al., 2020; Tsvetkov et al., 2018), and general machine learning/AI applications (Doshi-Velez and Kim, 2017; Dwork et al., 2012), researchers have made arguments for more equitable, fair, and inclusive practices. This includes verifying that the research approach is informed by ethical and human-centered principles, developing research methods that detect/mitigate unethical outcomes, and/or our aim of proposing that research methods should translate ethically when used in real-world contexts.

With the recent focus on equity and inclusion across our fields of interest, formal inclusion theories are stated as important to integrate as a future idealized goal, though we lack blueprints for what forms these integrations may take. Within research across learning analytics, NLP, and machine learning, formal experiments provide empirical support for those methods with the most promise for identifying and reducing unwanted societal bias, ambiguity, and exclusion in datasets and models (Caliskan et al., 2017; Dinan et al., 2020; Hutchinson et al., 2020; Sap et al., 2020), though there is less support for what works as an embedded practice within downstream tasks that utilize these algorithms, datasets, and platforms. This study considers ethical research approaches and outcomes, but primarily focuses on the stated areas of potential development - the ethical deployment of our NLP and learning analytics research methods in downstream tasks situated within actual learning settings by detecting lack of inclusion and intervening. Our focus is not yet to identify any causal relationship between one or more social identities and collaboration quality, but rather to detect inclusive collaboration of individuals and groups, and in the process identify any disparities in collaboration quality among individuals and within the group as a whole.

In this sense, our work advances the concept of inclusion (Mor-Barak and Cherin, 1998; Young, 1995), defined as the degree to which diverse individuals demonstrate that they are part of the collaborative process. We recognize that this study falls short of addressing equity since equity examines

outcomes at the societal rather than individual or group level, though we highlight that inclusion is an integral step on the way to equity and ethical treatment within collaborative experiences (Bernstein et al., 2020).

4 Methodology

We have sub-divided the planned methodology into multiple tasks as: Due to the multidisciplinary nature of collaboration, this study will incorporate methods that stem from four distinct fields - learning analytics, cognitive science, natural language processing, and inclusion theory - to create an inclusive view of learning to collaborate. From learning analytics, we get a roadmap for developing an automated feedback loop necessary for learning to collaborate, and a variety of methods for detecting collaborative behaviors and operationalizing them into signals for model building. From cognitive science, we have an example for linking psychological theory, model, and real world behaviors, as well as ongoing research on intelligent agents as used for understanding learning and adaptation through feedback. From natural language processing, we have access to the ability of large language models to parse and generate human language, as well as approaches for addressing inclusion in language model building. Lastly, we operationalize tenets of inclusion theory in order to build a learning to collaborate model that detects linguistic bias, thus working towards a more inclusive collaboration environment.

All aspects involving human subjects, including Phase 1 data collected via Amazon Mechanical Turk, Phase 2 large language modeling, and Phase 3 interventions will receive full approval of the University's Institutional Review Board (IRB) prior to launching the study. Datasets are either open for research use and cited, or collected and stored as part of the IRB approval process and regulations.

4.1 Phase 1: Multimodal collaboration detection and dataset creation

As part of Phase 1 (multimodal collaboration detection and dataset creation), we will (a) develop a rubric for inclusive collaboration; (b) finalize the process of capturing and preprocessing multimodal data (video and transcribed audio) from collaborative exchanges, and (c) create an evaluation dataset. The inclusive collaboration rubric pulls from existing research on collaboration quality that identifies four collaboration indicators (information sharing, reciprocal interaction, shared understanding, and inclusion) from participants' audio, text, and video data (Cukurova et al., 2018; Praharaj et al., 2021), and the technical feat of capturing and preprocessing collaborative exchanges is informed by previous scholarship in Multimodal Learning Analytics research (Ochoa et al., 2013; Worsley and Blikstein, 2015). Automatic distillation of raw data into collaboration features would include: automatic speech recognition, computational linguistic methods to clean, parse, and analyze transcribed dialogue (eg. word counts, duration, general content analysis, inclusive content analysis), detection of non-linguistic audio (speech prosody), and video signal filtering to detect person placement and basic gestures.

Following the general dataset collection procedures described in He et al. (2017), we will gather human annotations according to our collaboration rubric of transcribed audio at the sentence-level and video portions at the frame-level that is captured for collaborative exchanges. We will use representative samples of open source collaboration datasets and datasets collected as part of an approved IRB protocol that contain text-based dialogue, spoken dialogue, and/or video of multi-person collaborative exchanges, including the AMI Meeting Corpus (Carletta et al., 2006), D64 Multimodal Conversation Corpus (Oertel et al., 2013) How2 Dataset for Multimodal Language Understanding (Sanabria et al., 2018), Pragmatic Framework for Collaboration (Boothe et al., 2022), and MutualFriends Corpus (He et al., 2017). In addition to annotation of the four dimensions of interest, we also have annotators evaluate along the modality (text, image, and video). We integrate recent NLP crowdsourcing research findings (Nangia et al., 2021) by collecting expert annotations that will then inform guidance for generally skilled Amazon Mechanical Turk (MTurk) workers, and and will use the process outlined in Bowman et al. (2015), and the Fair Work tool (Whiting et al., 2019) to ensure a fair payment structure.

The contributions of Phase 1 are multiple: to expand beyond research that analyzes collaborative language at the surface level, such as looking at word counts or temporal durations, and support deeper content-level analysis (Praharaj et al., 2021); to map current trends in large language modeling to theoretically-sound learning and inclusion frame-

works that extend past pure performance measures and support responsible downstream usage of such models.

4.2 Phase 2: Multimodal large language models for measuring collaboration quality

Phase 2 focuses on formalizing the task specification for inclusive collaboration, a process in which we operationalize human-supplied descriptions into an inclusive collaboration quality classification model. Specifically, we will conduct finetuning experiments with large transformer models to detect collaborative language and behaviors of individual members of a 3-person group.

We will utilize several pretrained large language models accessible through HuggingFace ((Wolf et al., 2020)), including BERT-base (Sanh et al., 2020), GPT-2 (Radford and Narasimhan, 2018), GPT-J, the open source version of GPT-3 (Brown et al., 2020), and FLAVA (Singh et al., 2022), a recent multimodal language model pretrained on visual and linguistic data. We will also integrate lessons learned from education-specific research utilizing large language models (Clavié and Gal, 2019; Shen et al., 2021; Suresh et al., 2021). These pretrained models will be finetuned on a random sample of the multimodal collaboration data (audio, text, and/or video frames) that has been held out of the evaluation dataset step. We will generate finetuned models with unimodal and multimodal collaborative data, and learning rates and batch sizes will be determined according to standard task settings, and we follow the training-test splits and standards articulated by Guo et al. (2020) and Minaee et al. (2021). For this study, we will limit our datasets and modeling experiments to English-language text and dialogue datasets to supplement those pre-trained models primarily trained on English-language data.

We compare the performance of our finetuned models in terms of classification accuracy of our expert and general crowdworker classification scores on the 4 collaboration dimensions. The area under the receiver operating characteristic curve (AUROC) metric is used for each dimension. Following Pugh et al. (2022), we report the chance baseline as a random shuffling of labels within each collaborative session and thus computing accuracy. Comparing different unimodal and multimodal finetuned model performance will serve as an ablation approach to examine the role of additional data modalities in terms of overall model performance, as well as a comparison between unimodal and multimodal models (Singh et al., 2022). Additionally, we conduct an analysis of random examples to determine points of synergy with, divergence from, and bias markers that differ from human classification. This will serve as essential future directions to frame the use of automated collaboration detection using large language models.

Following the design-based protocol outlined in (Praharaj et al., 2018), we will complete a pilot study within a real classroom. Small groups (of 3 people) conduct a general collaborative task and we use the detection setup established in Phase 1 to detect multimodal signals (eg. speaking duration, pauses, large language model features) correlated to collaboration quality and use our multimodal models to assess quality. We will conduct an additional automated and human evaluation on this real-life scenario.

There are two novel aspects of this modeling of collaborative quality. One involves using the large language model to provide a nuanced view of collaborative linguistic exchanges at the content level. According to Praharaj et al. (2021) note that very few studies integrate an analysis of "verbal audio indicators or the content of the audio for the analysis of [in-person] collaboration quality" (pg. 2). We leverage the large language model to explore improvements in supervised dialogue detection tasks, and also unsupervised training strategies to explore emergent and content-specific cases of collaboration so that the model can learn without direct supervision. Additionally, we propose a measure on inclusive collaboration and evaluate its association on overall collaboration quality.

4.3 Phase 3: Language generation to support collaboration learning

Since we are ultimately concerned with learning to collaborate, we build a learning analytics cycle with the development of a robust feedback loop. The feedback system will take the form of an intelligent agent that can monitor and detect aspects of the collaboration process, focusing on the measurement of collaboration quality. The key behavior is for our model to detect differences in collaboration, in order to pinpoint disparities in inclusion. The inclusive collaboration models created by generative language models will drive generative behavior

of the intelligent agent, which will produce select audio-based feedback during the collaboration exchange based on detected features.

The study will take on an experimental setup for higher education course recitations that engage in collaborative problem solving. The three groups - the no feedback control group (i.e. those randomly assigned as the control group with no intervention), the manual feedback experimental group (i.e. those randomly assigned as the manual feedback group which entails an instructor offering general, preparatory guidance on quality collaboration), and the automated feedback experimental group (i.e. those randomly assigned as the automated feedback group) - will engage in a series of four collaborative sessions. During session 1, we will record collaboration exchanges between the randomly assigned groups in order to capture multimodal baseline collaboration data. During sessions 2 and 3, the control group will collaborate in the absence of any explicit feedback, the manual feedback group will collaborate with initial collaboration guidance by the instructor, and the automated feedback group will collaborate while the intelligent agent interjects in real time. Session 4 will record collaboration exchanges between the three groups in the absence of any intervention. The goal is to assess how well all groups perform on inclusive collaboration quality.

This study hypothesizes that feedback loops built on top of multimodal large language models will capture the most relevant information associated with collaboration due to their scaled representational qualities. We will extend progress - fine-tuning; masked language model prompting; contextual prompting; and case-based prompting - made in extracting relevant information from language models to serve as knowledge sources for cognitive agents, and identify the method that maps to encouraging collaboration quality (Huang et al., 2022; Wray et al., 2021; Yousfi-Monod and Prince, 2007). The development of the agent will use language and simple feedback to offer corrective and encouraging input to students.

5 Initial Results

An initial pilot focused on the language modeling portion, and uses IRB-approved data that takes place within recitations of a large, STEM class. Groups of 3 students participated in small group work for the duration of the 75 minute period, and were tasked with solving problems related to the lecture and readings. Audio and video recordings were captured, cleaned, and processed. Transcripts were generated by an Automated Speech Recognition (ASR) software and corrected by hand, and were then paired with video frames. A random sampling of the text-based dialogue and video frames were generated and then mapped to the inclusive collaboration framework by 2 expert annotators and an additional 5 general skill annotators. These data will serve as the evaluation set. BERT-base and GPT-2 were finetuned on a randomized sample (80%) of the AMI collaboration dataset, as well as dialogue (text-based) portions of the Multi-party Collaboration corpus. Results indicate some marginal improvement between the finetuned models, and between BERT and the larger GPT-2 model, but additional analysis and more thorough data preparation and testing are needed. The finetuned GPT-2 model performed better than chance on all except for the inclusion dimension. We anticipate that more thorough finetuning and integration of multimodal finetuning data should improve performance on multimodal classification tasks.

6 Conclusion

As an essential 21st century skill, our aim is to utilize the potentials of multimodal large language models to advance our ability to detect and model collaborative behaviors, with the ultimate goal being to offer feedback to learners as they develop these important skills. Importantly, we focus on the tenets of inclusive collaboration, so that collaborators are encouraged to have equitable and inclusive exchanges as they work with each other. This doctoral research project builds an automated end-to-end inclusive collaboration feedback loop, relying on advancements in large language modeling as it is used in downstream tasks, and grounding machine learning methods within theory and real-world practice.

References

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, Virtual Event, Canada. Association for Computing Machinery.

Dimension	BERT	BERT*	GPT-2	GPT-2*	Shuffled
Info sharing	.43	.51	.52	.59	.56
Reciprocity	.41	.49	.55	.61	.54
Understanding	.47	.49	.59	.64	.52
Inclusion	.37	.43	.45	.50	.53

Table 1: Mean AUROC score across 5 iterations on 4 collaboration dimensions. Asterisk indicates models finetuned on dialogue data only.

Ruth Sessler Bernstein, Morgan Bulger, Paul Salipante, and Judith Y. Weisinger. 2020. From Diversity to Inclusion to Equity: A Theory of Generative Interactions. *Journal of Business Ethics*, 167(3):395–410.

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. *arXiv:2004.10151 [cs]*. ArXiv: 2004.10151.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. *arXiv:2005.14050 [cs]*. ArXiv: 2005.14050.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Kohd, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the Opportunities and Risks of Foundation Models. *arXiv:2108.07258 [cs]*. ArXiv: 2108.07258.

Maurice Boothe, Collin Yu, Armanda Lewis, and Xavier Ochoa. 2022. Towards a Pragmatic and Theory-Driven Framework for Multimodal Collaboration Feedback. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, LAK22, pages 507–513, New York, NY, USA. Association for Computing Machinery.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv:1508.05326 [cs]*. ArXiv: 1508.05326.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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. *arXiv:2005.14165 [cs]*. ArXiv: 2005.14165.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. The AMI Meeting Corpus: A Pre-announcement. In Steve Renals and Samy Bengio, editors, *Machine Learning for Multimodal Interaction*, volume 3869, pages 28–39. Springer Berlin Heidelberg, Berlin, Heidelberg.

Benjamin Clavié and Kobi Gal. 2019. EduBERT: Pretrained Deep Language Models for Learning Analytics. *arXiv:1912.00690 [cs]*. ArXiv: 1912.00690.

National Research Council. 2011. *Assessing 21st Century Skills: Summary of a Workshop*. National Academies Press, Washington, D.C.

Mutlu Cukurova, Rose Luckin, Eva Millán, and Manolis Mavrikis. 2018. The NISPI framework: Analysing collaborative problem-solving from students' physical interactions. *Computers & Education*, 116:93–109.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. Multi-Dimensional Gender Bias Classification. pages 314–331.

Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608 [cs, stat]*. ArXiv: 1702.08608.

Nia Dowell, Yiwen Lin, Andrew Godfrey, and Christopher Brooks. 2019. Promoting Inclusivity Through Time-Dynamic Discourse Analysis in Digitally-Mediated Collaborative Learning. In *Artificial Intelligence in Education*, Lecture Notes in Computer Science, pages 207–219, Cham. Springer International Publishing.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. 2012. Fairness through Awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM.

Erik Ekstedt and Gabriel Skantze. 2020. TurnGPT: a Transformer-based Language Model for Predicting Turn-taking in Spoken Dialog. *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2981–2990. ArXiv: 2010.10874.

Robert L. Goldstone and Todd M. Gureckis. 2009. Collective Behavior. *Topics in Cognitive Science*, 1(3):412–438.

Mandy Guo, Yinfei Yang, Daniel Cer, Qinlan Shen, and Noah Constant. 2020. MultiReQA: A Cross-Domain Evaluation for Retrieval Question Answering Models. *arXiv:2005.02507 [cs]*. ArXiv: 2005.02507.

He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning Symmetric Collaborative Dialogue Agents with Dynamic Knowledge Graph Embeddings. *arXiv:1704.07130 [cs]*. ArXiv: 1704.07130.

Kenneth Holstein and Shayan Doroudi. 2021. Equity and Artificial Intelligence in Education: Will "AIEd" Amplify or Alleviate Inequities in Education? *arXiv:2104.12920 [cs]*. ArXiv: 2104.12920.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents. *arXiv:2201.07207 [cs]*. ArXiv: 2201.07207.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social Biases in NLP Models as Barriers for Persons with Disabilities. *arXiv:2005.00813 [cs]*. ArXiv: 2005.00813.

Md Mofijul Islam and Tariq Iqbal. 2021. Multi-GAT: A Graphical Attention-Based Hierarchical Multimodal Representation Learning Approach for Human Activity Recognition. *IEEE Robotics and Automation Letters*, 6(2):1729–1736.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv:2001.08361 [cs, stat]*. ArXiv: 2001.08361.

Emily Lai, Kristen DiCerbo, and Peter Foltz. 2017. *Skills for Today: What We Know about Teaching and Assessing Collaboration*. Pearson.

Mina Lee, Percy Liang, and Qian Yang. 2022. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. *arXiv:2201.06796 [cs]*. ArXiv: 2201.06796.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep Learning Based Text Classification: A Comprehensive Review. *arXiv:2004.03705 [cs, stat]*. ArXiv: 2004.03705.

Michal E. Mor-Barak and David A. Cherin. 1998. A Tool to Expand Organizational Understanding of Workforce Diversity. *Administration in Social Work*, 22(1):47–64.

Nikita Nangia, Saku Sugawara, Harsh Trivedi, Alex Warstadt, Clara Vania, and Samuel R. Bowman. 2021. What Ingredients Make for an Effective Crowdsourcing Protocol for Difficult NLU Data Collection Tasks? 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 1221–1235, Online. Association for Computational Linguistics.

Xavier Ochoa, Katherine Chiluiza, Gonzalo Méndez, Gonzalo Luzardo, Bruno Guamán, and James Castells. 2013. Expertise estimation based on simple multimodal features. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, ICMI '13, pages 583–590, Sydney, Australia. Association for Computing Machinery.

OECD. 2015. PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving.

Catharine Oertel, Fred Cummins, Jens Edlund, Petra Wagner, and Nick Campbell. 2013. D64: a corpus of richly recorded conversational interaction. *Journal on Multimodal User Interfaces*, 7(1-2):19–28.

Sambit Praharaj, Maren Scheffel, Hendrik Drachsler, and Marcus Specht. 2018. MULTIFOCUS - MULTImodal Learning Analytics FOr Co-located Collaboration Understanding and Support. *Proceedings of the 13th EC-TEL Doctoral Consortium co-located with 13th European Conference on Technology Enhanced*

Learning (EC-TEL 2018), Leeds, UK, September 3rd, 2018.

Sambit Praharaj, Maren Scheffel, Marcel Schmitz, Marcus Specht, and Hendrik Drachsler. 2021. Towards Automatic Collaboration Analytics for Group Speech Data Using Learning Analytics. *Sensors*, 21(9):3156.

Samuel L. Pugh, Arjun Rao, Angela E.B. Stewart, and Sidney K. D'Mello. 2022. Do Speech-Based Collaboration Analytics Generalize Across Task Contexts? In *LAK22: 12th International Learning Analytics and Knowledge Conference*, pages 208–218, Online USA. ACM.

Alec Radford and Karthik Narasimhan. 2018. Improving Language Understanding by Generative Pre-Training. *undefined*.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating Multimodal Information in Large Pretrained Transformers. *arXiv:1908.05787 [cs, stat]*. ArXiv: 1908.05787.

Joseph A. Rios, Guangming Ling, Robert Pugh, Dovid Becker, and Adam Bacall. 2020. Identifying Critical 21st-Century Skills for Workplace Success: A Content Analysis of Job Advertisements. *Educational Researcher*, 49(2):80–89.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: A Large-scale Dataset for Multimodal Language Understanding. *arXiv:1811.00347 [cs]*. ArXiv: 1811.00347.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv:1910.01108 [cs]*. ArXiv: 1910.01108.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. pages 5477–5490.

Jia Tracy Shen, Michiharu Yamashita, Ethan Prihar, Neil Heffernan, Xintao Wu, Ben Graff, and Dongwon Lee. 2021. MathBERT: A Pre-trained Language Model for General NLP Tasks in Mathematics Education. *arXiv:2106.07340 [cs]*. ArXiv: 2106.07340.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. FLAVA: A Foundational Language And Vision Alignment Model. *arXiv:2112.04482 [cs]*. ArXiv: 2112.04482.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2020. Energy and Policy Considerations for Modern Deep Learning Research. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(09):13693–13696.

Abhijit Suresh, Jennifer Jacobs, Vivian Lai, Chenhao Tan, Wayne Ward, James H. Martin, and Tamara Sumner. 2021. Using Transformers to Provide Teachers with Personalized Feedback on their Classroom Discourse: The TalkMoves Application. *arXiv:2105.07949 [cs]*. ArXiv: 2105.07949.

Yi Tay, Mostafa Dehghani, Jinfeng Rao, William Fedus, Samira Abnar, Hyung Won Chung, Sharan Narang, Dani Yogatama, Ashish Vaswani, and Donald Metzler. 2021. Scale Efficiently: Insights from Pre-training and Fine-tuning Transformers. *arXiv:2109.10686 [cs]*. ArXiv: 2109.10686.

Yulia Tsvetkov, Vinodkumar Prabhakaran, and Rob Voigt. 2018. Socially Responsible NLP. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 24–26, New Orleans, Louisiana. Association for Computational Linguistics.

Laura Weidinger, John Mellor, Maribeth Rauth, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sarah Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from Language Models. Technical report, DeepMind.

Mark E. Whiting, Grant Hugh, and Michael S. Bernstein. 2019. Fair Work: Crowd Work Minimum Wage with One Line of Code. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7:197–206.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv:1910.03771 [cs]*. ArXiv: 1910.03771.

Marcelo Worsley and Paulo Blikstein. 2015. Leveraging multimodal learning analytics to differentiate student learning strategies. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, LAK '15, pages 360–367, Poughkeepsie, New York. Association for Computing Machinery.

Marcelo Worsley and Xavier Ochoa. 2020. Towards collaboration literacy development through multimodal learning analytics. In *Companion Proceedings 10th International Conference on Learning Analytics & Knowledge (LAK20)*, volume 2610, pages 53–63.

I. I. I. Wray, James R. Kirk, and John E. Laird. 2021. Language Models as a Knowledge Source for Cognitive Agents. *arXiv:2109.08270 [cs]*. ArXiv: 2109.08270.

H. Peyton Young. 1995. *Equity: in theory and practice*, 1. princeton paperback printing edition. A Russell Sage Foundation book. Princeton Univ. Press, Princeton, NJ.

Mehdi Yousfi-Monod and Violaine Prince. 2007. Knowledge Acquisition Modeling through Dialog Between Cognitive Agents. *International Journal of Intelligent Information Technologies (IJIIT)*, 3(1):060.