Natural Language Processing in the Classroom

Nicholas Brown, Liliana Deonizio, Charles Lucas, Heather Rodney
Capstone 210
University of California, Berkeley | Berkeley, California
December 14, 2023

Abstract

Providing teachers with timely feedback is an issue especially in public school districts with schools that have multiple competing needs for administrators. One way to improve this feedback loop is by using transcripts of classroom recordings to look at the fine grain data that emerges from a classroom. Using Natural Language Processing, transcripts can be analyzed to measure various features of a conversation such as when students show reasoning and how often teachers follow up on a student response or how often teachers and students are on task with the lesson at hand. We use an anonymized dataset shared by Professor Dora Demszky from the Stanford Graduate School of Education which has been made available for research purposes. This dataset includes 1,660 classroom transcripts of lessons that have been anonymized and given labels by subject experts. Providing teachers with feedback about their instructional practice is important. To this end, our research will look at how Natural Language Processing tools, such as Large Language Models, can predict teacher discourse moves from classroom transcripts. We utilize the RoBERTa-large model fine-tuned with the Low-Rank Adaptation (LoRA) technique (Hu et al. 2021), GPT, and Llama 2. Our RoBERTa-large model with LoRA is able to perform the highest out of our three models, showing comparable results to the original Demszky and Hill (2023) study with fewer trained parameters. The potential impacts of applying these models in the classroom space would include improving learning and teaching outcomes.

Introduction

Providing teachers with feedback is critical, looking at patterns of participation can help teachers continue to encourage students to share their thoughts by making follow up questions or comments that stimulate conversation. Many teachers find themselves looking for feedback as formal evaluations are typically only done two to four times a year for 15-30 minutes, and often do not provide actionable feedback. What if instead teachers had the tools in their hands to gain insights from their teaching and make data-informed improvements daily? Building on existing studies, our research addresses this need by applying natural language processing techniques to classify when students and teachers are on or off task, when students are showing reasoning skills, and how often teachers are asking focusing questions or revoicing what their students say. Our research can be applied to automate the feedback process for teachers, empowering them to make data informed decisions and improve their practice with targeted discourse moves. It is important to note that these models should not be used to provide punitive measures nor should they be used as tools for evaluating overall teacher performance. The goal is for technology such as this to be available to teachers looking to improve their practice and who want to gather insights from their classroom data.

Related Work

Michaels et al. (2008) describe Accountable Talk as discourse moves that promote academic learning and equity. The TalkMoves study (Suresh et al. 2022) builds on the Accountable Talk work from Michaels et al. (2008) by using the talk moves to annotate a dataset composed of 567 K-12 mathematics classroom transcripts. In the TalkMoves study, Suresh et al.

look at talk moves which are various dialog acts that teachers and students can take part in to create a participatory classroom environment. Teachers could practice specific talk moves by asking questions to students that ask them to explain their learning or respond to a peer. Student talk moves could include making a claim, asking a question, or showing reasoning among others. Suresh et al. (2022) use three models, BERT-base, RoBERTa- base, Electra-base for separate teacher and student models. They achieve an F1 score of 76.32 with the RoBERTa-base model for their teacher model where they looked at six teachers moves and the input was a concatenated student-teacher pair of utterances. The output of the model was "a 7-way sequence classification (softmax) over the six teacher talk moves and 'None'" (Suresh et al., 2022). For the student model the input were student to student paired utterances where they looked at how students interacted with each other and the output was a "5-way sequence classification (softmax) over the four student talk moves and "None" (Suresh et al., 2022). They achieved an F1 score of 73.12 with the BERT-base model for their student model.

Demszky and Liu (2023) deployed an automated tool based on natural language processing called M-Powering Teachers. The tool provides feedback on dialogic instructional practices on mentor's uptake of student contributions, talk time, actionable advice for eliciting and building on students ideas and reflection opportunities. A randomized controlled trial was performed to evaluate the effects of M-Power in which two groups were created: treatment (n=192) and control (n=222). NLP was used to measure changes in mentor's instructional practices and used as dependent variables. A linear regression analysis was performed to estimate the effect of the treatment on the dependent variables (number of times mentors took up student ideas per hour, number of questions raised per hour, number of times mentors repeated student words per hour and proportion of mentor talk). The results from study indicated that the M-Power automated tool improved teacher practices for those in the treatment group. Compared to the control group, student contribution was 9% more ($p<0.5$), mentors asked 6% more questions ($p<0.1$), and repeated substantive words in student utterances 6% more ($p<0.05$). Additionally, treated mentors reduced their talk time by 69% ($p<0.01$) which was 5% compared to the control group.

Demszky et al. (2021), conducted a study using Pointwise Jensen-Shannon Divergence (PSJD) to measure uptake using student-teacher conversational data (n=2246) and fine-tuned a BERT-based model. The study chose to compare the results of PSJD to a baseline of the proportion of student words repeated by the teacher (%-IN-T). Using Spearman-Correlation to compare the models, the PJSD model (ρ=.540) was significantly better than %-IN-T (ρ=.523) at identifying uptake such as question and answering and reformulation.

The dataset used in our study comes from work done by Professor Dora Demszky and Professor Heather Hill in the article, "The NCTE Transcripts: A Dataset of Elementary Math Classroom Transcripts." The data includes 1,660 anonymized classroom transcripts that were collected for the National Center for Teacher Effectiveness (NCTE) Main Study. Demszky and Hill focus on select discourse moves: student reasoning, students on or off task, teachers on or off task, teachers asking focusing questions, and teachers showing high uptake. Demszky and Hill show a positive correlation between the model predictions of discourse moves and teacher and student outcomes through a linear regression model. The authors finetune RoBERTa "on turn- level annotations for each discourse feature," (Demszky, Hill, 2023). Working with binary labels the fine-tuned RoBERTa model, "performs best on classifying on vs off task instruction —

F1 score (harmonic mean of precision and recall) is .942 and .923 for student and teacher utterances, respectively. The model performs moderately well on higher-inference discourse moves, including Student Reasoning (F1 = .651), High Uptake (F1 = .688), and Focusing Questions (F1 = .501)" (Demszky, Hill, 2023).

Our work builds on these previous studies by using NCTE Transcripts but applying new models to replicate and build on the work done by Demszky and Hill. We use the labels of student reasoning, student on task, teacher on task, focusing question, and high uptake and run a model for each of these labels using either student, teacher, or paired utterances which are a combination of student and teacher utterances. We apply NLP techniques and utilize the RoBERTa-large model fine-tuned with the Low-Rank Adaptation (LoRA) technique from Hu et al. (2021), GPT, and Llama 2.

Ethics and Privacy Considerations

The use of technology in the classroom space has shifted greatly since the Covid-19 pandemic. Whereas before teachers might have been less familiar with technology tools that could support education, now, teachers are much more likely to have received training in technology tools due to distance learning and providing asynchronous and synchronous learning opportunities to their students. With the rise of technology usage in the classroom space, also comes the need to consider privacy and ethics.

In order to use the classroom transcripts from the NCTE study, our team had to request permission from Professor Dora Demszky and state our intended use for the data. This added layer of protection and data tracking helps keep student data safer and maintains accountability for its usage. Both parents and teachers gave consent for the de-identified data from the NCTE study to be retained and used in future research. None of the district and school names are disclosed, and the transcripts are fully de-identified. We will not be sharing the data publicly per the use guidelines set by Professor Demszky.

The NCTE dataset is specific to the participating students and teachers which means it may not be representative of all students and teachers. For example, some students may not have participated as often in their math class which means their response would not have been captured. Additionally, every teacher has their own unique style of teaching and the way the teachers in the study responded to students may be different than other teachers implementing different curriculums or with different training backgrounds. Within the data there is some homogeneity. The majority of students are students of color (43% African American, 23% Hispanic/Latinx, 8% Asian). The majority of teachers are female (84%). While there is homogeneity, the data also represents student populations that are historically underrepresented in machine learning models. People from black, indigenous, or people of color backgrounds and of female identities have often been missing from the training data for models. In this regard, having this specific composition of data is a positive feature.

Our team also aims to prevent misuse of our research by including the following use clause: The research and models from this report should not be used to punish teachers and/or students. We do not assume a universal application of results as we recognize each student and teacher community will be unique. We also acknowledge there may be bias in the accuracy of previously

assigned labels as they were assigned by human subject experts.

Data

The labeled dataset we use consists of 1,660 anonymized transcripts of 4th and 5th grade elementary math classrooms collected as part of the National Center for Teacher Effectiveness (NCTE) Main Study. The data represents 317 teachers in total with the majority identifying as female (84%). In terms of students, there are 10,817 students, the majority of whom participated in a free or reduced lunch program.

The transcriptions used for the analysis includes metadata on turn-level annotations for discourse moves that were classified. The following provides the description of the datasets and the discourse moves classification within each:

1. Single Utterance: Conversational exchanges between the student and the teacher.
2. Pair Utterance: Conversational exchanges between the student and the teacher for student and teacher on task, focusing question and high uptake labels
3. Student Reasoning: Student utterances labeled on student reasoning

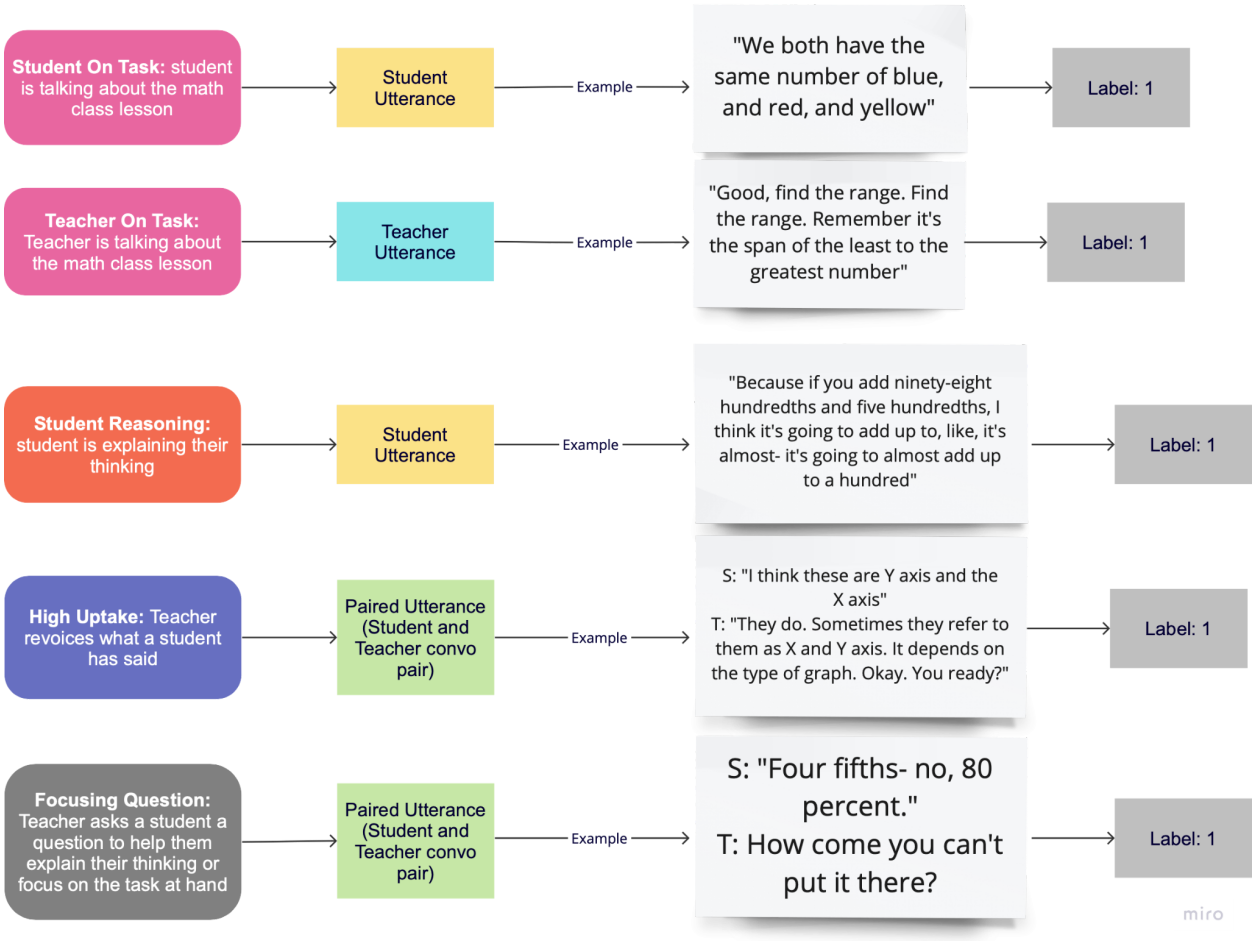

Figure 1: Shows a description and example of each of the five labels in the NCTE dataset.

Exploratory Data Analysis

An exploratory data analysis was performed on the NCTE dataset. Based on the results of the analysis, it was found that most of utterances between the student and teacher were indicated as on-task, however, there were fewer examples of high uptake for the teachers and student reasoning for the students. Figure 2 shows bar graphs revealing the class imbalances.

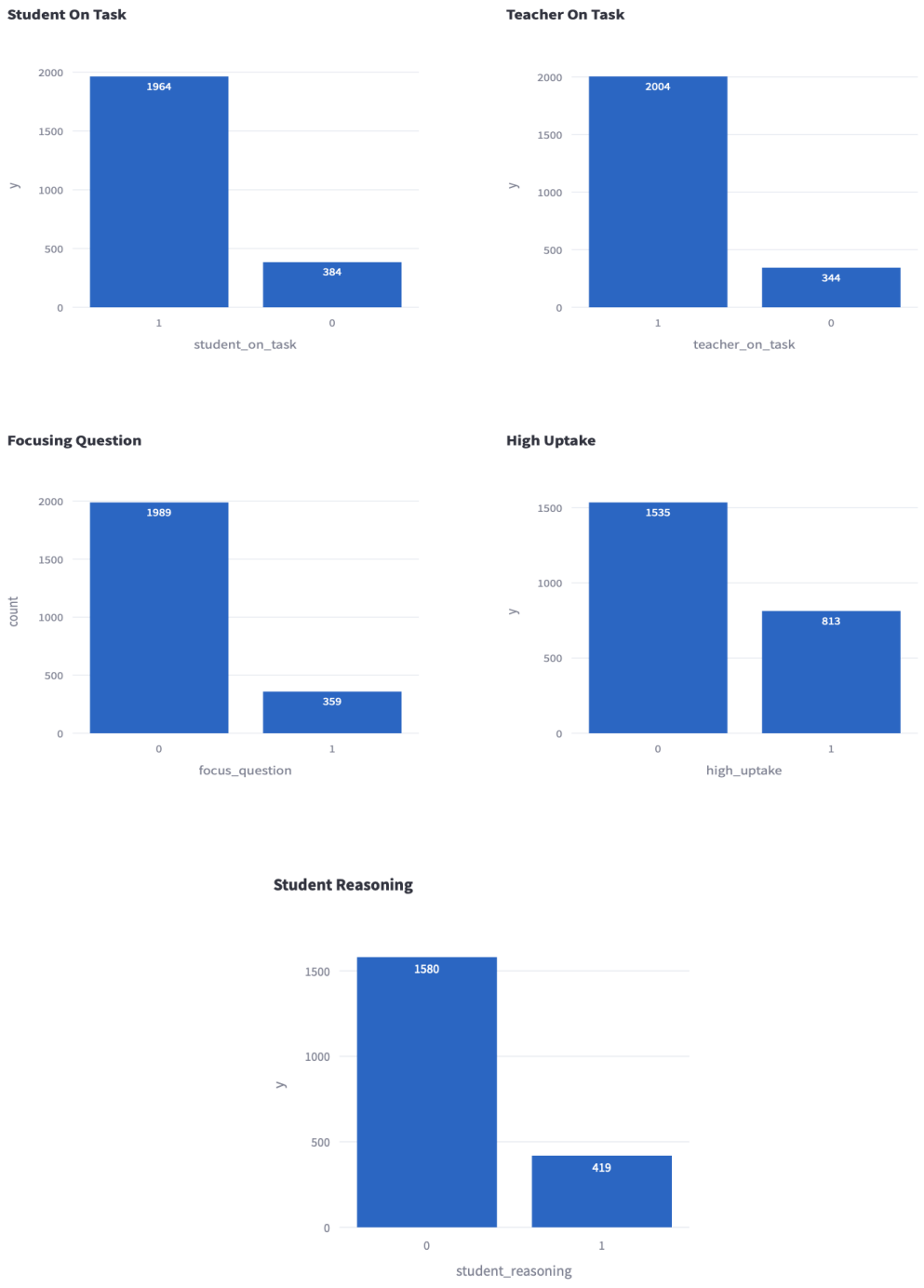

Figure 2. Bar graphs showing the number of examples for each label: student on task, teacher on task, focusing question, high uptake

Concerning text length, some of the student and teacher responses were very long, which substantially skewed the data. The data we used for our models included utterances that were below the 80th percentile of length. This helped us distinguish between the types of utterances that were most common and those that were outliers. Additionally, this helped with model performance as we were able to run models faster with the shorter text sequence.

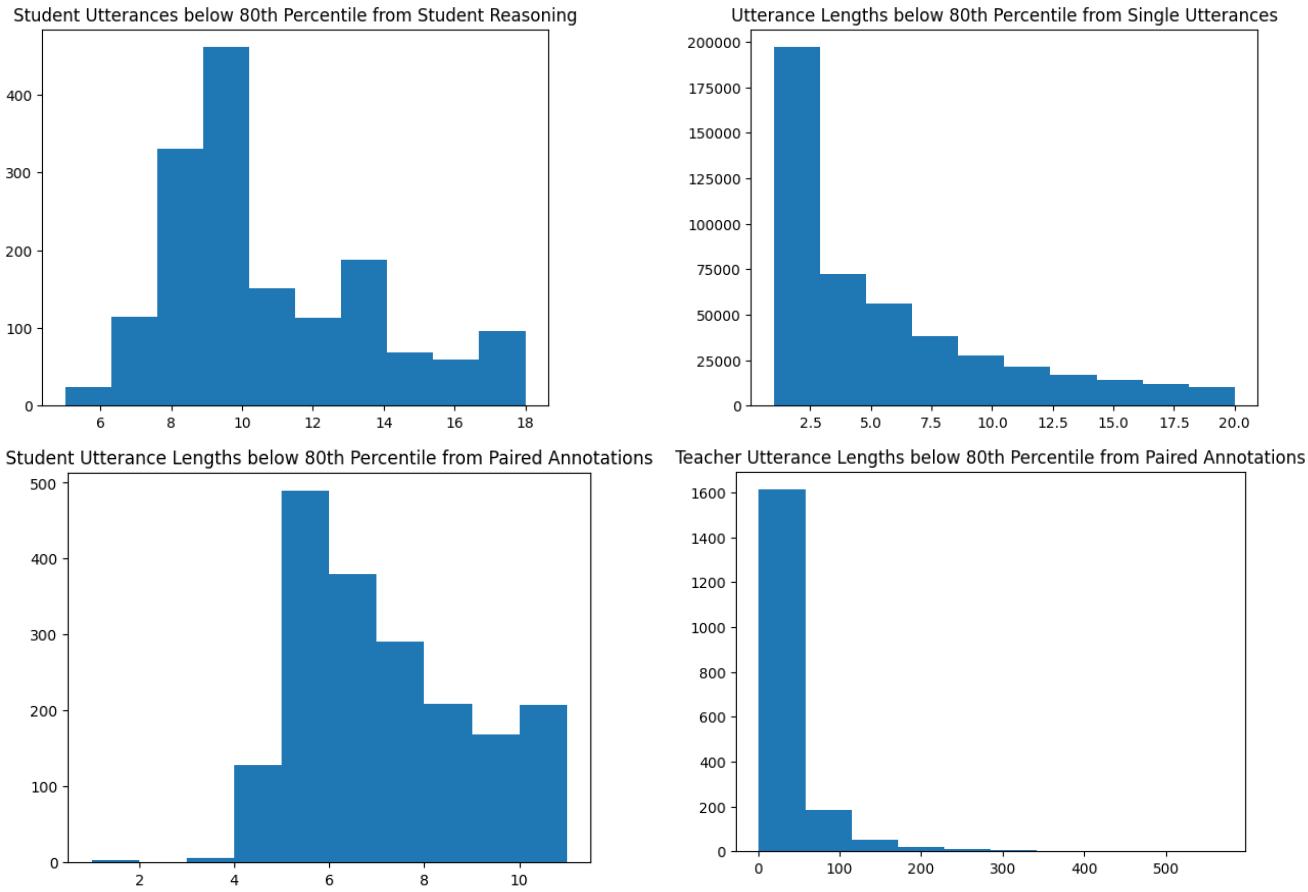

Figure 3. Bar graphs showing the length of utterances below the 80th percentile

User Interviews

User interviews were performed to gain an understanding of how teachers receive feedback and what methods to receive feedback would be helpful to them. A total of five teachers ranging from Elementary to High School, General Education to Intervention, private school to public, were interviewed who provided their responses to the questions.

Question 1: What are the current feedback practices at your school? How often do you receive feedback on your instruction and on what topics? How often would you like to receive feedback in an ideal scenario? And, in what ways?

Overall most teachers commented that they typically only received formal feedback two to four times a year. When they received the formal feedback from administrators it was often evaluative rather than constructive and did not provide actionable insights.

Question 2: Automated Feedback: How would you feel about receiving automated feedback on targeted discourse moves such as student/teacher on or off task, student reasoning, high uptake, focusing questions? How often would you see yourself using something like this? Do you feel like having automated feedback would improve your instruction? What other areas would you want automated feedback on?

Teachers responded that receiving automated feedback would be helpful and useful. Additionally, automated feedback could be seen as neutral and would be helpful to provide additional information such as how many times a student is called upon, being able to go back to what was said and review comments to improve their instruction.

Question 3: Have you heard about TeachFx, do you think it would improve your instruction?

Most had not heard of TeachFx, or any other automated feedback tools. From this we inferred that automated feedback tools have not been implemented as much in school districts. However, teachers still thought that an automated tool would be helpful overall in improving their teaching and allowing them to go back and review the data.

Overall, from the user interviews, feedback is important to teachers so they can improve on their own teaching practices, but most importantly to support their students. As such, they are supportive of an automated feedback tool, however the output should allow them to gain insights, receive positive feedback, and provide suggestions to take reasonable action.

Models

We used three different Large Language Models (LLMs) to achieve similar results as the original study. An LLM is a natural language model that, by processing large amounts of human-generated language, attempts to generate human-seeming language programmatically. Also called “generative AI”, these models excel at producing sensible sentences and paragraphs, and they also seem to “understand” a wide range of natural language prompts like “tell a joke” or “give me a vegetarian recipe”. We can think of them as highly sophisticated prediction models, allowing us to easily classify text input into different categories. We use Llama 2 (an LLM owned by Meta), GPT (an LLM owned by OpenAI and the “brain” behind ChatGPT), and

RoBERTa-large (Liu et al. 2019) (a variant of BERT, an LLM owned by Google) fine-tuned using LoRA (Hu et al. 2021).

GPT

OpenAI has an API for sending prompts to GPT via Python script. We used this API to send hundreds of requests to predict our outcomes. The results from GPT, without fine-tuning, were tepid at best, only barely outperforming "guessing the mean." We used five different prompts and conducted 0 shot, 1 shot, 2 shot, and 3 shot prompting. The prompts ranged in detail of the label description and context for the student and teacher utterances. Within the prompts we also tried asking for predictions for one label at a time and, for paired utterances, all labels at the same time.

Llama 2

For our analysis using Llama 2, we used SageMaker to run the Llama 2 model. We chose the 7-billion parameter model for Llama 2, which is the smallest of three models available. We performed three different analyses using Llama 2. For the first analysis, we implemented the few-shots prompting technique. For the second analysis, we fine-tuned five different models, one for each label. We chose to implement the instruction format option over the domain adaptation option when fine-tuning our models. The instruction fine-tuning option is a structured approach to fine-tuning, comparable to supervised learning. For the last analysis, we combined fine-tuning and prompting; we first fine-tuned the model and then used few-shot prompting on the fine-tuned model. Of the three methods implemented, we achieved the best results when performing fine-tuning only. The results we obtained from fine-tuning the models can be viewed in our "Results" section. Below, are some of the hyperparameters that we used for fine-tuning Llama 2:

Hyperparameters

add_input_output_demarcation_key: True

epoch: 4

int8_quantization: False

learning_rate: 0.0001

max_input_length: 256

RoBERTa and LoRA

In addition to the GPT API and Llama2, we tested a RoBERTa-large (Liu et al. 2019) model fine-tuned with Low-Rank Adaptation, LoRA (Hu et al. 2021). In their study, Liu et al. discuss a Robustly Optimized BERT pretraining approach, RoBERTa. Given that we had low compute resources we turned to the LoRA technique so that we could reduce the number of parameters we had to train to fine-tune the RoBERTa-large model to our specific task. LoRA allows us to freeze the pre-trained weights in a model and update a smaller amount by, "optimizing rank decomposition matrices of the dense layers' change during adaptation" (Hu et al. 2021). With this technique we fine-tuned a RoBERTa-large model on one GPU. With the LoRA technique we trained approximately 1.8 million of the 357 million parameters in the RoBERTa-large model which is about .5% of the total parameters. We found that the rank of 8 and alpha value of 16 was more effective for some labels such as Teacher On Task and for others a rank of 8 and alpha value of 32 gave better results.

Results

Our results using OpenAI's GPT were underwhelming, and could not outperform simply guessing the mean in most cases. Perhaps the results would have been improved with fine-tuning, but due to difficulty accessing the API, we were not able to fine-tune a GPT model. With the results that were obtained, regarding accuracy, the model performed poorly on High Uptake and Focusing Question compared to the other two models and the original study, but showed similar results for three of the discourse features. GPT API frequently timed out, which made predicting thousands of outcomes at once impossible. This is likely due to GPT's nascent stage. We additionally had issues accessing the API which prevented us from fine-tuning. We leave this to future work.

Running a separate LoRA fine-tuned RoBERTa-large model for each label we achieved better results than with Llama2 and GPT. The RoBERTa-large with LoRA showed comparable results to the original study from Demszky et al. 2023 in terms of accuracy on all labels; this was achieved with training on only a fraction of the parameters. This is promising for future work that may want to use Large Language Models without having to fine-tune all the existing parameters. For our task this was helpful since we also did not have a large amount of labeled data and by training less parameters we may also be keeping more of the benefits of the pre-trained RoBERTa-large model.

Models	Student Reasoning	Student On Task	Teacher On Task	High Uptake	Focusing Question
Llama 2	Accuracy- 84% Precision- 78% Recall-19% F1-30%	Accuracy- 84% Precision - 85% Recall- 98% F1- 91%	Accuracy- 84% Precision- 84% Recall-100% F1-91%	Accuracy- 69% Precision- 45% Recall- 31% F1-37%	Accuracy- 83% Precision- 33% Recall-06% F1-10%
GPT	Accuracy- 83%	Accuracy- 78%	Accuracy- 87%	Accuracy- 56%	Accuracy- 44%
LORA-RoBERTa-Large	Accuracy- 86% Precision- 89% Recall- 35% F1- 51%	Accuracy- 89% Precision- 94% Recall- 93% F1- 93%	Accuracy- 88% Precision- 91% Recall- 96% F1 Score- 93%	Accuracy: 76% Precision- 69% Recall- 58% F1- 63%	Accuracy- 90% Precision- 100% Recall- 6% F1- 12%
Original Study (Demszky & Hill)	Accuracy- 86% Precision-64% Recall-67% F1-65%	Accuracy - 90% Precision - 95% Recall - 93% F1 - 94%	Accuracy- 87% Precision- 93% Recall - 91% F1- 92%	Accuracy- 77% Precision- 72% Recall- 67% F1- 69%	Accuracy- 86% Precision-47% Recall- 54% F1- 50%

Figure 4: Shows the performance metrics of the three models

Discussion

We believe that LLMs are a valuable tool that can help automate the process of providing feedback for teachers. After fine-tuning our models, we were able to predict hundreds of discourse moves within seconds. The ability to predict hundreds of discourse moves within seconds, shows the feasibility of using LLMs as a tool to provide more frequent feedback.

We also believe that LLMs can provide useful and accurate information to help guide teachers in their practice. However, it is also important to note that LLMs have limitations, and they will not be able to solve every problem that exists. Through our analysis, we were able to reproduce results reported by Professor Dora Demszky and Professor Heather Hill in their NCTE transcripts report (Demszky, Hill, 2023). We achieved comparable accuracy scores for predicting all discourse moves when using the LoRA fine-tuned RoBERTa-large model. With this same model, we achieved comparable F1 scores for predicting teacher-on-task and student-on-task. Additionally, we produced lower scores than what was reported by Demszky and Hill (2023) for predicting focusing questions, student reasoning, and high uptake. This indicates two possibilities. First, the possibility that some discourse moves contain more of a subjective element than other discourse moves. Subjectiveness introduces an element of bias, which may make it difficult for an LLM to detect a consistent pattern. Second, the formulation of the instructions (prompting) provided to the LLMs. Though we did attempt to provide multiple prompting instructions to the LLMs, we did not observe a significant improvement in the metrics. However, we still believe that prompting can help to improve the results significantly.

Prompting may be optimized by implementing domain adaptation fine-tuning techniques. The domain adaptation fine-tuning techniques can help provide the LLMs with a more in depth context related to a specific domain, which in turn, can lead to a better understanding of the prompts being provided.

We also discovered in our analysis that the RoBERTa-large model consistently outperformed Llama 2 in predicting every discourse move. This came as a slight surprise. We thought because Llama 2 contained approximately 20 times more parameters, 7 billion parameters compared to 357 million parameters, that Llama 2 would outperform RoBERTa-large at predicting discourse moves. The RoBERTa-large model outperforming Llama2 indicates that a larger model does not necessarily equate to a better performance. Additionally, this indicates that some models may be better than others at achieving certain tasks. Considering that the RoBERTa-large model was fine-tuned with LoRA, meaning only .5% of the weights were updated, adds another layer of interest. This could indicate the power of using LoRA to train models for niche tasks with small datasets.

Another important consideration is that the LLMs used were trained on a corpus of text that includes natural language which is very different from the language encountered inside of a classroom. Most student utterances were short and followed a cadence that is starkly different from a wikipedia article about the same topic. Future work can look at how to train a large language model on a larger corpus of student speech and classroom speech.

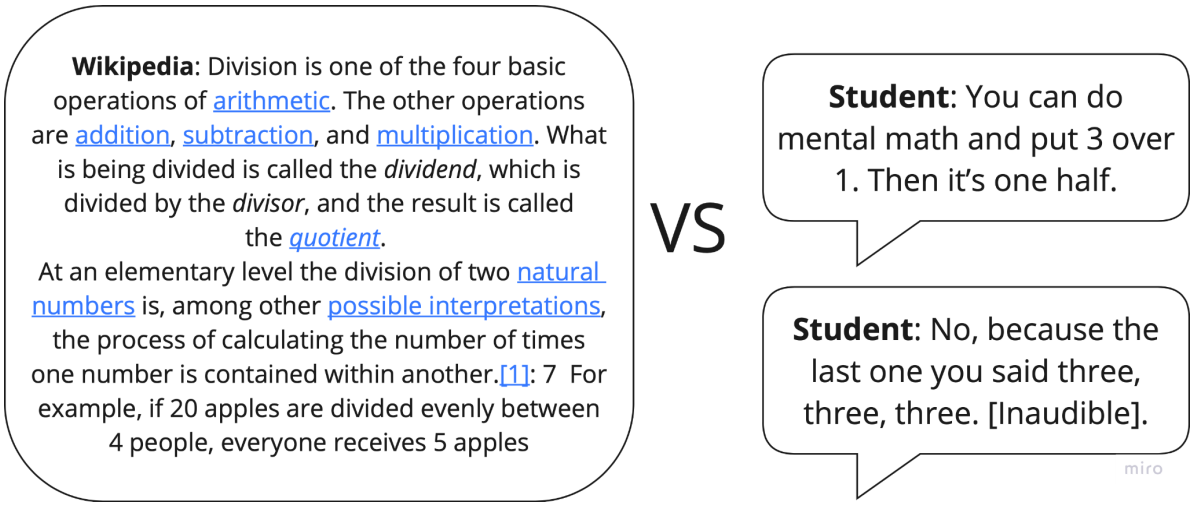

Figure 5: Shows the contrast between a Wikipedia article about division and two student utterances in the context of a math class about a similar topic

Implications and Significance

Text classification is a difficult and noisy problem, by virtue of the simple fact that natural human language is inherently ambiguous. Since different humans would disagree about different interactions counting as “on-task” or “off-task”, we cannot expect our models to perform perfectly or even exceptionally.

That said, we do observe accuracy better than “guessing the mean”. This implies that some models can “understand” important concepts like “on-task” or what a “focusing question” could look like. While some educators are bound to bristle at automated machine feedback, others may welcome the second opinion, which is practically free, except for the model training and processing. Motivated educators can review transcripts of their lessons and identify, with reasonable accuracy, positive interactions from class and areas they may want to improve on.

The National Center for Education Statistics (2023) shows math and reading declining since 2012, with the rate of decline increasing since 2020. We should not expect this trend to reverse itself without serious and diverse interventions. To improve learning outcomes for students, we should consider all options available, and with such impressive strides in machine learning and natural language processing in recent years, we should seek to apply these advances to teaching and learning. More work in this area would result in better models with better feedback, which could help teachers improve their teaching.

Limitations

The LLMs we used for our predictions are very much “generalists.” They excel or exceed expectations at a great number of diverse tasks related to natural language, including logic and reasoning, generation of “new” ideas, and prediction. But, like all generalists, they lack specialization. In particular, they are not specialized to predict the output class of text input. A more specialized model, even a specialized version of one of these LLMs, might be better suited to the task. That said, these tools are very accessible, easy-to-use.

Furthermore, these LLMs are popular and resource intensive, and OpenAI in particular had trouble serving all of our requests. For example, we ran a script to send a thousand prompts to GPT, but the script timed out after only a few hundred prompts, due to an inability from OpenAI to keep up. A search online revealed that this was a common phenomenon experienced by people working with OpenAI, and is likely a result of the nascent stage of corporate LLMs.

Another difficulty in using these LLMs was output processing. GPT, for example, is designed to reply with a natural language response to the user’s natural language prompt. This is starkly different from a logistic regression, for example, which mathematically outputs a number between 0 and 1. Although it was not terribly difficult to process GPT’s responses into vectorized data, it was a layer of friction which was a consequence of the tools we used.

Finally, the most difficult part of working with LLMs is prompt engineering. The difference in quality of output between a good prompt and a bad prompt is monumental. One of the major "developments" in prompt engineering is to use "few shot" prompting, in which the user feeds the LLM a small number of "correct" input-output pairs, which has been shown to improve LLM performance in a number of tasks.

Conclusion

We set out to help educators by classifying teacher and student interactions to provide high-quality, automatic feedback. Assuming high-quality transcription and an interface to allow educators to interact with LLMs, there is potential and promise in progress toward this goal. Large Language Models offer incredible capabilities, some of which we were able to utilize ourselves to achieve impressive results. As the capabilities of LLMs improve, we should expect their performance on a range of tasks to improve as well. Today's LLMs are accessible and powerful tools like LoRA are available for impressive fine-tuning, giving us F1 scores of over 90% on some tasks. That said, LLMs are also still very limited. GPT, for example, cannot outperform guessing the mean without fine-tuning, and OpenAI has problems accessing its API. Further work needs to be done to train an LLM to understand the types of student and teacher interactions in the classroom. Future studies can look at creating a larger corpus of classroom speech and training models on this data.

Acknowledgments

We acknowledge Professor Dora Demszky et al. for their foundational work on this problem and for providing easy access to such a huge dataset. Our project would not have been possible otherwise. We also thank the five teachers who participated in our user interviews for their thoughtful feedback and thought partnering. We thank Tim Slade and David Russell for thought-partnering with us, helping us ideate and considering ethics and privacy with us. Thanks as well to Google, OpenAI, Meta, and the countless data scientists, linguists, computer scientists, and statisticians who have worked to develop the field of natural language processing, and on whose shoulders our own work rests. Thanks as well to Professors Danielle Cummings and Fred Nugen as well as the University of California, Berkeley for providing the support and framework for this project.

References

1. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.https://arxiv.org/abs/2005.14165
2. Demszky, D., & Hill, H. (2022). The NCTE Transcripts: A dataset of elementary math classroom transcripts. arXiv preprint arXiv:2211.11772. https://arxiv.org/pdf/2211.11772.pdf
3. Demszky, D., & Liu, J. (2023). M-Powering Teachers: Natural Language Processing Powered Feedback Improves 1:1 Instruction and Student Outcomes. https://www.dorademszky.com/publications/19892-m-powering-teachers-natural-language-processing-powered-feedback-improves-1-1-instruction-and-student-outcomes
4. Demszky, D., Liu, J., Hill, H. C., Jurafsky, D., & Piech, C. (2021). Can Automated Feedback Improve Teachers' Uptake of Student Ideas? Evidence from a Randomized Controlled Trial in a Large-Scale Online Course. EdWorkingPaper No. 21-483. Annenberg Institute for School Reform at Brown University.?https://landing.teachfx.com/can-automated-feedback-improve-teachers-uptake-of-student-ideas
5. Demszky, D., Liu, J., Mancenido, Z., Cohen, J., Hill, H., Jurafsky, D., & Hashimoto, T. (2021). Measuring conversational uptake: A case study on student-teacher interactions. arXiv preprint arXiv:2106.03873.https://spaces-cdn.owlstown.com/blobs/x7sff7a8jjufhcflhuo008zy0588
6. Ford, K., & Welling-Riley, K. (2021). Student talk in science class. The Learning Professional, 42(3), 58-61.https://landing.teachfx.com/detroit-impact
7. Fried, J. (2021, February 1). 'Fitbit' for Monitoring Student Voice. AASA School Administrator, 78(2), 39. https://www.aasa.org/resources/resource/fitbit-for-monitoring-student-voice
8. Gewertz, C. (2019). How much should teachers talk in the classroom? Much less, some say. Education Week. Retrieved from https://www.edweek. org/ew/articles/2019/12/11/how-much-should-teachers-talk-in-the. html.
9. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.
10. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.
11. Michaels, O'Connor, C., & Resnick, L. B. (2008). Deliberative Discourse Idealized and Realized: Accountable Talk in the Classroom and in Civic Life. Studies in Philosophy and Education, 27(4), 283–297. https://doi.org/10.1007/s11217-007-9071-1
12. National Center for Education Statistics (2023.). Scores decline again for 13-year-old students in reading and mathematics. The Nation's Report Card. https://www.nationsreportcard.gov/highlights/ltt/2023/

13. Suresh, A., Jacobs, J., Harty, C., Perkoff, M., Martin, J. H., & Sumner, T. (2022). The talkmoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves.
.http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.497.pdf
14. T Kane, H Hill, and D Staiger. 2015. National center for teacher effectiveness main study. icpsr36095-v2.