Is ChatGPT a Good Teacher Coach? Measuring Zero-Shot Performance For Scoring and Providing Actionable Insights on Classroom Instruction

Rose E. Wang Dorottya Demszky
rewang@cs.stanford.edu ddemszky@stanford.edu

Introduction

- Coaching, which involves classroom observation and expert feedback, is a widespread and fundamental part of teacher training.
- However, the majority of teachers do not have access to consistent, high quality coaching due to limited resources and access to expertise.
- We explore whether generative AI could become a cost-effective complement to expert feedback by serving as an automated teacher coach.

Our research question

Can ChatGPT help instructional coaches and teachers by providing effective feedback, like generating classroom observation rubric scores and helpful pedagogical suggestions?

Contributions

- We propose the following teacher coaching tasks for generative AI:

Task A. *Score* a transcript segment for items derived from classroom observation instruments.
Task B. *Identify highlights and missed opportunities* for good instructional strategies.
Task C. *Provide actionable suggestions* for eliciting more student reasoning

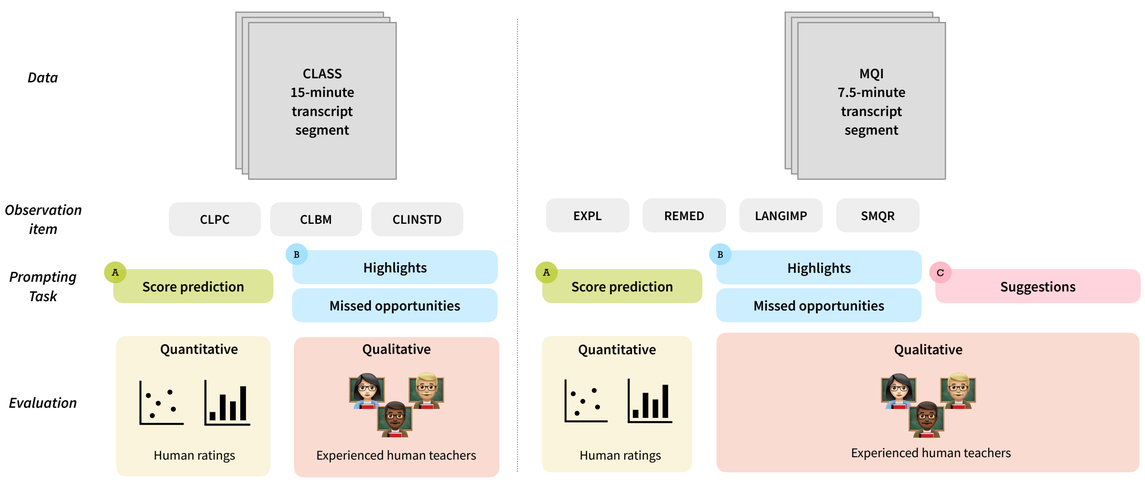


Figure 1. Setup for the automated feedback task.

- We recruit expert teachers to evaluate ChatGPT's zero-shot performance on these tasks.
- We demonstrate that ChatGPT is useful in some aspects but still has a lot of room for improvement.
- We highlight directions for future directions towards providing useful feedback to teachers.

Background: Automated feedback to educators

- Prior works on automated feedback tools provide analytics on student engagement and progress [8, 1, among others]. These tools enable teachers to monitor student learning and intervene as needed.
- Recent NLP advances are able to provide teachers feedback on their classroom discourse, promoting self-reflection and instructional development [7, 5, 3, among others].
- Altogether, these findings show a positive impact of cost-effective automated tools.
- They prompt further investigations into what other types of automated feedback are effective. Our work constitutes one exploration in this area.

Dataset and Methods

- We use the National Center for Teacher Effectiveness (NCTE) Transcript dataset [2]. It is the largest publicly available dataset of U.S. classroom transcripts (anonymized) linked with classroom observation scores.
- It consists 4th and 5th grade elementary mathematics observations collected by the NCTE between 2010-2013. It represents data from 317 teachers across 4 school districts that serve largely historically marginalized students.
- Expert raters annotate using the Classroom Assessment Scoring System (CLASS) [6] and Mathematical Quality Instruction (MQI) [4] instruments.

References

[1] Nathalie Bonneton-Botté, Sylvain Fleury, Nathalie Girard, Maëlys Le Magadou, Anthony Cherbonnier, Mickaël Renault, Eric Anquetil, and Eric Jamet. Can tablet apps support the learning of handwriting? an investigation of learning outcomes in kindergarten classroom. *Computers & Education*, 151:103831, 2020.

[2] Dorottya Demszky and Heather Hill. The NCTE Transcripts: A dataset of elementary math classroom transcripts. *arXiv preprint arXiv:2211.11772*, 2022.

[3] Dorottya Demszky, Jing Liu, Heather C Hill, Dan Jurafsky, and Chris Piech. Can automated feedback improve teachers' uptake of student ideas? evidence from a randomized controlled trial in a large-scale online course. *Educational Evaluation and Policy Analysis*, 2023.

[4] Heather C Hill, Merrie L Blunk, Charalambos Y Charalambous, Jennifer M Lewis, Geoffrey C Phelps, Laurie Sleep, and Deborah Loewenberg Ball. Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and instruction*, 26(4):430–511, 2008.

[5] E. Jensen, M. Dale, P. J. Donnelly, C. Stone, S. Kelly, A. Godley, and S. K. D'Mello. Toward automated feedback on teacher discourse to enhance teacher learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13, 2020.

[6] Robert C Pianta, Karen M La Paro, and Bridget K Hamre. *Classroom Assessment Scoring System™: Manual K-3*. Paul H Brookes Publishing, 2008.

[7] B. Samei, A. M. Olney, S. Kelly, M. Nystrand, S. D'Mello, N. Blanchard, X. Sun, M. Glaus, and A. Graesser. Domain independent assessment of dialogic properties of classroom discourse. 2014.

[8] Baruch B Schwarz, Naomi Prusak, Osama Swidan, Adva Livny, Kobi Gal, and Avi Segal. Orchestrating the emergence of conceptual learning: A case study in a geometry class. *International Journal of Computer-Supported Collaborative Learning*, 13:189–211, 2018.

Task A: Scoring transcripts

- We zero-shot prompt ChatGPT to predict observation scores according to the CLASS and MQI rubrics.
- Prompt techniques
  - *direct answer*, DA: prompting to directly predict a score with 1-2 sentence summary of the item
  - *direct answer with description*, DA$^+$: same as DA but with additional one-sentence descriptions for low/mid/high ratings
  - *reasoning then answer*, RA: same as DA, with asking the model to provide reasoning before predicting a score.

	CLPC	CLBM	CLINSTD
DA	0.00	0.35	−0.01
DA$^+$	0.04	0.23	0.07
RA	−0.06	0.07	−0.05

EXPL	REMED	LANGIMP	SMQR
0.02	0.05	0.00	0.17
0.12	0.06	0.02	0.17
−0.11	−0.06	0.04	0.06

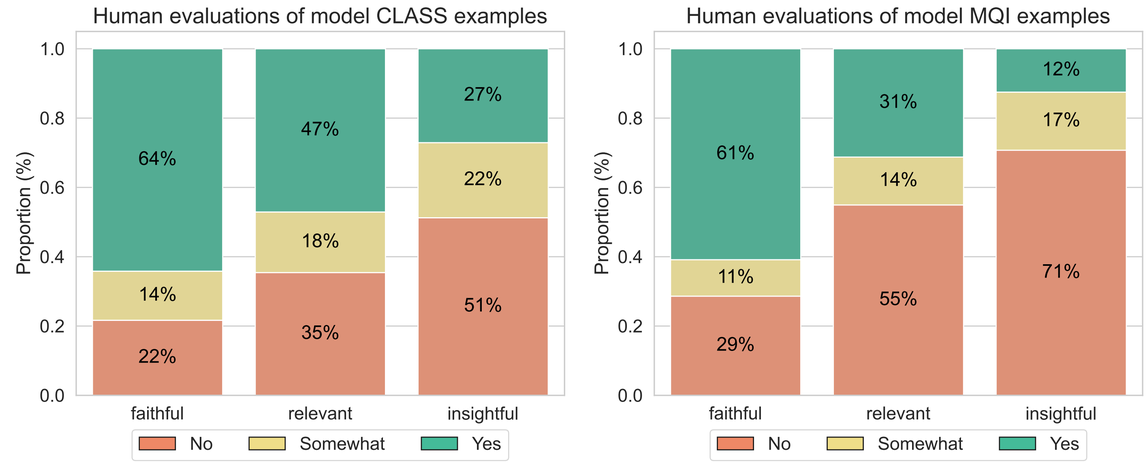
Table 1. Spearman correlation values between human scores and model predictions on the CLASS dimensions (left table) and MQI dimensions (right table). The columns represent the different dimensions and the rows represent the different prompting methods.

Takeaways:

- The model has low correlation with human ratings.
- Adding more information to the prompt like reasoning (RA) did not improve the correlation score—in some cases making the score worse, such as for CLBM.

Task B: Identify highlights and missed opportunities

- We prompt the model to identify highlights and missed opportunities per observation item in CLASS and MQI.
- Teachers are asked to rate each example along:
  1. *Relevance*: Is the model's response relevant to the CLASS or MQI item of interest?
  2. *Faithfulness*: Does the model's response have an accurate interpretation of the events that occur in the classroom transcript?
  3. *Insightfulness*: Does the model's response reveal insights beyond a literal restatement of what happens in the transcript?



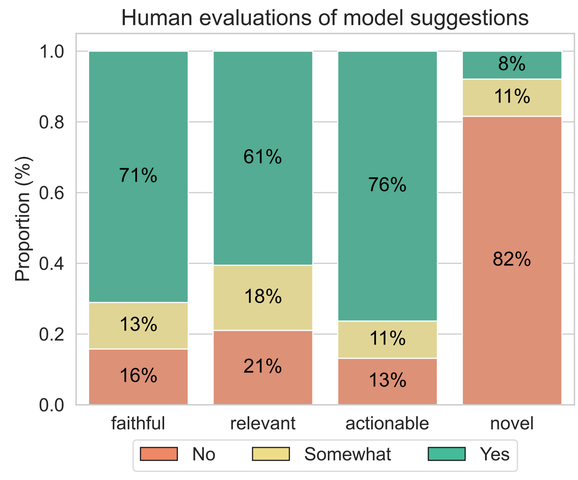
(a) CLASS (b) MQI

Takeaways:

- Teachers generally did not find the model responses insightful or relevant to what was being asked for both instruments.
- The MQI results are worse than the CLASS results across all evaluation dimensions.
- This suggests that the model performs relatively worse on interpreting and evaluating technical aspects of math instruction quality.

Task C: Provide actionable suggestions

- We prompt the model to generate suggestions for eliciting more student reasoning.
- Teachers are asked to rate each example along:
  1. *Relevance*: Is the model's response relevant to eliciting more student reasoning?
  2. *Faithfulness*: Does the model's response have the right interpretation of the events that occur in the classroom transcript?
  3. *Actionability*: Is the model's suggestion something that the teacher can easily translate into practice?
  4. *Novelty*: Is the model suggesting something that the teacher already does or is it a novel suggestion?



Takeaways:

- The model produces redundant suggestions, repeating what the teacher already does in the transcript 82% of the time.
- This may be due to ChatGPT not seeing examples of instructional feedback, given the scarcity of publicly available data in this area.
- Thus, it reproduces patterns in the text and does not produce out-of-the-box expert suggestions.