AI-enhanced coaching: What early studies show

BY HEATHER C. HILL, JIM MALAMUT, DORA DEMSZKY, JING LIU, SAMANTHA BOOTH, CLAIRE GOGOLEN, AND HANNAH ROSENSTEIN

Instructional coaching is one of the most effective ways to improve teaching, in part because feedback is grounded in teachers' own classroom practice and delivered in the context of collaborative relationships (Kraft et al., 2018). But the time- and resource-intensive nature of coaching often limits its reach and impact. A typical classroom observation and feedback cycle — the coaching component most associated with teacher growth — can take several hours per teacher and requires extensive coordination. On top of that, coaches have other responsibilities such as facilitating meetings and working with student assessment data. At over $3,000 per year per teacher, coaching is also one of the most expensive strategies for improving teaching (Blazar & Kraft, 2015).

Considering these challenges, many educators are now asking whether AI can make classroom observations and feedback more efficient. One option is to have AI generate written feedback, simulating the role of coach. Another is to use AI to augment the coaching process, for instance by analyzing classroom recordings and providing summary data to coaches who can then study it with teachers.

Our team has conducted several investigations to better understand the potential of these strategies to expand coaching's reach and impact. Using mathematics instructional improvement as a testing ground, we have found promising avenues as well as areas for improvement.

USING CHATGPT AS A COACH

One possibility is to use AI directly as a coach. Some teachers are already experimenting with this, either independently or alongside their school-based coaches. They record lessons, upload transcripts to ChatGPT or similar large language models (LLMs), and request feedback.

Studies show that early versions of ChatGPT were poor instructional coaches (Wang & Demszky, 2023). But models have improved quite a bit in the last few years. One of us (Heather) recently tried using ChatGPT 5.1 as a coach, asking it to provide feedback on anonymized transcripts representing low-, medium-, and high-quality mathematics lessons. Two mathematics education experts had rated these lessons using our Mathematical Quality of Instruction (MQI) observation instrument, and Heather had taken descriptive notes about the lesson. Our prompt to ChatGPT attempted to reproduce those notes, asking for an overall assessment of lesson quality, lesson strengths and weaknesses, and suggestions for improvement.

ChatGPT nailed its analysis of the low-quality lesson. Like the human raters, it noted that despite good classroom rapport and strong student engagement, the teacher's intended goal of students constructing pie charts was not realized by a task that asked the number of pizzas needed to feed a class of 17 students who each wanted two slices. Both our team and GPT also noted a lack of clarity in teacher speech and several instances when the teacher led students down unproductive solution pathways. ChatGPT's suggestions aligned to what we would have recommended based on the lesson.

GPT's analysis of a medium-quality lesson tracked our analysis in most places too, pointing out strengths such as the teacher conducting an error analysis on a student mistake but also noting weaknesses, including the slow pace of the lesson. Notably, both our and GPT's analyses identified teacher explanations for the conversion of percents to decimals as needing work. However, ChatGPT recommended the teacher voice a simple procedural explanation: "To convert a percent to a decimal, move the decimal point two places left." Our team, by contrast, preferred an explanation describing how percents and decimals can express the same part-of-a-whole value, just in different forms.

Surprisingly, GPT's analysis of the high-quality lesson was the furthest from our team's analysis. In this lesson, the teacher asked the class to identify appropriate graphs for several small sets of data, a cognitively demanding task that requires students to understand the nature of the data and match it to an appropriate display. Like us, GPT noted strong student engagement and productive math talk during the lesson. However, it noted that the lesson "drifted" at times — seemingly referring to mathematically rich student-to-student conversations during pair-share work. Further, it docked the teacher for accepting multiple correct answers even though there were, in fact, multiple correct answers to some of the tasks posed. GPT also recommended the teacher follow a series of actions that would have lowered the cognitive demand of classroom instruction, including "providing concise, step-by-step directions."

Better prompt engineering is one solution to this problem. Teachers can, for instance, ask GPT to analyze a lesson from the perspective of discourse- and discipline-rich models of mathematics teaching. Sometimes — but not always — GPT will even suggest doing so. In an unexpected twist, at the end of an analysis of the lesson on matching data to graphs, GPT offered to analyze the transcript with our own observational instrument, the MQI. Setting aside some wonderings about how the MQI ended up in GPT's training dataset, Heather took up its offer. She found GPT's analysis of the high-quality lesson using our tool made the analysis and recommendations more aligned to our own. However, it also resulted in GPT underestimating lesson quality relative to our own analysis, in some places harshly so. It also persisted in perceiving legitimate multiple correct solution methods as the teacher adding confusion to the lesson.

There are other downsides to using ChatGPT. Because it uses probabilistic sampling, it can give very different answers to the same prompt applied to the same transcript. LLMs also suffer from sycophancy, telling users what they want to hear. For many, uploading teacher and student data to a LLM is a nonstarter due to privacy concerns, especially when information in the data could be identifiable. And ChatGPT typically offered five to six suggestions for improvement per lesson, which could be overwhelming or confusing to teachers.

However, the upsides are many, including the fact that ChatGPT-type feedback is free aside from the costs of capturing instruction and running the recording through an automated speech recognition platform.

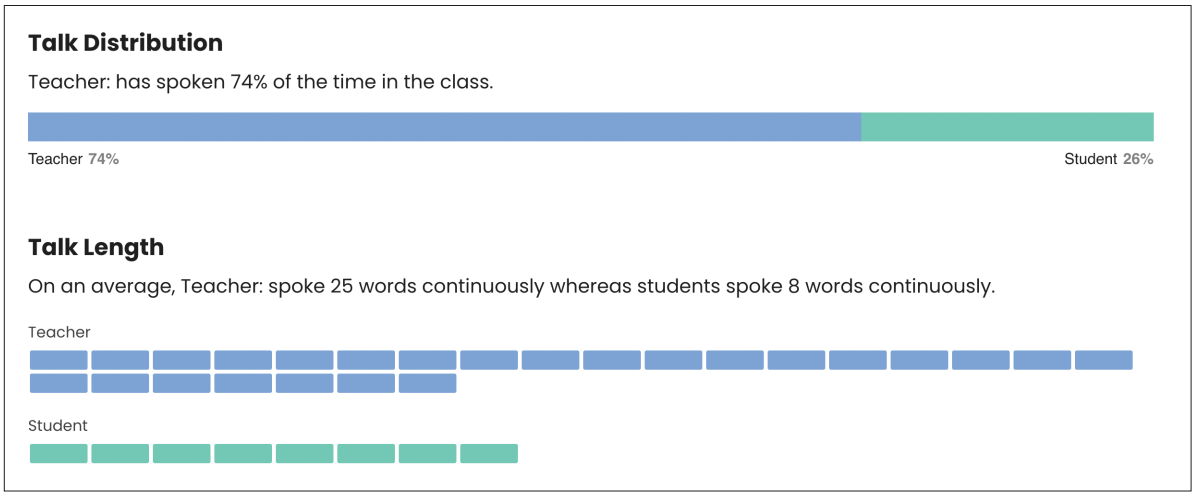
HUMAN-IN-THE-LOOP FEEDBACK AND COACHING

Other approaches to coaching with AI capitalize on the power of LLMs but provide more targeted analyses and feedback to teachers. These come in several forms.

One approach is using platforms designed for educators that provide teachers with private, automated feedback on their instruction. These platforms typically assist teachers in recording classroom instruction and then use machine learning models to transcribe and analyze those recordings. Platforms such as Edthena, IRIS Connect, TeachFX and Vosaic offer teachers feedback in consistent, structured categories, for instance reporting the amount of teacher wait time and open-ended questions, thus focusing attention on key, high-leverage teaching practices.

Teachers can then work with either human coaches — or, increasingly, AI-powered coach avatars — to analyze the lesson recording and arrive at concrete,

MPT DATA ON A TEACHER'S TALK PATTERN DURING CLASS



next-step improvement plans. This design allows for the benefits of human coaching, including support and collaboration over problems of practice, while also potentially improving the efficiency of coaching and ensuring teacher-coach conversations stay grounded in evidence.

A PILOT STUDY OF AI-ENHANCED HUMAN COACHING

We have been doing research for several years on such a coaching model, one that melds the automated feedback platform M-Powering Teachers (MPT) with an established program called MQI Coaching. A deeper dive into what we are calling MPT Coaching provides a peek behind the curtain into how a human coach plus an AI tool might help drive instructional improvement.

We began by working with Dora Demszky, a computational linguist at Stanford's Graduate School of Education, to brainstorm a set of measures of math instruction that could be automated by AI. We identified aspects of classroom mathematics instruction that were recommended as critical by researchers, measurable within teacher and student speech, and can, according to practitioners, provoke teacher insight and change. The chosen classroom phenomena include components of dialogic instruction — for instance, open-ended focusing questions, teacher revoicing of student thinking, and teacher affirmation of student competence (Michaels & O'Connor, 2015; Wilson et al., 2019). They also include how students engage with the mathematics content by communicating their strategies, explanations, and reasoning as well as by using academic language.

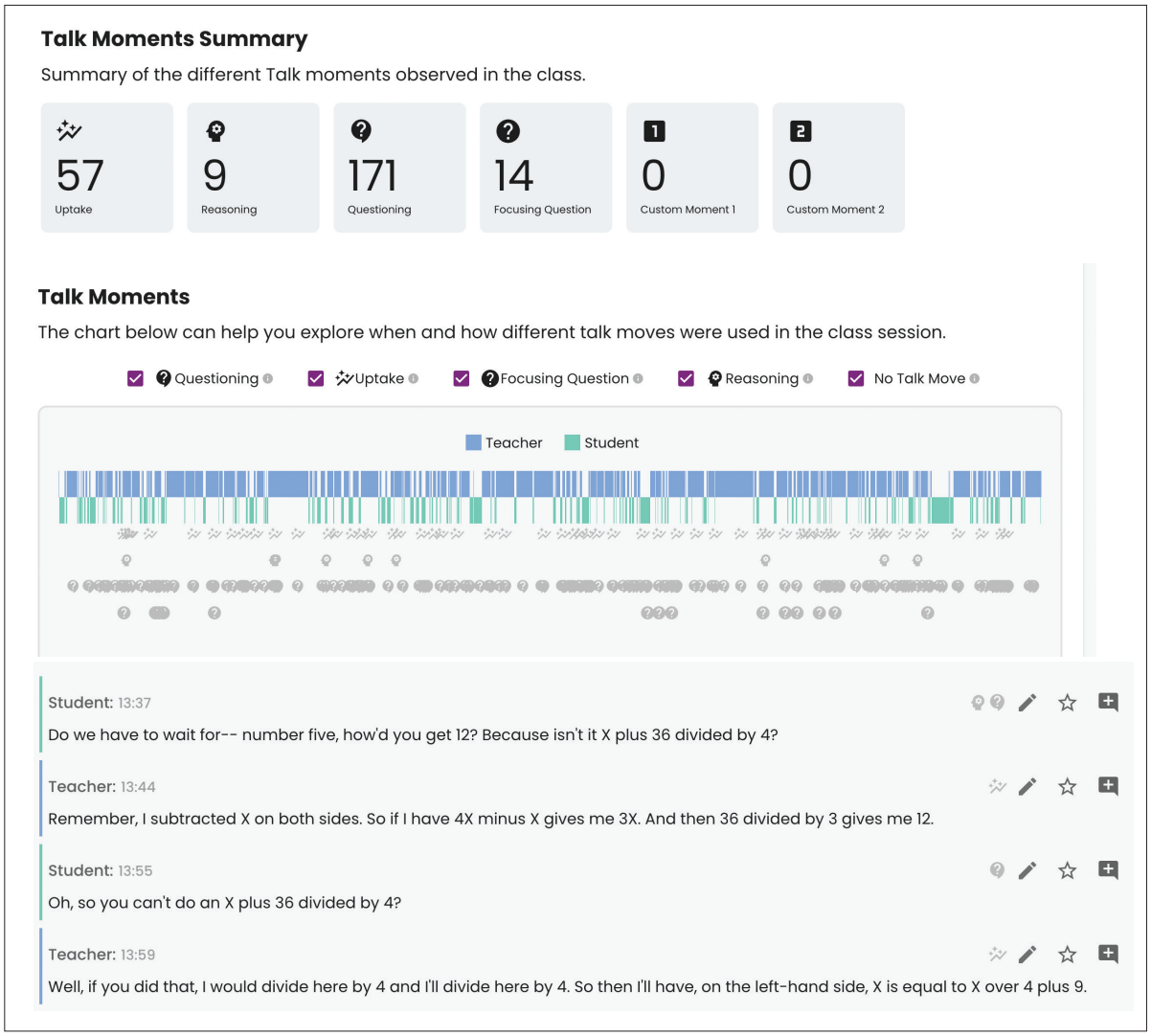
Next, we labeled line-by-line transcript data generated from hundreds of teachers' classrooms for the presence or absence of each practice. Then we used this labeled dataset to fine-tune existing large language models, asking those models to learn to identify instances of these practices in new data. Fine-tuning existing LLMs allows us to conduct analyses on our private, encrypted servers and thus avoid uploading teacher and student data to AI platforms.

A team at Stanford led by Demszky then built the MPT platform to convey this information to teachers and coaches. After recording and uploading a transcript, teachers log onto the platform and first see the ratio of teacher and student talk in their classroom, as shown in the figure above. Teachers are often surprised by how much they talk; 81% is about average for a typical transcript in our original datasets. Presenting a teacher with this information often sparks immediate self-reflection and motivates improvement.

As shown in the figure on the next page, the platform next provides teachers with a graph of their lesson, illustrating passages of teacher talk (in blue) and student talk (in green). Underlaid on the graph are icons for the classroom practices we want teachers to recognize and reflect upon. For instance, a teacher can see all the questions asked during a lesson (quote with question mark in it), focusing questions intended to uncover student thinking (circle with question mark in it), teacher uptake (revoicing) of student responses (hills and stars), and student reasoning (student head with gear in it). Clicking on any icon brings teachers directly to the portion of the transcript that shows the moment in which the practice happens.

In 2024 we engaged nine coaches who recorded 41 MPT-aided coaching conversations with teachers. Teachers had been trained in reflective coaching protocols, either by Harvard's MQI Coaching or Stanford's Center to Support Excellence in Teaching (CSET), but we also asked coaches to draw on their intuition and prior experience

DETAILED FEEDBACK ON TALK MOMENTS DURING CLASSROOM INSTRUCTION



while integrating coaching with the MPT platform feedback. We expected coaches to innovate in this space and wanted to learn from those innovations. We were also curious whether coaches could guide a rich conversation without actually spending time viewing video or conducting live observations.

We found they could. Coaches and teachers agreed on an MPT indicator to focus on, then dove into the information on the MPT platform, discussing the evidence most relevant to that area of focus. Links to the transcript allowed them to progress from pure numbers to deeper analyses of what teachers and students said.

In analyzing these coaching sessions, we found many areas of similarity with video-based coaching models. Coaches used the feedback to name and explore moments when teachers succeeded in fostering student thinking and reasoning. Coaches also used the talk moments and transcript to probe teachers' instructional perceptions and decision-making, often with the goal of helping teachers see connections between their instruction and student activity.

At the same time, coaches innovated in ways supported specifically by the MPT platform. One coach, for instance, pointed out that a teacher's focusing questions were not always followed by student reasoning, prompting the teacher and coach to look in the transcript for the cause of this disconnection. Some coaches extended this line of inquiry, disaggregating teachers' focusing questions into subcategories such as "why," "what," or "how" to examine how different question types elicited different amounts of student talk and reasoning.

Another coach asked a teacher to rephrase a closed-ended question to be more open-ended, working from the transcript to help that teacher prepare the new question. Additionally, several coach-teacher pairs started their

conversations by noting MPT's talk distribution visualization to remark on the overall the balance of teacher and student speech.

Over half of the coaches who participated in the pilot said MPT saved them time. Automatic detection of talk moves sometimes helped narrow the amount of video or transcript a coach needed to look through to find relevant examples. Coaches also appreciated the transcripts generated with automatic speech recognition and the ability to jump from a specific talk move to that spot in the transcript.

However, two coaches reported MPT did not save as much time as they expected. For example, one said, "It just minimizes … the time it takes to make notes. … You just reflect on what is recorded, which is still work, right?"

Coach interviews also pointed to the work we need to do to improve the model. Classroom transcripts often contained inaccuracies, leading to incorrect labeling of teachers' and students' talk patterns. Other inaccuracies were driven by classroom microphones failing to pick up student talk, pointing to the need for more sensitive recording equipment at a price point schools can afford. We noticed coaches working adroitly around these issues in their conversations with teachers, sometimes even using inaccuracies in the labeling of talk moves to help teachers internalize the meaning of specific measures.

WHERE AI COACHING MIGHT GO NEXT

Stepping back and looking at the broad picture of AI and coaching, we see other efficiencies AI can provide. Lindsay Clare Matsumura and colleagues (2025) have been experimenting with using generative AI to identify coachable moments in classroom recordings, saving coaches the work of watching entire lessons by recommending short segments for teachers to view. Using video rather than transcription also opens the door to new analyses, such as measures of teacher and student voice mirroring and conversational alignment.

Ultimately, however, the success of any of these tools depends upon whether and how teachers use them. In a study of automated feedback without coaching (Demszky et al., 2025), we found many teachers were unable to find time to read the feedback, as they were overwhelmed by meeting their classrooms' immediate needs. Blending human coaching with AI tools can help address these barriers while increasing coaches' reach and impact. This can allow human coaches, especially those who specialize in supportive persistence, to shine.

REFERENCES

Blazar, D. & Kraft, M.A. (2015). Exploring mechanisms of effective teacher coaching: A tale of two cohorts from a randomized experiment. *Educational Evaluation and Policy Analysis, 37*(4), 542-566.

Demszky, D., Liu, J., Hill, H.C., Sanghi, S., & Chung, A. (2025). Automated feedback improves teachers' questioning quality in brick-and-mortar classrooms: Opportunities for further enhancement. *Computers & Education, 227*, 105183.

Kraft, M.A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research, 88*(4), 547-588.

Matsumura, L.C., Walsh, M., Li, T., Correnti, R., & Litman, D. (2025). Exploring the use of generative AI to identify 'coachable moments' in classroom discussions. In A. Rajala, A. Cortez, R. Hofmann, A. Jornet, H. Lotz-Sisitka, & L. Markauskaite (Eds.), *Proceedings of the 19th international conference of the learning sciences-ICLS 2025* (pp. 2971-2972). International Society of the Learning Sciences. learnfwd.org/to4

Michaels, S. & O'Connor, C. (2015). Conceptualizing talk moves as tools: Professional development approaches for academically productive discussion. In L. Resnick, C. Asterhan, & S. Clarke (Eds.), *Socializing intelligence through talk and dialogue* (pp. 347-362). American Educational Research Association.

Wang, R. & Demszky, D. (2023). Is ChatGPT a good teacher coach? Measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. In *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2 023)* (pp. 626-667). Association for Computational Linguistics. learnfwd.org/vn7

Wilson, J., Nazemi, M., Jackson, K., & Wilhelm, A.G. (2019). Investigating teaching in conceptually oriented mathematics classrooms characterized by African American student success. *Journal for Research in Mathematics Education, 50*(4), 362-400.

•

Heather C. Hill is Hazen-Nicoli professor in teacher learning and practice at the Harvard Graduate School of Education. James Malamut is a postdoctoral scholar at the Stanford Graduate School of Education. Dorottya (Dora) Demszky is an assistant professor at the Stanford Graduate School of Education. Jing Liu is associate professor in education policy and the director of the Center for Educational Data Science and Innovation at the University of Maryland. Samantha Booth is director of the Mathematics Teaching and Teacher Learning Program at the Center for Education Policy Research at the Harvard Graduate School of Education. Claire Gogolen is senior manager of the Mathematics Teaching and Teacher Learning Program at the Center for Education Policy Research at the Harvard Graduate School of Education. Hannah Rosenstein is senior research project manager at the University of Maryland. ■