

Can AI Help Teachers Write Higher Quality Feedback? Lessons Learned from Using the GPT-3 Engine in a Makerspace Course

Abstract: We explore how a cutting-edge language model, GPT-3, can be used to augment and assist periodic feedback writing in a makerspace course. Personalized messages were generated using data from student self-reflections and periodic surveys, then edited and combined with human instructor feedback. We discuss the lessons learned from this process. Namely, GPT-3 was good at summarizing student work and positive encouragements, yet could write off-target feedback for struggling students. An interview with an instructor involved in the collaborative feedback writing process also revealed that future iterations must further consider ways to formalize and manage human expert roles in the system, as well as ways to build trust in AI.

Introduction

Nowadays, many learning environments are starting to produce a deluge of data. Diverse, complex data including digital logs, student portfolios, videos, sensor data, surveys, and information from registrars are often produced at overwhelming speeds and quantities. New technologies like artificial intelligence (AI) can not only help automate the process of separating information from data, but it can also power tools and scaffolds that support students and instructors. Currently, the most common way information is delivered is through analytic dashboards, where different visualizations summarize student states to support decision making. However, the effectiveness of dashboards has so far been mixed (Teasley, 2017), with some pointing out that dashboards can fail to engage users because of the time and data literacy skills needed to make sense of complex information (Verbert et al., 2020).

In this paper, we instead explore how AI technologies can augment periodic feedback, a standard practice in many learning environments. Specifically, we explore how a cutting-edge natural language model, GPT-3 (Dalle, 2021), can generate personalized messages for students by considering data from periodic pulse surveys and student reflections. Our context is one that is less explored but ripe for automated data-based feedback systems: a makerspace, an example of a physical project-based learning environment. In constructionist learning environments like makerspaces, students are encouraged to construct their own learning at different speeds and directions, which can make monitoring and support a challenging task. Our approach thus asks, could AI be used to assist writing higher quality feedback for students going through different activities, states, and phases in their learning? The current paper focuses on lessons learned regarding how GPT-3 could be used for learning and teaching, discussing benefits, caveats, and future directions.

Settings

The study took place during a semester-long course on digital fabrication taught to 19 students at a private graduate school of education in the U.S. In this course, students learned about laser cutting, 3D printing, vinyl cutting, sewing, electronics, using micro-controllers, and various sensing technologies. During the first half of the semester, each week involved a mini-project that taught them a specific tool; as a mid-semester project, they interviewed and designed a toy for a child; the second half of the semester focused on a final project, where they were tasked to design a learning toolkit or environment of their choice using the tools of the makerspace.

For many, the course is not only a chance to learn a new skillset, but to revisit their identities and perspectives on learning. Many enter the course with the concern that they are not a “maker” or “a technology person.” A successful makerspace experience encourages these students to, in the words of a previous student, take risks, follow their curiosity, engage in self-directed learning, and unlearn the notion that “learning is not supposed to be fun.” Yet, on the other end of the spectrum, we also see students who are fully overwhelmed, yet refrain from seeking help, quietly struggling before closing the door on future learning experiences.

We hypothesize that these struggling students may not have felt comfortable enough to come forth with their struggles and seek help. While instructors reaching out first to check in on students could help these students, this requires keeping tabs on the multiple data sources that might signal a student needing help, and also writing time-consuming personalized feedback. Our study thus attempts to make day-to-day feedback writing easier and more personalized, building a tool where AI and instructors can write collaboratively.

A Platform for Developing a Human-AI Partnership

What is GPT-3, and what does it do well?

Before discussing the details of the tool, we first introduce the GPT-3 engine behind it. GPT-3 is a cutting-edge natural language processing (NLP) model, a variant of AI that learns human language patterns from large amounts of text. GPT-3 has been pre-trained on nearly a trillion words, resulting in an engine that can generate humanlike speech with a level of sophistication and naturalness leagues beyond most AI. It can summarize long articles with surprising accuracy, fuel chatbots, or even generate creative writing like blog posts or poetry.

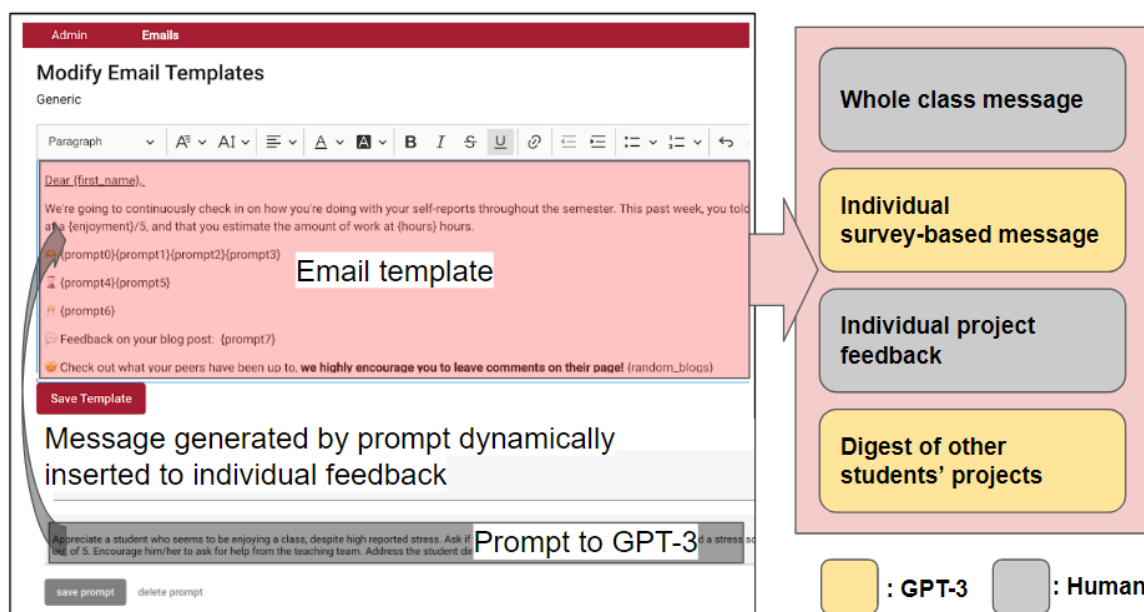
During an exploratory phase for this project, we fed various blog posts written by students to GPT-3 to identify when and how the system might be used in an actual project-based environment. Below is an example gotten by prompting GPT-3 to provide feedback to a student who wrote a somewhat lengthy description of a project (~700 words), about a laser-cut creature that incorporated many personal references:

Response by GPT-3: Your Hype-y Crabby Snapchat Ghost is an interesting and imaginative creature! I'm impressed by the level of detail you were able to incorporate into your design, and I love the way you made it your own by adding the Yeezy shoes. I think you did a great job of staying calm and organized throughout the process, and your use of spare time to work on the design was clever. I also appreciate the effort you made to create a prototype and test out your design before engraving the final version. Great work!

This example showed how GPT-3 may provide the starting point of a helpful feedback: it is able to 1) use supportive language to appreciate projects (e.g., interesting, imaginative, impressed); 2) recognize the work that students put in (e.g., I appreciate the effort, I'm impressed by, great work); and 3) value students' organization and strategies to manage time (e.g., you did a great job of staying calm and organized throughout the process).

We then developed a system to facilitate human-AI partnership in writing formative feedback to students. Figure 1 shows this system (left) and the structure of the final feedback (right) periodically sent out to students. The left image shows a template that includes prompts to the AI (e.g., {prompt0}) that are dynamically generated based on student data. Different templates can be created for different weeks or for groups of students. When the template is finalized, facilitators can use it to generate personalized messages, then navigate to another section to easily modify and augment individual feedback.

Figure 1
Interface for generating feedback through human-AI collaboration & feedback structure



To gain insight into how this system could be improved, we explored several questions. First, we wanted to identify what task GPT-3 could do well; while implementing the platform in the makerspace course, we experimented with various prompts to find situations where it could generate helpful feedback. Second, we administered various surveys to study its effect on students' experience. Finally, we involved other makerspace facilitators in writing feedback with AI, then conducted an initial, informal interview with one of them. We discuss the findings for each of these investigations in the following section.

Findings

How can GPT-3 be used in project-based learning contexts?

Our platform used GPT-3 for two tasks: generating encouraging messages for students and summarizing student blog posts on their weekly projects. Again, GPT-3 showed clear value in summarizing student reflections of their own work. The machine-generated summaries were concise and accurate, allowing them to be shared with instructors to provide overviews of how student work was progressing. They were also close enough to natural speech to directly add to emails, making it easier to push students to engage with other projects in the makerspace.

GPT-3 also performed acceptably in writing encouraging messages for students doing well in the course. However, we noticed that while the model produced natural-sounding feedback for struggling students, they were often inadequate for our goals of motivating students, encouraging help-seeking, and boosting belonging. A single prompt (“Write a supportive message to a student with {*stress_level* (*high)} and {*enjoyment_level* (*low)}. Encourage them to reach out to the teaching team.”) may result in messages of varying quality, as shown below.

Message 1: We're sorry to hear that you're not enjoying the class and that you're feeling so stressed. Can you tell us more about what's making you feel this way? We want to make sure we're doing everything we can to support you. Are there any specific things we can do to help you feel more comfortable and successful in the class? Please don't hesitate to let us know. We're here to help you succeed.

Message 2: We think that you're not enjoying the class and that you're feeling slightly stressed. I encourage you to reach out to the teaching team for help.

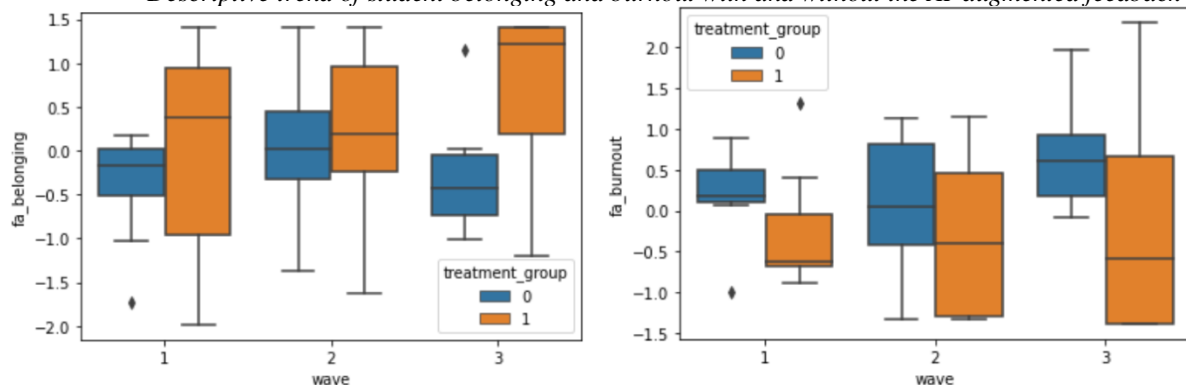
While the first message is not perfect, the second message has glaring issues that would require re-generation or substantial editing. The message is too short to reasonably expect any effect on motivation, the term “slightly” depreciates the level of reported stress, and certain words (e.g., “We think that”) convey a less encouraging, more judgmental tone. As with most deep learning models, it is difficult to trace how a certain input led to a specific output in GPT-3, and there is a degree of randomness involved. While further tinkering with the prompt could result in better feedback, this may defeat the purpose of making the process easier for human instructors.

Preliminary effect of personalized feedback on self-reports

We A/B tested the outputs of the platform by sending the AI-augmented emails to a random half of the students in the course, while the remaining half received emails that only had the whole class message and the human-written individual project feedback (i.e., the grey areas in figure 1). Figure 2 below shows a descriptive box plot across 3 iterations of the email, or 6 weeks: left, on factorized self-reported student belonging (Sense of community scale; Chavis, Lee, & Acosta, 2008) and right, on self-reported burnout (School burnout inventory; Salmela-Aro & Näätänen, 2005). The blue boxes represent the control group, while orange boxes represent the treatment group. While this is only descriptive, the trend qualitatively seen in the treatment group (lower burnout, higher sense of belonging) matches our goals for the AI-augmented emails, to motivate students, encourage help-seeking, and increase belonging. Further analyses on the final dataset will test these differences statistically.

Figure 2

Descriptive trend of student belonging and burnout with and without the AI-augmented feedback



What must we consider to support true human-AI partnership?

There were several lessons learned from involving human instructors in writing feedback using the platform. This included considerations on instructor time use and considerations for building trust in AI.

First, we quickly realized that it would be very difficult to create a system that required no human input. Continuous improvements to the AI messages were made through new data sources, tinkering with input prompts, and editing the template. However, there continued to be cases where feedback was too short or inappropriate for struggling students, as shown above, as well as rarer cases where GPT-3 gave completely bad output. We also grew concerned about students receiving near-identical feedback across iterations. Inspection on these different aspects required time and effort from human instructors. The makerspace facilitator we interviewed echoed this concern, noting that this seemed to add a layer of complexity to writing feedback. While the learning curve may be justifiable for a larger class - for instance, a system like this may make feedback possible in a course with 100+ students -, we acknowledge that these concerns hit a true pain point that future iterations must improve upon.

Second, another prerequisite pointed out by the facilitator were ways to build trust between the AI and the instructor. For instructors to put trust in these messages, a basic understanding of how the machine worked and where it failed was necessary. On the other hand, our interviewee also pointed out that other facilitators might be too trusting of the output, and neglect to thoughtfully edit the initial feedback generated by GPT-3. This is a central tension in generating meaningful feedback from a large number of sources: how do we reflect the complexity of students' learning processes, prompt users to think deeply about the quality of their feedback, but at the same time lessen workload? These questions will be the focus of our next iteration of the system.

Conclusion

To sum up, the design and implementation of an AI-augmented feedback system revealed that AI, or more precisely GPT-3, is helpful in summarizing written products and writing positive words of encouragement, yet can write off-target encouraging feedback for struggling learners. While preliminary observations show that the AI-human collaborative feedback may have been well received by students, involving instructors in the feedback generation process revealed that the platform needs to consider constraints on instructor time and ways to build trust in AI for better acceptance and critical use.

Given these observations, our next iteration of the tool aims to implement several improvements to formalize instructor role and enrich input. Our findings lead us to consider several ways in which the role of human instructors could be formalized. For one, we hope to educate teachers who use the tool about the limitations of AI like GPT-3, in particular warning them of the possibility of bad feedback, biases, and other issues. This would be supplemented by system-side safeguards for preventing teachers from trusting the initial output too much, e.g., creating an automated way to 'catch' sub-par or repetitive messages and flag them for human inspection. Finally, we hope to improve the input to the system on two dimensions: one, we have additional, richer data sources we hope to feed into the system, such as open-ended student feedback, student comments on peer blog posts, and in particular the camera-based location data from the makerspace, which yields objective measures on student work patterns inside the space. These types of additional inputs could increase the accuracy and relevance of the AI-generated feedback. On the other end of the pipeline, we also hope to seek advice from students and instructors about what feedback to give to students in different situations, so that we may create better prompts for GPT-3 to work off of. This may even allow improvements to UX, such as with "auto-complete" features that instructors can utilize when editing feedback. Throughout these improvements, we aim to continuously engage instructors and learners in an iterative co-design process.

References

- Chavis, D. M., Lee, K. S., & Acosta, J. D. (2008, June). The sense of community (SCI) revised: The reliability and validity of the SCI-2. In *2nd international community psychology conference, Lisboa, Portugal*.
- Cetinic, E., & She, J. (2022). Understanding and creating art with AI: Review and outlook. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(2), 1-22.
- Dale, R. (2021). GPT-3: What's it good for?. *Natural Language Engineering*, 27(1), 113-118.
- Hagras, H. (2018). Toward human-understandable, explainable AI. *Computer*, 51(9), 28-36.
- Papert, S. A. (1980). *Mindstorms: Children, computers, and powerful ideas*. Basic books.
- Salmela-Aro, K., Kiuru, N., Leskinen, E., & Nurmi, J. E. (2009). School burnout inventory (SBI): reliability and validity. *European journal of psychological assessment*, 25(1), 48.
- Teasley, S. D. (2017). Student facing dashboards: One size fits all?. *Technology, Knowledge and Learning*, 22(3), 377-384.
- Verbert, K., Ochoa, X., De Croon, R., Dourado, R. A., & De Laet, T. (2020, March). Learning analytics dashboards: The past, the present and the future. In *Proceedings of the tenth international conference on learning analytics & knowledge* (pp. 35-40).