

ALLEGHENY COLLEGE
COMPUTER AND INFORMATION SCIENCE DEPARTMENT

Senior Thesis

GPTA+

by

Stephen Palmer

ALLEGHENY COLLEGE

**DEPARTMENT OF COMPUTER AND
INFORMATION SCIENCE**

Project Supervisor: **Douglas Luman**
Co-Supervisor: **Emily Graber**

Abstract

An Artificial Intelligence training tool made in Python the fine tunes AI teacher responses to better encompass human teacher metrics.

Table of contents

1	Introduction	5
1.1	Motivation	6
1.2	Current State of the Art	8
1.3	Goals of the Project	9
1.4	Ethical Implications	9
2	Related work	11
3	Method of approach	13
4	Experiments	17
4.1	Experimental Design	17
4.2	Evaluation	18
4.3	Threats to Validity	19
5	Conclusion	23
5.1	Summary of Results	23
5.2	Future Work	23
5.3	Future Ethical Implications and Recommendations	29

List of Figures

List of Tables

1 Introduction

More and more in everyday life, Large Language Models(LLMs)and Artificial Intelligences(AIs) are becoming more and more relevant in a wide variety of fields. From banking to data analysis to video games, they are affecting almost every industry and expanding the capabilities of those who work in the field. One field that has been particularly slow to develop is education. Part of that is because LLMs are not as trusted by people because of their notoriety for being inaccurate, and also due to their lack of connection with students. There is plenty of academic research to show that LLM teachers are found to be less creative, engaging, and trustworthy compared to their human counterparts. This results in lower quality learning, as forming a connection with a teacher is a big part of what helps students learn. But, how specifically do most AI teachers lack these qualities? Well, to start, they're very impersonal, which is good for a wide audience, bad when students pay less attention because of it. There is also a distinct lack of humor, figurative language, or other common types of creativity that keep students interested in learning. Although it may seem like a new field to some, the subfield of Artificial Intelligence in Education(AIED) has existed and contributed a lot to education for over 60 years. Not in the sense of teaching a class necessarily, but it has had a wide range of applications. For instance, personalized learning paths are widely used nowadays Thus, GPTA+ was created to fix those problems with using AIs in the educational sector. It uses Open AI's Generative Pre Transformer(GPT) LLM in testing to produce an AI teacher that teaches at a level approaching the quality of a teacher.

The standard cycle of the tool is the AI agent generating a full lesson plan on the given study or textbook section, then GTPA+ generates a report card. This report card allows the user, an AI developer, to see what human attributes the AI teacher is struggling to achieve. Then, the user selects which one they want to improve, and give it a suggestion to improve, either pregiven or made by the user. Finally, that suggestion is given to the teacher, and it teaches the lesson again, in a process known as iterative development, scoring higher and higher on the next report card.

Given the scope of the problem and the limited success already achieved by LLMs like Chat GPT, why are they seen more already in colleges and schools? Well, a seriously contributing factor is that there's just not enough data available to prove their efficacy. Therefore, the goal of this project is to make an AI that has those human attributes, or at the very least, find out a method to test how much of these human traits an AI teacher has. Since the start GPTA+, the approach has differed multiple times. It began with training an AI teacher, but there was a need for an academically credible way to show that it had encompassed more human traits than standard AI teachers. To accomplish that, the first version of this project was to write up an IRB study focused on people learning from an AI teacher then having them take a survey to measure what Each of the participants thought of the AI teacher based on the previously mentioned human attributes. However, after some research into what had already been done in the field, a research paper focused on the subject

outlined that OpenAI has been carrying out this for over a year, having human participants grade each of these qualitative attributes. Since the AI teacher being tested is a GPT agent, not only would this not be new research, the work would've already been carried out on the specific AI this project is using. Thus there was a change in approach to finding quantitative ways on measuring human attributes of the AI teacher without the use of human input. Therefore, the project is now working quantitatively measure attributes such as creativity, coherence, and engagement. Because there is so many different possible methods to measure each one of these attributes, each with their own pros and cons, it is possible that this project might only be able to find a way to conclusively measure one, or it might be able to measure all of them. The goal at the end is to be able to measure as many attributes as possible, then to train a GPT agent off of their own results of these tests to improve the quality of it's output through newly measured metrics. The testing systems combined with custom prompting is called GPTA+, and the plan is to use the metrics to make a tool that generates personalized learning content including generating teaching material, key terms to focus on, and a tailored learning path with a heavy focus on reliability and creative content to keep students engaged. There is a priority of making it seem like humans are teaching it rather than an AI simply reciting a textbook.

Since its a Data Science project, the tool will be primarily interacted with via a Dashboard to see the various measured metrics, and ways to see the score for individual tests rather than the aggregated score for the attribute. The goal is for the AI teacher to show a marked difference in quality of content after applying the metrics to the agent to modify its behavior.

The AI teacher that is used to measure these metrics is a specifically trained GPT agent that teaches a sample paper to use as input for the testing data. Since then, the capability has been added for the AI teacher's lesson to be examined and rated via a batch of tests designed to measure creativity, coherence, and reliability. There is a system in place where when the AI teacher rates low enough on any of the tests, additional directions are added to its prompt that reflect its needed modifications.

1.1 Motivation

Due to the expanding field of AI in a multitude of fields, there are some unexplored niches that could greatly benefit the field they're in, such as agriculture and healthcare. In this case, education is the field where I felt that there needed to be more development. With the skills of LLMs like Chat GPT and some separate tools like GPTA+, plenty of people don't have the financial resources or availability of human teachers can still effectively learn new material in fitting environment, even if its taught by Artificial Intelligence. I felt that more people should benefit from the use of AI to improve their daily, and due to my minor in Psychology, I felt that I was fit to address that missing niche with GPTA+. The research and development for GPTA+ has been made easier given my background knowledge of psychology, emotions, and psychological studies. It also

helps with accessibility and ease of use for the users, such as improving visibility and simplicity so it's usable by a larger audience of prompt engineers.

With there being a teacher shortage both in America and internationally, there's a lot of areas such as rural or undeveloped areas that have little to no access to quality teaching without moving elsewhere. One of the goals of GPTA+ is to produce high enough quality AI teachers to give access to a decent education even with lacking local facilities or personnel. Even if there human teachers available, such AI teachers would be a great supplement. Sure, teachers are better at relating to students and making content more interesting, but AI teachers can take on the more strenuous and monotonous jobs to reduce the chances of burnout and allow the teachers to focus more on the human side of things. Stuff like grading assistance or making general lesson plans could help take some of the burden off teachers to perform overly repetitive tasks. I don't think this will necessarily replace all teachers if done right. Plenty of the capabilities that go into a teacher can't be fully emulated by an AI teacher. For for instance, physical activities are completely unable to be performed, whether a lab assignment with experiments or to move from table to table helping students personally. Thus, I believe a mix of both human and AI teachers would be ideal; AI teachers for grunt work while still retaining human attributes and human teachers for physical help, personal guidance for struggling students, and to appeal the content to students and break it down for them. However, that is and ideal case; it's still necessary to assume a purely Artificial Intelligence taught lesson for reasons of accessibility.

There is also the issue of increasingly expensive costs for education, so I wanted to help train AI teachers to make a low cost alternative that still has somewhat of a similar effectiveness of a human teacher. If the project meets its goal and significantly improves the AI teacher's human attributes, then affordable quality education would become much more accessible for the average person wanting to learn something. And this possibility is only going to grow more likely as AI development progresses at a rapid pace. It would also specifically benefit low-income families or those in developing countries without access to affordable education.

This extends beyond use in public schools. One of the most useful cases for a humanlike AI teacher is for working adults. If these people want a career change that requires an education, currently they can self teach it which is flawed or they can pay a large sum of money to become a college student or apprentice somewhere. If the long reaching goal of this project, a human level or close to AI teacher, then these adults will have a quick and easy path towards an alternative and effective education.

I've learned from Artificial Intelligences before, such as ChatGPT. And my problem with them is that its like a recitation of a textbook rather than a real teaching experience. It made me wish there were some better forms of alternative low cost education. My eventual goal was to have an AI teacher not only showing human attributes but playing the full role of a teacher: asking if anybody has any questions, relating course content to current events, simplifying the material or explaining it in creative ways. All of the non physical aspects of

teaching could eventually be feasible for Artificial Intelligence. Of course this project isn't that large in scope but it does play a role in bridging that quality gap.

If this goal of eventually fostering high quality AI teachers is met, then it could go a long way in improving the overall trustworthiness of Artificial Intelligence outside just the education sector. Increased human-like qualities play a large role in both education and in other fields. Even in education, trust is a critical factor; whether human or Artificial Intelligence, if the teacher isn't trusted, then the students' learning suffers, whether from not believing what is being taught or because the content is in fact inaccurate. This project has the potential to help address that skepticism surrounding AIs in education and more.

Finally, in a world with many new educational practices, such as fully remote or hybrid learning, GPTA+ can help to train AI teachers to better teach using new technologies. Although not fully demonstrating new human attributes, AI teachers are a large more adjustable when it comes to a rapidly changing learning environment. Plenty of human teachers struggled when COVID came because they couldn't adjust to the new system, so if something like that were to happen, having AI teachers as either a substitute or a good example for using the current technologies to their advantage would benefit a lot of the less tech savvy human teachers.

1.2 Current State of the Art

A GPT agent has been trained to serve as an AI teacher, and there is a series of tests that its lesson undergoes to gauge its creativity, coherence, and engagement. Based on those tests, there is additional directions added to the prompt so that the AI teacher fits those attributes more. This can either be done via pre generated messages that highlight things for the teacher to focus on, such as using more figurative language, or custom messages. These are entered by the user and are therefore much more specific to the problem. The current issue is that when the AI teacher receives feedback to its prompt via a pregenerated suggestion, some scores improve, but a large number actually decrease from these suggested improvements. That means one of three things; either the tests are not adequate, they don't measure the attribute I am looking for, or the text added onto the prompt doesn't serve its objective of improving the AI teacher's human attributes. That's why most of the time the custom messages are preferable to see better check scores.

Part of the reason for this project is to encompass multiple credible methods of testing human attributes and congregating them along with user feedback. Many of these tests didn't involve Artificial intelligence but rather simply analyzed the nature of human attributes. Since there is a collection of credible checks for these attributes, as well as a lower trust level of AI being teachers compared to real people, it is a necessary advancement to combine the checks to better the quality and trust of AI in the field of education.

There are plenty of organizations that are working to develop completely

independent AI teachers, such as David Game College. Plenty of these programs exist at a basic level, but are lacking much of what is required as a teacher. AI teachers can repeat textbooks, but a human teacher does much more than that. They relate the content to the student, use creativity to interest the student, make sure the lesson is clear or reworded in case the student doesn't understand. The programs looked at during research for GPTA+ are missing these human qualities due to the lack of conclusive methods to measure them that encompass all parts of the attribute.

There are also many text analysis tools out there, both for any text such as Grammarly or AI specific analysis tools like the text evaluation models on HuggingFace. Both of these point out mistakes and suggested improvements. Where they lack is iterative improvement such that the errors aren't even made in the first place. This tool affects the writing style ahead of time so there's no real time text suggestions but rather prompting suggestions for the teacher.

1.3 Goals of the Project

The goal of developing GPTA+ is to provide a platform for developers to improve their AI teachers to approach the level of human teachers via personalized adjustments and suggested fixes/suggestions to better teach the content. Thus, GPTA+ works to bridge that gap in education via quantitative tests to grade the AI teacher on its performance through 3 overall human metrics with multiple checks for each. Bridging this gap in educational quality will make quality learning much more easily accessible to the average person who wants to learn. These measurable, repeatable tests will allow GTPA+ to accurately measure and display the capabilities of the AI teacher when compared to a human teacher. Then, based off the score of these tests, GPTA+ will provide the various checks going into the final metric score and offer personalized optimization using customized prompts given by the user to raise the metric scores further and improve the quality of the AI teacher. The larger goal for this tool is to help AI serve as a substitute for human teachers when they are not available or convenient, such as someone who doesn't have access to a quality education. Currently, learning from an AI that is acting as a teacher is more like reading a textbook or being recited one rather than what actual human teachers do; break it down, make it easier to understand, and keep students engaged.

1.4 Ethical Implications

GPTA+ does have ethical considerations, as does almost any work related to AI. An obvious implication is reliability; no matter how reliable this project is there can always be inaccuracy of the resulting information. Whether by the actual training/testing tool or the teacher itself, the data can never be 100% trusted. There is also the chance that someone could end up using this tool to train an AI to act in an unforeseen way. That particular risk isn't as likely as the risk of inaccuracy but it's not impossible. This project also collects data; data from the AI teacher in the form of output, data from analyzing this

teacher’s generated lessons, and more. There isn’t much risk in the potential power difference or social imbalance due to the universality of this code; anyone who has access to the internet and an AI model like GPT. There are although of course still people who do not have access to those resources that will be unable to use this tool. There is also the risk that through prompt optimization for the human attributes, the AI teacher could get too personal or push the boundaries of what is acceptable as a teacher. This problem should also be possible to remedy through custom prompts to fix the behavior, but it must be considered nonetheless. If at any point in the training the AI teacher demonstrates these problems, they should not be used in their current state in an actual teaching environment. This issue is even more important when considering the audience of teachers, since many students are minors. Although all LLMs produced and publicly released have content filters to prevent profanity, AI teachers can still not use language fitting for a teacher because they don’t have the moral compass that human teachers have when it comes to what to say and what not to say to students. For instance, if an AI teacher is too harsh on a student for performing poorly, it might be able to tell when the message has been made and criticize to an overwhelming extent. Some of these issues can be fixed via optimization of the input prompt while other cases are just an example of the limitations of Artificial Intelligence in education. One of the larger ethical issues with GPTA+ is the risk of data bias, both in the data fed to the LLM used as the teacher and for the data received from each individual check. There are arguments for why each checks on why they don’t belong in a certain category of metric measurement, or why they qualify for a different metric. Each of these checks come from individual studies and projects that are specifically targeting a certain metrics, so there is a research basis for almost all of the checks used to prove the credibility of the placement in each category. In terms of data bias from the perspective of LLM data usage, the vast majority of the data does not come from the developer or individual teacher agent in question, so this risk is hard to mitigate. If a problem arises from a specific LLM’s data bias, it is recommended to either add a custom prompt to mitigate it as much as possible or switch to another model, whether a different LLM entirely or a different version of the same model. Aside from misuse or problems arising with the tool, the resulting effects of the tool being effective could possibly be a loss of jobs for teachers. If the standard for AI education is raised, not even to the level of a competent teacher, it could cost many their jobs without the position including the assistance of AI.

2 Related work

In terms of literature on the subject, there are several research journals in the field that help give insight onto what developing GPTA+. The first useful journal was written by Thomas K.F. Chiu finds via survey a high amount of interest among students in engaging in interactive ai learning[3]. Then, the subject of research became looking for how the learning process would differ from normal teaching from two journals; the journal written by Munir pointed out the different methodologies that one could employ for the AI learning tool while the other journal written by Bosede Edwards that goes through the different requirements that have to be carried out for AI learning and what different approaches are suited for such a learning environment[Edwards and Cheok [5]][7].

Along with literature on the theoretical makeup of an AI teaching tool, there are also already existing AI teaching tools to look into. There are many assistive tools that help teachers generate content to help teach their students such as Magic School AI, but less that actually carry out the entire teaching process. MIT also has their own guide to using AI in a teaching environment for material generation but not for a standalone project. There was one college, David Game College, that has an alternative program for GCSE(General Certificate of Secondary Education) students that teaches all subjects through an experimental AI tailored learning experience[6]. Although it's still in its early stages, it shows the recognized potential and limitations of AI in an educational context. Another source that contributed to that idea was McKinsey Global Institute's writings on the viability of AI in different fields including education, outlining where exactly the pitfalls of AI lie. It focuses a great deal on fields that are unlikely to be automated soon, and tasks that require human metrics like creativity are the largest roadblock for use of AI in fields that require creativity, reliability, etc[2].

Stanford's Pedagogical Abilities of GPT 3 helped to refine the human metrics that have solid quantitative measurements and the methods that have been used to achieve it[8]. They conducted a study consisting of a student learning from an AI teacher while measuring how the teacher performs in teaching different subjects as well as comparing models, outlining what each model specialized in. They also outline the 'AI Teacher Test' which set the standards for the minimum an AI teacher needs to approach human standards for teaching. It was a useful supplement, but used a decent amount of tests that weren't quantitative.

The need for alternative AI teachers to address the teacher shortage was discussed by Edwards in his article "Why Not Robot Teachers", which advocated for AI teaching assistants to be more widely used as well as suggested what human attributes are necessary to make that work[5]. He also suggested how important it is for student-teacher interactions for the use of audio in terms of a voice teaching the class as well as pointing out the necessity of human-like attributes for the teachers for any serious effectiveness.

There are also a multitude of studies looking into creativity and coherence, and part of the research has gone into the different tests of measurement for these statistics. One common method found in research is the use of other LLMs

to measure human attributes LLMs. For instance, many studies employ the use of ChatGPT to measure these attributes, because it has been trained on human data rating how much it embodies human attributes. This makes it usable as a replacement for a human, but since that isn't exactly true as ChatGPT still has some problems with creativity and other human attributes, a more suitable method is needed.

The last focus on research papers is differentiating between AI and human written papers and text to better measure human attributes. The focus The difference ranges from words that imply engagement to sentiment analysis and more. A great example is Sardinha's experiment on having people predict which text is human written and AI written. It was almost always obvious on which of the texts was AI generated due to the lack of human qualities inside the text, the impersonal nature of the writing, and overgeneralization of the material[1]. This proved the usefulness of this project in helping 'humanify' teaching material via human attributes.

3 Method of approach

To run this project, the user needs to have Python and Pip installed and needs to have cloned the repository. They also need to have access to an openAI API key, which will be put into the .env file. The following commands install the necessary dependencies and run the project. `cd src pip install openai pip install python-dotenv pip install streamlit pip install scikit-learn pip install textblob pip install textstat pip install nltk pip install spacy pip install tk`

At this point, you'll need to add your API key to the .env file to run the project. After you've done so, you'll need to run the following commands.

```
echo OPEN_AI_KEY=your_openai_api_key > .env      python -m spacy
download en_core_web_sm streamlit run webapp.py
```

Running the final command will locally launch the webapp, and will open up the window automatically and begin loading the first report card.

The original plan for this project was to use an IRB study to either predict AI teacher metrics as recorded by human participants and writing about the results or a study about repeatedly surveying students on AI teaching metrics and feeding those results back into the AI teacher, using existing datasets on the subject as a backup plan. The plan changed somewhat drastically after discovering a paper that outline how training an LLM model on human survey feedback is already being done by OpenAI with GPT, and using a self-made GPT for this projects that do the same thing seemed unnecessary and not unique whatsoever. Existing datasets have resulted in multiple clear ways to measure certain metrics. Thus, the project pivoted to measure the same human metrics but only through quantitative testing. A great example is creativity, where GPT-4 and up has been trained on human feedback enough to where studies have used GPT to grade the creativity of other LLM models with similar grading results to humans. Thus, the modified goal is to adjust the project to use multiple datasets and different strategies to generate test cases that can comprehensively grade each of the previously mentioned metrics.

This project uses OpenAI's GPT-4o model in order to generate teaching material. The prompting for this material includes instructions for the AI to act like a college professor who is passionate about teaching, specifically mentioning positive human attributes to show in their teaching. The model might change throughout the project, as it began in the GPT-4 model, especially due to the recent release of new models by OpenAI. GPT models are used for this project as they have been shown to score the highest on surveys and tests regarding human attributes commonly associated with teachers.

Since this project uses so many different checks to measure the various human metrics, each check is individually advocated for why it is useful, and its origin or use in similar projects. To begin, one of my most important individual modules is readability, a compound check with over 7 different checks inside of it[PyPI].Originally they were all used for creativity, but upon further examination and constructive criticism by my reader, they have been split up to better reflect what they measure. The first checks is the Flesch Reading Ease check

which measurements readability in the context of sentence length and syllable count. This obviously qualifies to measure coherence, and upon further thought creativity as well as multiple studies have concluded that syllable count is tied at least at a basic level to creativity. The second test is the Flesch-Kincaid Grade Level, which approximates the text into a rough U.S. grade level, which is a clear metric to show coherence. The third check is the Gunning Fog Index which estimates how many years of education is needed to understand the text. Due to its similarity to the Flesch-Kincaid Grade Level check, it is also used exclusively for coherence. The SMOG index is the fourth test; it is designed for scientific writing and measures how many words are used with three syllables and puts that number into an equation in comparison to sentence count to output a score. This is used for both coherence and creativity checks in GPTA+. The Automated Readability Index(ARI) is the fifth check which looks for character count per sentence rather than syllables and output a higher score for more complicated text. This is being used for the coherence check. The sixth check is the Coleman-Liau Index which functions very similarly to the ARI, so it is used in the same metric testing. The final check in the readability module is the Linsear Write Formula, which compares the amount of low syllable words with high syllable words with sentence size and returns a score. This is being used for coherence and creativity. As for other modules used in the metric grading process, the sklearn module was used for checking lexical diversity by comparing the ratio of unique words to total words, which is applicable to creativity since word variability is one of the few clear indicators of creativity according to a study done by Brochhagen in 2023. The re module, which uses regular expressions for string manipulation and analysis, was used in this project to find the amount of figurative language in the given text, compare it to the total number of words, and output a score for it. This was also used in the creativity metric's checks. The last check used for measurement of creativity currently is the unique word count function, which does not use an external module but rather uses sets to keep track of the total number of unique words divided by the overall number of words to output a score between 0 and 1, like the other checks.

In terms of checks related to coherence other than those in the readability module, there are three other checks. The transition word check, which is also imported using the re module. It looks for transitional words in AI generated sentences and counts the total amount. After comparing that count of transitional words with total words, the check returns a score between 1 and 0. Next, the sentence length variation checks for difference in sentence size without using an external module, only finding the range of the variation and dividing it by the largest sentence length to return a score between 1 and 0. Finally, the last coherence based test is the well structured paragraph test, which checks to see if the paragraphs are about 50 words in length which is the correct average length for paragraphs. The test finds the mean length of paragraphs and returns a score between 1 and 0 based on it.

The last human metric that GPT seeks to measure is perceived reliability of the AI teacher. The first test for reliability is the nltk's module to measure

sentiment analysis of given text to see the sentiment polarity of the text via a score between 0 and 1. This is used to measure reliability based on the fact that teachers are shown to be more effective if they speak positively while teaching. The second and final current check that is being used for reliability in GPTA+ is the direct address check, imported via the re module. This check looks for words that directly address the reader/student such as ‘you’ or ‘our’ and keeps track of the count of those words to return a score between 0 and 1.

Although for most of the project’s duration the output has come from the command terminal after running the main project file, eventually a front end window pop up was added where users can fine-tune what they want to correct about the teaching method. Currently, the front end of the project is actually a locally hosted Web App to better visualize the results of the tool and to improve the ease of use for users. The webapp is made via Streamlit which replaced the prior local window popup. On the webapp, each of the overall metric scores are displayed. Then, when the user specifies which of the metrics they want to target improving, a submenu pops up that allows the user to see individual check scores and to select one. This opens another submenu that shows two options; prewritten prompt or custom prompt. Pressing the prewritten prompt button will add about a paragraph worth of prompt suggestions that were made to targetedly improve that check in specific. All the checks in this project have a single prewritten prompt suggestion so this feature is always available. The user should keep in mind though that since this project is meant for iterative improvement of an AI teacher over time, premade prompt suggestions have an almost nonexistent effect after the first use, since adding a duplicate suggestion does not significantly change anything.

The AI teacher gets better iteratively using a set route of improvement. It begins by the teacher generating the lesson’s content based on the input study paper or textbook section. Then, the report card is generated showing how well the teacher did on each metric and check. Then, the user adds a prompt suggestion that is either created by them at the time of improvement or prewritten to target that specific score. Then, the suggestion or suggestions are given to the AI teacher, and it remakes the lesson with the suggestion having been added to its prompt. Then the report card is generated again, and this continues until a satisfactory AI teacher is produced that encompasses these human-like attributes.

This tool up until around the end of March was a local application, opening a window containing metric scores with subwindows showing graph results and individual check scores for the metric the user wants to improve upon. However, the separate nature of each of these windows led to difficulty during iterative development due to not being able to go between the pages quickly. So there was a need to pivot to something that visually is easier to understand and use rather than a simple application. Thus GPTA+ was converted from a local application to a webapp, specifically using Streamlit for design and hosting. It can display dynamic graphs and one of its the biggest selling point is ease of use, and considering the personal lack of experience creating webapps, it was the ideal choice. There was a problem during conversion where the torch module that had

been used since the graphs were first created isn't compatible with any publicly available version of Streamlit. This resulted in a pivot to a different module, , for dynamic graphing without causing incompatibility errors. However, this change did cause several bugs to exist inside the new webapp version of the tool, such as duplicate graphs or not all metrics/checks being shown infrequently.

During the implementation of this project, Copilot was used for a variety of uses to assist development. That includes text generation, module explanation and recommendations, docstring generation, and error correction/suggested fixes, and suggestions writing structure.

GPT 4 is used throughout this project, and although GPT 3.5 was used initially, most of training was done with GPT 4. Although new models show more human attributes such as GPT 4o, the variance from model to model can impact the effectiveness of the prompt, the suggestions, and the metric measuring wildy. Swapping up to more recent models would be a threat to validity without restarting a lot of the framework of the project.

4 Experiments

4.1 Experimental Design

Since a large part of GPTA+ is experimenting with the output of an AI teacher, most of the data produced by this tool comes from experiments. Currently, there is two different main categories of experiments going on. First is the regular tests that gather the metrics and are ideally improved over time. This category of experiments differs entirely in execution and results based on the fact if the user is adding to the prompt for the teacher or if they instead chose the pre written suggestions. An example of this first output is below:

```
--- Evaluation Results ---
Creativity Score 0.43
Coherence Score 0.44
Engagement Score 0.46
Reliability 0.18
Overall Score 0.38 (The range is 0 to 1)

--- Low Scoring Metrics ---
1. Creativity: 0.43
2. Coherence: 0.44
3. Engagement: 0.46
4. Reliability: 0.18
```

Which metrics would you like to address(Enter comma-separated list of numbers, e.g., 1,2)

The only way to show an experiment for this type of prompt addition to the AI teacher would be to use pre generated messages, but each check only has one pre generated prompt suggestion. Thus, without changing the method, only adding a suggestion once will cover all the available improvement without customized responses.

Each of the tests for these experiments is explained in the Methods chapter. The second category is the experiments testing if the checks are working properly. These are tested by giving each test a control that should score a maximum score for each metric. A standard output for testing each of the metric tests looks like this:

```
Flesch Reading Ease is proven to work! Score: 1.00
Flesch-Kincaid Grade is proven to work! Score: 1.00
Gunning Fog is proven to work! Score: 0.88
SMOG Index is proven to work! Score: 1.00
Automated Readability Index is proven to work! Score: 0.85
Coleman-Liau Index is proven to work! Score: 1.00
Dale-Chall Readability Score needs improvement. Score: 0.34
Lexical Diversity is proven to work! Score: 0.90
Figurative Language needs improvement. Score: 0.29
```

Unique Words is proven to work! Score: 0.91
Transition Words needs improvement. Score: 0.00
Sentence Length Variation needs improvement. Score: 0.25
Well-structured Paragraphs needs improvement. Score: 0.00
Polarity is proven to work! Score: 0.92
Direct Address needs improvement. Score: 0.22
Entity Consistency needs improvement. Score: 0.50

As mentioned earlier, pregenerated prompts are ineffective for continued improvements over time for the AI teacher using GPTA+. However, due to the controlled nature of experiments, custom prompts, which show the best results, are unable to be used for experiments. And since all improvements that could be added for suggestions, the iterative improvement is hard to show in experiments. Thus, the only way to circumvent this is to use specific quantitative tests and add strict requirements of what must be added. For instance, when during the final presentation, the Figurative Language check was used to show this. The first prompt suggestion as the normal pregenerated one targeted at improving the score through 3-5 bullet points, then each prompt suggestion after that point is introducing stricter and stricter requirements. The second prompt suggestion requires at least 3 examples of figurative language per paragraph, the third prompt suggestion requires 5 examples of figurative language per paragraph, the fourth prompt suggestions requires 7 examples of figurative language per paragraph, and the final prompt suggestion requires 10 examples per paragraph. Though these prompt suggestions do lower the score of other checks such as the Flesh-Kincaid grade, they do improve the score of the Figurative Language check. This improvement, however, is within the range of 2% total which is not a significant enough improvement in score to make this project a qualified success at the moment.

4.2 Evaluation

The experiments have shown three obvious results this far; first is that several of the tests are not working as intended. This necessitates improving the test to accurately measure it's given text or to swap out the check for a similar one that's better. The second result is that custom responses have shown more effect than standard pregenerated suggestions that are added to the prompt.

The final result is that in a controlled experiment environment, the tool is not a qualified success. Although one of its strengths is iterative improvement, GPTA+ does not conclusively and significantly raise scores in a controlled environment. This ignores two factors; the type of messages and the amount of iterations. These messages are premade and don't adapt to the differing scores but rather only a single check. It makes them less adjustable and therefore less reliable. Also, the iteration amount is quite small considering other training styles of AI agents like GPTA+. Five iterations showing about a 2% improvement isn't significant, but what about 10 iterations? 20 iterations? The limited nature of experiments leads to bad results for trying to push premade mes-

sages that far via increasingly strict prompts, hence my experiment only uses 5 iterations.

Below is an outlined list of current strengths and weaknesses of the tool.

Strengths: * iterative improvement * academically sourced checks * Almost all checks are quantitative and don't require human input * Scalability for metrics, languages, and fields of Artificial Intelligence * 3 metrics with numerous checks each * User friendly web app interface * Low computational requirements for local hardware * Prewritten prompts for less skilled prompt engineers * Custom prompts for more skills prompt engineers * Graphs displaying change in metrics over time * Graphs that display change in individual checks over time * Tests the effectiveness each of the checks using control messages * Modular design for checks, allowing easy addons of more checks and metrics

Weaknesses: * runtime of 20 to 40 seconds per iteration; varies depending on openAI's traffic * requires an api key to use the tool * Webapp has to be locally hosted before the tool can be used * Need for dependencies to run the web app, testing, and queries * Dependent on OpenAI due to Chat GPT usage * Costs money for regular queries, which can add up with iterative development * Difficult for a user to switch AI teachers without substantial code changes * Inability to reload previous states of prompt suggestions in case of a bad or ineffective suggestion * Limited token count means lessons grow smaller the more iterations are done * Has a bug where the creativity metric is the only option available to improve * Only effective for English speakers * No human scores to relate check scores to * Lack of any real world testing aside from personal use * Currently limited to specifically education not Artificial Intelligence in general * No offline capabilities; needs a stable source of Internet access * Steep learning curve for the user; needs some level of prompt engineering knowledge for best effects

4.3 Threats to Validity

One threat to validity is in the effect of the pre generated messages. While the experiments have shown improvement in the tests following adding a pre generated message to target a certain metric(s), there is little to no effect if the user still wants to improve that metric. This is due to the fact that there is only one pre generated message for each metric, and thus if the user wants to improve the metric twice the prompt will only get a second copy of the original pre generated text. This threatens the validity of any user of this tool who is unfamiliar with prompt engineering and thus can't create effective custom prompt suggestions.

Another threat to validity is the risk of the AI teacher overfitting to meet certain metric scores while ignoring others. It could be overall a worse teacher if the user only focuses on one specific check which worsens the score for other checks. This is very possible and even happened during the Figurative Language experiment. While the Figurative Language check increased iteratively in score, almost every other check suffered as a result which lowered the overall scores not counting Figurative Language. This is mitigated by the use of custom prompts;

they allow the user to keep other check scores in mind while also improving the targeted check, which is the best of both worlds. This doesn't work in a controlled experiment of course but in regular usage the risk is overall minor as long as the user knows how to make fitting prompt suggestions.

Along with overfitting to the metric scores, there is also a risk of overfitting to the specific use case used in testing. All of the training has been done with a college teacher teaching Psychology study papers, and it could be much less effective in for example a 3rd grade class setting. Mitigating this requires changing the AI agent's role to fit whatever the use case for the teacher is.

One possible threat to validity is misinterpretation for what each check and metric represents. There is no explanation inside the tool on what each of the checks do, so without prior knowledge it's possible that the user might misinterpret a check or metric and make a custom prompt that doesn't actually help improve the score at all. To mitigate this, the simplest way would be to supplement the in app text to include drop down descriptions of what each metric and check represents. This is less necessary on certain checks, like Well Structured Paragraphs and Figurative Language, but very necessary for checks like Flesh-Kincaid grade or other less self-explanatory checks.

Also, the fact that the experiment specifically used Figurative Language is a threat to validity. It was chosen at random from the possible checks where adding strict limits would work to iteratively improve the score in a controlled environment, and the improvement was present but minor across 5 iterations. But what if that was the least effective or most effective check for artificially improving metric scores for experiments? If the experiment used the Well Structured Paragraphs check instead, the experiment results could've been very different, in the sense that either the improvement is more drastic or even the score decreases with strict prompt suggestions across multiple iterations. This threat can be limited by carrying out experiments for all the quantitative checks that can be used with strict limits in the prompt suggestions.

A particular threat to validity is what the measure of success is for the effectiveness of each of the checks in GPTA+. Although they all are used for a specific academically credible reason, its hard to measure if they do what they're supposed to do. The only two ways to tell if each check is working effectively is either the use of test messages that are meant to score low and high or having a human grade the generated lesson plan personally rather than using checks. The latter method isn't very feasible, as it would be difficult and unfeasible to have a bunch of people regularly grading AI lesson plans to test the effectiveness of the checks. Even if I am the only one grading the resulting lesson plan, that introduce bias with me being the only person measuring the lesson through opinion. My idea of creativity might be wildly different than what other people think of it as. Thus the best way to lessen this risk is to use a large amount of test messages to routinely make sure that the checks are working as shown in academic sources.

One future threat to validity is emerging technologies and pedagogical research rendering current checks obsolete. This is a fairly easy fix; in the future when emerging research finds better checks or completely new metrics, they can

be merged with the existing project so it retains its function and becomes more effective.

A further threat to validity is the risk of bias in training data. If a historical event is told from the perspective of a different country than where the students reside, it could have unintended results and less effective teaching. It could also have lasting impacts if some stereotypes or incorrect facts based on substandard input academic sources. Even if the material is changed afterwards, if the prompt is not reset then any custom prompts related to such skewed information could cause it to repeatedly make the same inaccurate claims or opinions.

A key possible threat to validity is the generalization of the various checks into singular overall scores. Although all the checks are related to their accompanying metric, are they all exactly equal in importance? That what GPTA+ does, average out the scores equally into one overall metric. The problem with preventing this risk is the lack of research or evidence to suggest what checks are important to prioritize and which are not important to prioritize in the weighing of each score. It could be completely possible that one of the checks for creativity, Flesh-Kincaid grade for example, is actually one of the most important scores and should make up 50% of the overall metric score. Thus, there is no clear cut method or research to support the split of credit for the checks into the overall metric score.

A threat to validity that is pretty important in this project is the limitations of the current tests. They all do have an academic basis for being used where they are in each metric, but they all have limits. There could always be more checks added that brings the AI teacher closer to a human level of education. There's no clear limit to how many tests must be added to reach a human level. This is also impacted by the lack of quantitative testing for several different key human attributes. There's no consistent quantitative way to measure emotional intelligence from a lesson plan for instance. Since that's a key part of the human attributes that human teachers possess, these checks will never perfectly measure what humans can measure.

Another threat to validity is the variability of Artificial Intelligence, Chat GPT 4 for this version. Even when given identical prompt suggestions, the variability each text generation can change the scores such that seemingly significant scores can be skewed because of a lucky or unlucky generation. So although overall the score will still change, this variability threatens the validity of the experiment and to a larger extent prompt suggestions as a whole. This is not currently a preventable risk, but mitigating is possible with more stringent and strict suggestions. However, as the experiment showed, this carries the cost of potentially lowering all the other scores that don't have strict requirements. As Artificial Intelligence develops further perhaps this risk will go away, but given that variability to a certain extent is a priority, its uncertain when if ever this will happen. However, aside from the check scores, it would be better to have more varied lessons. If all lessons are taught the exact same way with the same amount of jokes, tone, and overall word choices the lesson plans will become much more monotonous and boring.

One concern about validity is the variability of custom prompting. Even if

the user is knowledgeable on prompt engineering, there is always the possibility that the AI teacher completely misunderstands the prompt suggestion, leading to an unexpected result.

Another threat to validity is the heavy focus on text in this project. It helps for fields that primarily focus on text, but for those that a lot more visual tools to teach material such as physics or art for instance, metrics such as coherence might be artificially lowered due to missing visual context, even if the AI teacher demonstrates reasonable human attributes.

Even for text based fields, there isn't enough training in edge cases which is a threat to validity. Some subjects that are highly technical such as quantum physics could lead to unexpected results from the training process due to how unconventional they are.

Focusing on just GPT+ to produce a humanlike AI is a threat to validity. Along with human metrics that can't be measured quantitatively like emotional intelligence, trying to strictly score high on all checks could backfire. If strict limitations were set to make all paragraphs the same number of words, the same number of figurative language uses, and so on, they could score high at the cost of actual lesson quality. Even if the score is high, using 5 examples of figurative language in 6 sentence doesn't contribute to an effective learning environment. The risk is mitigated by trying to not set strict numerical limits on all prompt suggestions to not have this problem.

This threat to validity is one of the unintended consequences for improving AI teachers through iterative development. Depending on the content in the custom prompts, the prioritization of these textual analysis checks could lead to overly wordy, hard to comprehend lessons.

Although earlier in the project, there was an upgrade of the AI teacher from GPT 3.5 to GPT 4, regular upgrades to the latest model can cause a threat to validity. New models can have wildly different behavior, so maybe the pre generated prompt suggestions will have less effect than before.

5 Conclusion

5.1 Summary of Results

Overall, there were two major goals of this project. First, accurately measuring the human attributes of an Artificial Intelligence teacher, and second is iteratively and significantly improving that Artificial Intelligence using the measurements as a basis for the improvements. The first goal was overall a success. Large changes in the teaching style do adequately affect the report card scores given to the user. Where this project lacks is in significant improvement iteratively of the human attributes of the AI teacher. The experiments were limited in the sense that each improvement made by the user had to be controlled and decided beforehand which limited its effectiveness greatly, but that doesn't change the fact that the experiment only showed about 2% improvement over 5 iterations. This would be higher if custom prompting were allowed, having the user adjust to how each score changes iteration to iteration. Thus, although this project isn't a qualified success, it is a success with limited results.

5.2 Future Work

This subfield of using AI models like Chat GPT in the education sector has a broad scope of future possible work. From designing a visual representation of the teacher to adding relatability to nailing down more stringently the quantitative ways to measure human attributes. In terms of the 'next step', it is reworking how feedback is given back to the AI teacher. Adding to the prompt is crude compared to fine tuning the newer agents, so it would be interesting to see if the training process is more efficient for the average user. Also, creating a guide for users to write their own customized prompt addition. The pre generated prompt suggestions are on average less effective than customized prompt suggestions by users who know what to tell the teacher.

This project uses GPT 4 for the custom AI agent used in this project and originally used GPT 3.5 before GPT 4 was released. Using either different models from OpenAI such as 4o or entirely different companies' LLMs like Claude or Opus could greatly affect the efficiency of GPTA+'s training and effectiveness. The training style used in this project might be much more effective on a different model and is therefore a useful avenue for future work. Another version of this improvement would be to always include multiple different queries to different LLMs that could serve as teachers such that they all receive any prompt suggestions and generated their own lesson plans and were graded separately. This could then compare each of the metric scores in a color coded line graph to show what metric applied the prompt suggestions and improved the most. Perhaps a study could be made comparing their specializations and adaptability based on feedback from GPTA+'s report card.

This idea could be taken further by having multiple different LLMs working as teacher collaboratively, sharing insights from refinement in terms of prompt suggestions as well as other improvements throughout iterative development.

There could be a system of each different LLM taking a different part of the prompt suggestions and comparing the results with the other LLMs' resulting report cards. Then, if one of the LLM's is showing much better results with human attributes, that LLM's suggestions could replace the existing suggestions for the other LLMs as a form of cross-training. Having a network of AI teachers collaboratively working might produce much higher quality lessons with better demonstrations of human attributes, with each covering individual drawbacks of other LLMs.

For OpenAI use in this project, an account was made to pay for the repeated queries over the process of iterative development. This account uses a 5000 token limit of input plus output, which limits the lesson plan and prompt size making larger lessons or more suggestions difficult to implement in the long term. Using an API key with an extended plan or a different AI without those limitations could greatly expand the effectiveness and capabilities of this project.

If the current token limits were not an issue, it would make sense to expand upon the actions of the AI teacher. Rather than just teaching a lesson plan about the given academic source, the AI teacher could make quizzes, study guides, classwork activities, and flash cards. It could even stop periodically to ask questions and receive responses from students, introducing more variability in its responses and giving the AI teacher more opportunities to display human attributes and attempt to build trust with students.

In terms of the subfield rather than my project specifically, a stringent way to measure many of these human attributes like creativity would greatly help the development of AIs that need interpersonal skills, like a teacher would. Another missing niche is a credible source of human data for data usage purposes that doesn't require a IRB. Even a credible method of creating synthetic data that is close enough to the real thing would greatly help the development of research in Computer Science and Psychology. There is existing AI-generated student teacher discourse but the credibility of the data is questionable.

Even without having credible synthetic data, the existing sources of synthetic data could be used in the future to simulate full class sessions to measure the adaptability and response to questions and unexpected scenarios to make the training process much more dynamic and well rounded, beyond quantitative checks, though they would still be used.

An alternative to credible synthetic data generation is to actually demo this project in a sort of test class to gather more data on real students with their consent. A similar program was run at David Game College; an alternative program with a normal class also available so they can collect data and refine their teacher[4]. With the use of quantitative checks and actual student and teacher feedback; the accuracy of the scores and refined suggestions could be monumentally improved. A lot more metrics would need to be measured every iteration, including student performance, how engaged they are, and retention of the knowledge and application in other parts of the class. This would add a great deal of credibility if the scores were similar to the quantitative check scores, as well as helping develop GPTA+ via showing which of the checks is more important to human students so that they can have greater weight in the

final metric score. In terms of teacher feedback, asking the AI teacher post lesson how it thought it did at the lesson could provide another layer of finetuning; one for the student feedback, one for the check feedback, and one for the teacher self-reflecting on the class and improving its own behavior in terms of human attributes.

If this was implemented in a classroom, rather than feeding in the feedback after class, it could be worth trying for a real time feedback iterative improvement. Especially if there is a large problem due to some past suggestion, this would allow for quick fixing and improvement throughout the class session. There would be risk of students purposely giving bad feedback for entertainment, but depending on the level of supervision that might not be too big of an issue.

If an AI teacher was considered trained well enough to teach an actual class, there should be a certification system added to test if the AI teacher is fit for whatever culture and level of school it's going to be teaching. A survey, maybe, asking for how the AI teacher would respond to various situations happening in the classroom, and if it doesn't meet some minimum score, it has to be retrained. It would be too risky for all AI teachers trained by GPTA+ to teach in real class settings without other forms of ensuring that they can serve as a viable human teacher alternative.

Of course, with many other companies working on AIs that could serve as teachers, there is more work to be done using other AIs. This wouldn't be a monumental add-on but a comparison between different AI's results might find certain strengths and weakness in their teaching that was not previously seen.

This project originally began with measuring more human metrics than the 4 currently being measured, mostly due to the goal of fine tuning a smaller amount of metrics vs. working less on more total metrics. Many of these affect teaching and could improve the teaching further. In particular, reliability also known as trust in the teacher is a key factor that would be a very useful addition to the current report card generated at the end of GPTA+ grading. There are also some human attributes that go beyond the limit of what is currently measureable, such as grading emotional intelligence. Future research could provide an avenue to measure and therefore improve more complex metrics such as emotional intelligence. These improvements could be added in the same way that all of the other checks and metrics were added, or a front end way of implementing them could be created. This would further improve the ease of use for AI developers and allow for dynamic pivoting to focus hard on 1 or 2 tests and add a ton of checks to them for maximum accuracy and tuning. This could either be added by adding a button for a code window popup for the user to put in a new check in Python and for it to be added to the bottom of main.py, or it could possibly be done through a second api query. A code specific openAI query could take in textual input for what check the user wants and then return an implemented version which would then be added to the bottom of main.py.

This project is made in Python due past experience with it as well as ease of use, but there are some other viable options. Switching to a more data focused language such as r could greatly improve the analysis tools available as well as

extensive libraries for data manipulations. This makes it a viable direction to consider working towards.

Aside from just running the checks, there are other language other than R that are tempting additions to this project. A more efficient language like Rust could vastly improve the runtime speed. This would help address one of the more taxing problems with GPTA+, which is a fairly poor runtime at 20-40 seconds every iteration of the tool. Since the priority is iterative development that time adds up and is a significant obstacle to quick development.

Another tool that could be possibly switched out in some future work on this project would be Streamlit. It is used in GPTA+ due its easy visuals and ease of use. Due to personal lack of experience working with webapp creation, it was a desirable option. However, in the future, working with a more robust and scalable framework like Flask or Django could improve the visual representation of the report card as well as having more complex customization options.

Due to the fact that this project works entirely with text and graphs, a more visual learning system could significantly improve the results. generating images to demonstrate the material better beyond graphs would be a useful addition to this project. This could include helpful charts, prompting guides, and other ease of use items to make custom prompts more effective. Along with images, having the AI teacher vocally teaching the lesson would add another layer of complexity to the teaching and the implementation, as vocal tone and style would impact the teaching a lot. The same goes for giving the AI teacher a face; which has been shown to increase human attributes like reliability. However, there are risks of achieving the opposite result if the face isn't well made enough or synced to voice improperly. There would also be the concern of what the AI teacher would look like and how that would affect teaching efficiency based on different audiences.

The training cycle could in the future, handle improvement of visual media as well. If the AI teacher also generated images and graphic related to the lesson, they could also be tested for the various metrics. However, many of the checks are text exclusive so beyond analyzing the text inside the images it would be difficult to fine tune the image generation to demonstrate human attributes.

There isn't any flexibility in teaching depending on the student beyond culture and language. For instance, if the student is a auditory or kinaesthetic learner than they will have less effective learning if the lesson is purely text based. Ideally in the future there would be a audio component such as a voice saying the lesson content, a visual component, either a face to look at during the lesson or generations of charts and graphics based on what is currently being said by the teacher, and a kinesthetic element. In terms of helping kinaesthetic learners, beyond class activities and calling upon students to demonstrate their knowledge individually, there aren't too many ways that a fully virtual teacher can facilitate learning.

This idea of making AI teachers more appealing to students using GPTA+ could be expanded upon using a number of emerging technologies. One of the greatest setbacks for using Artificial Intelligences as fully functioning teachers is the lack of a physical presence worsening the 'humanness' of the teaching.

Generating a face and voice could help that, but so could using Virtual Reality(VR) or Augmented Reality(AR) to make the teaching much more personal, as well as giving more opportunities for displaying human attributes. Facial expressions, use of the new technologies to teach more visually interesting classes than standard in person teachers, or even a haptic a system for actual simulated contact with students would help personify the teacher in the students eyes as well as give more data for GPTA+ to analyze and use for improvement in the next iteration. If the other future suggested addon was implemented, which was a greater variety of generated lesson content including labs, quizzes, and in class activities, then adding either of these technologies would improve the standard quality of all of these additions to almost balance out the AI teacher's lacking human attributes when compared to a human teacher. Although GPTA+'s goal is to bring AI teachers as close to these human attributes as possible, much of it isn't testable or recreatable at the moment. So, even ideally with all check scores being high, there will very likely still be some level of lacking compared to human teachers. Higher quality teaching material and activities could mitigate that difference.

In terms of benefitting the AI trainers, a tutorial system or automated prompt suggestion generation would greatly expand the target audience for users of GPTA+. Whether simply a series of videos describing the iteration system used in GTPA+ and a prompt engineering tutorial or an actual guided tutorial showing the training process, either would help less experienced LLM users to effectively improve their AI to show human attributes. Currently, anyone who isn't knowledgeable about prompting styles will have trouble showing consistent improvements beyond the premade suggestions. To that end, auto generating prompt suggestions would also be a future possible solution to the small user base. GPTA+ could stage another, smaller query, showing the scores for each of the checks and asking the AI teacher to find faults in its lesson plan, basing the next prompt suggestion on what the AI teacher thinks specifically is lowering its scores. These might not be as effective as a prompt engineer, but it would certainly helped any who are not experienced in prompt engineering use this tool to near full effect. This automated feedback loop does have potential downsides, though. It could lead to errors in teaching getting larger and larger through multiple iterations, which could greatly lessen the validity of the training and lead it away from human-like teaching attributes.

This could be taken a step further by having the AI teacher refine the metrics rather than just respond to their scores via prompt suggestions. If a pre generated prompt suggestions has little to no effect, the individual check that failed to function properly could be removed or swapped out depending on the AI teacher's assessment of it. There is the risk of losing the credibility of the scores if the AI teacher ends up getting rid of checks that it consistently scores low on, but on the other hand if it worked correctly it could ensure stable scalability of metrics and checks by automatically filtering out the bad checks and improving existing checks and metrics. This would result a faster training time with hopefully more accurate results from the checks.

This project is meant for training Artificial Intelligence to be more competent

teachers, but a lot of the methods for that aren't education exclusive. This training method, with some relatively minor adjustments, could be converted to a multitude of different fields that use AI text generation. For instance, having GPTA+ grade a corporate HR AI would be well within the overall scope of this tool. The primary things that would have to be changed would be the core prompt describing the position of the AI as well as the input material for the AI to generate text about. For this specific example, something like an HR handbook describing how to deal with questions and new situations could always be added as a part of the input, while the specific problem or question would be the actual data input that would cause a need for the AI to react to a specific case. A job like is still able to be done by an AI, and in that subfield as well, human attributes like creativity and coherence would be very helpful for the AI to have more of. Even within the field of education, adjustments could be made to better fit the situation. This current GPT agent is told that its teaching a college class and acts accordingly, but a corporate trainer would act very differently. This would require even less modifications, only giving the material traditionally taught as textual input to the teacher and add on the functionality to generate story problems for better corporate training. The basic checks used in GPTA+ aren't education specific, but rather patterns in text that imply certain human attributes, so little to no changes would have to be made to them. This also addresses part of the Motivation section of GPTA+ leading to AI teachers that allow adults to switch career paths. Having the AI teacher work from a corporate perspective would help fulfill that goal in the long term. This would however worsen the ethical implications of costing people jobs beyond just the academic world.

This project is meant for English learners, the applicability of its use to different cultures is very limited. Adding the capabilities for other non-English users would be a useful addition. This would be done by translating via either an LLM, asking a native speaker of whatever language is being added, or a standard translation tool like Google Translate to convert all writing on the webapp into the target language. The same would have to be done for the prompt query and all text involved including the agent custom role text and the pre generated prompt suggestions. Then, other than changing the input lesson content from what was there before to academic data from that language and culture, the tool is fully capable of working in multiple languages

One of the most useful additions that could be added to this project in the future is the ability to reload past states. Whatever suggestions for improvement are given throughout the training, they are permanent. They could be invalidated via a future prompt suggestion, but if the user makes a prompt suggestion and the scores they want to improve and go down, they can't choose to immediately reload to the previous state. Adding this would help the tool's ease of use a lot and make iterative development much smoother.

Another beneficial addon would be improving the runtime for this tool. GPTA+ takes about 20-40 seconds per iteration to generate a new lesson plan and report card, different depending on how busy OpenAI's web traffic is at the time of regenerating the lesson plan.

There are currently several small bugs that can show up during the use of GPTA+, and are therefore a target for future development. For instance, sometimes there is a duplicate graph for individual check scores after selecting which overall metric the user I want to improve. This doesn't affect the functionality per se but does affect how long each iteration takes by making the user have to scroll down further to improve the AI teacher. One bug that does affect the tool's use is the toolbar displaying overall metrics to grade, sometimes only showing creativity as an option. All graded metrics and check should be displayed, and the checks themselves always come up but occasionally there is only creativity available. This is an avenue of future work that directly affects functionality and is therefore high priority.

Another potential limitation of the current setup regarding the use of OpenAI's GPT 4 is the token limit for queries. The output and input must be 5000 tokens or less total, and with each input growing gradually via prompt suggestions, there might be a problem of output. After reaching a sufficiently high iteration, the output lesson script that is generated by the AI teacher will grow gradually smaller to fit inside the token limit. This can be fixed via a more expensive subscription or use of a different AI teacher with a larger token limit.

This project GPTA+ is currently run via a locally hosted webapp using a streamlit structure. This means that it is not constantly available without executing the code for launching the webapp. Publicly hosting it would increase its accessibility to people who aren't familiar with python or using code directly and just want access to the interface

Depending on the audience using this tool, which is developers who want to train an AI to have more human attributes, there is a possible need for offline capabilities. If the model being trained is run locally then there isn't a need for an internet connection which improve flexibility, at the cost of runtime. Though it differs from computer to computer, the computer that is currently used for this project would take much longer than 20-40 seconds to run every iteration.

While this project does aim to produce a fully independent AI teacher with human attributes, combining this project with other educational technologies to serve as more of a management or teacher assistant system could positively affect the education sector while no longer carrying the risk of eliminating jobs for human teachers. However, that subfield of Artificial Intelligence is much more saturated; there are a lot more teacher assistant Artificial Intelligences than there are fully Artificial Intelligence teachers.

Collaborating with educational institutions other than Allegheny College could greatly benefit this project. Although I believe myself capable enough to accomplish what I've accomplished, I am not perfect. Those with different skill sets, knowledge bases, and style of problem solving could greatly expand the capabilities of this tool in the future.

5.3 Future Ethical Implications and Recommendations

The main ethical implications of the public use for this tool is the lack of universality that comes from many people being unable to use this tool, whether

because they don't have access to the technology required or due to lack of knowing English. This tool is meant to improve AI teachers to serve to substitute for real teachers, and this if developed further could put those who don't have access to this tool at a learning disadvantage. Therefore, adding language capabilities beyond just English would solve one of this project's ethical implications. In terms of misuse of the tool, since its function is to adapt a ChatGPT agent with personalized or prewritten prompts, the only misuse that could occur is training it outside the scope of this project, or maybe doing the opposite, training the AI teacher to be worse at its job. There is also an ethical risk behind the api key used to query the AI teacher. This key if found by someone else could be exploited by spending all of the available money or allotted tokens that the user has. Currently in this version the api key is accessed via a GIT excluded environment file, and the risk would only grow greater if the project was ever publicly hosted on a website. Either using this project api key or having the user input an api key carries the ethical risk of the api key being apprehended by someone else and emptied.

Another ethical risk is the potential for bias that comes with the training data(studies and textbook sections), which could introduce cultural bias that damages the effectiveness of the teacher when teaching different cultures. This could be prevented via adjusting the training data to what culture the students consist of or finding as unbiased training data as possible. For instance, say an AI teacher trained using GPTA+ is teaching the history of the Revolutionary War of the USA to a group of British students via a section from an American textbook. The perspective of the lesson will be completely different and not as effective in teaching compared to using a British textbook as training data.

A key ethical implication of this project is if it is too successful or future research improves the technology significantly, this could displace the jobs of many qualified teachers via having competent AI teachers. This cannot be wholly prevented but can be somewhat mitigated by keeping several things in mind. First, human teachers do a lot more than just output several paragraphs of lesson material. They make studies guides, relate class material to current events, appeal to the students via various methods such as humor, and much more. There is also a physical aspect to teaching; field trips, breaking up disputes, and in class labs in activities are a key part of a quality education. GPTA+ is not at a level where those capabilities can be displayed. Maybe with the future development in AI this might be possible but for now, even if the AI teachers score very high on all metrics, they're still a very limited and substandard alternative to human teachers. Even if that level was reached, human teachers possess the same potential for improvement or better than AI teachers. They take feedback both from students and from the school administration and use it to refine their teaching.

Another reason that Artificial Intelligence won't beat humans out of the education sector is the environmental impacts. LLM usage already uses a very substantial amount of energy, and is greatly contributing to environmental risks like global warming and the thinning of the ozone layer. Now imagine that instead of the millions of teachers in the United States alone, every single one was

replaced by an AI teacher. How large would the carbon footprint resulting from that action be? Thus there is not only ethical concerns in the sense of destroying jobs but also destroying the environment that goes into such a decision.

There is an ethical concern with having an eventually unsupervised artificial intelligence teaching a bunch of students, especially since many students are children. Although there is the same risk with human teachers, the overall untrustworthiness and lack of common sense that comes with Artificial Intelligences leads to a potential risk of students being told inappropriate information or jokes, not adjusting for the audience of students. AI is much more susceptible to this risk than human teachers are. To mitigate this risk, custom prompting to clearly let the AI teacher know what the audience is and what is appropriate and what is not appropriate would go a long way in lowering the risk as much as possible. And although most large LLMS have this, adding a custom list of banned words that can't be included in the lesson plan could also help ease that ethical concern.

References

- [1] Tony Berber Sardinha. 2024. Ai-generated vs human-authored texts: A multidimensional comparison. *Applied Corpus Linguistics* 4, 1 (Apr 2024). <https://doi.org/10.1016/j.acorp.2023.100083>
- [2] Jacques Bughin, Eric Hazan, James Manyika, Michael Chui, and Peter Dahlström. 2017. The promise and pitfalls of AI. <https://www.mckinsey.com/mgi/overview/in-the-news/the-promise-and-pitfalls-of-ai>
- [3] Thomas K. F. Chiu. 2024. Using Self-Determination Theory (SDT) to explain student STEM interest and identity development. *Instructional Science* 58 (2024), 89–107. <https://doi.org/10.1007/s11251-023-09642-8>
- [4] David Game College. 2023. David Game College’s experimental class taught entirely by AI. <https://www.davidgamecollege.com/courses/courses-overview/item/102/gcse-ai-adaptive-learning-programme> GCSE AI Adaptive Learning Programme.
- [5] Bosede I. Edwards and Adrian D. Cheok. 2018. Why Not Robot Teachers: Artificial Intelligence for Addressing Teacher Shortage. *Applied Artificial Intelligence* 32, 4 (2018), 345–360. <https://doi.org/10.1080/08839514.2018.1444611>
- [6] Linda A. Meyer, Michael Wess, and Aidan Young. 2014. USING CARNEGIE MELLONS OPEN LEARNING INITIATIVE (OLI) TO SUPPORT THE TEACHING OF INTRODUCTORY STATISTICS: EXPERIENCES, ASSESSMENTS, AND LESSONS LEARNED. In *Proceedings of the 9th International Conference on Teaching Statistics (ICOTS9)*. https://icots.info/9/proceedings/pdfs/ICOTS9_9D3_MEYER.pdf
- [7] Hussan Munir, Bahtijar Vogel, and Andreas Jacobsson. 2022. Artificial Intelligence and Machine Learning Approaches in Digital Education: A Systematic Revision. *Information* 13, 4 (2022), 203. <https://doi.org/10.3390/info13040203>
- [8] Anaïs Tack and Chris Piech. 2022. The AI Teacher Test: Measuring the Pedagogical Ability of Blender and GPT-3 in Educational Dialogues. <https://educationaldatamining.org/edm2022/proceedings/2022.EDM-short-papers.54/>