The Future of Engineering Education
2024 Annual Conference & Exposition
Oregon Convention Center
Portland, OR . June 23 - 26, 2024
ASEE
Paper ID #44419

Scaffolding AI research projects increases self-efficacy of high school students in learning neural networks (Fundamental)

S. Shailja, University of California, Santa Barbara

Shailja is an incoming post-doctoral fellow at Stanford University. She completed her Ph.D. in the Electrical and Computer Engineering (ECE) Department with interdisciplinary emphasis on College and University teaching at the University of California, Santa Barbara (UCSB) in 2024. She graduated with a bachelor's degree from the Electrical Engineering Department at the Indian Institute of Technology, Kharagpur in 2016. Shailja has been awarded the Fiona and Michael Goodchild best graduate student mentor award during her PhD. She has also been named an NSF iRedefine ECE Fellow for leadership potential among underrepresented graduate students across US/Canada. Shailja's research vision is to develop AI methods for healthcare that "close-the-loop" between surgeons, research scientists, educators, and engineers.

Mr. Satish Kumar, University of California, Santa Barbara

Machine Learning Researcher-PhD student at UC Santa Barbara with 10+ years of research experience building advanced algorithms for large-scale solutions. It includes 6+ years in computer vision and machine learning algorithms and infrastructure at Vision Research Lab at UCSB. The current research is on multi-spectral image analytics, and I lead the project BisQue, an open source ML platform for data storage, AI/ML analysis, and visualization.

Arthur Caetano, University of California, Santa Barbara

Arthur Caetano is a Human-Computer Interaction Ph.D. student at the University of California, Santa Barbara, researching generative user interfaces in Extended Reality at the Human-AI Integration Lab under Prof. Misha Sra. With a Bachelor of Science in Computer Science from Universide Federal Fluminense (2017), he brings 5 years of experience in Product Management within the financial industry, focusing on internal technical solutions for data scientists and data platform regulators. Arthur also mentors high-school and undergraduate students in research and has 2 years of teaching assistant experience in Human-Computer Interaction and Computer Graphics.

Dr. Ayush Pandey, University of California, Merced

Ayush Pandey is an Assistant Professor of Teaching in the Electrical Engineering department at the University of California, Merced. Before joining UC Merced, he completed his Ph.D. in Control and Dynamical Systems at the California Institute of Technology working with Dr. Richard M Murray. He is interested in increasing access to computational tools for inclusive data analysis education and research. His dissertation research develops computer-aided design tools for the engineering of biological systems to address sustainability and health challenges at scale. In 2022, he was appointed as an Adjunct Professor at the Harvey Mudd College where he explored the role of mathematical modeling and analysis tools in interdisciplinary computational biology education. Prior to that, in 2018, he obtained his master's degree in Electrical Engineering from the California Institute of Technology. In 2017, he obtained dual bachelor's and master's degrees in the Electrical Engineering department from the Indian Institute of Technology (IIT), Kharagpur.

To appear in the proceedings of the 2024 American Society for Engineering Education (ASEE)'s Annual Conference & Exposition

Scaffolding AI research projects increases self-efficacy of high school students in learning neural networks (Fundamental)

Abstract

With the rise of Artificial Intelligence (AI) in the mainstream and the impending need for an AI-trained workforce, we must devise strategies to lower the entry barrier to AI education. Advanced mathematical preparation and computational thinking skills are two major barriers in imparting a rigorous AI course at the high school level. Consequently, many existing AI-focused educational programs for high school students are basic primers and lack technical depth. In this paper, we assess two pedagogical instruments for increasing the self-efficacy of students in learning neural networks at the high school level. The first research question is whether high school students learn the basics of neural network design through scaffolded AI research projects. We also explore whether a dual advising structure with a research mentor and a communication teaching assistant enhances student's self-efficacy in computing. For both of these questions, we define key variables to quantify student mastery and their computational thinking using qualitative student feedback and student reflection using GPT-3. We provide a reproducible blueprint for using large language models in this task to assess student learning in other contexts as well. We also correlate our results with a pre- and post-course Likert survey to find significant factors that affect student self-efficacy and belonging in AI.

With our course design and dual advising mentoring model, we find that students showed a significant improvement in their ability to articulate technical aspects within the AI domain and an increase in their confidence in speaking up in the AI field. Two out of the ten research projects applied AI techniques beyond classroom teachings, yielding original research contributions, and another six showcased students' capabilities in building neural networks from scratch. Our study has a strong selection bias since it focuses on top-performing students. However, the exploration of the two pedagogical instruments (scaffolding research projects and dual advising structure) aimed at high school students provides promising insights for future AI curricula design at the high school level.

1 Introduction

Artificial Intelligence (AI) education conventionally starts at the undergraduate upper-division level with courses in Computer Science (CS) and CS-adjacent disciplines. Topics such as convolutions, kernels, backpropagation, and gradient descent are logically introduced after students have attained a requisite level of expertise in calculus, programming, data structures, linear algebra, probability, and optimization. This foundational knowledge enables students to not only apply AI but also contribute to AI research and theoretical development. However, the evolution from a predominantly research-intensive field to one that emphasizes application and product development has led to a significant role of AI in everyday life. This shift has created an unmet educational need for data science and AI instruction at the high school level [1]. Beyond the societal fascination with AI, its introduction at the high school level is anticipated to correlate positively with the retention of students from diverse backgrounds in CS and related fields. Research has consistently shown that the field of computer science experiences high attrition rates among women and people

of color. This limits the participation from underrepresented minorities and perpetuates underlying biases among decision-makers and leaders. The scarcity of diverse role models in CS remains a persistent concern as well. These issues lead to secondary yet serious consequences. Biased and under-scrutinized AI models in practical applications such as AI-based law enforcement [2], voter profiling using AI [3], pervasive and skewed product marketization [4] are a few examples.

Our goal is to broaden the participation of students from all backgrounds in AI fields by lowering the entry barriers. We hypothesize that if AI is taught in a hands-on manner following a data-science workflow [5] high school students can comprehend complex AI topics as well. They can grasp not only the broad principles but also the technical topics even without relying on advanced math expertise or complex data structures. To facilitate this, we propose a design for the course and use a research mentoring structure that scaffolds the AI education. Additionally, we appeal to a combination [6] of students' intrinsic value in fascination with healthcare and utility value [7] in expecting rewarding AI careers. Therefore, we propose healthcare-motivated research projects in our course design to foster challenging yet rewarding experiences in learning and applying AI methods.

1.1 Related work

AI education at the high school level has diverse goals — building foundational knowledge, stimulating student interest in technology, broadening participation in CS, and as an alternative way to develop problem solving and critical thinking skills. Successful precedents from other areas exist for these educational aims at the high-school level. Many robotics and system design competitions, such as the FIRST Robotics Competition, the Solar Car Challenge, and the International Genetically Engineered Machine (iGEM), have spurred student interest in hands-on building and development. Such projects and competitions have been shown to have long-term positive effects on student engagement with advanced engineering education. They play a crucial role in enhancing students' career post high school [8]. Therefore, it is logical to anticipate similar learning benefits from piloting AI education programs, workshops, and competitions at the high school level.

The need for AI education for young students is clear from studies that show systemic inequities in CS education [9]. Often, students with informal preparatory privileges benefit from and continue to participate in CS programs while stringent undergraduate curricula create hurdles for underprivileged learners. Recent surveys have reported low percentage of high schools who offer CS courses. For example, in California [10], only 39% of the high schools offer CS courses. Some reasons for this include the lack of rigorous CS teacher preparation [10], limited resources to expand CS education capacity [11], and research gaps in analyzing efficacy of CS education at the high school level. Initiatives like "AI4K12" and "AI4ALL" are aligned with this need.

AI programs for high school students: Educators and professionals have taught AI to high school students at various levels, scales, and duration [12, 13]. Here, we review a selected number of such programs and refer the reader to other systematic reviews for more details. Notably, we do not review educational programs at the undergraduate level and the online courses in this area since the goals for massive open online courses (MOOCs) are tangential to the research in this paper. A notable exception is Code.org — an educational platform that is geared towards K-12 CS education. It

provides a path towards structured integration of computer science into school curricula to broaden the participation of underrepresented minorities in CS. The AI-focused courses on Code.org are GUI-based and use Google Teachable Machine while depending on some prior programming experience with Java. Focusing on the inputs and outputs of an AI model with a GUI enhances accessibility but obscures the technical details of neural networks. This is similar to most other in-person efforts including AI education at the high school level. Short bootcamp or workshop-style programs are common where educators and mentors work with high school students for 1-2 weeks. Similar to code.org, the learning goals in such programs are skewed toward the input data (usually images) and the output results (predictions shown in a GUI) while skipping the technical details due to lack of student preparation and time. Nevertheless, the high school students go out of these workshops with increased motivation for studying CS [13], increased preparation for college [14], and a positive outlook of AI [15].

Educational programs that include technical details have also been organized. In 2019, a workshop [16] was conducted using Scratch (`scratch.mit.edu`) to teach concepts like k-means clustering and neural networks using functional programming. Students faced different situations and were able to demonstrate their understanding successfully as evaluated via pre- and post-course surveys. The students in this program could create models from data to observe patterns by interactively creating data [17] and training models on it. On similar lines, a workshop that uses an image classification task as the main educational goal [13] was organized. The paper reports the data from more than 100 students to show successful engagement in the technical understanding of k-means clustering. This approach is also geared towards broadening the participation in AI, as the module design is in Brazilian Portuguese. For both of these workshops, the learning goals are limited to topics in machine learning. The technical details of topics in deep learning were not taught, as they are usually more challenging due to the requirement of advanced mathematical background.

Benefits of AI education at the high school level: To broaden the participation from underrepresented groups in CS fields, in 2019, a week-long data science workshop [18] was organized in the Chicago Public School system. The goal of this workshop was to increase the participation of underrepresented minorities (Hispanic, African American, and Women) in CS and data science fields. The workshop facilitators reported that student experiences were largely positive with the highlights being the students' enjoyment in industry talks, data visualizations, and working with real-world datasets. Another aspect of broadening the participation in AI is to reduce the entry barrier through accessible education material towards realizing "CS for All". It has been argued that a data science approach can lower the math-heavy AI education barrier [5]. Instead of depending on prior expertise in calculus, probabilities, and linear algebra, the AI course design in this paper also uses a data science workflow. Concepts of model learning, training, and predictions are taught using data manipulation and data as inputs and outputs of projects. However, in contrast with existing approaches, our educational program focused on using research projects as scaffolds to also teach technical topics so that students were able to design their own neural networks from scratch.

Teaching AI in high school has also been shown to broaden participation in CS from girls [15]. Hands-on and project-based learning increase enthusiasm and confidence, also aiding in intuition

building. Game-based approaches [19] are also often used. A biomedical AI applications-oriented course has been attempted for high school students [20], yielding positive results in student enthusiasm and learning. The reader is referred to the systematic reviews on teaching machine learning to high-school students [13, 21] for a detailed comparison of various AI education attempts at the high-school level. This review compares the technical content, the pedagogy, and the technology utilized among various education programs.

1.2 Research problems

To summarize the related work, we note that in most instances of AI education for high school students, an increased enthusiasm, interest in career in CS and data science, and learning were positively impacted. But some topics are difficult to teach like societal implications of AI tools and the ethics in AI. Thus, we recognize the following challenges with AI education at the high school level:

1. Technical topics are often skipped entirely, briefly mentioned, or taught unsuccessfully in a passive learning style.
2. AI projects are often limited to image classification tasks or use game-fied interfaces. Natural language processing tasks like sentiment analysis from text are also common outcomes. Such projects do not sufficiently explore the technical challenges in training neural networks, do not highlight the limitations or the scope of AI models, and may not provide enough motivation to students at the high school level.
3. A common challenge is that coming up with project ideas for high school students is difficult [22].
4. Successful group work requires similar prerequisite preparation whereas group composition, team dynamics, and learning styles are challenging to navigate with high school students.

To study some of these challenges, we formulate two research questions that evaluate the two pedagogical instruments employed in our AI course design: (1) scaffolds with healthcare-focused AI projects, and (2) a dual advising structure with a research mentor and a communication TA. So, we study: **(RQ1)** Can high school students comprehend the technical details of neural network design through scaffolded AI research projects?, and **(RQ2)** Does the dual advising structure enhance student self-efficacy and confidence in computing?

1.3 Main contributions

Our main contributions are highlighted below:

1. We define categorical variables to quantify student comprehension and self-efficacy by analyzing qualitative textual data from student self-reflection and open-ended survey questions to study RQ1 and RQ2. We develop and provide a GPT-3 based pipeline for reproducible analysis of qualitative data in other contexts.

2. We use a project facilitating mentor and a research mentor to enhance the success of group work, implement active learning in labs, and discussion sections that are focused on team building and on learning how to do research. This approach is akin to the recently proposed Early Research Scholars Program for undergraduate research [23]. We effectively use and extend this approach to the high school level and for AI education.

3. With data from pre- and post-course Likert surveys and our GPT-based categorical analysis of qualitative data, we report significant increase in the students' technical articulation and fidelity in defining a neural network and their outlook on the ability of AI to solve complex problems in the world.

4. By scaffolding research projects in the students' learning process, we observe a significant increase in the students' outlook on AI being able to solve complex problems. Further, we observe a significant increase in students believing that they have advisers or role models in AI and CS.

2 Research Methods

2.1 Course design

The course was proposed under the "Pre-College Programs Office" with qualified students in the 9th, 10th, and 11th grade. The course was developed over 8 months (from student selection to material development) under the guidance of the Director of pre-college programs in UC Santa Barbara's Summer Research Academies (SRA). This course was offered as a research track among 10 other STEM and non-STEM tracks in the SRA 6-week summer program. The title of the course was "Diagnostic AI — Transforming Healthcare Using Image Processing and Learning from Biomedical Images". The course was delivered and mentored by a faculty in the pre-college program, the pre-college program director, and three graduate students (an instructor-of-record, a teaching assistant, and a communication teaching assistant). In this interdisciplinary course, students were exposed to traditional image processing and advanced deep learning methods for biomedical applications.

The course design goes beyond image classification tasks using GUIs, game-based learning, and other conventional high school-level learning modules that do not expose the students to the technical topics in AI. We teach technical topics in deep learning using applied research projects to real-world healthcare datasets (such as tumor growth, cancer, and more). Other than teaching technical skills, the program was also aimed to enable students to produce novel contributions to the domain of diagnostic artificial intelligence. The program required students to present their research projects in a capstone seminar and submit a research report similar to a conference paper. To accomplish these goals, discussions on research practices and academic communication were included in the course design, facilitated by a dedicated communication TA for these aspects.

The 5E approach used in course: The 5Es approach [24, 25] to course design was crucial in designing the course as shown in Figure 1. To "Engage", we incorporated seminar talks, memes, game-ified activities. Students "Explore" hands-on and research skills through labs, mentoring sessions/team building process. To "Explain", we used lectures and lab sessions. To "Extend",

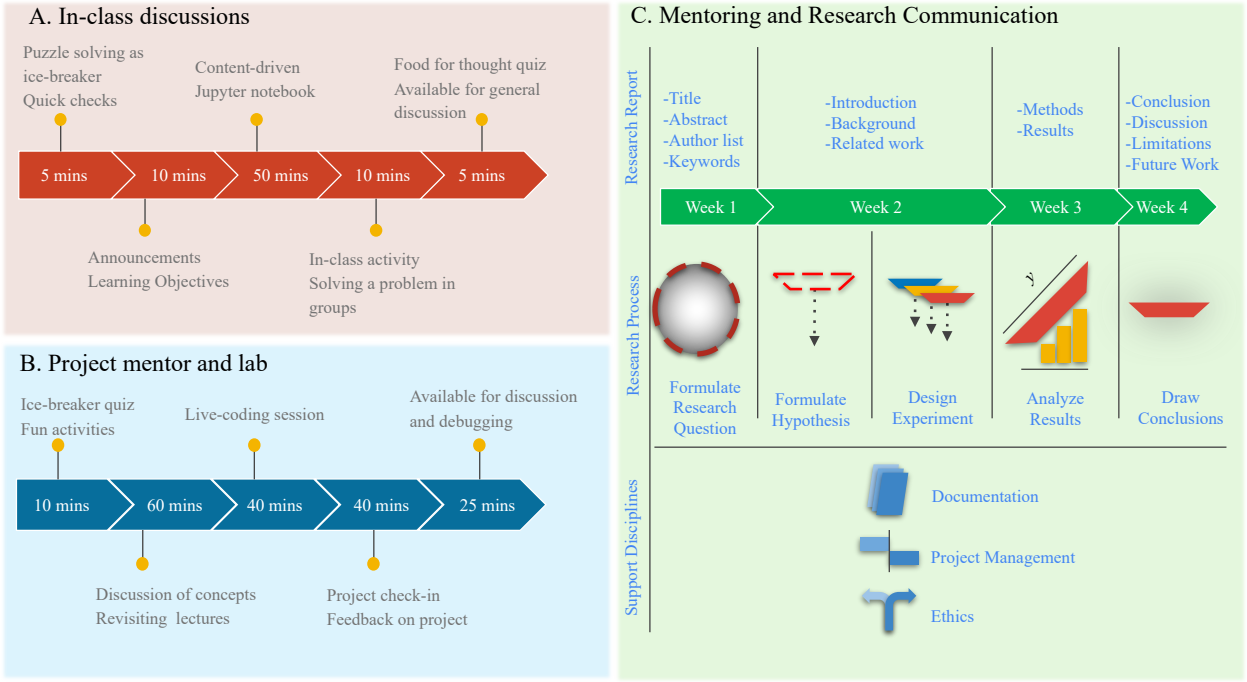


Figure 1: **Course structure with dual-advising structure for pre-college students.** (A) The role of the instructor in the pre-college summer research program (UC Santa Barbara's Summer Research Academies). (B) The project mentor and lab structure led by a teaching assistant who mentors the research projects and demonstrates required technical skills in hands-on sessions. (C) Research project structure and research communication strategies taught to the students in the course, describing the role of the communication mentor.

we provided ample opportunities for students to work on projects. Finally, to "Evaluate", we used project presentations, reports, and reflections.

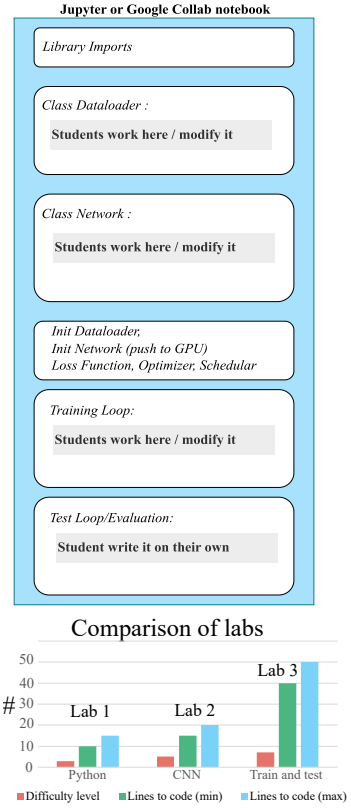
Role of communication TA in the dual-advising structure The communication TA in our pre-college course played an important role in teaching the basic principles of a research project. This mentor presented a simplified research process composed of five sequential activities: research question formulation, hypothesis formulation, experiment design, data analysis, and conclusion (see Figure 1C). The communication TA emphasized the role of supporting elements in the process of a research project: documentation, project management, and ethics.

Scaffolded AI projects: Ten groups of three students each worked on ten different projects. Each group's self-reported Python proficiency before the course is shown in the Figure 2B. We expected that these self-reported ratings would be slightly exaggerated in accordance with performance avoidant behavior [26]. So, the course design relied only on basic Python proficiency. Since the pre-college students participating in this program were new to research, three scaffolds were used in administering the research projects. First, the instructor's in-class discussion involved solving a tangible problem directly related to all projects irrespective of the choice of dataset – neural network building blocks using the Python pytorch library, statistics of model accuracy, and the data science pipeline (see Figure 1A). Next, in the lab, the TA used jupyter notebooks and live coding

sessions to implement the concepts from the lecture for each team's particular application. The TA worked with each team to develop project milestones and offered debugging sessions to achieve these milestones (see Figure 1B). Finally, the communication TA offered guidance on structuring the research project along with specific achievable goals on the report and the presentation (see Figure 1C).

Majority of the groups built a neural network from scratch for their final projects. Group 6 chose to use an off-the-shelf neural network model with some modifications while three groups created an ensemble model. The neural network models for 7 out 10 groups were independently built. Each group worked on a project with an application in healthcare with varying levels of difficulty (shown in the table in Figure 2B). The students were offered some ideas by the instructors but 6 out of the 10 groups selected their own Capstone projects. All of the 10 projects address a diverse range of biomedical applications — brain scans, cell microscopy images, X-rays, tissue imaging, and so on.

A. Formative assessments



B. Student teams and summative assessments

Group	Python proficiency (1-5)	From scratch	Off the shelf	Ensemble	Instructor driven	Difficulty level (1-10)
1	3	5	1	✓			✓	7
2	4	4	2	✓				6
3	3	5	3	✓			✓	5
4	2	4	4			✓		8
5	4	2	5	✓				4
6	3	3	4		✓			8
7	4	3	4			✓	✓	5
8	1	5	4	✓				7
9	4	1	2	✓				7
10	5	3	1			✓		7

C. Diversity of student selected Capstone projects

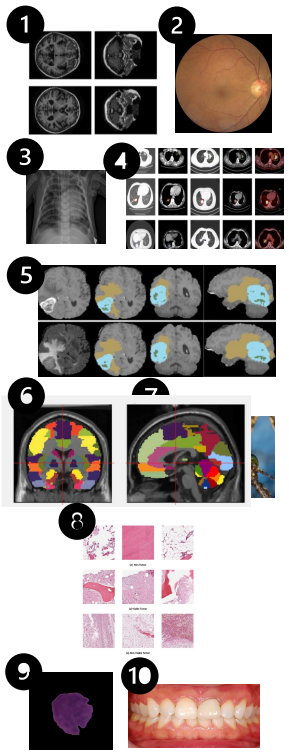


Figure 2: **Assessment design.** (A) Design of formative assessments to promote computational thinking without relying on advanced Python proficiency. Gradual increase in difficulty level and lines of code required for assignments is also shown. (B) Student teams Python proficiency and the properties of their summative assessment projects. (C) Diversity of student projects in the healthcare domain.

2.2 Selection of high school students

High school students were formally recruited by the summer program office in coordination with the instructor at the university for this research-based course. Priority was given to students with prior background in programming, junior or senior year high school students, and students who explicitly discussed healthcare motivation in their personal statements. Thus, the student population in this study reflects a strong selection bias. The course was offered under the direction of the summer program office. The summer program office also obtained consent from the students' parents for the participation of their children in this research-based education program. No prospective educational research project was designed. The data shown in this paper was extracted retrospectively from regular education activities designed for the course — anonymous surveys on outlook with a question on consent of use of the student's responses, technical project reports, in-person labs, and final presentations.

2.3 Course data

During the course, a pre-program survey, a pre-course survey, and a post-course survey was administered. The pre-program survey was used in assessing student preparation and in forming student teams. The groups were constructed such that each group's average Python proficiency was balanced. In addition, we asked questions akin to a "Values Affirmation Survey" [27] to promote a growth mindset in the class.

The pre- and post-course survey consisted of the same questions so that we can analyze any significant changes. Students were asked to rate each statement according to the Likert scale with strongly agree, agree, neutral, disagree, and strongly disagree options. In addition to the objective questions, we also asked a few open-ended questions to gauge the satisfaction of learning outcomes of the course. We obtained a total of 30 pre- and 25 post-survey responses. We highlight a few relevant survey questions here and point the reader to Appendix A for the list of all questions.

(a5) [Likert scale] I am confident in the ability of AI to solve the most complex problems in the world in the future.

(a10) [Likert scale] I have advisers and/or role models in AI and CS (other than my parents).

(b1) [Open-ended] What do you know about neural networks (write in one sentence without looking it up)?

(b2) [Open-ended] What kinds of problems do people in AI work on? What kinds of medicine and healthcare problems do you imagine can be solved with AI?

To facilitate binary classification analysis and simplified visualization, we clubbed the "Strongly Agree" together with "Agree" and the "Strongly Disagree" with "Disagree" responses. We note the effect of this data reduction wherever it led to a change in the conclusion.

Table 1: Examples of Prompt Engineering for Different Learning Paradigms

Type of prompt settings	**Example Prompts**
In the pre-survey and the post-survey, the question is: "What do you know about neural networks (write in one sentence without looking it up)?" Students have responded to this question. The task is to score each response based on the following criteria on "Fidelity".	
Zero-shot	Title: Technical Depth & Accuracy. Description: Evaluates the richness and correctness of the technical details provided in the answer.
One-shot	Example ratings: - Input text: "Neural networks are a series of hidden layers (Conv2D, BatchNorm, MaxPooling, etc.) that each contain a certain number of trainable parameters (neurons), which are updated by using gradient descent after each iteration when the loss is calculated." Rating: 0.95
Few-shot	Example ratings: - Input text: "Neural networks are a series of hidden layers (Conv2D, BatchNorm, MaxPooling, etc.) that each contain a certain number of trainable parameters (neurons), which are updated by using gradient descent after each iteration when the loss is calculated." Rating: 0.95 - Input text: "Neural networks is the structure behind artificial intelligence that make it able to "think" and operate the way it does." Rating: 0.2 - Input text: "A neural network is a way for AI to simulate the human brain through interconnected nodes with many different hidden layers." Rating: 0.6

2.4 GPT codebook and the categorical scoring system

For the open-ended questions, the text in responses was analysed using the `gpt-3.5` model provided by Open AI [28]. We designed a categorical scoring system [29] for this text analysis relevant to our research questions. Three distinct categories were defined for each question. Each category was characterized by specific high-scoring and low-scoring phrases that represented the desired response. The students' answers were then evaluated against these criteria. A continuous score ranging from 0 (weak alignment) to 1 (strong alignment) was assigned to each response in each category. For example, to assess students' understanding of neural networks, responses were evaluated on expression, fidelity, and applicability. Expression measures the students confidence and clarity when defining neural networks. Fidelity measures technical depth of the student definition. Applicability scores the students' definitions on whether they can comprehend the applications and limitations of neural networks. Similar criteria were designed for other questions that were concerned with the student outlook and the general scope of AI. A detailed description of the criteria is given in Appendix B.

We develop a qualitative coding method to identify patterns in the data (open-ended survey responses). We explore a novel approach that leverages large language models (LLMs) to facilitate

this qualitative coding. We develop prompts that include specific instructions and codebooks to increase LLMs' performance in a new task with unseen data. We chose `gpt-3.5` model during the prompting process. We designed different example-centered prompt settings, specifically, zero-shot, one-shot, and few-shot settings as shown in Table 1. For the one-shot setting, we provided only one example for each code. For the few-shot setting, we provided three examples for each code. An expert developed a codebook and coded responses to one of the survey questions "What do you know about neural networks (write in one sentence without looking it up)?" on three criteria: "Expression", "Fidelity", and "Applicability". The expert is a senior author in this paper who was separated from all elements in the course design, delivery, and assessments, that is, the expert rater was not biased with pre- and post-survey questions or their mapping and any other details about the students. This removes the bias of any hypothesis in the expert rater. The prompt follows the structure of [Question/Criteria/Title/Example ratings] as shown in Table 1. For all prompt variants, we included an identity modifier and a custom instruction to constraint output space (see Appendix B).The code utilized in this study is accessible through GitHub [30].

2.5 Statistical data analysis

We measured the performance of our LLM-based approach for categorical scoring of qualitative data with Pearson correlation coefficient (PCC) that measures linear correlation between multi-shot rating and expert rating [31]. Higher PCC indicates how two coders agree with each other as shown in Table 2. Our results suggest that it is feasible to use the GPT-based codebook method described earlier for deductive coding. When analyzing questions (b1) and (B1), our LLM-based approach achieved strong agreement with the expert rater in the categories of fidelity and expression with PCC > 0.7. For applicability, we observed moderate agreement, PCC > 0.5 with expert rating, see Table 2.

We perform t-test to examine the presence of statistically significant differences in ratings from pre- versus post-survey on paired responses. For objective data, we conduct an independent two-sample t-test on the response distribution of pre-course and post-course Likert surveys. For subject data, we conduct t-test on the scoring obtained from few shot learning. A p-value < 0.05 demonstrates potential significance of the observed differences between the two response distributions answering **RQ1** and **RQ2**.

Table 2: Pearson correlation between expert ratings and `gpt-3.5` ratings

Experts	Few Shot	One Shot	Zero Shot
Expression	0.73	0.78	0.71
Fidelity	0.78	0.85	0.79
Applicability	0.52	0.49	0.47

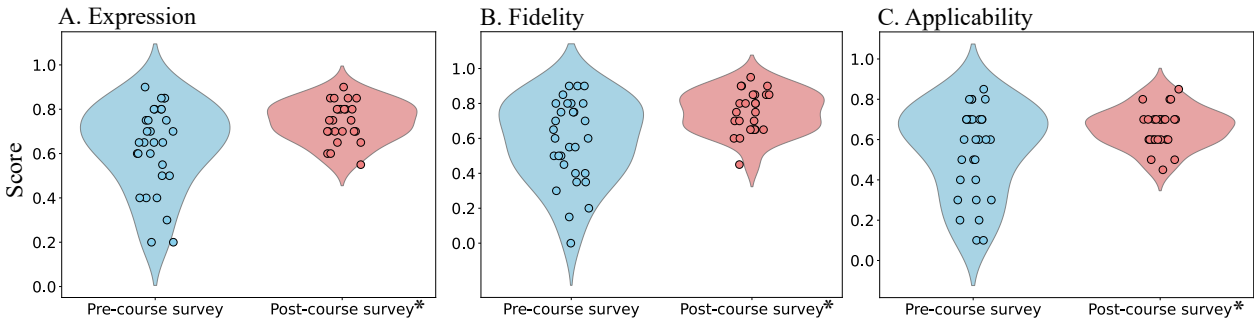


Figure 3: **Quantifying neural network comprehension with GPT-based codebook.** (A-C) Few-shot learning based GPT scores given to student responses under the "Expression", "Fidelity" and "Applicability" categories respectively shown in violin plots. * denotes statistically significant.

3 Results

We analyzed the course surveys to study the change in student self-efficacy, satisfaction of learning expectations, enthusiasm, and their outlook on AI. To get further data on these evaluation categories, we analyzed the text from open-ended questions, student evaluations, and student projects (as described in research methods above).

Increase in students' comprehension of neural networks. Students were asked to define a neural network in pre- and post-surveys. Using the GPT-based categorical scoring method that we developed, we scored students' answers on their 1) accuracy and fidelity, 2) their understanding of applications, and 3) their confidence and expression. In all three categories, we observe a statistically significant improvement as shown in Figure 3.

The course employed two pedagogical instruments — healthcare-focused AI research scaffolds to teach technical neural network concepts, and a communication TA for a dual advising structure. In response to these, we observe that the students did not simply gain a superficial understanding of the technical topics in AI. Instead, they were able to accurately articulate neural networks in technical depth and contextualize that information with applications. The p-values are shown in Table 3. On the other hand, no significant change was observed in students' reflections on ethical considerations in the field of AI. We hypothesize that due to the short-time and fast-paced nature of the course, not much time could be devoted to discussions on ethical considerations in applying AI models to healthcare data. This is a crucial area and has been highlighted as one of the topics that is hard to teach at the high school level [32]. From our data analysis, we cannot conclude any

Table 3: p-values for pre- vs post- analysis

Pre vs post	Expression	Fidelity	Applicability
What do you know about neural networks?	7.94×10^{-3}	2.31×10^{-3}	2.35×10^{-2}
What kinds of problems do people in AI work on?	1.84×10^{-1}	4.20×10^{-1}	7.43×10^{-1}

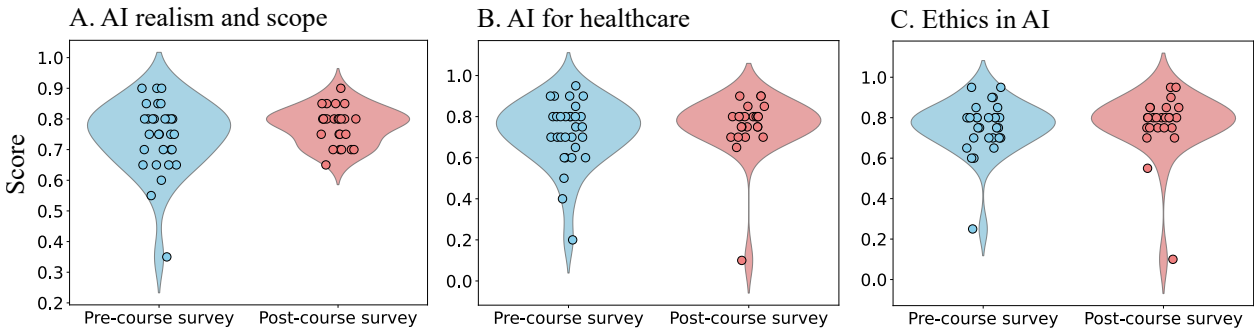


Figure 4: **Quantifying student outlook towards AI using GPT-based codebook**. (A-C) Violin plots show GPT scores categorized under AI realism/scope, AI for healthcare, and ethical reflections respectively.

significant change in how students perceive the limitations and scope of AI (see Figure 4).

Before the program, approximately 50% of students lacked confidence in Python programming as shown in Figure 2B. The lab instructor offered pre-requisite resources beforehand and directed students in the labs who struggled with Python in a project-based learning fashion. By the end of the course, 60% constructed neural networks from scratch in Python, demonstrating the improvement in students' increased proficiency. Moreover, 90% of the students developed models either from scratch or by ensembling multiple models. This involves significant coding in Python (Figure 2A).

Increase in student self-efficacy. We report the change in student self-efficacy measured using three related variables: (1) student confidence on speaking up about a technical area like AI, (2) student self-assurance and positive outlook for success in an AI career, and (3) outlook towards the field of AI. First, we observe an increase in the students' ability to understand and communicate AI research. As shown in the post-survey results (see Figure 5A), students' showed a significant increase in confidence in speaking up about topics in AI. The students' ability to handle technical questions from the audience when they presented their research projects reflects this reported increase in confidence on speaking about AI in the survey. A central element of the course structure was the research mentoring and team building guided by a communication TA. The process of research mentoring creates a supportive environment with new advisers and role models. We observe a statistically significant change in students' self-assurance in identifying advisers and role models in AI and CS (see Figure 5B). Finally, being able to follow the latest advances in the field is an essential skill to any practitioner in a fast-paced field such as AI. We observed an increase in the grasp of fundamental concepts of AI after this course, as discussed in the previous section. Moreover, we also note a significant increase in students' self-belief on AI being able to solve complex problems in the future (see Figure 5C).

Related to self-efficacy, we also find that there were no statistically significant changes observed in questions on the following topics: (1) student's belief in pursuing a career in AI, and (2) students' confidence in finding support among peer group. These two observations highlight the selection bias in our study where the selected students' were likely already interested in AI careers, hence, no significant change was observed.

The diversity of student projects All ten projects addressed unique biomedical image modality and research question (see Figure 2). This highlights the potential of healthcare application as not only invigorating but also providing a wide variety of diverse examples to facilitate projects. The GitHub repositories for some of the projects and selected arXiv papers for the teams are publicly available [33, 34, 35].

4 Discussion

4.1 Future research directions on self-efficacy

The structure of the program was unique and it afforded a communication mentor beyond the technical lab leading mentor. This setup helped all students to develop soft skills that accelerate learning. However, our research did not collect any data that directly measures the impact of the communication TA. Student interviews and focus groups may be helpful in this regard, where qualitative data on teaching strategies, course design, and research project design could be gathered. Further, student evaluations may be used to correlate the findings of our paper.

For the construction of prompts used in the GPT categorical scoring system, we used a codebook developed by experts. This provides transparency but may limit the model's performance as indicated by a decline in Pearson correlation from one-shot to few-shot. In response to this finding, an interesting future direction could be to design an effective codebook by analyzing transformer model performance as well. Another interesting future direction is to study the interplay between the qualitative GPT scoring with the quantitative Likert survey responses. This exploration could lead to more sophisticated strategies for integrating these two forms of analysis using GPT technologies for advanced mixed methods.

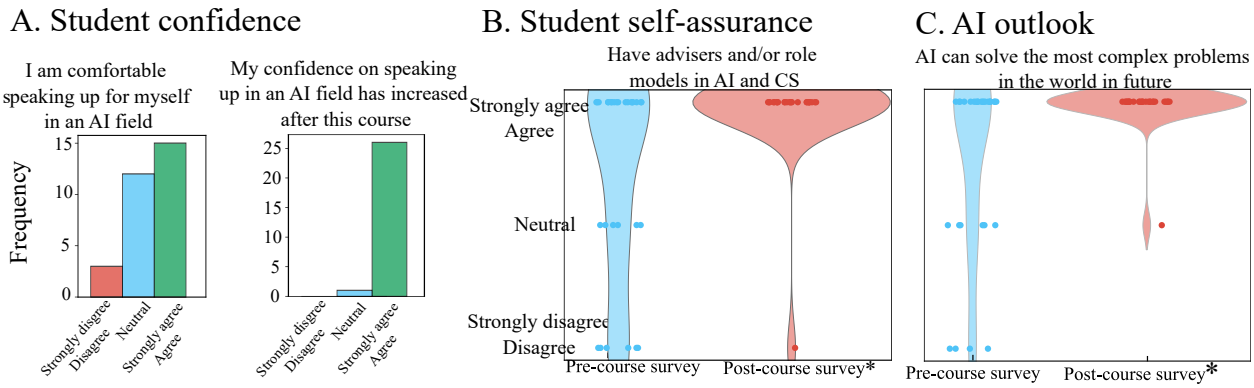


Figure 5: **Quantification of student self-efficacy using quantitative surveys** (A) Pre- and post-course survey responses on questions related to speaking up about AI. (B) Student self-assurance increased significantly in post-course survey as they reported having advisers or role models in CS. (C) Students' outlook towards the field of AI being transformative in the future increased significantly.

4.2 Limitations

We note the following limitations of our study:

1. Selection bias: The students enrolled in the program were top-performing students of their class. A majority of the enrolled students already displayed a passion for AI before joining the program. Both of these factors limit our findings.

2. Confirmation bias: Some survey questions used language that may have contributed to a confirmation bias — a commonly acknowledged issue with Likert surveys.

3. Retrospective study: The research presented is retrospective, hence, there was no prospective course design to act as a control experiment. The number of students in the class ($n = 30$) is low to draw any strong statistical conclusions.

4. Lack of diverse population of students: Although the student population represented in our paper is gender-balanced (17 girls and 13 boys), there is no representation of African American or Native American students who belong to historically underrepresented minorities in CS. Further, the reported literature and findings in this paper are strongly centered on the American education system, limiting its wider applicability.

5. The evidence presented for increased self-efficacy as a result of the dual-advising structure is indirect. Further research must be conducted with focus groups and interviews of past participants to gather direct evidence.

5 Conclusion

In this paper, we study the educational benefits of a dual advising structure and AI research scaffolds for high school students. The course design studied is a pre-college program at the University of California, Santa Barbara. We collected both quantitative and qualitative survey data to analyze student self-efficacy and technical comprehension of neural networks. We propose a LLM-based categorical scoring system and codebook for qualitative data to interpret open-ended survey responses. Using this method, we find that students' self-efficacy and comprehension significantly improved. We provide a publicly available blueprint to extend this GPT codebook to analyze other qualitative data in the future. The fundamental exploration of student efficacy and learning of AI at the high school level provides promising insights for future AI curricula design at the high school level.

Acknowledgement

The authors would like to thank Dr. Lina Kim for providing crucial logistical support in running the mentoring program with high school students that led to the research in this paper, the Summer Research Academies program at the University of California, Santa Barbara for supporting the instructors throughout the program, and Dr. B. S. Manjunath for his help with the course design.

References

[1] D. Touretzky, C. Gardner-McCune, C. Breazeal, F. Martin, and D. Seehorn, "A year in k–12 AI education," *AI Magazine*, vol. 40, no. 4, pp. 88–90, 2019.

[2] M. Kusak, "Quality of data sets that feed AI and big data applications for law enforcement," *ERA Forum Springer*, vol. 23, no. 2, 2022. [Online]. Available: https://link.springer.com/content/pdf/10.1007/s12027-022-00719-4.pdf

[3] B. Brubach, A. Srinivasan, and S. Zhao, "Meddling metrics: the effects of measuring and constraining partisan gerrymandering on voter incentives," in *Proceedings of the 21st ACM Conference on Economics and Computation*, 2020. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/3391403.3399529

[4] C.-C. Cheng, C.-C. Wei, T.-J. Chu, and H.-H. Lin, "AI predicted product portfolio for profit maximization," *Applied Artificial Intelligence*, vol. 36, no. 1, p. 2083799, 2022.

[5] M. Neumann, "AI education matters: a first introduction to modeling and learning using the data science workflow," *AI Matters*, vol. 5, no. 3, pp. 21–24, 2019.

[6] A. Valle, R. G. Cabanach, J. C. Núnez, J. González-Pienda, S. Rodríguez, and I. Piñeiro, "Multiple goals, motivation and academic learning," *British Journal of Educational Psychology*, vol. 73, no. 1, pp. 71–87, 2003.

[7] A. Wigfield and J. S. Eccles, "Expectancy–value theory of achievement motivation," *Contemporary Educational Psychology*, vol. 25, no. 1, pp. 68–81, 2000.

[8] E. Baran, S. Canbazoglu Bilici, C. Mesutoglu, and C. Ocak, "The impact of an out-of-school stem education program on students' attitudes toward stem and stem careers," *School Science and Mathematics*, vol. 119, no. 4, pp. 223–235, 2019.

[9] J. Margolis, *Stuck in the shallow end, updated edition: Education, race, and computing*. MIT Press, 2017.

[10] S. Koshy, L. Hinton, L. Cruz, A. Scott, and J. Flapan, "The california computer science access report," *Kapor Center, Oakland, CA*, 2021.

[11] J. Flapan, J. J. Ryoo, R. Hadad, L. Aranguren, S. Kong, and S. Mendoza, "Guide on the side: School leaders' case studies facilitating equitable computer science education in california," in *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 2*, 2022, pp. 1202–1203.

[12] M. Yue, M. S.-Y. Jong, and Y. Dai, "Pedagogical design of k-12 artificial intelligence education: A systematic review," *Sustainability*, vol. 14, no. 23, p. 15620, 2022.

[13] R. M. Martins and C. Gresse Von Wangenheim, "Findings on teaching machine learning in high school: A ten-year systematic literature review," *Informatics in Education*, 2022.

[14] N. Wang and M. Johnson, "AI education for K-12: Connecting AI concepts to high school math curriculum," in *Proceedings of the Workshop on Education in Artificial Intelligence K-12, 28th International Joint Conference on Artificial Intelligence*, 2019.

[15] M. E. Vachovsky, G. Wu, S. Chaturapruek, O. Russakovsky, R. Sommer, and L. Fei-Fei, "Toward more gender diversity in cs through an artificial intelligence summer program for high school girls," in *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, 2016, pp. 303–308.

[16] J. Estevez, G. Garate, and M. Graña, "Gentle introduction to artificial intelligence for high-school students using scratch," *IEEE Access*, vol. 7, pp. 179 027–179 036, 2019.

[17] M. H. Kaspersen, K.-E. K. Bilstrup, and M. G. Petersen, "The machine learning machine: A tangible user interface for teaching machine learning," in *Proceedings of the Fifteenth International Conference on Tangible, Embedded, and Embodied Interaction*, 2021, pp. 1–12.

[18] B. Mobasher, L. Dettori, D. Raicu, R. Settimi, N. Sonboli, and M. Stettler, "Data science summer academy for chicago public school students," *ACM SIGKDD Explorations Newsletter*, vol. 21, no. 1, pp. 49–52, 2019.

[19] A. Alam, "A digital game based learning approach for effective curriculum transaction for teaching-learning of artificial intelligence and machine learning," in *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*. IEEE, 2022, pp. 69–74.

[20] C.-J. Huang, T. Wu, J.-T. Lu, B. Lin, D. Chang, P. Wang, M.-C. Wang, P. Lee, and W. Wang, "Developing a medical artificial intelligence course for high school students," in *International Forum on Medical Imaging in Asia 2021*, vol. 11792. SPIE, 2021, pp. 103–108.

[21] L. S. Marques, C. Gresse von Wangenheim, and J. C. Hauck, "Teaching machine learning in school: A systematic mapping of the state of the art," *Informatics in Education*, vol. 19, no. 2, pp. 283–321, 2020.

[22] I. Voulgari, M. Zammit, E. Stouraitis, A. Liapis, and G. Yannakakis, "Learn to machine learn: designing a game based approach for teaching machine learning to primary and secondary education students," in *Interaction Design and Children*, 2021, pp. 593–598.

[23] K. Redfield, S. Sidhu, and C. Alvarado, "The early research scholars program: Analyzing correlation with academic outcomes in computer science students," in *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 2*, 2022, pp. 1409–1409.

[24] R. W. Bybee, *Achieving scientific literacy: From purposes to practices.* ERIC, 1997.

[25] L. B. Duran and E. Duran, "The 5e instructional model: A learning cycle approach for inquiry-based science teaching." *Science Education Review*, vol. 3, no. 2, pp. 49–58, 2004.

[26] T. Urdan, A. M. Ryan, E. M. Anderman, and M. H. Gheen, "Goals, goal structures, and avoidance behaviors," in *Goals, goal structures, and patterns of adaptive learning*. Routledge, 2014, pp. 55–83.

[27] B. L. Bayly and M. F. Bumpus, "An exploration of engagement and effectiveness of an online values affirmation," *Educational Research and Evaluation*, vol. 25, no. 5-6, pp. 248–269, 2019.

[28] OpenAI, "GPT-4 technical report," *arXiv*, pp. 2303–08 774, 2023.

[29] Z. Xiao, X. Yuan, Q. V. Liao, R. Abdelghani, and P.-Y. Oudeyer, "Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding," in *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, 2023, pp. 75–78.

[30] S. Shailja and A. Pandey, "https://github.com/pyEdTools/ASEE-2024," Accessed May 1, 2024.

[31] I. Cohen, Y. Huang, J. Chen, J. Benesty, J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," *Noise Reduction in Speech Processing*, pp. 1–4, 2009.

[32] S. Akgun and C. Greenhow, "Artificial intelligence in education: Addressing ethical challenges in k-12 settings," *AI and Ethics*, vol. 2, no. 3, pp. 431–440, 2022.

[33] A. Garg, J. Lu, and A. Maji, "Towards earlier detection of oral diseases on smartphones using oral and dental rgb images," *arXiv preprint arXiv:2308.15705*, 2023.

[34] A. Agrawal, C. Hsu, and S. Tan, "A comparative analysis of supervised and semi-supervised learning models for malaria classification using explainable AI techniques," *International Journal of Scientific Research and Engineering Development*, vol. 6, no. 4, 2023.

[35] A. Paleczny, S. Parab, and M. Zhang, "Enhancing automated and early detection of alzheimer's disease using out-of-distribution detection," *arXiv preprint arXiv:2309.01312*, 2023.

Appendix A

The Likert survey questions are shown here. For each question, the student was asked to choose their opinion from Strongly Disagree, Disagree, Neutral, Agree, and Strongly Agree. The pre- and post-course survey questions were:

(a1): I am comfortable with advanced mathematical concepts such as differential equations, probabilities, matrix algebra.

(a2): I believe that understanding advanced mathematics is critical for learning AI.

(a3): I am comfortable speaking up for myself in an AI field.

(a4): I am confident in my ability to grasp complex mathematical concepts.

(a5): I am confident in the ability of AI to solve the most complex problems in the world in the future.

(a6): I am excited about the possible practical applications of the mathematical concepts in AI.

(a7): It is likely that I will have a career in AI in the future.

(a8): How likely do you think it is that you attain that position [your dream AI position]?

(a9): I am confident that I can easily find support from peers as I learn topics in AI and CS.

(a10): I have advisers and/or role models in AI and CS (other than my parents).

The post-survey repeats most of the questions above. In addition, we added the following questions in the post-survey.

(A1): I am now more comfortable with advanced mathematical concepts, such as differential equations, probabilities, and matrix algebra, as a result of this course.

(A2): This course has enhanced my self-awareness regarding my comfort and understanding of mathematical concepts like calculus (differentiation, integration), matrix convolutions, and proofs of optimizations.

(A3): I found that my learning of AI topics was limited due to my prior understanding and comfort with advanced mathematical concepts.

(A4): While this course did not concentrate on imparting the necessary mathematical concepts, I was still able to learn about the critical concepts and principles of AI and learning models.

(A5): My personal learning goals were met in this course.

(A6): My confidence in speaking up in an AI field/on a topic in AI has increased as a result of this course.

(**A7**): My confidence in my ability to grasp complex mathematical concepts has increased after this course.

(**A8**): I am confident in the ability of AI to solve the most complex problems in the world in the future.

(**A9**): It is likely that I will have a career in AI in the future.

(**A10**): How likely do you think it is that you attain that position [your dream position]?

(**A11**): I am confident that I can easily find support from peers as I learn topics in AI and CS.

(**A12**): I have advisers and/or role models in AI and CS (other than my parents).

The subjective questions in pre-survey include the following:

(**b1**): What do you know about neural networks (write in one sentence without looking it up)?

(**b2**): What kinds of problems do people in AI work on? What kinds of medicine and healthcare problems do you imagine can be solved with AI?

The subjective questions in the post-survey are as below:

(**B1**): What do you know about neural networks (write in one sentence without looking it up)?

(**B2**): What do you know about deep learning (write in one sentence without looking it up)?

(**B3**): What kinds of problems do people in AI work on? What kinds of medicine and healthcare problems do you imagine can be solved with AI? How have your views changed after this course?

Appendix B

Description of categorical coding of qualitative data to study the comprehension of neural networks:

(**C1**): **Expression:** Clarity & Confidence. Description: Assesses the answer's coherence, understandability, and the assertiveness with which a statement is made.

(**C2**): **Fidelity:** Technical Depth & Accuracy. Description: Evaluates the richness and correctness of the technical details provided in the answer.

(**C3**): **Applicability:** Utility, Application, & High-level understanding. Description: Assesses the student's understanding of how neural networks and deep learning align with core concepts and their utility in applications. Student's high-level understanding of the overall modeling is also rated higher.

Description of categorical coding of qualitative data to study student outlook in AI:

(**D1**): **AI Realism & Scope:** Reflects the student's understanding of the breadth of AI applications, awareness of its limitations and failures, and appreciation for the detailed technical aspects. This category seeks to gauge a realistic and comprehensive perception of AI.

(**D2**): **AI for Health:** Assesses the student's insight into the specific applications of AI in healthcare, recognition of its limitations within this field, and an understanding of the more technical aspects. This category aims to evaluate a well-rounded view of AI's role in healthcare.

(**D3**): **Ethical Reflection:** Gauges whether the student reflects on ethical considerations related to the application of AI, especially in fields like healthcare, which can directly impact human lives.

Prompts for GPT to rate each of the criteria are composed of three parts — (1) a custom instruction text that remains the same throughout the study, (2) a description of criteria to rate (as given above), and (3) example ratings, 3 examples for few-shot learning, 1 example rating for one-shot learning, and no example ratings for zero-shot learning.

The common preamble instructions given to GPT are: *About you: I am an instructor for a four-week course on "Diagnostic AI" with 30 high school students. I am a researcher on pedagogy and currently working on a research paper to quantify the AI comprehension in this course that uses limited math and calculus background (because they are high school students).*

For this research, I have conducted a pre-survey and a post-survey. This survey has subjective questions that I would like to classify among different categories and quantify the subjective sentences into these categories.

How to respond:

This is a research task. Stop responding in sentences. Stop responding in sentences.

I will give you a list of student responses that are answers to questions in a pre-survey and a post-survey from students. I will also give you the criteria according to which you will score the text. Based on the criteria, score each answer for a given question in the continuous range of 0-1, where the weaker answer will be near 0 and the stronger near 1.

Stop responding in sentences.

Then, for each criteria the prompt followed the structure below. We give one of the prompts here. The other 5 prompts followed a similar structure with different expert ratings for each category given as examples. For one-shot, only one expert rating was added in the prompt, while in zero shot no expert rating was added.

Prompt:

This is a research task. Stop responding in sentences. This is a research task. Stop responding in sentences. This is a research task. Stop responding in sentences.

In the pre-survey and the post-survey, the question is: "What do you know about neural networks (write in one sentence without looking it up)?"

Students have responded to this question. The task is to score each response based on the following criteria on Students have responded to this question.The task is to score each response based on the following criteria on "Expression":

Criteria: Expression

Title: Clarity & Confidence

Description: Assesses the answer's coherence, understandability, and the assertiveness with which a statement is made.

Example ratings:

For the input text: "A model inspired by the interconnected neurons in the human brain that trains on data to learn patterns and increase its accuracy for classification, segmentation, or detection tasks" The rating is: 0.85

For the input text: "If I remember correctly, neural networks use many layers of convergance(?) to arrive at a result." The rating is: 0.15

For the input text: "A neural network is a way for AI to simulate the human brain through interconnected nodes with many different hidden layers." The rating is: 0.55

Each new line consists of a new student response. You must respond with scores between 0 to 1 for the criteria on Expression defined above.

This is a research task. Stop responding in sentences.

This is a research task. Stop responding in sentences. You must respond with scores between 0 to 1 for the criteria on Expression defined above.

This is a research task. Stop responding in sentences. If you understand the task, your next response will read: "Yes, give me the student responses separated by a new line and I will score them based on the criteria on Expression."

The reader may observe that some repetitive instructions were included in the prompt to prevent the random hallucinating behavior of the gpt-3.5 model.