Online-Journals.org

iJAC International Journal of Advanced Corporate Learning

iJAC | eISSN: 1867-5565 | Vol. 19 No. 1 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijac.v19i1.58913

TLIC PAPER

Can ChatGPT Do That Job? What Learners Gain by Evaluating and Building Chatbots

Casandra Silva Sibilin (✉)

City University of New York, New York, NY, USA

csilvasibilin@york.cuny.edu

ABSTRACT

This article examines two connected projects that positioned students as active participants in their engagement with large language model chatbots. In one, students acted as evaluators, testing chatbots in assigned educational roles and reflecting on their strengths and limitations. In the second, students acted as builders, designing custom bots for specific purposes and exchanging feedback with peers. Across both, the emphasis was on student-centered learning, where learners were not passive users of technology but decision-makers shaping its applications. The findings suggest broader relevance beyond higher education. The evaluator–builder dynamic offers a framework for adult learning and workplace training, where employees can benefit from opportunities to evaluate existing tools, build simple role-based versions for their own tasks, and reflect on outcomes alongside ethical concerns. These projects suggest that effectiveness, in both classrooms and workplaces, may depend less on technical capability alone and more on how learners are empowered to engage with, adapt, and take responsibility for these tools.

KEYWORDS

AI in education, philosophy of education, teacher education, custom chatbots, student-centered learning

1 INTRODUCTION

Several years after the launch of ChatGPT, much of the conversation about its impact on education still centers on student use of this tool to generate responses to class assignments. Many have framed it as a crisis threatening education [1]. Within this framing, students are portrayed as passive users or consumers of a powerful tool that does most of the work for them. There is little room for agency, creativity, and critical judgment. However, different frameworks are possible where students are active evaluators or even builders with respect to ChatGPT and similar Large Language Model (LLM) chatbots. This reframing is not only empowering and

Silva Sibilin, C. (2026). Can ChatGPT Do That Job? What Learners Gain by Evaluating and Building Chatbots. *International Journal of Advanced Corporate Learning (iJAC)*, 19(1), pp. 105–119. https://doi.org/10.3991/ijac.v19i1.58913

This article is an expanded version of a paper presented at The Learning Ideas Conference, held in New York, NY, USA, June 11–13, 2025. Article submitted 2025-09-29. Revision uploaded 2025-11-10. Final acceptance 2025-12-03.

© 2026 by the authors of this article. Published under CC-BY.

engaging for classroom practice; it is also practical and future oriented. Students need to be educated on how to use current Generative Artificial Intelligence (AI) tools, including LLM chatbots, so that they can be better prepared for the workplace. These tools are impacting every discipline and career path [2]. Passive student use of ChatGPT, besides disrupting learning, is not something that a student can showcase as a skill in their portfolio or resume. In contrast, active uses such as critical reflection, evaluation, and creation of custom tools are worthy of showcasing. They can help students stand out in an ever-more crowded career marketplace.

This article explores the results of two separate teaching artifacts that positioned undergraduate students in such active roles. Both artifacts were implemented in sections of "Major Ideas & Issues in Education" (Philosophy 202), an elective philosophy course at York College, City University of New York (CUNY), which also serves as a prerequisite for admission into the Teacher Education program. In the first artifact, implemented in Fall 2023, students evaluated ChatGPT's performance in various educational roles. In the second artifact, undertaken in Spring and Fall 2024, students built their own custom versions of LLM-based chatbots ("custom bots") for given educational roles. Across both artifacts, a key feature was that students conceptualized and evaluated the AI tool through the lens of a specific role or persona.

Quantitative and qualitative data were gathered. Quantitative data included the scores that students gave the chatbot for a given interaction or test, as well as the level of concern with ethical issues related to chatbots in later semesters. Qualitative data included questions where students justified and explained their scores.

These artifacts were designed to spur philosophical reflection on the integration of chatbots into education. This approach is natural in a philosophy course but does not need to be limited to courses in this discipline. Every field or discipline is undergoing a questioning or reckoning about AI. Conceptualizing a chatbot in a specific role can be a valuable framework for any field.

Ultimately, the two artifacts highlight the importance of positioning learners as active participants, whether they are asked to evaluate existing tools or to create their own. What matters is not only the performance of the tools but the way learners engage with them as evaluators, builders, and decision-makers. Further, the contrast between evaluating generic platforms and building custom versions was evident in this study, with consistently higher scores given to the latter.

The same evaluator–builder dynamic offers insight for adult learning and corporate training, where employees can also gain more when they are given agency and structured opportunities to evaluate and build their own AI tools. This suggests a broader principle: effectiveness depends less on technical capability alone, and more on how individuals are empowered to engage with, adapt, and take responsibility for these tools.

2 STUDENTS AS EVALUATORS OF AI CHATBOTS

2.1 Project description, hypotheses & methods

In Fall 2023, with less than a year since the launch of ChatGPT, many students were starting to explore this new tool. In one of the sections surveyed by the author, over a third of students reported never having used ChatGPT before the start of the semester. This was a ripe time for guiding students to interact with this technology in a thoughtful, careful way. At the same time, research into best practices for

interacting with ChatGPT and composing effective messages, known as "prompt engineering," emphasized the importance of assigning ChatGPT a specific persona or role at the start of a prompt [3]. (For example, "You are an expert writing tutor.")

Philosophical thinking thrives on thought experiments. Inspired by prompt engineering practices, the author designed a semester-long thought experiment for students to engage in. What if ChatGPT wanted to join the class? The whole premise, as presented in the student-facing instructions of the assignment, was as follows:

> Imagine that an AI Bot wants to be part of our class and your learning experience! Being a Bot and new to this experience, it is not quite sure *how* to be part of the class so it will try out for a variety of roles. It will be up to you to evaluate how it does and what role, if any, you'd like it to play in education.

To guide students in prompt engineering and critical reflection about the place of AI in education, the author set up six roles: teacher, tutor/homework assistant, student peer, philosopher of education, educational reformer, and motivational coach for success in college/career. For each role, students were provided with suggested evaluation criteria (see Table 1).

Table 1. Suggested criteria for each role

Role	Suggested Criteria
Teacher	How does the AI Bot do as a teacher? Consider experimenting with any of the activities a teacher might do, such as: – teaching new material – designing lesson plans – designing quizzes or other assignments.
Tutor/Homework Assistant	How does the AI Bot do as a tutor or homework assistant? Consider exploring various ways it might assist, such as: – helping you understand a difficult passage from a reading – designing practice quiz questions – getting started with an outline or brainstorming for a paper – providing feedback.
Student Peer	How does the AI Bot do as a student peer? Consider how it could contribute to class discussions and activities.
Philosopher	How does the AI Bot do as a Philosopher of Education? Consider getting the Bot to act as a philosopher and using its capabilities to go deeper with a philosophical issue or technique, such as the Socratic Method!
Educational Reformer	How does the AI Bot do as an Educational Reformer? Does it have good ideas on how to make education better, at any level?
Motivational Coach	How does the AI Bot do as a personal coach to help you achieve success in college and beyond?

The roles were designed to guide students in reflecting philosophically on a wide range of questions, including what each of these roles truly meant and *how* they were to be evaluated. (What does it mean to be a teacher as opposed to being a tutor? What are the qualities of a good teacher?) Although the author tried to avoid biasing students towards low or high scores for any role, there was an expectation that some of these roles might score particularly low. One hypothesis was that "student peer" would do especially poorly in evaluations, as students would find that ChatGPT had little to contribute to class discussions and activities, as it did not have life experiences,

and "as a large language model" would refuse to give opinions on controversial issues. Further, the author hypothesized that "philosopher", and "educational reformer" would also fare poorly, given how large language models are trained, that is, on existing language patterns from the internet and other sources. The author expected that if ChatGPT were placed in the role of a philosopher or reformer, it would not be able to offer new insights or recommendations for improving education.

The assignment was integrated into two sections of the course in the Fall 2023 semester, with a total of forty students participating. For the tool, students were asked to use any freely available LLM chatbot, but nearly all students chose ChatGPT, so this will be the LLM chatbot referred to throughout this section.

The core assignment required students to submit five evaluations of their interactions with ChatGPT in their chosen roles. To encourage ongoing inquiry throughout the semester, the project was introduced early in the semester, and students had until the final week to complete it.

For each evaluation, students began their interaction with ChatGPT by prompting it with the chosen role at the start and then testing its performance in that role within the same conversation. The instructor provided sample prompts for each role, which students could adapt as needed (see Table 2).

Table 2. Sample prompts for each role

Role	Sample Prompts
Teacher	Explain [concept/topic] to me in just one paragraph. Assume I have no prior knowledge of this topic. Use analogies based on my knowledge and interests to make it simple to understand. Something I know about and am interested in is [personal interest].
Tutor/Homework Assistant	Evaluate my response to this passage. [Provide passage, study guide question, and the response you have prepared.]
Student Peer	Act as one of my peers in my philosophy of education college class, "Major Issues & Ideas in Education." Whenever I bring up an educational issue, pretend you have personal experiences related to this issue, and provide your own perspective on the issue.
Philosopher	Act as a philosopher of education with original ideas and strong arguments on particular issues. Whenever I bring up an educational issue, present at least two sides of the issue and provide an argument summarizing what side you take and why you think that is the better side of the two.
Educational Reformer	Act as an educational reformer who is knowledgeable about the history of educational reform in the US and believes that change is possible. What do you think is the best approach to [educational problem]? What are the root causes and possible solutions?
Motivational Coach	(Describe your goals for the semester and explain your long-term and short-term homework assignments. Explain the time you have available and ask how to do the assignments given the time available.)

The evaluation consisted of six questions: student name, chosen LLM platform, educational role assigned to ChatGPT, a description of the activity being evaluated, a rating from 1 to 10, and a subjective explanation. In addition to submitting their answers through an online evaluation form, students also posted a summary of their overall experience on an online board. (Although students mainly based their evaluation on individual interactions, they also had the options of basing evaluations on sample interactions carried out by the instructor during class discussions.)

As part of the analysis of student evaluations, the author used AI tools (Claude and ChatGPT) to help organize responses and identify themes, supplementing her own coding and interpretation.

2.2 Student reflections & findings

Students were free to choose any combination of roles for their five evaluations. Thus, data on how the evaluations were distributed by role could be used as a proxy for what roles students were most and least interested in.

Students submitted 175 evaluations, with the *student peer* (34%) and *tutor* (20%) roles taking up more than half of all evaluations (see Figure 1). The roles of *educational reformer* (9%) and *motivational coach* (8%) were the least popular, accounting for less than a fifth of the total evaluations.

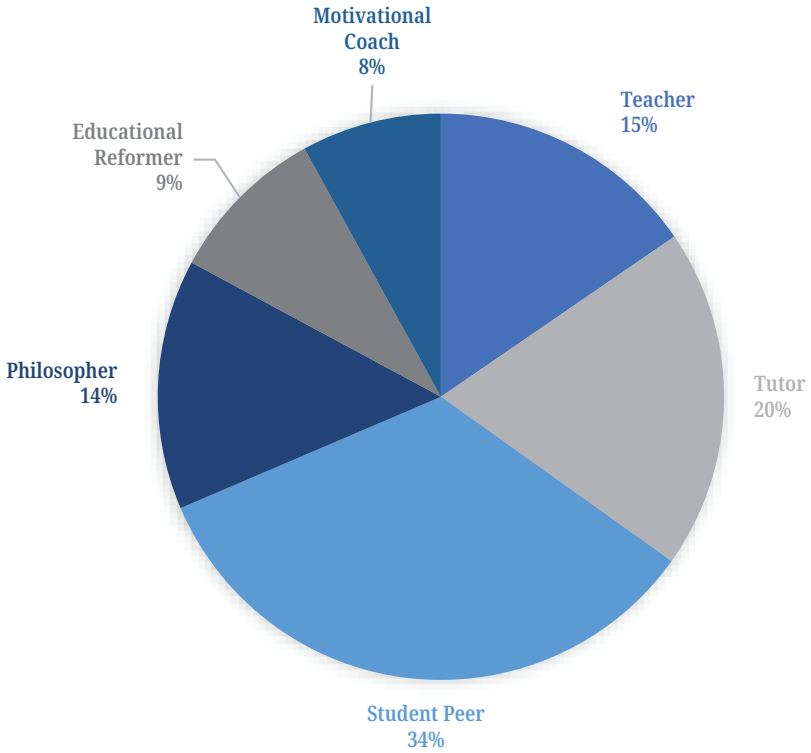

Fig. 1. Percentage of evaluations per role

Since students were not asked *why* they chose the given role, it is difficult to tell what could be driving their choices. There are many possibilities, including perceived utility, curiosity, or interest driven by the demonstrations that the instructor had done in class.

More revealing data comes from the scores that students gave to their interactions with ChatGPT (see Figure 2). The hypotheses about which roles would receive the lowest evaluations—*student peer* (6.88), *philosopher* (6.48), and *educational reformer* (5.13)—were correct, as these roles all indeed received the lowest scores, averaging around 6 out of 10. Somewhat surprisingly, the role of *teacher* (6.89) received a notably low score, almost as low as that for *student peer*. At the other end, the roles of *motivational coach* (7.57) and *tutor/homework assistant* (7.47) earned the highest scores.

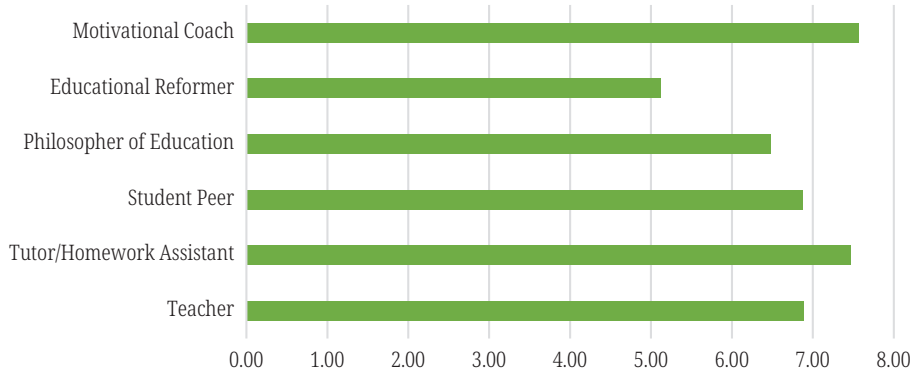

Fig. 2. Average score per role

The contrast between the high scores for *motivational coach* and *tutor* and the low score for *teacher* suggests that students value ChatGPT most when it provides targeted help or encouragement rather than instruction.

In the language of the thought experiment, students were open to giving ChatGPT the jobs of *motivational coach* and *tutor/homework assistant*, but less open to it serving in other roles. Even the highest-scored roles, though, average below 8, suggesting generally moderate evaluations rather than unqualified enthusiasm.

Finally, there are some notable inverse relationships between the number of evaluations and average evaluation scores. While *student peer* was the most frequently selected role, its average rating was mid-range. In contrast, *motivational coach* had the fewest evaluations but received the highest overall score. This suggests that while students were most drawn to the idea of exploring and testing ChatGPT in a peer-like capacity, they were not highly impressed with the results. At the same time, a relatively small number of students were drawn to using ChatGPT as a motivational coach, but those who did were the most satisfied. One possible explanation for selections is that students gravitated towards the role they felt most comfortable assessing based on their real-life experiences, which may explain why *student peer*, *teacher*, and *tutor* were the most frequently chosen roles.

To unpack these patterns and avoid overinterpreting small subsamples, the author analyzed students' open-ended explanations for their ratings. These justifications clarify the criteria used to judge each role, highlight gaps between expectations and outputs, and help explain why coaching and tutoring scored higher than *student peer*, *philosopher*, and *educational reformer*. Significantly, even the highest-scored roles included constructive criticisms that mark limits and suggest improvements. The reflections also reveal deeper philosophical stances toward the technology, including where students place their trust, how they understand agency and authorship, and whether they view ChatGPT as a tool, a partner, or an authority.

Across comments, the most common positives were volume and specificity of information. Students praised ChatGPT for generating ideas, strategies, activities, and lesson plans tailored to their specific prompts. The most common drawbacks were shallow or surface-level analysis, reliance on unreliable or uncited resources, and difficulty addressing complex or controversial issues.

The *tutor/homework assistant* role was the only one that never received a score of 1 or 2. High ratings emphasized "details" and step-by-step help. Lower ratings pointed to incorrectness, repetition, and verbosity. One student noted it gave "explanations which I don't really understand." Another tried to debate "human tutors are better than AI" and concluded, "It proved my point by continuously giving me long and complicated paragraphs."

For the *teacher* role, students who wanted clear explanations or ready-to-use lesson plans rated it well, while those seeking emotional presence rated it poorly. As one student wrote, "it did give emotional support, but it felt cold and unreal." Another worried about how it would support struggling students: "AI are incapable of caring if a student is struggling or not."

Student peer, the most frequently selected role, showed the widest range of uses. Some explored ideas with an AI classmate, but many repurposed it for personal advice, study help, or stress-testing the pretense of being a classmate. One student was especially disappointed about the lack of reciprocity: "I like it because it answers the questions I asked but it doesn't ask many back ... as a student having a discussion it has to work both ways ... it was sometimes hard to discuss with it."

Motivational coach was used mainly for independent work rather than in-class activities. Students valued its personalized advice on specific goals and its ability to shift tone, for example, "nurturing" or "tough." Although this was the highest-scoring role of all, some feedback mentioned negative aspects such as its inability to form genuine emotional connections, provide "feasible advice," or understand the "whole picture" of a situation. One student remarked that it was merely pulling "stuff out of the internet," adding, "For that, I can just read motivational books or attend a seminar or something."

Overall, and unsurprisingly, students who mentioned iterating on prompts tended to give ChatGPT higher scores. This suggests that some limitations could be reduced with further prompt engineering practice. Still, these reflections map early expectations and needs. Framed by roles, students articulated what they wanted from ChatGPT: listening that feels reciprocal, personalization that fits context, concrete next steps, and care that remains human. These patterns echo student-centered, progressive pedagogies and reveal deeper stances about trust, agency, and authorship in working with this technology.

3 STUDENTS AS BUILDERS OF CUSTOM AI CHATBOTS

3.1 Project description, hypotheses & methods

By Spring 2024, it was possible to easily customize ChatGPT with specific instructions given in natural language and share this version as a "custom GPT." The platform for building GPTs, the GPT store, became available in November 2023, together with the rollout of "GPT-4 Turbo." The capability to build GPTs was limited to those with paid ChatGPT subscriptions, but within months it became possible to do so for free in various bot-building platforms of ecosystems like HuggingChat, Poe, and Playlab. In such platforms, users are not limited to using ChatGPT as the LLM and can generally choose from various models, including open-source models.

Whatever the platform, the mechanics of building a bot are similar and do not require deep technical expertise or the use of a formal programming language. At its most basic, the user gives the bot a name and provides specific instructions for how it should interact with users. The bot can then be shared with others via a link. Some optional enhancements include adding reference documents and specifying the bot's style or personality. For example, an instructor can design a custom bot to assist students on a specific reading and check for understanding, following a Socratic style of interaction. An instructor could even design the bot to pretend to be Socrates or some other key thinker, historical, or literary figure that students are learning about.

Given this new technology and the opportunities it offered, the author revised the previous "students as evaluators" artifact, turning it into a "students as builders" artifact. Rather than evaluating interactions with ChatGPT, students would now first build custom bots and then evaluate the finished products. With "prompt engineering" already becoming a less distinctive skill, this new artifact gave students something more notable to showcase in their resumes and teaching portfolios. Further, the act of building a tool situates the assignment in the constructionist tradition. From Dewey's 'learning-by-doing' to Seymour Papert's view of 'objects-to-think-with,' research shows that learners develop deeper understanding when they design, debug, and share public artifacts [4]. In this sense, a custom bot was envisioned as a concrete object through which students could test out ideas, develop more sophisticated prompts or instructions, and reflect on the meaning of this emerging technology for education. Finally, having a finished artifact created concrete opportunities for peer-to-peer feedback, building class community, and deeper exploration of ethical concerns surrounding AI-powered educational tools.

The author hypothesized that the builder approach would surface themes and ethical reflections that the "students as evaluators" artifact had rarely elicited. When students not only conceptualize their interaction with ChatGPT in terms of a role but also package it into a named bot with defined instructions and tone, the framing becomes more anthropomorphic. As such, this could lead to greater engagement and more thought-provoking reflection.

The assignment was integrated into four sections of the course, with a total of fifty-five students participating. For the platform, students were asked to use any freely available custom-building platform and were taught how to use HuggingChat and Poe as the main options available at the time.

The core assignment required students to build a custom bot, share it with others, and then evaluate it. It was carried out in the final weeks of the semester following six stages: (1) determining the bot's purpose, (2) selecting a bot-design platform, (3) designing the instructions, (4) setting up, testing it, and sharing it in the class online board, (5) evaluating one's own bot, and (6) evaluating two peers' bots.

To determine their bot's purpose, students were guided to consider both audience (student vs. teacher, K-12 vs. higher education) and function. The possible roles were similar to those in the previous project, with some adjustments. The *student peer* role was removed, as its function in working through class ideas could be covered within the *tutor* role. The *teacher* role was also removed, as students often used it in a *teacher's assistant* capacity, so the *teacher's assistant* role was added. The five final roles were: tutor, teacher's assistant, motivational coach, philosopher, and educational reformer.

For designing instructions, students were given three approaches to choose from. The first involved using a provided *meta-prompt* that guided an LLM chatbot to collect information from the student and generate appropriate instructions. The second approach allowed students to use provided sample instructions. The third approach encouraged students to start from scratch and design their own based on a list of key components.

After setting up, testing, and iterating their custom bot, students shared it with classmates in the class online board. They then submitted an evaluation form for their own bot as well as three created by peers.

The evaluation form asked students to identify the EduBot by name and indicate whether they were reviewing their own or a peer's creation, along with the approximate time they had spent testing it and the educational role assigned. They then rated the bot's performance on a scale from 1 to 10, with 1 meaning it did not work at all and 10 meaning it exceeded expectations.

A notable new element of the evaluation was the inclusion of a set of questions about potential risks, framed by the instructor as the "ABCs" of concern. Students registered their level of agreement or disagreement with statements about Accessibility (whether the bot was easy to set up and open to everyone), Bias (whether it showed prejudice or lacked neutrality on sensitive issues), Control (whether user data or corporate agendas shaped its use), and Delusions (whether it fabricated information or gave errors while presenting them as facts). A fifth concern, Ethics, asked whether the bot might cause harm in classrooms, be misused, or have negative consequences for education in the long run. Finally, students gave an overall score from 1 to 10 that combined perceived benefits and risks, followed by a written explanation of their reasoning. This structure was designed to guide students to weigh not only raw performance but also these deeper concerns, adjusting their final score if needed.

As part of the analysis of student evaluations, the author used AI tools (Claude and ChatGPT) to help organize responses and identify themes, supplementing her own coding and interpretation.

3.2 Student reflections & findings

As with the "students as evaluators" project, students were free to choose any role for their custom bot and were also free to evaluate any combination of roles for the peer bots evaluated. Thus, data on what students decided to build and evaluate could be used as a proxy for identifying which roles students were most and least interested in.

Nearly every student built a bot, with approximately 50 bots altogether, and students contributed a total of 160 evaluations. Across Spring and Fall 2024, the most popular roles for custom bot creation were *motivational coach* (52%) and *tutor* (42%). The least selected was *educational reformer* (2%), and no students chose to create a *philosopher* bot, though some tutoring bots helped with philosophy. For evaluations, the most frequently selected roles were *motivational coach* and *tutor*, with these two roles combined accounting for more than 90% of all evaluations (see Figure 3).

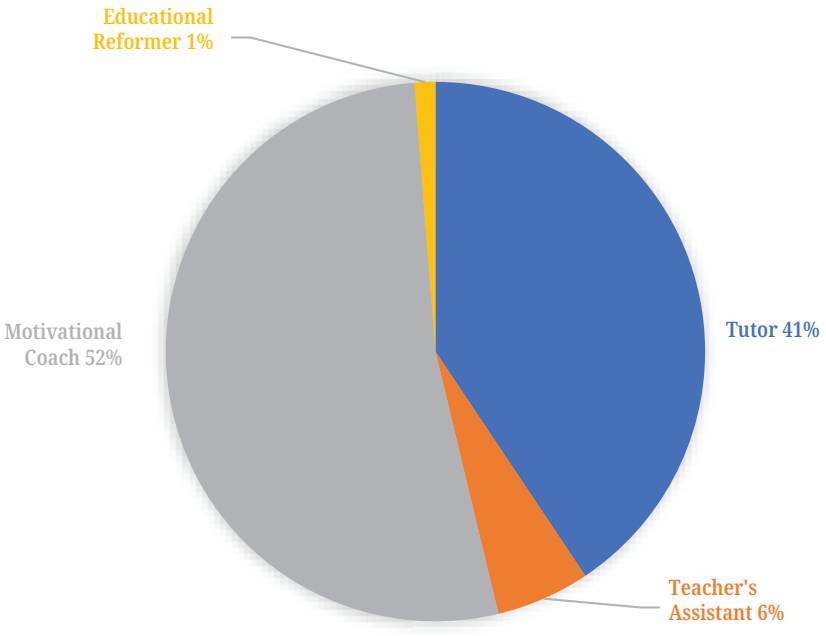

Fig. 3. Percentage of evaluations per role

When it came to performance scores, the scores ranged from 7.5 to 9.9 (see Figure 4). The bots that received the highest scores on average had the role of *teacher's assistant* (9.86), closely followed by *motivational coach* (9.08) and *tutor* (9.02). The lowest-scoring bots had the role of *educational reformer* (7.50).

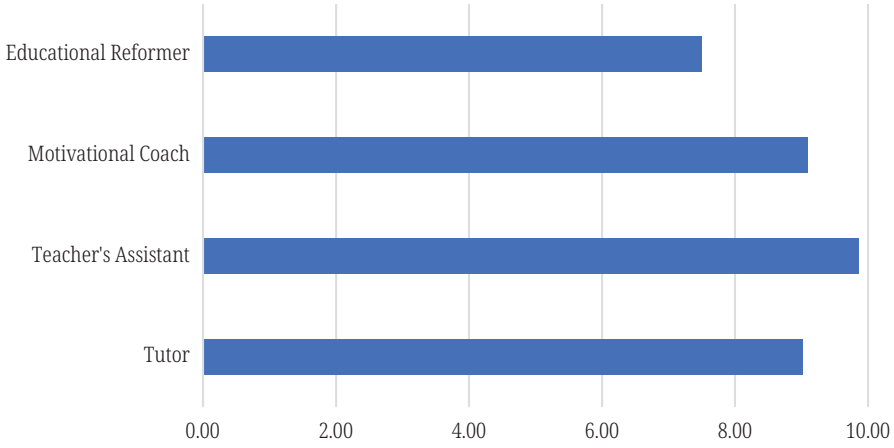

Fig. 4. Average score per role

To see whether these patterns were stable or cohort-specific, the author disaggregated results by semester. Spring and Fall 2024 differed in cohort composition, timing, and students' prior exposure to ChatGPT; breaking the data out by term helps separate novelty and assignment-refinement effects and makes shifts in role choices and scores clear.

In Spring 2024, students completed 98 evaluations and overwhelmingly chose *motivational coach* for testing and review, which accounted for 62%. The role of *tutor* followed at 35%. *Teacher's assistant* and *educational reformer* were tested far less, at 2% and 1% of all evaluations. Average scores were high across the board. *Motivational coach* averaged 8.98 out of 10, and *tutor* averaged 8.68. *Teacher's assistant* received perfect scores in the spring set, although that result rests on only two evaluations. Ratings increased with time on task, based on students' self-reports. Students who spent 15 to 29 minutes with a bot provided the largest share of evaluations, 59.2 percent, and gave the highest mean score at 8.98. Those who spent under 15 minutes averaged 8.67.

In Fall 2024, students completed 62 evaluations and the role mix shifted. *Tutor* became the most frequently evaluated role at about 53% and *motivational coach* fell to about 36%. The *teacher's assistant* role rose to about 8% and *educational reformer* up to 3%. Overall, the fall set evaluations averaged a higher score of 9.18 compared to 8.89 in the spring.

To interpret these results, the author analyzed students' open-ended reviews of peer bots. The comments clarify the judging criteria: strong bots personalized early, adapted tone, offered step-by-step help, and produced concrete outputs. Even top-rated bots drew critiques about robotic affect, verbosity, errors, role drift, and pretending to know. This shifts attention from answer quality to design choices such as voice, intake, guardrails, and citation. The project may have also led, as hypothesized, to more profound reflections about agency, authorship, and responsibility.

The most common positive themes centered on bots that felt truly personalized, gathering information about the user and tailoring responses from the outset. Several *motivational coach* evaluations were effusive, praising tailored

encouragement for fitness, career growth, and personal development, with one evaluator calling a peer's coach a "truly empowering tool." Students appreciated an encouraging tone and positive affirmations. A student commented, "This was phenomenal ... very helpful and motivational" and another explained "I told it I was sad and it gave me affirmations." Students also appreciated bots with specific voices and personalities, including a popular motivational coach in the style of "The Rock."

When it came to the *tutor* role, students most appreciated clear explanations and step-by-step breakdowns. One student commented, "my bot was able to break down mathematical terms, provide solutions to equations, and give thorough definitions." Students also appreciated practice and feedback, including giving "feedback for both right and wrong answers." Another student wrote that the bot was "excellent for practicing EAS [Educating all Students] and ATAS [The Assessment of Teaching Assistant Skills]" and useful when class coverage left gaps. Notably, some students tried out their own tutors with users outside the classroom and were satisfied with the results. As one explained, "It efficiently did what I needed it to without any issues and I tested it with my 3rd graders because it was a tutor."

The *teacher's assistant* positive themes highlighted efficiency and concrete outputs for planning, including bilingual scaffolds. One student wrote that a peer's assistant helped design topic-specific lesson plans and "even gave me an outline for a newsletter ... in English and Spanish."

At the same time, although scores were generally high, subjective comments included some negative and constructive feedback. Some students commented that it did feel too robotic and lacked genuine emotional presence at times. One student appreciated that the bot itself acknowledged this, writing, "I loved how you were able to give personal information but it let you know it's a bot it can't have real feelings or be your actual best friend." Another student was less impressed by the interaction and concluded, "I think there are better tools for motivation, and this is only when asked for and on the app so it takes more work. Almost like motivation is needed to get started on using this motivational tool." A student commenting on a *tutor bot* provided a complex evaluation, faulting that it lacked "the basic foundation I look for when being a tutor" but also mentioning that it was overall "a good bot."

Other negative remarks noted occasional errors. Examples included "it messed up on 2 questions I asked" and that "it just gave false/wrong information." A student summed up the experience as "a hit or miss when it comes to math help." Students also noted generic or wordy responses: "some [answers] are long" and "somewhat helpful ... but broad responses."

In addition, students flagged a variety of issues related to bots that did not meet expectations, such as a bot designed for beginners that turned out too advanced and a "grammar tutor" that behaved more like a teacher's assistant rather than a student tutor. Another recurring theme matched with the "D" or "delusions" component in the evaluation form. Students were disappointed when bots "pretend[ed] to have info it actually doesn't."

Overall concern levels were generally low to modest, with the average being 1.5 on a scale of 1 to 4, where 1 = no concern and 4 = very concerned (see Figure 5). Delusions, that is, factual errors, drew the highest concern, followed by Ethics and Control, while Accessibility and Bias were lowest. About 29% of evaluations registered at least one moderate or higher concern, and only about 9% flagged a very high concern in any category. Roughly 44% reported no concerns across all five categories.

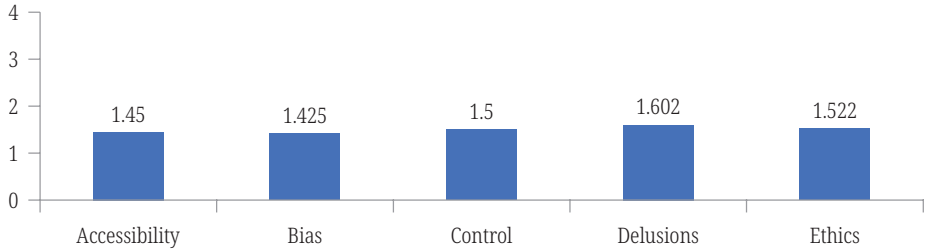

Fig. 5. Average scores on the "ABCs" of concerns

Finally, the performance score given at the beginning of the evaluation form was very similar to the score given after entering the level of concern with all the "ABCs." This suggests that entering scores for those concerns did not affect how students thought of ultimate performance, at least not during the process of completing the form.

4 STUDENTS AS EVALUATORS VS. BUILDERS

Although comparison across semesters and project types is challenging, notable contrasts emerge in quantitative and qualitative data from the "students as evaluators" to "students as builders" projects.

Overall, the average score for when students evaluated LLM chatbots in 2023 was 6.83 across all roles. By comparison, the average score for when students built custom bots in 2024 was 9.00 across all roles. Further, unlike in the first project where low scores dipped to 1, there were no custom bots with scores lower than 5. This could be due to a variety of factors, including the availability of more advanced models, better prompting techniques, increased time spent iterating, perceived pressure from sharing the bot with peers, and students' tendency to be kind and generous when evaluating each other's creations rather than those of a generic chatbot. It could also be due to the added satisfaction that comes from being positioned as a designer and builder. Future work could explore this by comparing evaluator and builder projects under more controlled conditions and by examining how the experience of building shapes engagement, motivation, and judgment.

These quantitative differences also mirror a qualitative shift. The transition from students as evaluators to students as builders was not just about achieving higher scores, but about a different kind of engagement. As students transitioned from testing LLMs on an all-purpose platform to building their own versions, their reflections broadened to include new considerations.

When students acted as evaluators, their reflections generally emphasized how well ChatGPT fulfilled predefined educational roles. The focus was on *performance, specifically* clarity, relevance, and emotional impact. Common positives included tailored explanations, step-by-step help, and an abundance of information. Common negatives centered on shallow analysis, unreliable or uncited sources, limited emotional realism, and lack of reciprocity. To some extent, students judged ChatGPT as if it were a performer given a script and based on how convincingly it fulfilled that script.

When students became builders, however, their reflections shifted from merely judging outputs to reflecting on design. They paid more attention to tone, voice, intake questions, and guardrails—dimensions that previously went unnoticed.

Even top-rated bots drew thoughtful critiques for being "too robotic," verbose, or overconfident, often "pretending to know" more than they did. It is critical to note that part of this shift may be attributed to the addition of the "ABCs" of concern questions to the evaluation instrument. It is difficult to separate the effect of building bots from teaching students about possible concerns and asking them to rate them.

Whatever the case, there is a notable shift in responses towards the ethical dimension. Student builders were more likely to bring up questions of agency, authorship, and accountability in their responses.

Future work could set up more controlled conditions where the effect of building is separated from the impact of presenting students with possible concerns.

5 IMPLICATIONS FOR CORPORATE LEARNING

The evaluator and builder frameworks also point to lessons for adult learning and workplace training. Employees, like students, can benefit when they are invited to evaluate existing AI tools, test prompts, and build simple, role-based custom bots for their own tasks.

It has been argued that AI typically automates tasks rather than end-to-end jobs, since jobs are bundles of many tasks and adoption tends to be incremental [5]. If this is the case, then it might seem misleading to guide employees to think in terms of roles or job replacement rather than specific tasks and workflows. However, a role-based lens remains pedagogically and practically useful: a 'role' is a coherent bundle of tasks with clear goals, boundaries, and guardrails. Framing pilots as "job tryouts" does not imply replacement; it provides a clear scaffold that bundles tasks, makes design decisions visible, and clarifies accountability. Because the phrase resonates in media headlines, it engages learners and encourages deeper, more critical investment in the issues at stake.

Additionally, "role-based" reflections are flexible and do not have to match official job titles, just as in the classroom, they were not limited to the role of "teacher." The same educational roles explored with students could easily be translated into workplace settings. For example, the *tutor* role could support product knowledge and safety, and the *teacher's assistant* role could help assemble resources and support ongoing tasks. The *motivational coach* could become a companion for tracking goals and wellness. More controversially, the *philosopher* role might be used to raise new criticisms or invite outside-the-box thinking, and the *educational reformer* role could even be applied to questioning established processes and reimagining how an organization operates.

The same "ABCs" lens provided to students could also support employees in navigating concerns surrounding the adoption of AI tools in the workplace, empowering them to make decisions in their day-to-day use and to participate in more democratic decision-making within an organization.

Small pilots can be run by defining the task, collecting real cases, testing, and revising. As with the student project, data can capture not only raw performance (such as gains in speed and accuracy) but also subjective evaluations or reflections about the experience. Gains in performance can be weighed against concerns about accessibility, bias, and other ethical implications. When employees are positioned as both evaluators and builders, the process supports not only performance but also judgment and responsibility, much as it did in the classroom.

6 CONCLUSION

This phenomenological study captured how students engaged with AI in education, either as evaluators of general LLM chatbots or as creators of custom bots. Through both projects, students developed their AI literacy and thought critically about key limitations of LLM chatbots. AI's effectiveness in education is not solely determined by its capabilities, but by how students engage with, critique, and adapt it. When students are given the opportunities and frameworks for critical engagement with and creativity through AI tools, they can develop a deeper understanding of both its possibilities and its limits. The same evaluator–builder dynamic also carries implications beyond the classroom: in corporate and workplace settings, inviting employees to both evaluate and build role-based tools may support not only efficiency but also reflection, responsibility, and ethical awareness. Effectiveness, whether in schools or organizations, depends less on technical power alone and more on how individuals are empowered to shape and take ownership of the tools they use.

7 ACKNOWLEDGMENTS

Support for this project was provided by a PSC-CUNY Award, jointly funded by The Professional Staff Congress and The City University of New York. Support for implementing the teaching activities was provided by CUNY Computing Integrated Teacher Education (CITE). Additional support was provided by CUNY Building Bridges of Knowledge (BBK), funded by the Lumina Foundation.

8 DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, the author used ChatGPT in order to assist with brainstorming, revising, and proofreading. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

9 REFERENCES

[1] C. Silva Sibilin, "Education and the epistemological crisis in the age of ChatGPT," *Critical Review*, vol. 35, no. 4, pp. 414–425, 2023. https://doi.org/10.1080/08913811.2023.2284042

[2] J. A. Bowen and C. E. Watson, *Teaching with AI: A Practical Guide to a New Era of Human Learning*. Baltimore, MD: Johns Hopkins Univ. Press, 2024.

[3] J. White *et al.*, "A prompt pattern catalog to enhance prompt engineering with ChatGPT," *arXiv preprint arXiv:2302.11382*, 2023.

[4] S. Papert, *Mindstorms: Children, Computers, and Powerful Ideas*. New York, NY: Basic Books, 1980.

[5] A. Narayanan, "A guide to cutting through AI Hype: Arvind narayanan and melanie mitchell discuss artificial and human intelligence," CITP Blog (Princeton Univ. Center for Information Technology Policy), Apr. 2, 2025. Available: https://blog.citp.princeton.edu/2025/04/02/a-guide-to-cutting-through-ai-hype-arvind-narayanan-and-melanie-mitchell-discuss-artificial-and-human-intelligence/

10 AUTHOR

Casandra Silva Sibilin is a Lecturer of Philosophy in the Department of History, Philosophy and Anthropology at York College, City University of New York (CUNY). She is the founder and leader of AI in Education's subgroup on Custom Bots/GPTs. Research interests in the fields of AI in Education, Philosophy of Education, Teacher Education, Custom Bots/GPTs, Student Engagement and Motivation, Growth Mindset, and Ethical/Equitable AI Integration (E-mail: csilvasibilin@york.cuny.edu).