International Journal of Social Science and Economic Research

ISSN: 2455-8834

Volume:10, Issue:12 "December 2025"

An Adaptive Self-assessment Chatbot Greatly Outperforms Chat GPT in Teaching Students About the Health Topic of Stress

Yuqi Wang, John Leddo, Zarah Koroth, Srishti Kallu, Riya Biswas, Anika Dubey, Shree Garg, Dhairya Garg, Ishwarya Ramineni, Anvi Allada, Amisi Ghaus, Kiana Ghaus, Deetya Hamilpur, Rishik Vanga

MyEdMaster, LLC

DOI: 10.46609/IJSSER.2025.v10i12.039 URL: https://doi.org/10.46609/IJSSER.2025.v10i12.039

Received: 10 December 2025 / Accepted: 24 December 2025 / Published: 30 December 2025

ABSTRACT

Chatbots, personal assistants, and large language models (LLMs) have become pervasive in our world. A major function they serve is to provide information and answer questions. Collectively, these technologies typically have two major weaknesses: they provide general answers to questions they are asked without regard to the specific knowledge needs of the user and they do not assess whether the user actually understood the answers or information provided.

Previously, we addressed the first weakness by creating a self-assessment chatbot that helps a user to self-assess what s/he knows about a topic and then uses that knowledge when answering the user's questions with an eye toward filling in the knowledge gaps (Wang and Leddo, 2025). This self-assessment chatbot was presented to college students learning Algebra II and was compared to Chat GPT, which simply answers questions without such self-assessment information. Results showed that students using a self-assessment chatbot scored 10 points or the equivalent of a full letter grade higher on a posttest than those using the standard Chat GPT. Maviti and Leddo (2025) extended these results by having high school students learn calculus using either Chat GPT or a self-assessment chatbot. Those results showed that students using a self-assessment chatbot scored, on average, 44 points higher on a posttest. Maviti, Leddo and Prakash (2025) replicated the Maviti and Leddo (2025) study by comparing the effectiveness of a self-assessment chatbot to that of Gemini. There results showed that those using Gemini scored, on average, 68% on a posttest, while those using the self-assessment chatbot scored, on average, 92%. The present study extends these previous findings, in which learning math was the focus, to learning a new topic: stress. 25 middle and high school students were taught about stress and how to cope with it. 13 students used Chat GPT-5-2 to help them understand the topic and 12 used a personalized, self-assessment chatbot. All took a posttest on the topic afterwards and

ISSN: 2455-8834

Volume:10, Issue:12 "December 2025"

those using Chat GPT-5-2 scored, on average, 15.29 on the posttest, while those using the personalized chatbot scored, on average, 20.94. Results suggest that incorporating the self-assessment feature into the chatbot boosts performance in subject areas other than math.

INTRODUCTION

Across more than forty years of studies, researchers have shown that individualized instruction can improve learning outcomes far more effectively than whole-class teaching. Bloom's well known "2-sigma" finding showed that one-to-one tutoring can raise student performance by as much as two standard deviations (Bloom, 1984). Based on this foundation, research on intelligent tutoring systems (ITS) shows that well-designed computer tutors can closely approximate the effectiveness of human tutoring. Van Lehn (2011) showed that such systems work by modeling learners' cognitive states and adjusting both feedback and problem selection.

For example, cognitive tutor frameworks use domain modeling and step-level feedback.

According to Koedinger and Corbett (2006), such designs contribute to gains in both procedural fluency and conceptual understanding. More recently, the advent of large language models (LLMs) has renewed interest in natural language–based tutoring tools. These tools are valued for their conversational flexibility. Reviews, however, emphasize that their educational impact depends on alignment with learner needs, along with transparency and accuracy (Kasneci et al., 2023). Building on this concern, our study tests whether a self-assessment-informed conversational agent can outperform a general-purpose LLM in supporting undergraduate mathematics learning.

Adaptive learning systems—particularly intelligent tutoring systems (ITS)—have demonstrated consistent effectiveness in improving learner outcomes by tailoring instruction to individual cognitive states (Létourneau, 2025). Létourneau's (2025) systematic review of AI-driven ITS across K–12 schools found mostly positive results. Students using ITS showed stronger learning gains than those in non-intelligent environments, though the extent of improvement depended on design and duration. Central to these systems are mechanisms such as model tracing and knowledge tracing. These techniques allow real-time monitoring of students' problem-solving steps and estimation of skill mastery, which in turn support dynamic task selection and just-in-time feedback (Koedinger and Corbett, 2006). Moreover, modern AI-based ITS structures often incorporate natural language processing modules and real-time assessment pipelines. Through these components, the system can adjust how it delivers content in response to students' ongoing performance (Villegas-Ch et al., 2025).

Researchers are now exploring how LLMs, including GPT-5, might be applied in adaptive learning. Early findings suggest that such models could act as tutoring systems that are both

ISSN: 2455-8834

Volume:10, Issue:12 "December 2025"

flexible and sensitive to context. Although LLMs have demonstrated strong capabilities in generating explanations and scaffolding problem-solving (Kasneci et al., 2023), concerns remain about their tendency to produce overly general responses when not anchored in student-specific data. Emerging work in personalized educational agents suggests that coupling LLMs with diagnostic or self-assessment modules may provide a pathway toward more learner-sensitive feedback (Zawacki-Richter et al., 2019). Nevertheless, rigorous controlled experiments validating the effectiveness of such hybrid systems in real classroom settings remain scarce. This gap emphasizes the need for empirical studies to test the effectiveness of personalization mechanisms built on LLM infrastructure. The present experiment addresses this need. It tests whether such mechanisms improve learning outcomes more effectively than conventional chatbot interactions.

One way to achieve personalization is to adjust instruction to what the learner already knows. Indeed, the traditional ITS model contains a student model for that very purpose (Greer, 1995; Brna, Ohlsson and Pain, 1993). The lack of a student model represents a fundamental weakness in mainstream LLMs, which are geared towards answering questions without regard to who is asking them. This makes sense since LLMs are, by their very nature, language models not teaching models. Therefore, they are not constructed to strategically assess what knowledge learners have and what they are missing, so that these gaps can be used in the process of generating answers.

Once solution is to create an independent assessment system and link it to an LLM. This is labor intensive. Another solution is to allow a learner to enter his or her own existing subject matter knowledge into LLM and have the LLM use that information when answering a learner's questions. Given that learners may not be skilled in assessing their own knowledge, a self assessment LLM-based chatbot needs a reliable and easy to use self-assessment method.

Our previous work has been devoted to developing such a method. Given that the goal of the proposed self-assessment chatbot is to fill in knowledge gaps, traditional assessment methods that focus on whether users can correctly answer questions are inadequate since these methods do not diagnose knowledge but performance. The assessment method used in the present project is called Cognitive Structure Analysis (Leddo et al., 1990).

Cognitive Structure Analysis or CSA is based on decades of cognitive psychology research that have illustrated that people possess various knowledge types, each of which is organized and used differently in problem-solving. Since people possess different types of knowledge, our framework integrates several prominent and well-researched formalisms. These include semantic nets, which organize factual information (Quillian, 1966); production rules, which organize concrete procedures (Newell and Simon, 1972); scripts, which are general goal-based problem-

ISSN: 2455-8834

Volume:10, Issue:12 "December 2025"

solving strategies (Schank and Abelson, 1977; Schank, 1982); and mental models, which explain the causal principle behind concepts (de Kleer and Brown, 1981). Because our framework integrates these four knowledge types, it is called INKS for the INtegrated Knowledge Structure.

The INKS framework is based on research by John Leddo (Leddo et al., 1990) which shows that true mastery of a topic or subject requires all four knowledge types. The framework also brings helpful implications for instruction. For example, in John Anderson's ACT-R framework, people initially learn factual/semantic knowledge that is later operationalized into procedures (Anderson, 1982). Research by Leddo takes this one step further showing that expert knowledge is organized around goals and plans (referred to in the literature as "scripts" – Schank and Abelson, 1977; Schank, 1982) and abstracted into causal principles (referred to in the literature as "mental models" – cf., de Kleer and Brown, 1981) that allow people to construct explanations and make predictions/innovations in novel situations.

To identify the root cause of the mistake, the query-based assessment framework CSA incorporates principles from the INKS knowledge representation framework. CSA is chosen because previous research describes a strong correlation between user knowledge — as assessed by CSA — and performance practical problem-solving. In one previous research project, we found that using an automated multiple-choice CSA system to access student learning produced measures of knowledge that correlated .88 with student problem-solving performance and measures of change of knowledge as a result of the instruction that correlated .78 with change in performance from pretest to post test. Moreover, at risk students who had their learning needs diagnosed using CSA performed at a mainstream level three grades higher than their own after a 25-hour tutoring program in science (Leddo and Sak, 1994). Leddo et al. (2022) extended these findings. Students were given open ended questions to assess their factual (semantic), strategic (script-based), procedural, and rational (mental model) concept, knowledge of Algebra 1. The total INKS knowledge and individual component knowledge scores were correlated with the total number of correctly solved problems. Results showed correlations of .966 between problem-solving and total knowledge, .819 between problem-solving and strategic knowledge, .866 between problem-solving and factual knowledge, .937 between problem-solving and procedural knowledge and .788 problem-solving and rational knowledge. These findings were extended to pre-calculus (Zhou and Leddo, 2023), biology (Ahmad and Leddo, 2023), and elementary school math (Bekkari and Leddo, 2023). In two other projects, assessments of students' knowledge produced using the CSA methodology agreed with teachers' assessments approximately 95% - 97% of the time which was statistically equal to teachers' assessments with each other (Leddo et al., 1998, Liang and Leddo, 2020).

Our previous work shows that CSA can be a powerful tool in helping educators assess what students do and do not know. CSA has been presented as an alternative to the classical test

ISSN: 2455-8834

Volume:10, Issue:12 "December 2025"

theory approach of measuring learning as a function of the number of correct answers students give. However, it could be reasonably argued that the purpose of education is to improve student performance, and, therefore, replacing an assessment system with one that directly measures underlying knowledge but does not raise student performance would be less appropriate. Leddo and Ahmad (2024) addressed that issue directly. In that study, high school students were initially assessed in their knowledge of logarithms. Half were assessed using CSA and the other half were assessed by asking them to solve problems and show all work. After each problem, students received remediation on either their knowledge concepts (in the CSA condition) or in their problem-solving steps (the "show all work" condition). Results showed that remediating problem-solving steps raised student performance from an average of 68% on the pretest to 75% on the posttest, a statistically significant increase. However, those who had their knowledge assessed and remediated scored 85% on the posttest, a statistically significant, full-letter grade higher performance than those in the "show all work" condition. The Leddo and Ahmad (2024) was replicated in a follow-up study with middle schoolers that also showed that students who were assessed using CSA and had their knowledge remediated performed, on average, a full letter grade higher than those whose step-by-step procedures were assessed and remediated (Challagulla and Leddo, 2025).

Showing that assessing and remediating INKS-based knowledge improves performance addresses only half the issue. The previously-cited research involved learners being assessed using external means. For a self-assessment chatbot to work, the question remains whether learners can be taught to reliably assess their own knowledge and, equally importantly, whether learning to self-assess can be done quickly and easily so as to be practical to implement.

It turns out the answer to each of these questions is yes (Cynkin and Leddo, 2023; Dandemraju, Dandemraju and Leddo, 2024). In these two studies, we showed that learners can be trained to accurately assess what they do and do not know, and that this process takes about 10 minutes. To train a person to self-assess, s/he is shown a sample of what a self-assessment for a topic area looks like. The learner is then asked to use the sample as a model for generating a self assessment for a new topic. A template is provided for filling in the factual (semantic), strategic (scripted-based), procedural (production rule) and rational (mental model) knowledge.

To ensure that remediation of self-assessed knowledge also leads to improvement in performance, we have also taken the next logical step in that area to see if students can not only assess their knowledge gaps but also then remediate these gaps. It turns out that students can do so very successfully. To address this issue, Ravi and Leddo (2024) conducted a study in which high school students learned an advanced topic in chemistry by watching a video. Half the students were told to rewatch the video to fill in any knowledge gaps, while the other half were taught to self-assess their knowledge using CSA and then told to rewatch the video to fill in any

ISSN: 2455-8834

Volume:10, Issue:12 "December 2025"

assessed knowledge gaps. The group that was taught to self-assess scored 15 points or 1.5 letter grades higher on a posttest than students who simply rewatched the video without self assessment. Nehra and Leddo (2024) replicated the Ravi and Leddo study to the learning of Spanish. They found that high school students performing self-assessment plus remediation scored, on average, 25 percentage points or 2.5 letter grades higher than those re-reading the material without performing a self-assessment. Prakash and Leddo (2025a) extended the Ravi and Leddo (2024) and Nehra and Leddo (2024) findings to another subject area: high school reading comprehension. The results revealed a mean posttest score of 8.3 out of 12 (69.17%) for the control group and 11.2 out of 12 (93.33%) for the experimental group. This difference in averages is statistically significant (t = 3.75, df = 11.07, p < .01). Notably, individual scores further illustrate the disparity: the lowest score in the control group was 41.67%, whereas the lowest in the experimental group was 83.33%. This is the difference between an F letter grade and B letter grade. Following this, another study conducted by Prakash and Leddo (2025b) examined CSA's effectiveness in teaching math, specifically, the topic of Bayes' Theorem, and found a 27-point improvement. Individual scores also highlighted the disparity. The control group's lowest score was 6/20 (30%), whereas the experimental group's lowest score was 15/20 (75%). Following this, a history assessment revealed that students who utilized CSA for self-assessment and remediation significantly outperformed their peers in the control group (Prakash and Leddo, 2025c). Posttest results demonstrated that the experimental group achieved an average score of 87.5%, whereas the control group scored 65.8%, indicating a substantial difference in comprehension and retention of historical concepts.

These results on high school students were further extended by Leddo, Clark and Clark (2025) in their investigation of middle school math. Leddo, Clark and Clark found that middle school students who self-assessed using CSA and then remediated their knowledge gaps scored 18 percentage points higher on a posttest than those who relearned material without first performing a self-assessment. Following this, Prakash and Leddo (2025d) conducted a study on middle school students' reading comprehension, specifically through an analysis of To Kill a Mockingbird, a novel that explores complex themes of ethics and social structure. Students in the experimental group were trained to evaluate their own knowledge gaps and use targeted remediation strategies, while those in the control group engaged with the text without structured self-assessment. Results showed that students in the self-assessment group scored 16 points higher on a posttest than those who re-read the material without self-assessment. This was followed up with a study on middle school science (Prakash and Leddo, 2025e), in which students learned about topics in ecology. Results showed that students who used the self-assessment technique plus remediation scored on average 98% on a posttest, while those who simply reread the material without self-assessment scored on average 77.5%.

ISSN: 2455-8834

Volume:10, Issue:12 "December 2025"

Finally, Sathiyamoorthy and Leddo (2025) showed that college students who used CSA to self assess and then remediate knowledge performed 13 percentage points higher on a college psychology posttest than those who simply reread the material after initially learning it. Taken together, these results suggest that regardless of whether the students self-assess and remediate knowledge or the assessment and remediation is mediated by technology, assessing and remediating knowledge greatly improves student performance compared to traditional methods of assessment. This indicates that student achievement could be increased systemically and cheaply by introducing CSA-based knowledge assessment into educational practices.

Given that self-assessment enables students to remediate their own learning needs, Wang and Leddo (2025) explored the question of whether a chatbot could use the results of a user's self-assessment when answering the user's questions. These researchers constructed a chatbot that first had a user self-assess his/her knowledge of a math topic (Algebra II). The self-assessment was used by the chatbot to identify knowledge strengths and deficiencies that were then addressed in answers to users' questions. This was tested on college students, and results showed that students who used the self-assessment chatbot scored, on average, a full letter grade higher than those who used Chat GPT.

Since the Wang and Leddo (2025) explored the effects of a self-assessment chatbot on students learning relatively easy math topics (Algebra II is a high school level subject), Maviti and Leddo (2025) studied whether these results would hold up with more difficult subject matter. In their study, high schoolers learned calculus by using either Chat GPT or a self-assessment chatbot. In this case, the disparity was even stronger. Students using Chat GPT scored, on average, 48% on a posttest, while those using a self-assessment chatbot scored, on average, 92%. Maviti, Leddo and Prakash (2025) followed up this study and compared the effectiveness of the self-assessment chatbot to Gemini in teaching high school students calculus. Once again, those using the self-assessment chatbot scored, on average, 92%, while those using Gemini scored 68%.

The present study investigates whether the relative superiority of the personalized chatbot to Chat GPT in learning a completely different topic area: one related to health. Other chatbots have been developed that focused on health-related topics. For example, a review looked at studies where mental-health chatbots were tested on both high-school and college students, specifically in the subject of student mental-health support. The chatbots could hold real conversations, give coping tips, and offer supportive, empathetic responses. Most studies found they helped lower stress and boost confidence and engagement, even though the results weren't always strong in every single study (Konadu and Kusi, 2025).

With regards to general health, one study tested the use of a chatbot with patients with specific health conditions, primarily focusing on individuals seeking medical advice or support (Hasanein

ISSN: 2455-8834

Volume:10, Issue:12 "December 2025"

and Sobaih, 2023). The chatbots features were natural language understanding, personalized responses based on user input, and the ability to handle a variety of health-related questions. The findings showed that the chatbot engaged users and provided correct information, improving the satisfaction of users. Participants with access to the chatbot described it as helpful and accessible; this supports its potential utility in healthcare settings to help in traditional patient-provider interactions. The present study seeks to extend the previous work on health chatbots by incorporating self-assessment into the chatbot to allow for responses that are tied to the users' existing level of understanding of the topic the chatbot is being queried on.

METHOD

Participants

25 middle and high school students participated in this study. All Participants were from the United States.

Personalized Chatbot Technology

In this study, we developed a lightweight AI-powered personal agent designed to provide adaptive guidance based on students' inputs. The agent is built upon the GPT-5-2 large language model which incorporates a reasoning framework that supports dynamic prompt chaining and reflection processing. Meanwhile, through integrating a self-assessment step, the agent evaluates and captures the student's current understanding level of the subject before initiating personalized tutoring conversations. This pre-processing step allows the agent to generate instructional responses that are more aligned with each student's cognitive level. Although the infrastructure utilizes the common LLM API, the customized personal agent built on it can be more adaptable to the individual level of students than the common chatbot.

Materials

Two versions of a Google Form were created. These were identical, differing only in which chatbot was made available to the Participants. Both Google Forms began with instructional material that covered the topic of stress, including the biological basis for stress and how to cope with excess stress in one's life. These instructional materials can be seen by clicking the link: https://docs.google.com/document/d/1xRtNmBDl3DCKk31kwIUjMIk

The next part of each Google Form involved the chatbot. For the control condition, the Google Form contained a link to Chat GPT-5-2. For the experimental condition, the Google Form contained a self-assessment form which consisted of instructions on how to conduct a self-assessment, followed by space for the Participant to enter his/her own self-assessment. The

ISSN: 2455-8834

Volume:10, Issue:12 "December 2025"

results of the self-assessment were passed to Chat GPT-5-2 with instructions to use the self-assessment when answering the user's questions. The instructions for the self-assessment are shown below:

"I want to teach you to assess your own knowledge that you have about a health affliction and how to determine if this is impacting YOU and what to do about it. Suppose you wanted to assess your own knowledge about colds. If I want to be able to assess my knowledge about health conditions, I need four types of knowledge. These are facts, strategies, procedures and rationales. Facts are concepts or definitions you have to describe objects or elements. For example, in the case of a common cold, I would need to know what exactly colds are, how they spread, when symptoms typically appear, how they impact different age groups, and how long colds typically last. Strategies are general processes I would use to cure a cold or alleviate its symptoms. For colds, this would be things like resting, drinking plenty of fluids, hygiene, etc. Procedures are specific steps that you would use in processes or the strategies said earlier (think of them like mini-building blocks that make up strategies). So, in the example of hygiene, this would be things like washing your hands, avoiding close contact with people, properly disposing of things like used tissues, sneezing into your elbow, etc. Finally, I need to know the rationales, or the reasons why strategies and procedures work the way that they do. In the example of hygiene, this could be that the procedures above like washing your hands and avoiding close contact help to stop the spread of germs which could make you sick and give you a cold. Or, resting reduces the body's need to use its resources for other things and frees up resources for the body to use to fight the cold. Think of facts as the "what", strategies and procedures as the "how" and rationales as the "why".

Keeping these things in mind, this is how I might assess my own knowledge on colds. Practice writing what you know about each of these things (facts, strategies, procedures, and rationales) to figure out where your personal strong points are and which points need more work. Let's practice together using the example of colds.

For facts, I know that the cause of the common cold is viruses like rhinoviruses, coronaviruses, and adenoviruses. You know that colds spread through coughing and sneezing, but they also spread through touching contaminated surfaces, which is common to forget. I remembered that symptoms usually appear a few days after exposure, but forgot the specific number of days (typically 1-3 days). I know that most colds last about 7-10 days, though I didn't realize mild symptoms can linger for up to 2 weeks. Things you don't know: Are there any supplements/vitamins that could help me fight this cold? Can I still exercise when I have this cold? What is the point at which I am no longer contagious and can continue doing normal things?

ISSN: 2455-8834

Volume:10, Issue:12 "December 2025"

For strategies, I know that you can rest and allow your body to heal, drink warm liquids such as herbal teas and warm water, stay very hygienic (sanitize and wash hands frequently) , and you should isolate yourself to prevent spreading the cold to others. Things I don't know: Are there things I can do to reduce my symptoms/suffering while I'm healing?

For procedures, I know that you can monitor your temperature daily to see if your health is getting better and identify things that would need medical attention such as a high fever or difficulty breathing. I know I should drink a lot of fluids. Things I don't know: Are there any fluids that I might drink that might actually just be making me more dehydrated? Is it better to take my temperature on my tongue, forehead, armpit, etc.?

For rationales, I know that resting allows your immune system to focus on the energy fighting the cold, hydration keeps the mucous membranes moist, monitoring ear pain helps differentiate between cold and ear infection. Things I don't know: I kind of forgot how exactly humidifiers work, although I recognize they help cure/prevent colds. Why does it take a few days before cold symptoms begin to appear? Why do I get so much congestion when I have a cold?

After writing whatever you know for each of these points, look over what you have written to see which points you know, what you possibly got wrong, which points you did not know at all, etc. Then add what you do know and what you think you are still missing based on your self-assessment. When I look over what I wrote, I see that I am generally good with facts, but may have forgotten some specifics like how long it takes before symptoms show up, additional things I could do (like supplements/vitamins), when it is okay to continue doing normal activities, etc. For strategies, I can see that I'm generally good at remembering effective strategies for preventing and recovering from colds. I realize that I don't know how to reduce symptoms while healing, so that is a point that I need to do some more research on. For procedures, I see that I'm good at remembering specific procedures to implement strategies. I realize that I don't know which place is best to use to measure temperature and if there are any fluids I should actually avoid drinking. For rationales, I can see that I'm generally good but I did forget how humidifiers work. I don't know why it takes a while before symptoms appear and why I get as much congestion as I do, so that is something I need to look into further. Using those gaps in the writing that I did earlier, I can do some research to find out the things I didn't know earlier. For example, I can do research to find out that humidifiers work to cure/prevent colds by preventing the airway from drying out, as that can worsen coughing and irritation.

Use the above example of a self-assessment for colds and complete one for yourself on what you know about stress and how it affects people. Be sure to include facts, strategies, procedures, and rationales. Once you are done, review your self-assessed knowledge and summarize what things

ISSN: 2455-8834

Volume:10, Issue:12 "December 2025"

you know and don't know. Then use our health chatbot to help you fill in the gaps of what you don't know and to learn more things about stress."

The finaI part of each Google Form was a posttest that assessed how well Participants learned the material. The posttest consisted of the following questions:

1. What function does stress serve as a survival mechanism?

2. What is the purpose of cortisol in reallocating energy during a threat?

3. What does allostasis allow the body to achieve during changing stress conditions?

4. What does allostatic load indicate about long-term stress exposure?

5. Why does chronic stress increase vulnerability to cardiovascular disease?

6. Why can stress trigger or worsen respiratory problems such as asthma or hyperventilation?

7. Why does chronic muscle tension commonly lead to headaches or musculoskeletal pain?

8. Why does chronic stress weaken immune defenses over time?

9. Why do negative thinking patterns intensify physiological stress responses?

10. Why does social support reduce stress levels?

11. Why does physical activity help lower stress chemically in the body?

12. What purpose does mindfulness or visualization serve in regulating stress?

13. Why is professional support recommended when stress disrupts daily functioning?

Procedure

Participants were randomly assigned to either the control or experimental condition. They were then given the appropriate Google Form. The Google Form instructed them to read the instructional materials and then use the chatbot to ask whatever questions were required to ensure the Participants understood the material. Then, they took the posttest. Participants had unlimited time to go through the Google Form, including time with the chatbot. However, when they took the posttest, Participants were not allowed to review any material, use the chatbot, or have access to outside sources.

ISSN: 2455-8834

Volume:10, Issue:12 "December 2025"

RESULTS

Responses to the 13 posttest questions were scored on a 2-point scale with partial credit given. Therefore, the maximum possible posttest score was 26. Results showed that Participants in the control condition (using Chat GPT-5-2 alone) scored, on average, 15.29 points or 58.8% on the posttest. This is equivalent to getting a letter grade of F, based on a standard US grading scale. Participants in the experimental condition (using the self-assessment chatbot) scored, on average, 20.94 points or 80.5% on the posttest. This is equivalent to getting a letter grade of B, based on a standard US grading scale. This difference was statistically significant, $t(23) = 2.14$, $p < .05$.

DISCUSSION

The results showed that Participants who used the self-assessment chatbot, scored on average, 21.7 percentage points higher on the posttest than those who used Chat GPT alone. This suggests that providing chatbot with learners' knowledge needs and directing them to use those needs when answering questions (as opposed to giving general answers) can improve learning and this effect is not limited to just math, which was the topic area of our previous studies. Future research can compare self-assessment chatbots to other LLMs and subject areas to bolster the claim that having a chatbot use self-assessment information improves learning. Moreover, implementing the knowledge assessment and transfer to the chatbot through the self-assessment paradigm allows this enhancement to be implemented in a way that involves minimal updating to chatbots (and without changing their fundamental makeup) and minimal intrusion on the user. This provides the best of both worlds.

Given that this one intervention can dramatically improve chatbot effectiveness, the question arises as to whether there are other interventions that can also improve learning. (We use the term learning because it seems that chatbots focus on delivery of information, but information is useless until it is learned.) Variables that could be examined include learner variables (e.g., learning style, age), characteristics of the material (how advanced it is, is it abstract or sensory based), type of question (is it asking for facts, procedures, reasons?) and types of answers (informational, analogical). In a previous study, we found that how one answers a question can double how well the questioner learns the material (Leddo et al., 2021).

A final area of investigation is one that appears to be universally neglected by chatbots and even search engines. Chatbots, LLMs and search engines dutifully respond to queries by providing relevant (in most cases) information. However, information is useless until it is learned. Current chatbots, LLMs and search engines do not check to see if the recipient of the information understood the information/answer that was delivered. This may not be as easy as it seems.

ISSN: 2455-8834

Volume:10, Issue:12 "December 2025"

While humans frequently as each other “Did you understand what I just said?”, research by Leddo, Clark and Clark (2021) suggests that people are not always accurate in knowing whether they did or did not understand something. In their study, both middle schoolers and adults were given an algebra lesson and then asked if they understood what they were taught. They were then given problems to solve based on the taught topic. Both middle schoolers and adults missed a third of the problems that tested what they said they understood. Interestingly, while adults proved relatively accurate in determining when they did not understand something, correctly answering only 10% of the problems that tested what they said they did not understand, middle schoolers actually correctly answered a third of the problems that tested what they said they did not understand. In cases such as these, (self)CSA might serve as a means to measure how much people understand the answers given by chatbots.

Even then, checking for understanding will not matter unless the chatbot can adjust how it answer questions to improve that understanding. A solution to this may be to create a feedback loop in which answers to questions are given, users are assessed for understanding and then feedback from the assessment is used in a machine learning program to update the effectiveness of types of answers to types of questions for types of users.

REFERENCES

Ahmad, M. and Leddo, J. (2023). The Effectiveness of Cognitive Structure Analysis in Assessing Students’ Knowledge of the Scientific Method. International Journal of Social Science and Economic Research, 8(8), 2397-2410

Anderson, J.R. (1982). Acquisition of cognitive skill. Psychological Review, 89, 369-405.

Bekkari, V., and Leddo, J., (2023). Cognitive Structure Analysis: Assessing Elementary School Students in Math to Determine the Types of Knowledge They Have. International Journal of Social Science and Economic Research, 8(10)

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educational researcher, 13(6), 4-16.

Brna, P., Ohlsson, S.and Pain, H. (1993) (Eds.). Proceeding of Artificial Intelligent in Education ‘93. Charlottesville, VA: Association for the Advancement of Computing in Education.

Challagulla, M & Leddo, J. (2025). Assessing and Remediating Knowledge Improves Problem-solving Performance in Middle School Math Compared to Remediating Problem-solving Steps. International Journal of Social Science and Economic Research, 10(8), 3618-3629.

ISSN: 2455-8834

Volume:10, Issue:12 "December 2025"

Cynkin, C. and Leddo, J. (2023). Teaching Students to Self-Assess Using Cognitive Structure Analysis: Helping Students Determine What They Do and Do Not Know. International Journal of Social Science and Economic Research, 8(9), 3009-3020.

Dandemraju, A., Dandemraju, R. and Leddo, J. (2024). Teaching students to self-assess their own chemistry knowledge. International Journal of Social Science and Economic Research, 9(2), 541-549.

De Kleer, J. and Brown, J.S. (1981). Mental models of physical mechanisms and their acquisition. In J.R. Anderson (Ed.), Cognitive Skills and their acquisition. Hillsdale, NJ: Erlbaum

Greer, J. (1995) (Ed.) Proceeding of Artificial Intelligent in Education '95. Charlottesville, VA: Association for the Advancement of Computing in Education.

Hasanein, M., & Sobaih, A. (2023). Conversational AI chatbots for health-related support: User engagement, accuracy, and satisfaction. *European Journal of Investigation in Health, Psychology and Education.*

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F. and Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and individual differences, 103, 102274.

Koedinger, K. R., & Corbett, A. (2006). Cognitive Tutors: Technology Bringing Learning Sciences to the Classroom. In R. K. Sawyer (Ed.), *The Cambridge handbook of: The learning sciences* (pp. 61–77). Cambridge University Press.

Konadu, E., & Kusi, A. (2025). Effects of conversational mental-health chatbots on stress, confidence, and engagement among high-school and college students. *American Journal of Education and Learning.*

Leddo, J., Chen, T., Menachery, A., Agarwal, J. and Agarwal, T. (2021). Towards Improving Personal Assistants and Educational Software: How Questions Are Answered Affects Learning. International Journal of Social Science and Economic Research, 6(2), 696-705.

Leddo, J. Clark, A. and Clark, E. (2021). Self-Assessment of Understanding: We Don't Always Know What We Know. International Journal of Social Science and Economic Research, 6(6), 1717-1725.

ISSN: 2455-8834

Volume:10, Issue:12 "December 2025"

Leddo, J. Clark, E. and Clark, A. (2025). Using self-assessment and remediation to raise middle school student achievement in math. International Journal of Social Science and Economic Research, 10(3), 1083-1092.

Leddo, J., Cohen, M.S., O'Connor, M.F., Bresnick, T.A., and Marvin, F.F. (1990). Integrated knowledge elicitation and representation framework (Technical Report 90-3). Reston, VA: Decision Science Consortium, Inc..

Leddo, J., Li, S. and Zhang, Y. (2022). Cognitive Structure Analysis: A technique for assessing what students know, not just how they perform. International Journal of Social Science and Economic Research, 7(11), 3716-3726.

Leddo, J. and Sak, S. (1994). Knowledge Assessment: Diagnosing what students really know. Presented at Society for Technology and Teacher Education.

Leddo, J., Zhang, Z. and Pokorny, R. (1998).Automated Performance Assessment Tools. Proceedings of the Interservice/Industry Training Systems and Education Conference. Arlington, VA: National Training Systems Association.

Liang, I. and Leddo, J. (2020). An intelligent tutoring system-style assessment software that diagnoses the underlying causes of students' mathematical mistakes. International Journal of Advanced Educational Research, 5(5), 26-30.

Létourneau, A., Deslandes Martineau, M., Charland, P., Karran, J. A., Boasen, J., and Léger, P. M. (2025). A systematic review of AI-driven intelligent tutoring systems (ITS) in K-12 education. Science of Learning, 10(1), 29.

Maviti, A. and Leddo, J. (2025). A Self-assessment Chatbot Greatly Outperforms Chat GPT in Teaching High School Students Calculus. International Journal of Social Science and Economic Research, 10(10), 5519-5531.

Maviti, A., Leddo, J. and Prakash, P. (2025). A Self-assessment Chatbot Greatly Outperforms Gemini in Teaching High School Students Calculus. International Journal of Social Science and Economic Research, 10(11), 6073-6085.

Nehra, P. and Leddo, J. (2024). The effects of Cognitive Structure Analysis in self-assessing and remediating knowledge gaps in introductory Spanish. International Journal of Social Science and Economic Research, 9(12), 5956-5964.

Newell, A. and Simon, H.A. (1972). Human problem solving. Englewood Cliffs, NJ: Prentice

ISSN: 2455-8834

Volume:10, Issue:12 "December 2025"

Prakash, P. and Leddo, J. (2025a). Using Self-Assessment and Remediation to Raise Student Achievement in Reading Comprehension. International Journal of Social Science and Economic Research, 10(1), 277-286.

Prakash, P. and Leddo, J. (2025b). Using Self-Assessment and Remediation to Raise Student Achievement in Mathematics. International Journal of Social Science and Economic Research, 10(1), 447-456.

Prakash, P. and Leddo, J. (2025c). Using Self-Assessment and Remediation to Raise Student Achievement in History. International Journal of Social Science and Economic Research, 10(3), 650-659.

Prakash, P. and Leddo, J. (2025d). Using Self-Assessment and Remediation to Raise Middle School Student Achievement in Reading Comprehension. International Journal of Social Science and Economic Research, 10(3), 1130-1140.

Prakash, P. and Leddo, J. (2025e). Using Self-Assessment and Remediation to Raise Middle School Student Achievement in Science. International Journal of Social Science and Economic Research, 10(4), 1471-1482.

Quillian, M.R. (1966). Semantic memory. Cambridge, MA: Bolt, Beranek and Newman.

Sathiyamoorthy, S.S. and Leddo, J. (2025). Using Self-Assessment and Remediation to Raise College Student Achievement in Psychology. International Journal of Social Science and Economic Research, 10(5), 1859-1869.

Schank, R.C. (1982). Dynamic Memory: A theory of learning in computers and people. New York: Cambridge University Press.

Schank, R.C. and Abelson, R.P. (1977). Scripts, Plans, Goals, and Understanding. Hillsdale, NJ: Erlbaum.

Van Lehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational psychologist, 46(4), 197-221.

Villegas-Ch, W., Buenano-Fernandez, D., Navarro, A. M., and Mera-Navarrete, A. (2025). Adaptive intelligent tutoring systems for STEM education: analysis of the learning impact and effectiveness of personalized feedback. Smart Learning Environments, 12(1), 1-31.

Wang, Y. and Leddo, J. (2025). Improving the Effectiveness of Chatbots by Incorporating Self-Assessed User Knowledge into the Question-Answering Process. International Journal of Social Science and Economic Research, 10(8), 3574-3586.

ISSN: 2455-8834

Volume:10, Issue:12 "December 2025"

Zawacki-Richter, O., Marín, V. I., Bond, M., and Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators?. International journal of educational technology in higher education, 16(1), 1-27.

Zhou, L.N. and Leddo, J. (2023). Cognitive Structure Analysis: Assessing Students' Knowledge of Precalculus. International Journal of Social Science and Economic Research, 8(9), 2826-2836.

