ACQUISITION INNOVATION RESEARCH CENTER

Cognitive Assistant for Training Cost Estimators

EXECUTIVE SUMMARY AND REPORT

JULY 2024

PRINCIPAL INVESTIGATOR

Daniel Selva, *Texas A&M University*

CO-PRINCIPAL INVESTIGATOR

Theodora Chaspari, *Texas A&M University Affiliated*

Alejandro Salado, *The University of Arizona*

SPONSOR

Ms. Jennifer Bowles, SES, Director, Land and Naval Warfare Cost Analysis Division, Office of the Secretary of Defense (OSD), Cost Assessment and Program Evaluation (CAPE)

Mr. Bryan Coots, Operations Research Analyst, OSD CAPE

DISCLAIMER

The Acquisition Innovation Research Center (AIRC) is a multi-university partnership led and managed by the Stevens Institute of Technology and sponsored by the U.S. Department of Defense (DoD) through the Systems Engineering Research Center (SERC)—a DoD University-Affiliated Research Center (UARC).

This material is based upon work supported, in whole or in part, by the U.S. Department of Defense through the Office of the Under Secretary of Defense for Acquisition and Sustainment (OUSD(A&S)) and the Office of the Under Secretary of Defense for Research and Engineering (OUSD(R&E)) under Contract HQ0034-19-D-0003, TO#0285.

The views, findings, conclusions, and recommendations expressed in this material are solely those of the authors and do not necessarily reflect the views or positions of the United States Government (including the Department of Defense (DoD) and any government personnel), the Stevens Institute of Technology, Texas A&M University, or the University of Arizona.

No Warranty.

This Material is furnished on an "as-is" basis. The Stevens Institute of Technology, Texas A&M University, and the University of Arizona make no warranties of any kind—either expressed or implied—as to any matter, including (but not limited to) warranty of fitness for purpose or merchantability, exclusivity, or results obtained from use of the material.

The Stevens Institute of Technology Texas A&M University, and the University of Arizona do not make any warranty of any kind with respect to freedom from patent, trademark, or copyright infringement.

SYSTEMS ENGINEERING RESEARCH CENTER

AIRC ACQUISITION INNOVATION RESEARCH CENTER

STEVENS INSTITUTE OF TECHNOLOGY 1870

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

RESEARCH TEAM

Name	Organization	Labor Category
Daniel Selva	Texas A&M University	Principal Investigator (PI)
Theodora Chaspari	*Texas A&M University Affiliated	Co-Principal Investigator
Alejandro Salado	University of Arizona	Co-Principal Investigator
Gabriel Apaza	Texas A&M University	Graduate Research Assistant (PhD)
Aman Tutul	Texas A&M University	Graduate Research Assistant (PhD)
Joanna Joseph	University of Arizona	Graduate Research Assistant (PhD)
Abhishek Maurya	Texas A&M University	Graduate Research Assistant (MS)

ACKNOWLEDGEMENTS

In addition to the sponsors, Ms. Jennifer Bowles and Mr. Bryan Coots, the team would like to thank Mr. Greg Hogan from Air Force Cost Analysis Agency (AFCAA) for his useful feedback on the work and all the personnel from CAPE and AFCAA who volunteered their time to test the tool.

ACRONYMS AND ABBREVIATIONS

3PL	3-Parameter Logistic
AFCAA	Air Force Cost Analysis Agency
AI	Artificial Intelligence
AWS	Amazon Web Services
CA	Cognitive Assistant
CAPE	Cost Assessment and Program Evaluation
DoD	Department of Defense
GPT	Generative Pre-trained Transformer
IRB	Institutional Review Board
ITS	Intelligent Tutoring System
LLM	Large Language Model
ML	Machine Learning
NASA	National Aeronautics and Space Administration
OHRO	Office of Human Research Oversight
PL	Parameter Logistic
QA	Question Answering
RQ	Research Question
TAMU	Texas A&M University
TLX	Task Load Index

EXECUTIVE SUMMARY

The research team has developed a cognitive assistant to support the training of cost estimators in the Department of Defense (DoD). A Cognitive Assistant (CA) is defined here as an Artificial Intelligence (AI) tool, usually with a natural language interface, that augments human intellect in a specific task by retrieving and processing relevant information from multiple information sources and providing it to the user at the right time. It also has the capability to learn and adapt to the user and the problem at hand.

Cost estimation is a complex iterative process consisting of various steps: gathering the required information, selecting an overall strategy and one or more existing models, developing new models if needed (including calibration and validation), performing the estimate, and conducting sensitivity analyses as appropriate. There are challenges for beginner cost estimators in each of those steps, including dealing with incomplete datasets, appropriately assessing the performance of new models, projecting beyond historical ranges of validity, adequately reporting the level of uncertainty around a point estimate, understanding how to use joint cost-schedule distributions, etc. More experienced cost estimators may also struggle to learn the specifics of a new commodity (e.g., the cost drivers, relative orders of magnitude, etc.).

Currently, the training of new cost estimators is done primarily through traditional instruction in live classrooms, and thus it is a time-consuming process. Traditional instruction typically implies reduced opportunities for hands-on learning opportunities, which are known to improve learning. This type of instruction is also not tailored to each individual, so the pace can be too fast for some trainees and too slow for others. The use of CAs can allow for more interactive and tailored instruction for each individual and area, as demonstrated with intelligent tutoring systems in other areas of education (Corbett et al., 1997).

The idea of using AI tools to enhance the learning of trainees is not new and has been studied for decades (Ong & Ramachandran, 2003). However, in the DoD Acquisition context, we are still in the early stages of incorporating advanced AI tools into workflows and, in particular, CAs have not yet been adopted as training tools. Previous attempts to adopt this technology in the workplace failed because of a combination of insufficient performance of the underlying machine learning (ML) models and lack of familiarity of the users with this mode of interaction. With CAs now being ubiquitous in our daily lives and the significant recent advances we have seen in machine learning, the time is now ripe for infusion of this technology in the workplace.

In an Incubate Phase I of this project ($100k, Sep 2021—Jun 2022), the research team worked with the sponsor and other stakeholders to define the use case for the CA. We decided to focus on a user that is already familiar with cost estimation methods, but wants to learn a new commodity, namely space systems. The tool was to help the user learn the new material in an individualized way. We developed an initial version of the CA based on an existing agent developed by the team called Daphne. This allowed us to make fast progress as some of the software infrastructure was reused. In addition, we demonstrated the ability to do individualized training in the context of selecting questions for the various learning assessments and learning opportunities that best address the user's needs (e.g., reinforcing weaker areas). An initial estimate of the resources that would be needed to develop and maintain such a tool in the DoD was provided.

The project was approved for a Phase II with the goal of further developing the agent and validating it with real users. A previous report described the activities conducted during the first half of the Phase II effort ($134k, Sep 2022—Sep 2023). In that first year of the Phase II effort, we developed a second version of the agent that leverages Large Language Models (LLMs) to make the system more flexible, extensible, and easier to maintain. In addition, we developed a 3-module online course on Space Systems with slides, example questions, and quizzes and we started thorough testing of the effectiveness of the tool at Texas A&M University (TAMU).

This report documents the second half of the Phase II effort ($120k, Oct 2023—Jul 2024). During these 8 months, we have refined the software tool, added new instructional materials, and validated the effectiveness of the tool in the lab with 51 student subjects and with 22 real users at the Office of Cost Assessment and Program Evaluation (CAPE) and the Air Force Cost Analysis Agency (AFCAA). Results from the lab experiment showed that students who used the cognitive assistant scored 6 points higher in the test than students without access to the assistant, suggesting that this technology is promising to improve the efficiency and effectiveness of workforce training. This improvement is likely due at least in part to the assistant simply increasing the time the student is interacting with the material. The software tool and accompanying documentation including a transition plan document have been delivered to the sponsor for transition into production.

BACKGROUND

There is relevant literature in the use of cognitive assistants (CAs) and other intelligent agents for educational purposes. Much of this literature is contained under the umbrella term of intelligent tutoring systems (ITS). ITS are intelligent systems that help students master a subject by providing them with learning opportunities that are tailored to their specific needs.

Following the success of expert systems and other kinds of intelligent decision support systems in the 1980s, ITS were proposed as a method that could radically improve student outcomes in education by providing unprecedented ability to adapt to individual differences (Corbett et al.)

Key to this adaptation was the ability of these systems to estimate the skill level of a student for a number of areas based on the student performance in some learning opportunities provided by the system, using Bayesian algorithms among others (Mayo). These skill levels could then be used to select the next learning opportunity to provide the student given some goal, such as to reinforce the weaker areas. Theoretical frameworks and algorithms were developed and successfully deployed based on Partially Observable Markov Decision Processes (Folsom-Kovarik et al.) among others.

Educators were especially excited about the potential of this technology to democratize education and improve student outcomes for populations that needed it the most (Benjamin D. Nye). They were deployed in various educational centers with some success (Koedinger et al.). Specifically, it was observed that using ITS, student learning outcomes and student engagement could be improved (Kim et al.).

The basic rigid systems developed in the 1990s evolved into more advanced systems including mixed-initiative interfaces with question answering systems (A.C. Graesser et al.) and affective computing technologies (D'Mello, Craig, Gholson, et al.; D'Mello, Craig, Witherspoon, et al.).

While the initial emphasis of ITS was on K-12 education, the technology has also been applied to adult education (Cheung et al.) and training in the workplace (Ong and Ramachandran). In the latter case, it was found that using ITS could improve training performance and return on investment.

While the potential of these technologies is important, some barriers have also been identified for their implementation and widespread adoption. These include limitations in their performance (Sarrafzadeh et al.) and high development and maintenance costs among others (Benjamin D Nye).

Finally, Large Language Models (LLMs) (Zhao et al.) have recently emerged as a game-changing technology in artificial intelligence (AI) due to their abilities to reason (Kojima et al.) and learn without providing examples (Wei et al.). Beyond the widespread interest this technology generated, it has tremendous potential to improve question answering (QA) both in open and restricted domains (Singhal et al.). Its potential to improve (and also hinder) education has also been identified and discussed (Kasneci et al.).

DESIGN OF THE COGNITIVE ASSISTANT

USE CASE

Different use cases were identified in Phase I for the CA that concerned different types of users (e.g., novice vs expert in either cost estimation in general or a particular commodity) and specific tasks (e.g., going through instructional materials, example cost estimation tasks.) Specific example use cases included a standard **tutor** agent to support the user while going through instructional materials, an **assistant** agent to support the user while performing example cost estimation tasks, and a **trade space exploration** agent to support the user while exploring a space of alternative designs with different levels of performance and cost.

In Phase I, we narrowed the use case down to a tutor agent to support users who are already familiar with the basics of cost estimation but are learning a new commodity. Space systems was chosen as the commodity since it is aligned with the background and expertise of the team. The specific use case was slightly refined in Phase II and is provided in Figure 1.

Tutor agent for an expert cost estimator learning a new commodity (space)

The name of the CA is Daphne Academy. Daphne Academy looks and feels like an AI assistant combined with an online learning tool like Coursera. The vision is similar to that of Intelligent Tutoring Systems: to provide personalized training.

The intended user has some experience with cost estimation, but is not familiar with space systems and needs to learn fundamentals of such systems so they can understand cost drivers and how to apply standard cost estimation methods.

Daphne Academy has a web-based front end that allows the user to navigate a number of learning modules on various aspects of space systems (space mission payloads, architectures, space environment, orbits…). The learning modules contain traditional slides but also other things like videos, short exercises, etc. After each learning module, the user completes a short quiz to assess their progress. As the user is going through the learning modules or doing the practice quizzes, they can ask any questions to the agent in natural language through a chatbox, such as "*What is Delta-V?*"

The user can also take longer practice tests that serve the dual purpose of a learning opportunity for the user and helping the CA estimate the user's skill level across various areas. During these tests, the CA selects the next question to show to the user based on those skill levels, to maximize some objective such as reinforcing weaker areas.

Finally, Daphne Academy can also perform more formal learning assessments (tests) where the user does not have access to the CA and the goal is to assess the learning outcomes and determine if the trainee has mastered the material. At any point, the user can access a window where they can see their progress in their assigned learning modules, their estimated skill level in each area, and final grade on each module.

Figure 1. Use Case for the Cognitive Assistant

SOFTWARE ARCHITECTURE

The software architecture of the agent as implemented is shown in Figure 2. It is implemented as a web application hosted in Amazon Web Services (AWS). Early in the project, the repercussions of using AWS were discussed and it was decided that it was acceptable to use. However, how to implement this in an in-house cloud environment is explained in the transition plan document.

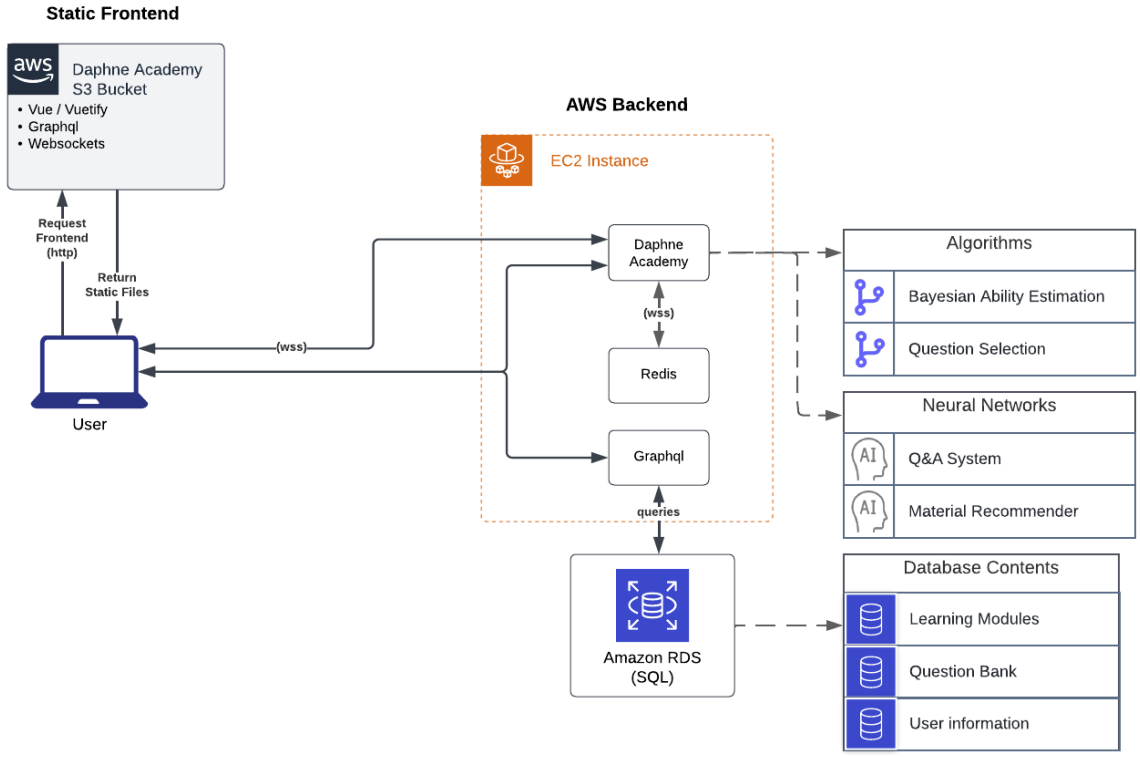


Figure 2: Software Architecture

QUESTION ANSWERING (QA) SYSTEM

The QA System developed in Phase I was a standard template-based restricted domain question answering system. First, a model based on convolutional neural networks attempted to classify the user question into one of N known types of questions. Then, parameter extraction was performed. Then, a query to the databases was generated based on the question type and extracted parameters. Finally, the answer was inserted in an answer template and returned to the user. This approach worked well but was not very scalable since answer templates need to be provided for each type of question.

The new system developed in Phase II leverages the Generative Pre-trained Transformer 4 (GPT-4) LLM to provide template-free QA capabilities. To make sure the LLM answers the question based on the content provided in the learning modules as opposed to general knowledge of the agent, the learning modules are provided as context to the LLM. The architecture of the Phase II QA system is shown in Figure 3.

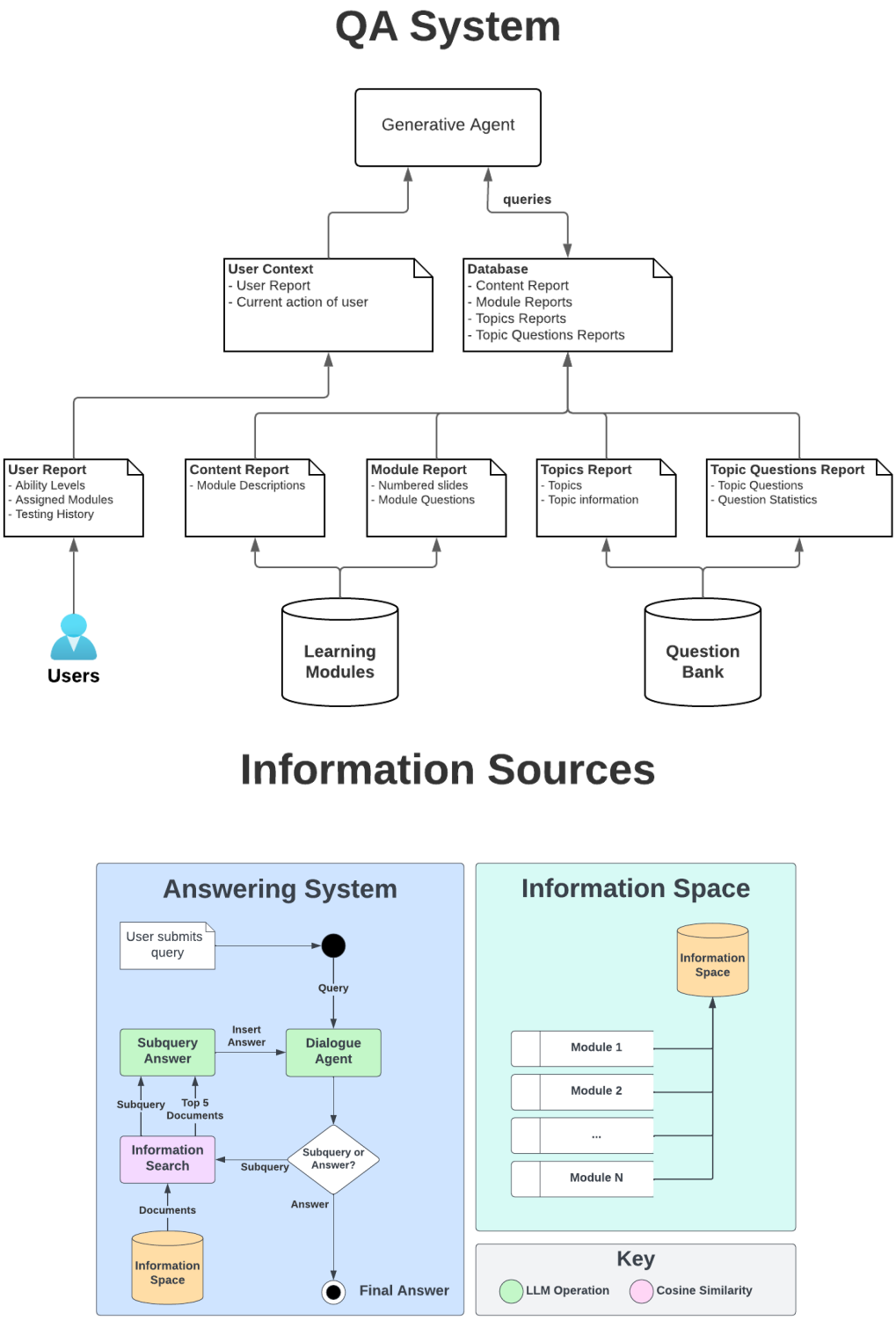


Figure 3: Architecture of the new Question Answering System leveraging a generative pre-trained transformer model (GPT-4)

The QA system utilizes GPT4 with the backend database to fuse data sources that encompass information about a given user (e.g., assigned learning modules and current ability estimates) including the material content they have been assigned (e.g., textual content from learning module slides). As this informational content is typically too large to provide to GPT4 in a single prompt, a dynamic inner-dialogue system is utilized to automate the retrieval of relevant information from the backend database. This dialogue systems consist of a dialogue agent (for determining and querying the necessary information to answer a question) and an answering agent (for answering queries posed to it from the dialogue agent). These two agents work together to answer a user query, while also having the ability to cite the source of the information from which it synthesized an answer (e.g., citing a specific slide in a learning module). Cosine similarity is used to compute document relevance with the query and retrieve the top N most relevant items from the relevant modules.

The resulting QA system has performed very well in testing, and it completely eliminates the need to create individual answer templates for different questions, thus substantially increasing the flexibility and scalability of the system.

LEARNING MODULES AND DATABASES

The main database contains all the information from the learning modules and the question bank for the tests. The schema of the database is provided in Figure 4 below. While the schema is largely unchanged from Phase I, the content of the database is completely different since it now contains the new learning modules.

All the reports used for question answering are automatically generated from the database when a user profile is created. They are also regenerated anytime there is a relevant change (e.g., a new learning module is assigned to a user, the skill level of the user is updated, etc.). The database is also directly queried by the other elements of the program that are not the QA system, such as when the user navigates to the window to see their current skill levels.

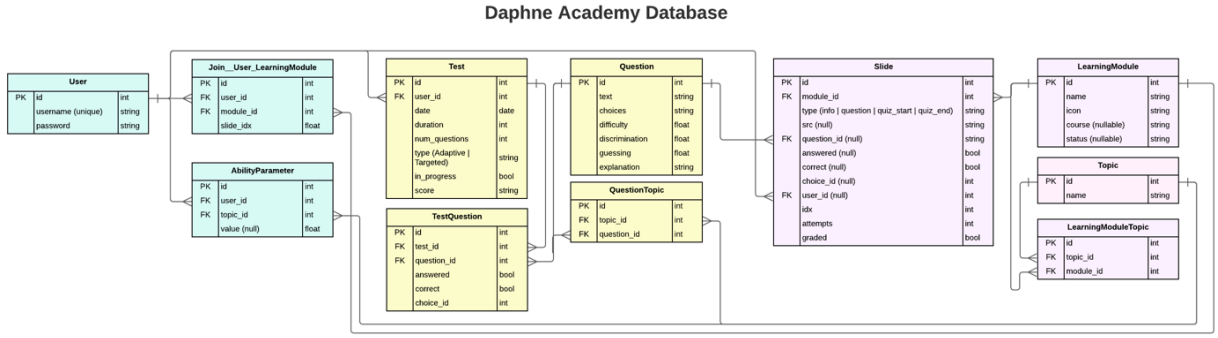


Figure 4: Schema of the Daphne Academy Database

The learning modules developed in Phase II are as follows:

1. Introduction to Space (3 topics, 30 slides, 54 questions)
2. Space environment and orbits (2 topics, 50 slides, 60 questions)
3. Spacecraft subsystems (8 topics, 129 slides, 80 questions)
4. Space-based remote sensing payloads (4 topics, 100 slides, 60 questions)

ADAPTIVE QUESTION SELECTION

An important feature of the agent is its ability to provide individualized training that adapts to the needs of an individual user. To do this, the agent estimates the skill level of the user across a number of areas and uses those estimates to select the questions or learning opportunities that are more likely to benefit the user, e.g., reinforcing their weaker areas.

The skill estimation algorithm is based on performing maximum a posteriori estimation of the skill parameter of the user assuming the probability that the user will answer a question correctly follows a 3-parameter logistic (3PL) model. Specifically, the model is as follows:

$$p_j(\theta) = c_j + (1 - c_j)\frac{\exp\{a_j(\theta - b_j)\}}{1 + \exp\{a_j(\theta - b_j)\}}$$

Where $p_j(\theta)$ is the probability that the user will answer the jth question correctly, θ is the skill level of the user, and are a_j, b_j, c_j question-specific parameters, namely the discrimination parameter, the difficulty parameter, and the guessing parameter for the jth question.

In Phase I, $\theta \in [0,1]$ was estimated from the user's responses ($u_j = 0,1$ if incorrect/correct) to a sequence of K questions using maximum a posteriori estimation assuming the question-specific parameters were known:

$$\theta^* = \arg\max_\theta p(\theta|u) = \arg\max_\theta \left[\prod_{j=1}^{K} p_j(\theta)^{u_j}\left(1 - p_j(\theta)\right)^{1-u_j}\right] p(\theta)$$

However, setting a value for the question-specific parameters (mostly a and b) a priori is hard. Therefore, in Phase II, we have extended the model to be able to estimate those question-specific parameters jointly with the skill level. This requires answers from a set of users as opposed to a single user. The method to do this is illustrated in Figure 5.

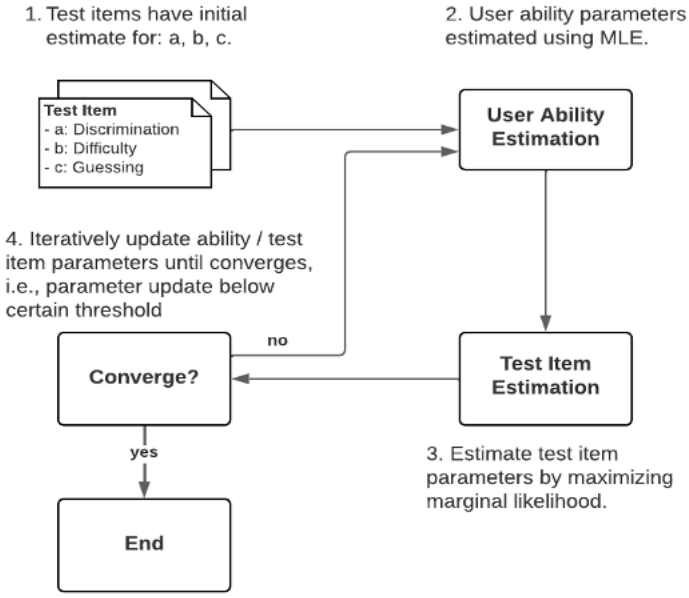


Figure 5: Method to jointly estimate user skill level and question-specific parameters

Using this approach, the agent can select questions based on which ones are more likely to help the user learn. The idea is that the next question a user sees in an exam can be optimally selected on the basis of some objective function which could relate to reinforcing the user's weaker areas, or simply obtaining as accurate an estimate as possible of the user's skill levels. Note that there can be multiple conflicting objectives.

The question selection algorithm is unchanged from Phase I. It selects the next question in two steps: 1) determine the user's weakest topic area, and 2) select a question from the determined topic area to maximize information gained with respect to the user's ability level. Step one is completed by comparing the user's ability levels across topics, selecting the topic with the lowest corresponding ability level, and querying the Question Bank database for questions in this topic. Step two is completed by calculating the information gained for each question using the Item Information Function and selecting the question j^* that produces the highest item information value:

$$j^* = \arg\max_j I_j(\theta) = \arg\max_j \left[a_j^2 \frac{1 - p_j(\theta)}{p_j(\theta)} \right] \left[\frac{\left(p_j(\theta) - c_j\right)^2}{\left(1 - c_j\right)^2} \right]$$

Other question selection strategies will be explored in future research.

FRONT END

The following figures (Figures 6-9) contain snapshots of the updated front end of Daphne Academy. Figure 6 shows a slide of the space environment learning module. The user asks the agent to explain the slide for them – a question that has never been seen by the agent before. The agent provides a reasonable answer, showcasing the summarizing abilities of the agent.

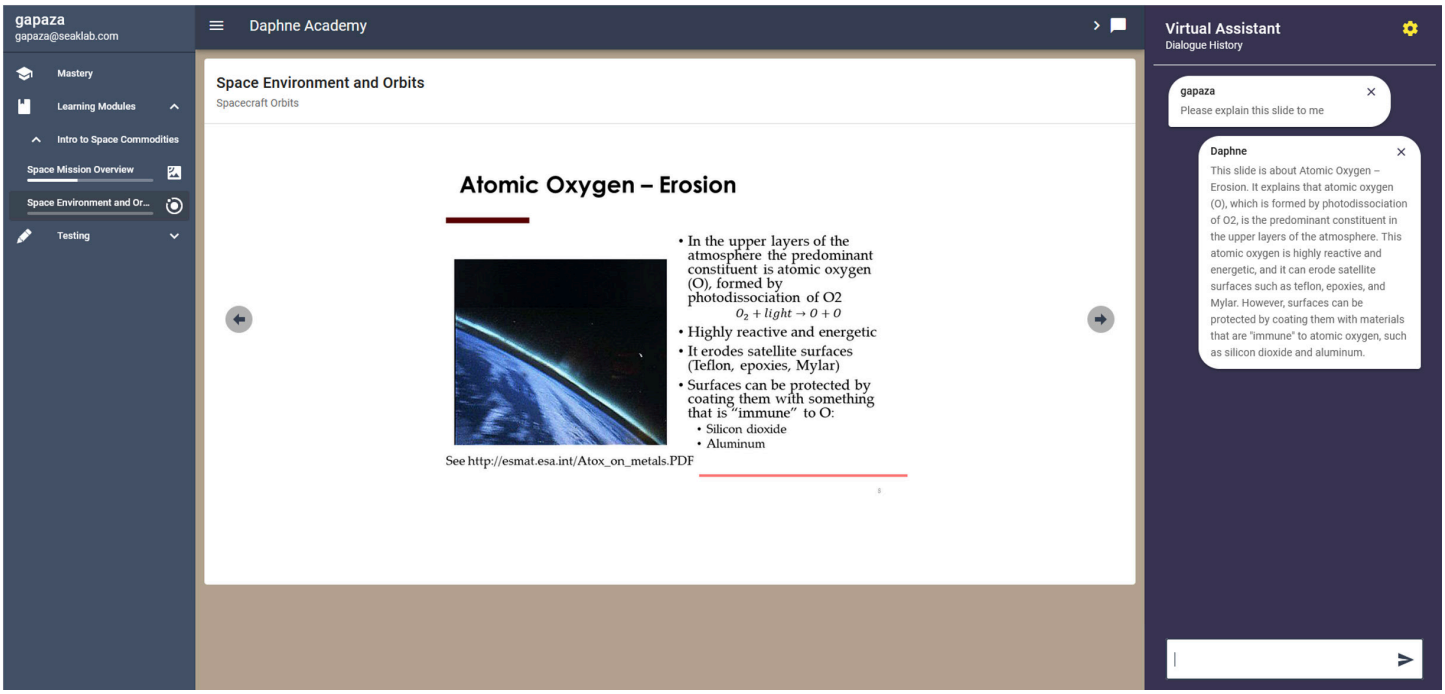


Figure 6: Snapshot of Daphne Academy's front end - the user asks the agent to summarize a slide

Figure 7 shows another snapshot of Daphne Academy in which the user asks the agent to explain “the equation in this slide”. The agent infers the equation the user is referring to and proceeds to explain the equation.

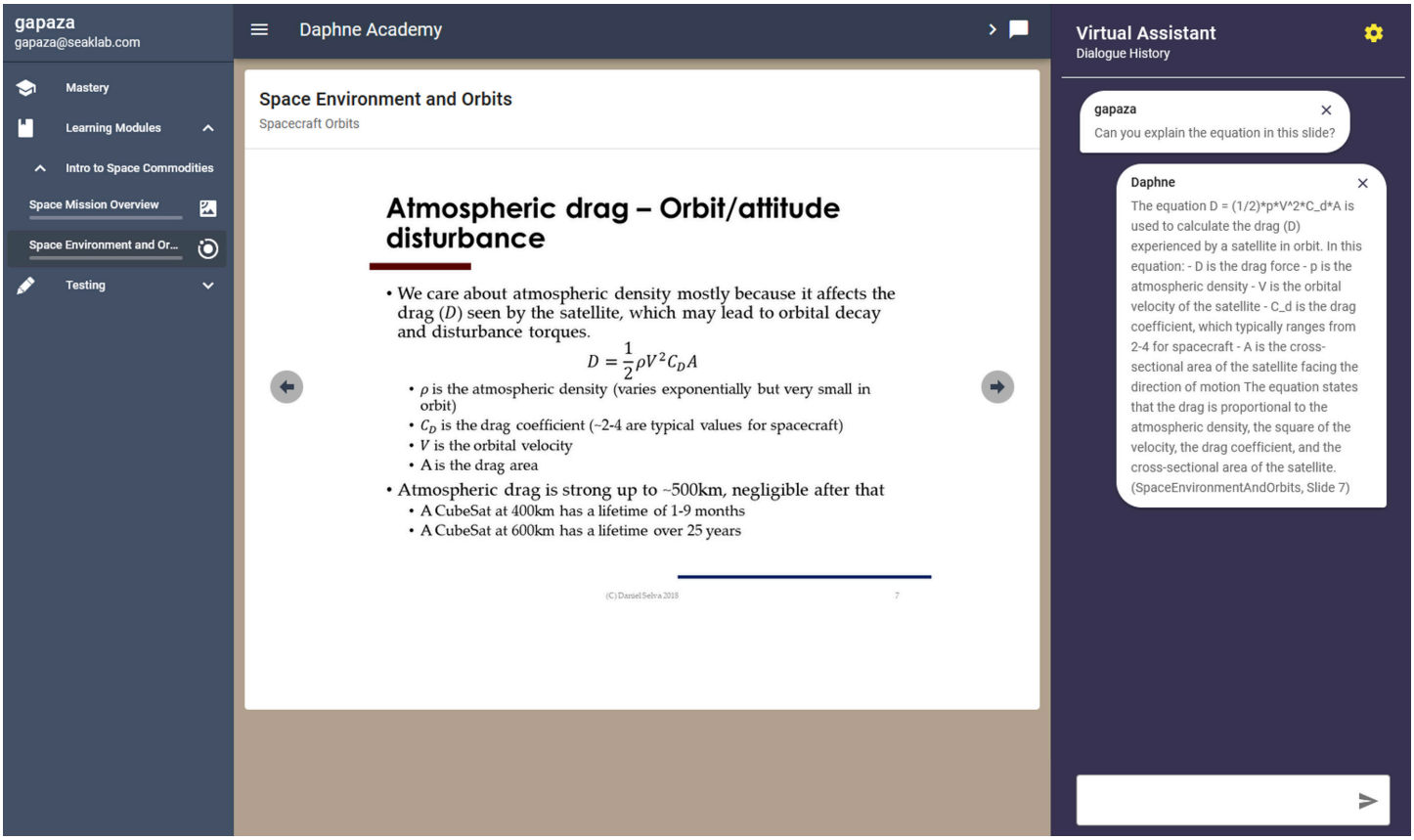


Figure 7: Snapshot of Daphne Academy’s front end - the agent explains an equation to the user

Figure 8 shows another snapshot in which the user asks an open-ended question not directly related to the current slide. The agent provides an answer and a link to the relevant slide.

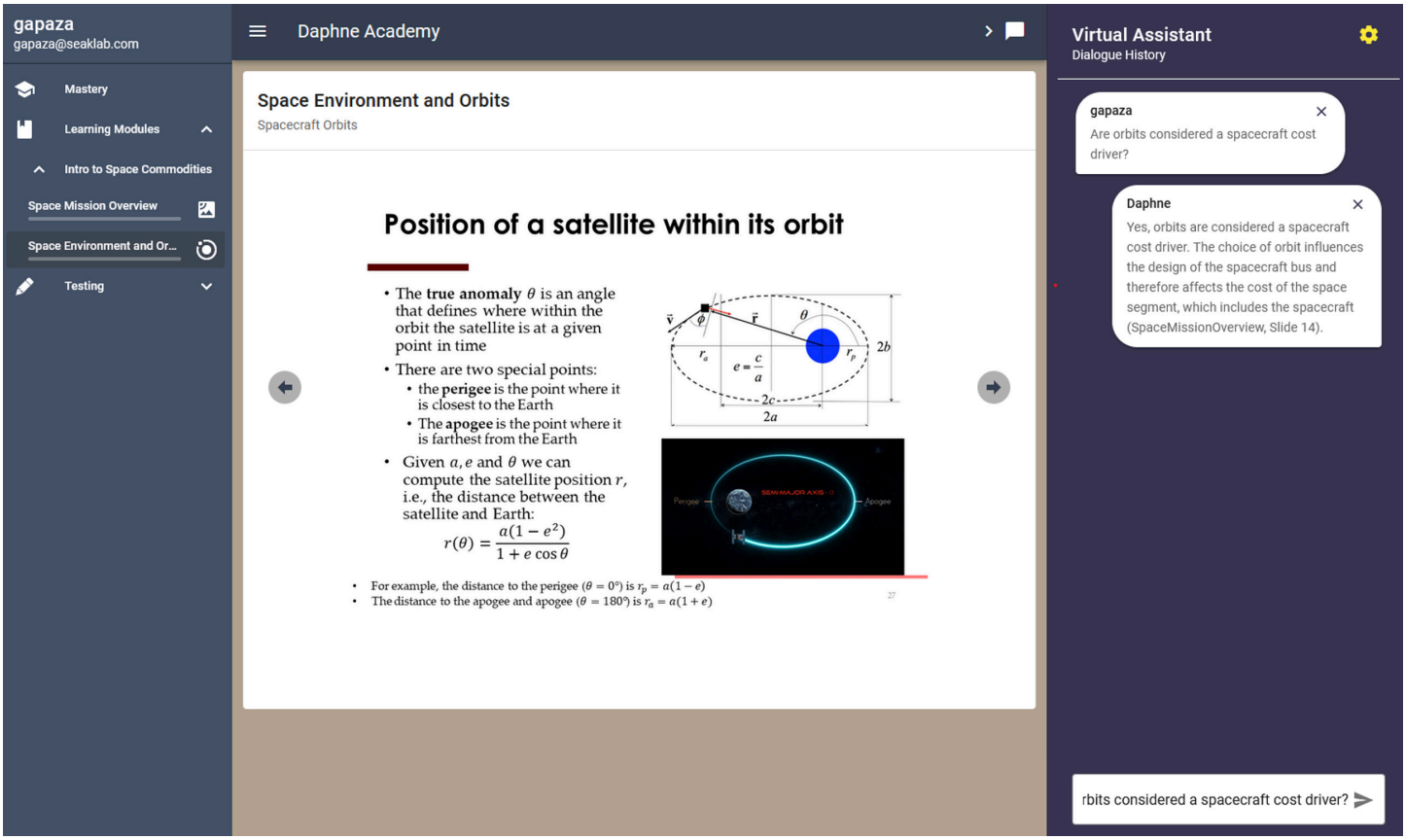


Figure 8: Snapshot of Daphne's Academy front end - the user asks an open-ended question. The agent provides an answer and a link to the relevant slide

Figure 9 shows an example of a user attempting to use the agent during a test. The agent detects that the user is trying to cheat and declines to provide an answer.

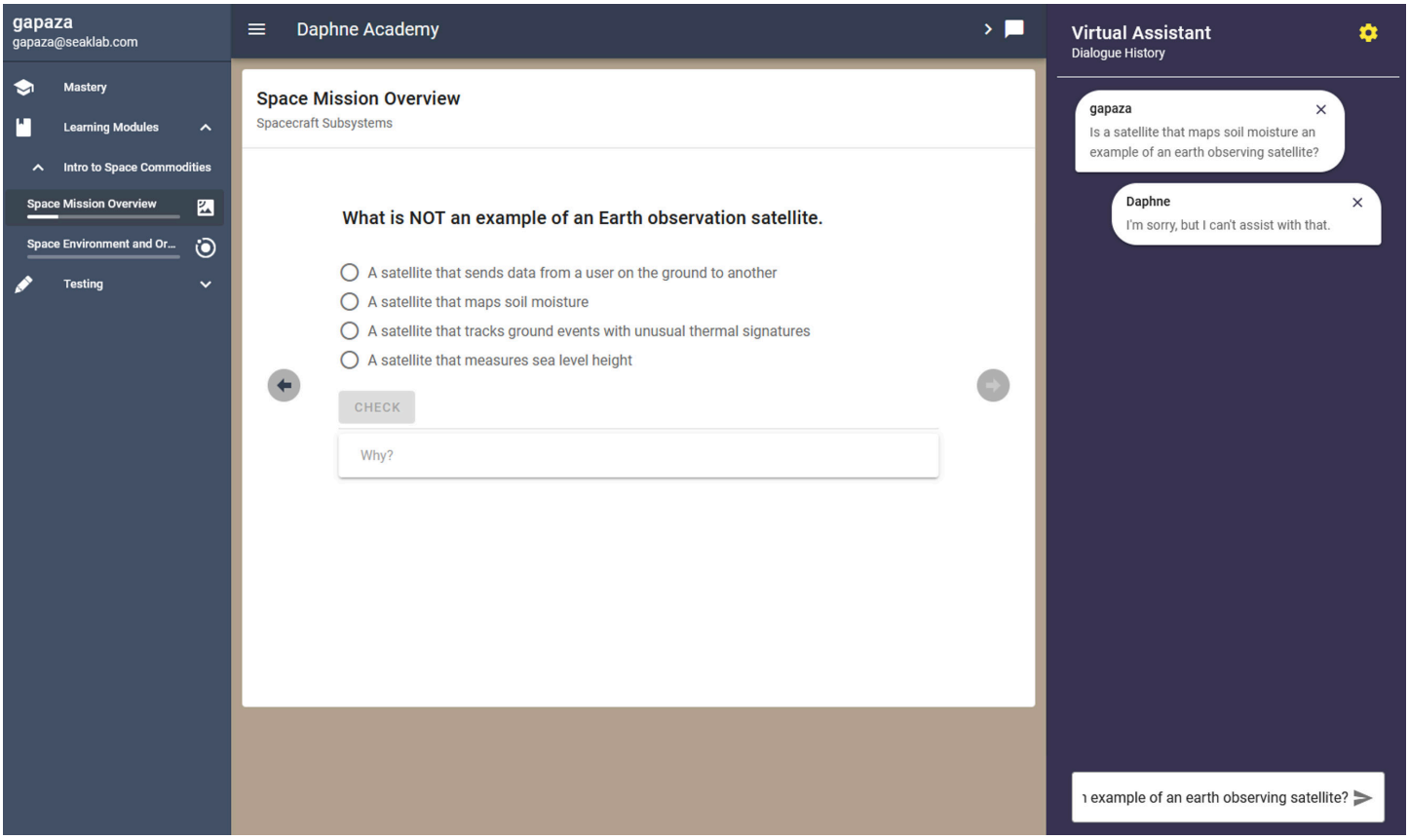


Figure 9: Snapshot of Daphne's Academy front end – the agent declines to answer a user question during a test

VALIDATION OF THE COGNITIVE ASSISTANT

OVERVIEW

An important part of the research plan was to validate the effectiveness of the CA. In Phase I, the research team got some initial feedback from the sponsors and potential users. In Phase II, we performed an exhaustive validation effort, with two different communities of users:

1. A small number of highly representative users (actual cost estimators from Cost Assessment and Program Evaluation (CAPE) and Air Force Cost Analysis Agency (AFCAA)) using the tool with realistic learning modules over a period of 1-2 weeks; and
2. A larger number of less representative users (students from Texas A&M University (TAMU)) using the tool with simplified learning modules over a period of a few hours.

While the Department of Defense (DoD) deployment has more external validity than the lab experiment thanks to the more realistic users and environment, the lab experiment has more internal validity thanks to the more controlled environment. The lab experiment allowed us to obtain additional data on user background and measurements of user cognitive state and to perform a rigorous statistical analysis whereas the data analysis for the CAPE/AFCAA users was largely exploratory since the dataset was not complete (e.g., many users only partially completed the tasks).

LAB EXPERIMENT

Protocol: The protocol for the lab experiment is provided in Table 1. Initially, each participant completed a set of questionnaires that included the Big Five Inventory (John and Srivastava, 1999) recording personality traits (i.e., extraversion, agreeableness, conscientiousness, openness, neuroticism), the Trusting Intention to AI survey defined as the level at which individuals feel they can rely on AI (McKnight et al., 2002). Out of 4 questions in the original Intention to Trust survey, we included two questions that suit the best to our particular use case for AI Chatbot. They also answered the question “How long have you been using conversational agents (e.g., ChatGPT, Google Home, Alexa, Siri)?” The response to this question was used as a measure of users’ past experience with Chatbot. Besides, they also answered the perceived ease of use of chatbot questionnaire on ChatGPT (Davis, 1989).

After finishing the pre-survey questionnaires, a research assistant calibrated the Tobii Eye-Tracker X5. Then, the users spent a maximum of 30 minutes to go through the first learning content, then they solved 5 quiz questions, and later solved 16 test questions. While answering the quiz questions, users had the option to review the slides and use the chatbot in the with chatbot condition, but during test questions they were not allowed to review the slides or use the chatbot in either condition. After the test of the first module, the users worked through a second learning module for a maximum of 30 minutes, and then answered 5 quiz questions and 16 test questions afterwards. The protocol is exactly the same for the second module. The difficulty level of the 16 test questions for both learning modules is equivalent. All the questions were multiple choice questions having a variable number of options and the difficulty level of the question was adjusted based on the number of options they have.

In one of the two modules, the users had access to our chatbot where users had the option to ask any question to the chatbot while reading the slides. However, the users did not have access to this chatbot while answering the test questions. The order of the modules, and the order of AI assistance were both randomized, therefore, we expect there should be no bias from the ordering of modules or the difficulty of the learning content of the modules to the user performance.

After going through the learning content and answering the questions for each module, the users were asked to complete a post-module questionnaire. This questionnaire includes the National Aeronautics and Space Administration (NASA) Task Load Index (TLX) survey (Hart et al., 1988) and some additional items on engagement, usability, and trust in automation. The NASA TLX survey captures user workload across different dimensions (e.g., mental demand, physical demand, temporal demand, effort, performance, and frustration level). The users also answered a question 'How confident do you feel about your understanding of the material you have learned?' that captures their self-confidence in their understanding of the learning content. In addition, for those modules for which users had access to our chatbot, they answered this question 'To what extent did you use AI chatbot in this module?' which captures how much they used the chatbot for understanding the learning contents. Apart from that, they answered three questions related to the usability of the chatbot out of six questions from the usability questionnaire (Davis, 1989), one question related to the trust in AI, and one question related to their engagement with the chatbot.

Table 1: Protocol for the lab experiment at TAMU

Page	Scheduled Activity		Time	Cumulative Time
--	1	Informed Consent and explain the experiment	15 min	15 min
--	2	Tobii Eye-Tracker X5 setup (blinks, pupil diameter, gaze position, etc.)	5 min	20 min
--	3	Individual Differences Measures *Demographic questionnaire (including education, work experience, and expertise in space)* *Perceived ease of use of ChatGPT, Trust Intention to AI and Experiences with ChatGPT* *Big Five Inventory*	2 min 3 min 10 min	35 min
--				
--	4	Training Overview	5 min	43 min
--	5	*Task 1 (with/without AI Assistant)* NASA TLX Survey Engagement Survey Usability Survey Trust Perceived confidence in material *Task 2 (without/with AI Assistant)* NASA TLX Survey Engagement Survey Usability Survey Trust Perceived confidence in material	1 hour 1 hour	2 hour 43 min
--	6	Participant compensation	2 min	2 hour 45 min
	TOTAL TIME			~ 2 hr 45 min

The TAMU Institutional Review Board (IRB) and the US Army Office of Human Research Oversight (OHRO) both approved the initial version of the protocol. An amendment to the IRB application was submitted and approved in Phase II to change the source of funding and to incorporate minor changes in the protocol resulting from a pilot study conducted in the first half of Phase II.

Subjects: A total of 51 users (32 graduate students and 19 undergraduate students) participated in our lab experiment, not counting the users from the pilot studies described in earlier reports. They were recruited from STEM, Business, Architecture, Liberal Arts, and Public Health departments. Participants with a background in Aerospace Engineering were excluded from the study due to the possibility that their prior knowledge could influence the tool's learning effect. The pool of participants included 29 students from Engineering Major, 7 from Business Major, 4 from Science Major, 1 from Liberal Arts Major, 1 from Public Health Major, 1 from Agricultural and Life Sciences, 1 from Architecture, 3 from Mathematics Major, 3 from Veterinary Medicine & Biomedical Sciences, and 1 from Geoscience Major. The distribution of academic majors of these participants are shown in Figure 10. Among the users, 28 were female, and 23 were male (i.e., 45% male and 55% female). Their age was 24.41 ± 4.79 years. Among the 51 users, 28 were Asian, 13 were White, 4 were Hispanic, 3 were Black, 1 was Middle Eastern, 1 included two or more races, and one user did not specify their race. The distribution of academic major, gender, race, and academic level of the users are shown in Figure 10.

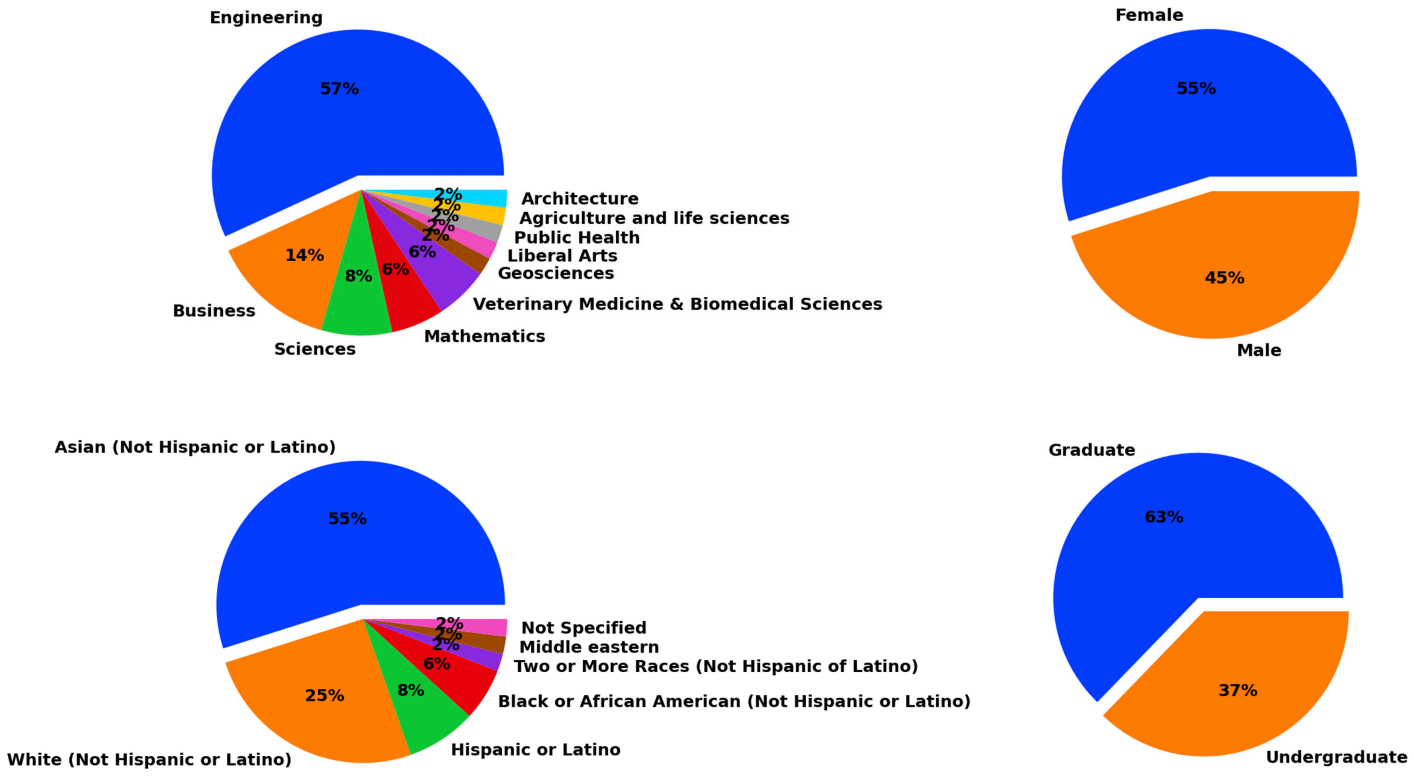


Figure 10: Distribution of academic majors, gender, race, and academic level of the users who participated in our lab experiment

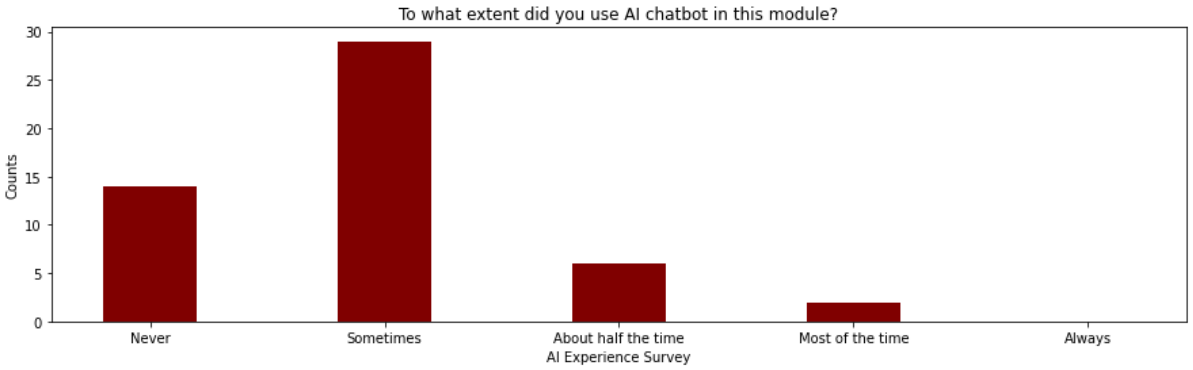
The results of the personality questionnaires, trust intentions to AI, and perceived ease of use of AI are shown in Table 2.

Table 2: Average, Standard deviation and Range of personality traits, Trust Intention to AI and Perceived ease of use of chatbot for 51 users

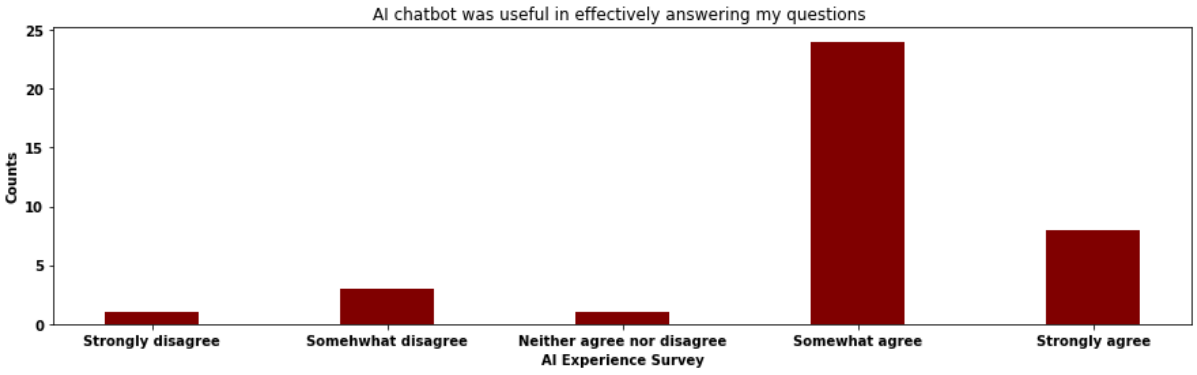
Trait	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness	Trust Intention to AI	Perceived ease of use of AI
Mean ± SD	25.51 ±5.87	35.61 ±4.57	33.35 ±5.78	24.31 ±6.16	37.43 ±4.90	6.80 ± 2.32	26.59 ± 3.23
Range	[8-40]	[9-45]	[9-45]	[8-40]	[10-50]	[2-10]	[6-30]

User experience: The distribution of user responses related to their experience with chatbot are shown in Figure 11(a-f). In addition to that, we show the correlation matrix between these measurements using Spearman's correlation in Figure 11g. We used Spearman's correlation since it captures the monotonic (i.e., rank) relation between two variables and we converted user responses to all these AI experience related questions to ordinal variables (i.e., 1-5 scale) rather than continuous. Our cross-correlation results show that all the AI experience related responses in Figure 11(a-f) are significantly positively correlated (i.e., $p<0.05$), except for the correlation of user response to the questions "To what extent did you use AI chatbot in this module?" The response to this question does not have any significant correlation with any other AI experience related surveys, except for the user engagement survey response (Spearman's$(r)=0.52, p<0.01$, $N=37$) that asks the users "How engaged did you feel when using AI chatbot?" (See Figure 11(g)).

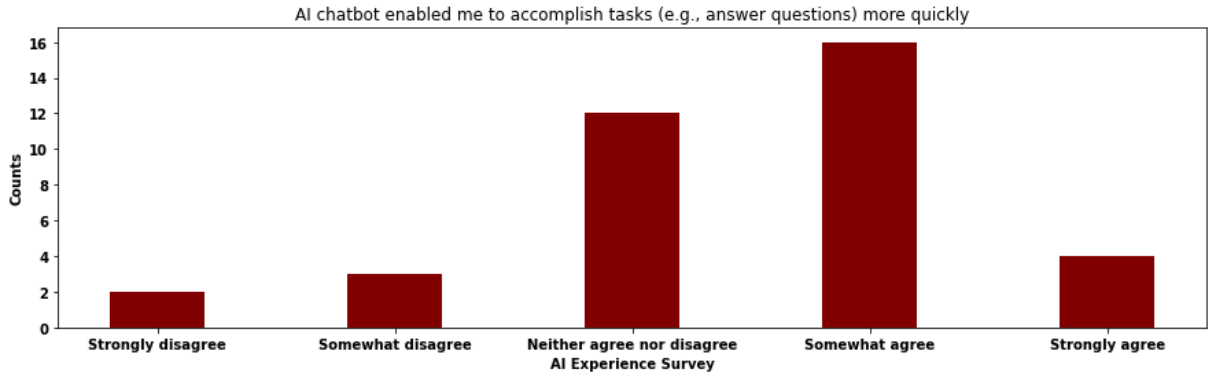
(a)



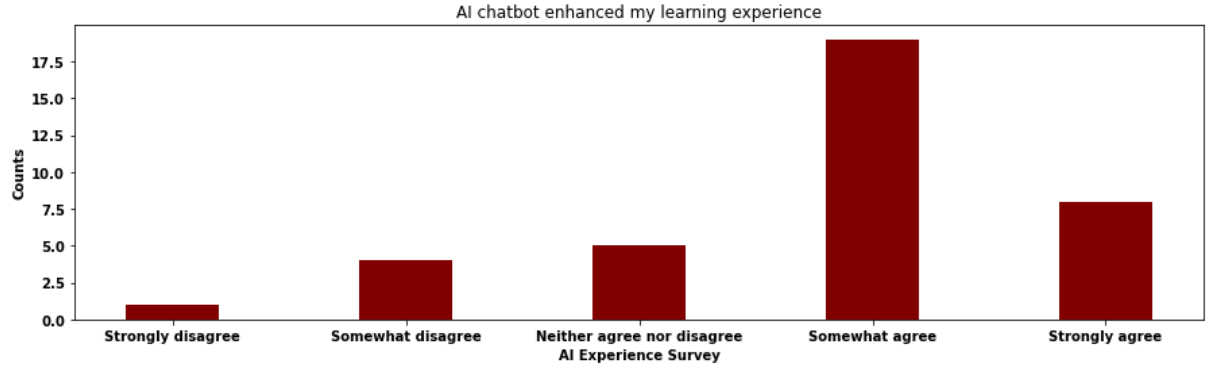
(b)



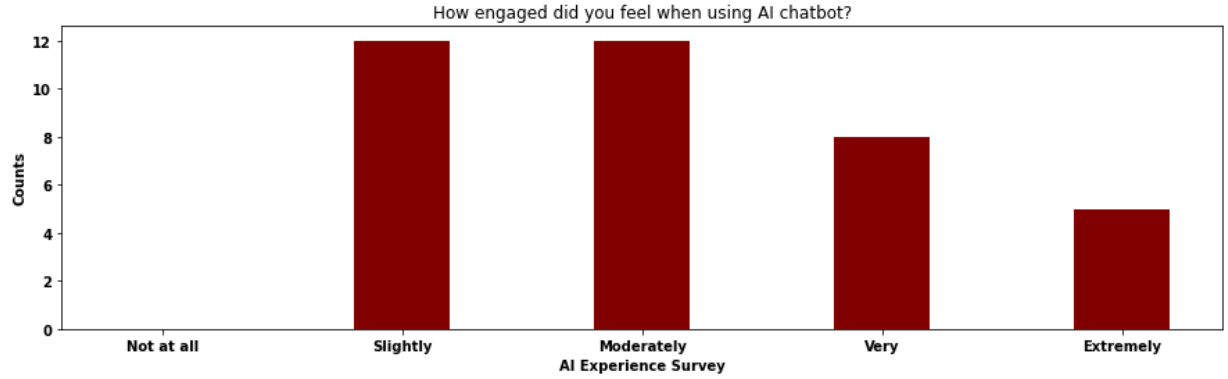
(c)


AI chatbot enabled me to accomplish tasks (e.g., answer questions) more quickly
Counts
0
2
4
6
8
10
12
14
16
Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree
AI Experience Survey

(d)


AI chatbot enhanced my learning experience
Counts
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree
AI Experience Survey

(e)


How engaged did you feel when using AI chatbot?
Counts
0
2
4
6
8
10
12
Not at all
Slightly
Moderately
Very
Extremely
AI Experience Survey

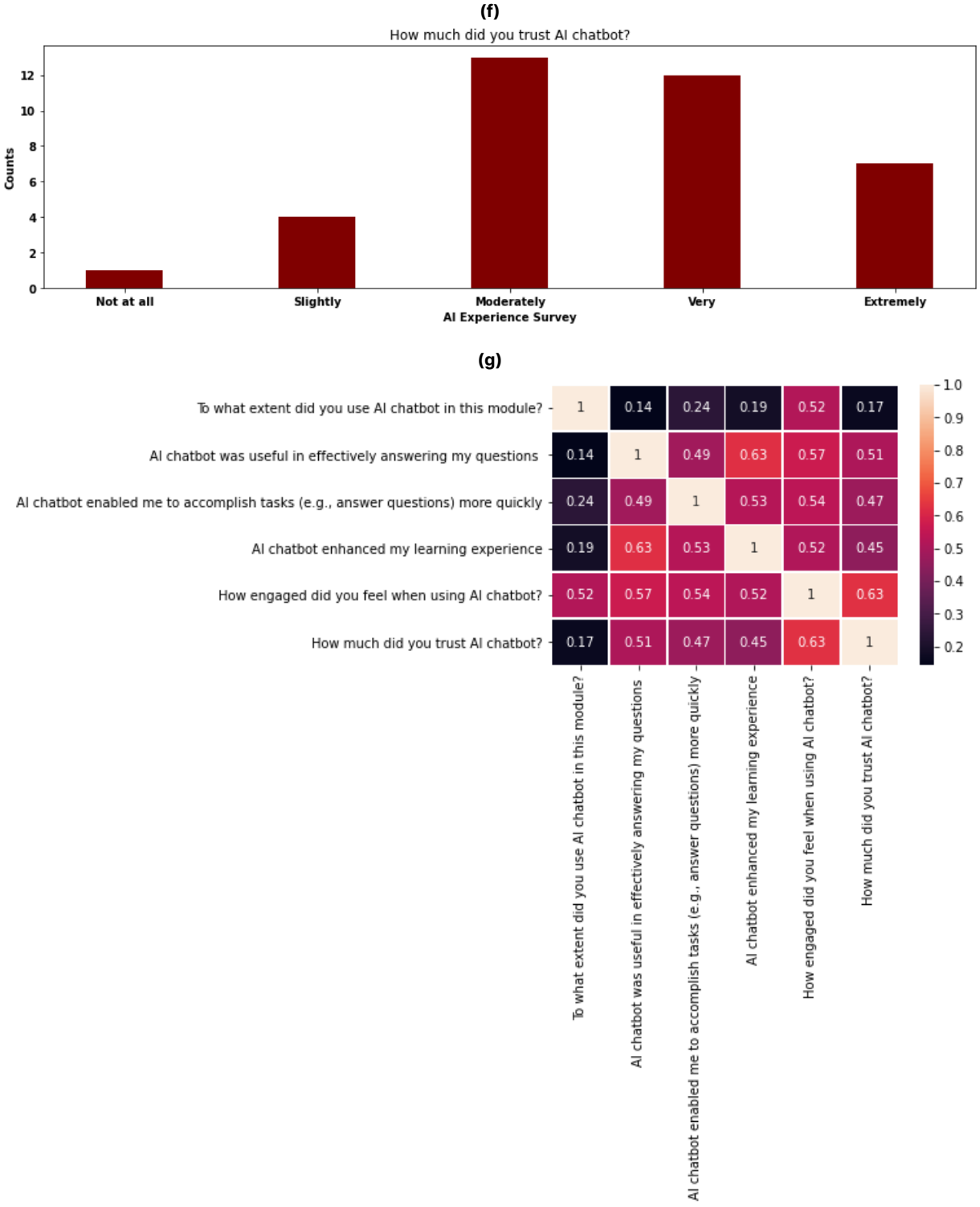


Figure 11: Distribution of user experience with AI related survey responses in post module surveys. Even though we have a total of 51 users, we have a total of 37 user responses for (b)-(f) since 14 users did not use AI at all based on their response. Also, the Spearman's correlation among all these post study AI experience related measures are shown in (g)

Difficulty of the learning modules: We used two of the four learning modules for the lab experiment (Introduction to Space and Space Environment/Orbits) which we adapted to the shorter duration, mostly by eliminating content. Our chatbot was integrated in one of the modules randomly for each user in the final user study. Since our ultimate goal is to find the learning effect of using AI, the difficulty of these two modules needed to be the same. Since the difficulty level of a particular content is subjective, it was hard to decide a priori which questions to include in each module so as to use questions of similar difficulty level across the two modules. To do this, we conducted two pilot studies each with 11 users. In both pilot studies, users went through the same two learning modules as in the final user study. However, the questions from the 1st pilot study were different from the questions from the 2nd pilot study for both learning modules, even though the contents were the same for both modules in these pilot studies. In the first pilot study, there was a total of 10 test questions for Introduction to Space and a total of 11 test questions in the Space Environment/Orbits module. While in the second pilot study, there was a total of 14 test questions in the Introduction to Space module and a total of 17 test questions in the Space Environment/Orbits module. In total, there were 24 test questions in the Introduction to Space module, and 28 test questions in the Space Environment/Orbits module for both pilot studies. The protocols for these pilot studies were similar to our final user study protocol (e.g., the users could not see the learning contents while answering the test questions).

There were multiple choice questions with two or four options in all test questions for both pilot studies. Since it is easier to just guess the correct answer for questions with two answers than with four, we computed an adjusted difficulty level for each question as follows. For multiple choice questions with four options, we defined the difficulty as difficulty=1-(performance/100), and for multiple choice questions with two options, we measured the difficulty as difficulty=0.5-(performance/200) where performance ϵ [0,100]. So, the difficulty level ranges from [0-0.50] for multiple choice questions with two options and from [0-1.0] for multiple choice questions with four options.

We removed any question with difficulty level >=0.9 or difficulty level<=0.1, since we considered these questions as potential outliers. This removed 5 questions from one module and 2 questions from the other module from the questions from both pilot studies combined. The distributions of the questions' difficulty for both modules in the two pilot studies are shown in Figure 12 (top), which shows that the distributions of the difficulty level of the questions from both modules are not comparable. Therefore, for the lab experiment, we selected questions from the two pilot studies to make the distribution of question difficulty as similar as possible across modules. Specifically, we selected pair-wise questions from both modules so that these questions from either module by pairs so that they are within 0.05 difficulty range and have the closest difficulty level. This way, we selected 16 questions from each module and the distribution of difficulty level for these selected questions from both modules are shown in Figure 12 (bottom). From Figure 12, we can see that the difficulty level of these selected 16 questions from each module follow roughly the same distribution; therefore, we included these 16 test questions in the final user study.

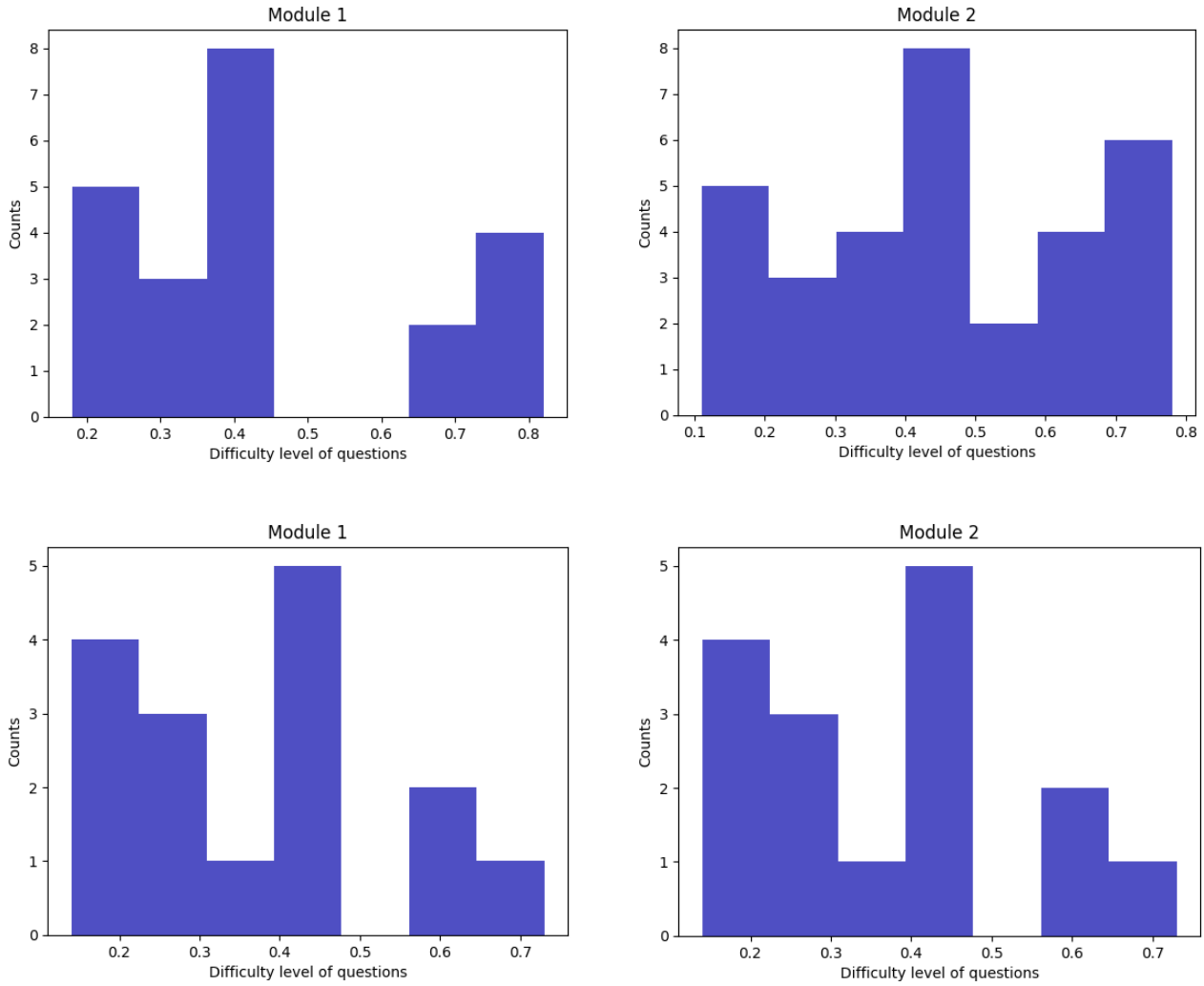


Figure 12: Top: Difficulty level distribution for module 1 (left) and module 2 (right) from pilot studies. Bottom: Difficulty level distribution for selected questions from module 1 (left) and module 2 (right) as included in the lab experiment

Impact of AI assistant on user performance: Users achieved test scores of 56.76 ± 19.02 points using AI, and 50.53 ± 17.29 points without using AI. The distribution of user performance using and without using AI is shown in Figure 13. We conducted a one tailed paired t test for user performance with and without AI and the t-test results indicate that user performance significantly improves when they use AI (i.e., T(50)=2.25, p=0.014). This result is encouraging as it suggests that the AI chatbot can potentially be effective in improving user performance.

We also found that users spent more time for the chatbot-assisted module compared to the module without chatbot (26.80 ± 6.6 min compared to 24.33 ± 7.52, N=51). The one-tailed paired t-test results show that users spend significantly more time for the chatbot assisted module compared to the other module (T(50)=1.79, p=0.03), thus suggesting that the improvement in performance might be partially because the chatbot made the users spend a longer time in the module, perhaps because it was more engaging. The distribution of the time spent with and without the AI module is shown in Figure 13.

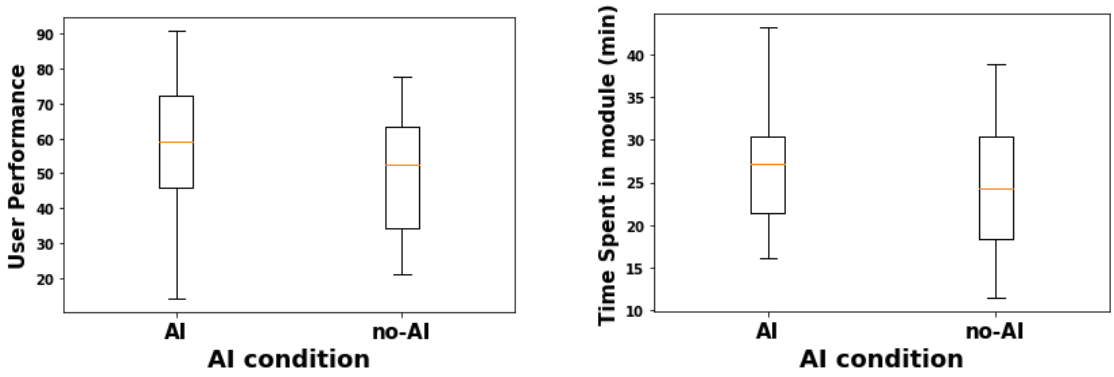


Figure 13: Left: Distribution of user performance for with versus without AI where user performance [0-100] is normalized by question difficulty level; Right: Distribution of time spent for the modules with and without AI

Impact of AI assistant on user perceived workload: Each user completed one module with the use of chatbot and another module without the chatbot. We measured user workload load in 5 different dimensions (i.e., mental demand, temporal demand, effort, performance, and frustration level) using NASA TLX questionnaires. We have excluded the physical workload from the NASA TLX question since our study does not involve any physical work, all the users completed the study sessions in a lab. We conducted a one-tailed paired t test to see if self-perceived workload of users captured by NASA TLX significantly increases using AI compared to when they don't have access to the AI. The one tailed paired t test results for each NASA TLX workload and overall NASA TLX workload are shown in Table 3 below. As we can see from Table 3, users' temporal workload significantly increases when they use AI. This is somewhat expected since the users have the same 30 min time to read each of the slides having a similar number of pages, so the interaction with the chatbot takes away from reading time and makes them feel hurried or increase their temporal workload. We show the distribution of temporal workload measures for the users with and without using AI in Figure 14, which further demonstrates that the workload measures for users with AI are higher than that of users without AI. As a lesson learnt for a future study, the research team should perhaps not limit the time to take each module, but measure the time it takes with each.

Table 3: Distribution of workload measures for each workload dimension and overall workload dimension by NASA TLX with and without using AI

NASA TLX Questions	Using AI (M ± SD)	Without using AI (M ± SD)	AI>no AI
Mental Demand: How mentally demanding was the task?	11.92 ± 5.113	12.51 ± 4.18	T(50)=-0.82, p=0.79
Temporal Demand: How hurried or rushed was the pace of the task?	8.72±5.36	7.10±5.03	T(50)=1.78, p=0.04
Performance: How successful were you in accomplishing what you were asked to do?	11.15±4.85	11.11±4.76	T(50)=0.05, p=0.479
Effort: How hard did you have to work to accomplish your level of performance?	9.94±5.28	10.19±5.62	T(50)=-0.33, p=0.63
Frustration: How insecure, discouraged, irritated, stressed and annoyed were you?	5.25±5.35	5.61±5.68	T(50)=-0.55, p=0.708
Overall Workload	44.68±17.72	44.29±17.69	T(50)=0.149, p=0.441

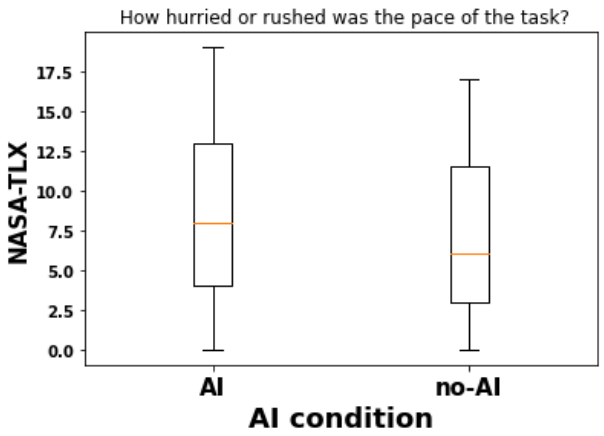


Figure 14: Distribution of temporal workload measures for using and without using AI

IMPACT OF PERSONALITY TRAITS AND TRUST INTENTION TO AI ON ACTUAL USER TRUST IN AI AND FREQUENCY OF USE OF AI

We captured the user's trust in AI and frequency of use of AI via post module questionnaires (e.g., 'To what extent did you use AI chatbot in this module?', 'How much did you trust AI chatbot?'). The Pearson's correlation between each personality trait and user trust in AI and also between personality traits and frequency of AI usage is shown in Table 4.

Table 4: Pearson's correlation between trust in AI and personality traits, trust in AI and trust intention to AI, frequency of AI usage and personality traits, frequency of AI usage and trust intention to AI

Pearson's correlation	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness	Trust Intention to AI
Trust in AI	R=-0.0048, P=0.977, N=37	R=0.1311, P=0.4392, N=37	R=0.429, P=0.008, N=37	R=-0.057, P=0.7357, N=37	R=0.162, P=0.339, N=37	R=0.291, P=0.079, N=37
Frequency of AI usage	R=-0.144, P=0.311, N=51	R=0.0025, P=0.986, N=51	R=-0.283, P=0.043, N=51	R=0.096, P=0.501, N=51	R=0.151, P=0.291, N=51	R=0.196, P=0.166, N=51

As we can see from Table 4, conscientious people trust the AI more and use our chatbot more compared to their counterparts. The scatterplots based on Trust in AI and conscientiousness and based on frequency of AI usage and conscientiousness are shown in Figure 15.

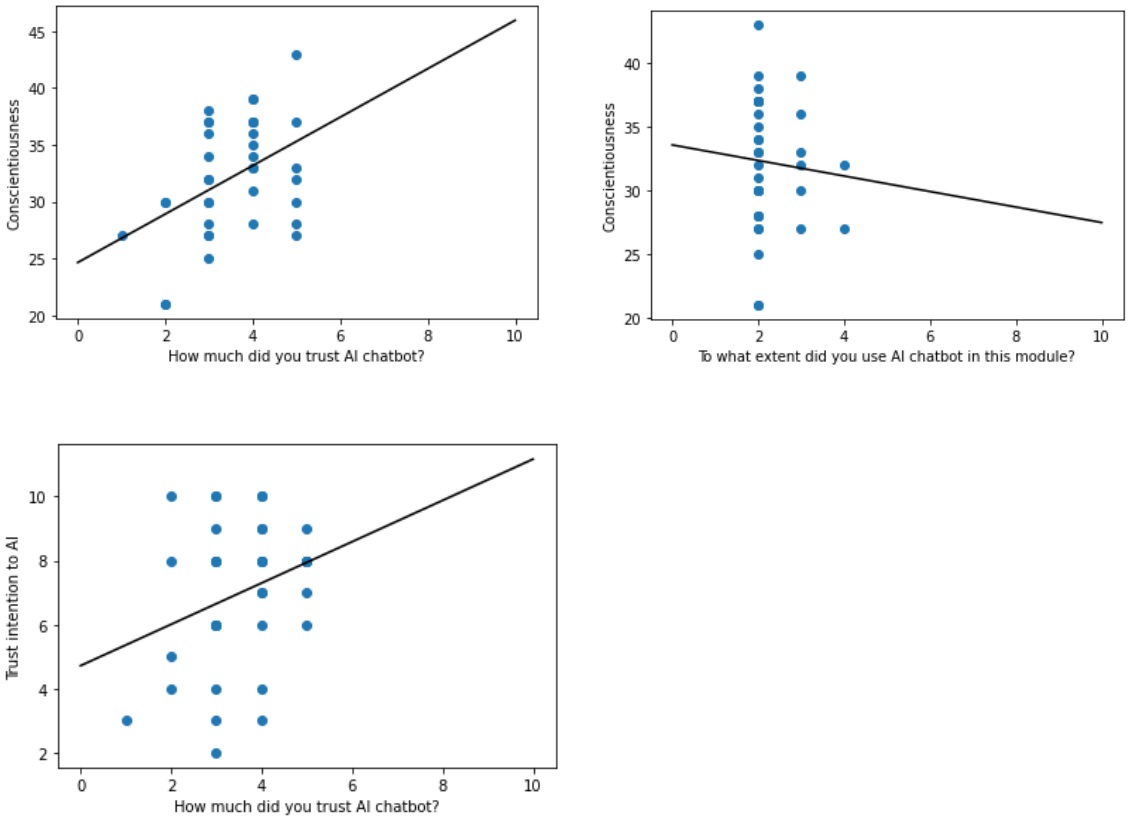


Figure 15: Upper left: Scatterplot showing the relation between user trust in AI and conscientiousness; Upper right: Scatterplots showing the relation between frequency of AI usage and conscientiousness; Lower left: Scatterplot showing the relation between user trust in AI and trust intention to AI. The regression lines are also drawn for all plots

From Figure 15, we can see that the positive correlation between user trust in AI and conscientiousness is stronger than the positive relation between the frequency of AI usage and conscientiousness. We also computed Pearson's correlation between trust intention to AI captured by pre-survey questionnaires and user trust in AI and also, between trust intention to AI and frequency of AI usage in Table 4. As we can see from Table 4, users with higher trust intention to AI tend to trust the AI chatbot more and this relation is approaching significance. We have added the scatter plot showing the relation between trust in AI and trust intention to AI in Figure 15, which also demonstrates the positive correlation between trust in AI and trust intention to AI. This is expected as users with high trust intention to AI are supposed to trust the chatbot more.

RELATION BETWEEN USER PERFORMANCE AND USER EXPERIENCE

We have captured the user experience with our chatbot via three different measures: 1) Usability measured by user response to these three questions, "To what extent did you use AI chatbot in this module? AI chatbot was useful in effectively answering my questions? AI chatbot enabled me to accomplish tasks (e.g., answer questions) more quickly"; 2) Engagement captured by the question, "How engaged did you feel when using AI chatbot?"; and 3) Trust quantified by the question, "How much did you trust AI chatbot?". We conducted Pearson's correlation between each of these three measurements and user performance with the AI-assisted module. We also showed the Pearson's correlation between trust intention to AI and user performance with AI, and between the previous experience with AI (i.e., "How long have you been using conversational agents (e.g., ChatGPT, Google Home, Alexa, Siri)?") and user performance with AI in Table 5. We note a negative trend in users that engaged more with the tool tended to perform lower. This is possibly an artifact of the fact that the time was limited so users who went slowly because the assistant forced them to spend more time on the material possibly ran out of time towards the end. We have seen similar phenomena in time-controlled experiments where interacting with highly capable AI assistants degrades performance in some tasks or metrics due to time trade-offs with those other tasks (Viros-i-Martin and Selva, 2022).

Table 5: Pearson's correlation between different measures of user experience with AI and user performance with AI

Pearson's correlation	User performance with AI
Usability of chatbot	R=-0.175, p=0.299, N=37
Engagement with chatbot	R=-0.345, p=0.0365, N=37
Trust in chatbot	R=-0.302, p=0.0692, N=37
Trust intention to AI	R=-0.239, p=0.09, N=51
Previous experience with AI	R=-0.015, p=0.91, N=51
Perceived ease of use of conversational agent	R=-0.19, p=0.19, N=49

USER CONFIDENCE IN THEIR OWN LEARNING

After the end of each module, the users respond to the question, "How confident do you feel about your understanding of the material you have learned?". The answer to this question is used as a measure of the users' self-confidence in their understanding of the learning module. The distribution of this response for both the chatbot assisted module and the module without chatbot for 51 users is shown in Figure 16. We can see that users show slightly more confidence using the chatbot. The average confidence score for users using chatbot is slightly higher compared to the confidence scores without using the chatbot (2.94 points versus 2.76 points) and based on the one-tailed paired t test, this result is marginally significant (T(50)=1.29, p=0.100). However, we did not find any significant correlation between the users' perceived confidence in their understanding of the material and their actual performance (Pearson's(r)=0.050, p=0.614, N=102) which is interesting.

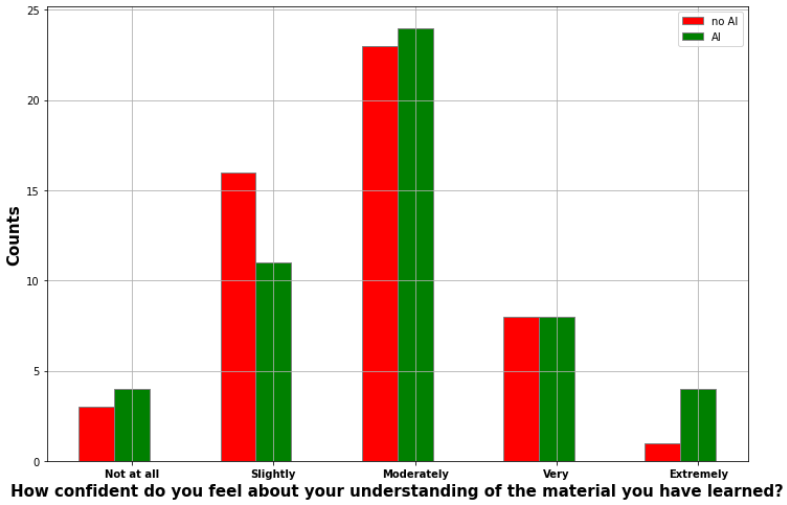


Figure 16: Distribution of user response to their self-confidence in their understanding of the learning materials per module for both with and without chatbot assisted modules

GENDER DIFFERENCES

Our pool of users was almost balanced in terms of gender (i.e., 23 male, 28 female). The user performance using chatbot and without chatbot for both male and female users are shown in Table 6. A one-tailed t test shows that the scores are significantly different for male and female.

Table 6: User performance based on gender and chatbot assistance

	Male	Female	Male>Female One tailed paired t test
Performance using chatbot	61.92 ± 12.71	52.52 ± 22.06	T=1.77, p=0.04, N_{male}=23, N_{female}=28
Performance without chatbot	52.106 ± 16.04	49.24 ± 18.15	T=0.577, p=0.283, N_{male}=23, N_{female}=28

Digging deeper, we discovered that this is due to females seeing a reduced benefit from using the assistant, as they perform about the same as males without the assistant but worse than males with the assistant. This is shown in Figures 17 and 18.

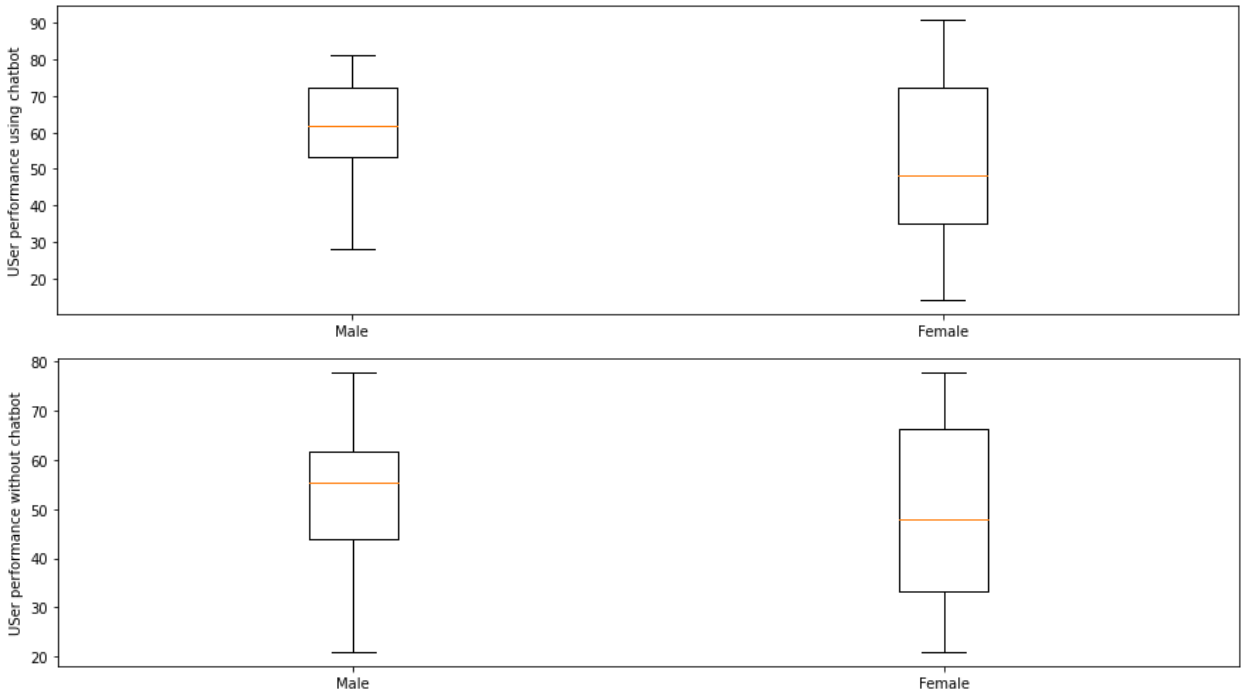


Figure 17: Performance distribution of male and female users for both using the chatbot (top figure) and without using the chatbot (bottom figure)

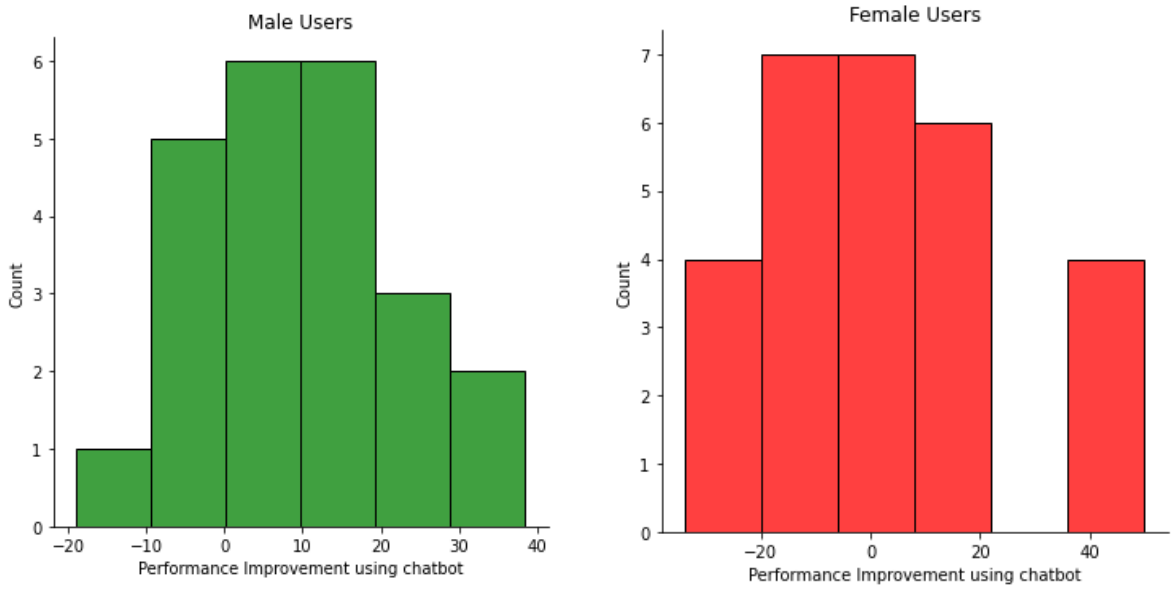


Figure 18: Distribution of performance improvements using chatbot for both male users (left) and female users (right)

Figure 18 clearly shows that males improve their scores more than females when using the chatbot. This difference is approaching significance (9.81± 14.45 points improvement compared to 3.27 ± 22.41 points, T=1.18, p=0.12, N male=23, N female=28). To be more exact, 78.2% male users have higher performance using the chatbot and total 50% female users have higher performance using the chatbot. Based on self-reported values on AI usage, male users significantly used the chatbot more than female users (2.26± 0.792 points compared to 1.64 ± 0.548 points; One tailed t test: T(49)=3.21, p<0.001).

SUMMARY DISCUSSION OF LAB EXPERIMENT RESULTS

In this work, we explored 5 research questions (RQs):

- **RQ1:** Does AI chatbot help users to improve their performance?
- **RQ2:** Does the use of AI chatbot have an effect on the workload of the users?
- **RQ3:** How do the users' personality traits (e.g., extraversion, agreeableness, conscientiousness, openness, neuroticism) and trust intention to AI impact the user trust in AI and frequency of AI usage?
- **RQ4:** Does users' perceived experience with AI impact their performance?
- **RQ5:** Is there any gender bias in learning effect with chatbot?

In response to **RQ1**, our results show that user performance significantly improves using AI based on one tailed paired t test (i.e., T(50)=2.25, p=0.014) Also, users show slightly more confidence in their understanding of the learning content using the chatbot (T(50)=1.29, p=0.100). Apart from that, our findings show that users spent more time understanding the learning module with chatbot, which justifies their encouragement to use our chatbot (AI (i.e., T(50)=1.79, p=0.03). This result is encouraging as it suggests that AI chatbot is effective in improving the user performance. Based on our understanding, the learning effect can be because of the chatbot or the time that they dedicated to learning using chatbot. In either case, our results show that they are more eager and effective in learning the materials using the chatbot. However, it is important to note that the overall performance of users with or without the assistance of AI chatbot is very low (56.76 ± 19.02 points and 50.53 ± 17.29 points out of 100 points using and without using the chatbot respectively), which makes sense since none of our users have prior knowledge on aerospace engineering contents. We have provided some sample questions that the users asked to our chatbot, and the responses given by chatbot in supplementary materials.

In response to **RQ2**, we captured the workload in five different dimensions including mental workload, temporal workload, frustration level, performance level, and level of effort via NASA TLX questionnaires. Each user completed these NASA TLX questionnaires twice, once based on their learning experience for the module without the chatbot, and another time based on their learning experience for the AI-chatbot assisted module. Based on one tailed paired t test, our results suggest that users felt more temporal workload using the chatbot (T(50)=1.78, p=0.04). Since users were given the same amount of time to go through the slides with and without chatbot and the length of the contents were almost same, it justifies that users spending time interacting with the chatbot feel the time pressure to understand the learning concepts in due time.

In response to **RQ3**, our results suggest that conscientious people trust our chatbot more (Pearson's(r)=0.429, p=0.008, N=37) and use our chatbot more (Pearson's(r)=-0.283, p=0.043, N=51) compared to their counterparts. Apart from that, we found that users who perceived high trust intentions to AI end up putting more trust in our chatbot based on Pearson's correlation result (Pearson's(r)=0.291, p=0.079, N=37) which approaches significance. Prior studies show conflicting results for the relation between conscientiousness and trust in AI. Some studies found positive correlation between them (Bawack et al., 2021; Chien et al., 2016; Rossi et al., 2018), while others found negative correlation between them (Aliasghari et al., 2021, Oksanen et al., 2020).

In response to **RQ4**, user performance degrades when they trust our chatbot more (Pearson's(r)=-0.302, p=0.0692, N=37) and when they engage more with our chatbot (Pearson's(r)=-0.345, p=0.0365, N=37). Users who perceive higher trust in AI depict lower performance and this relation approaches significance (Pearson's(r)=-0.239, p=0.09, N=51). These results suggest that users who interact with our chatbot more may have less time to interact with the actual learning content, or may have run out of time towards the end, as shown in the temporal demand results.

In response to **RQ5**, we observe that male users perform significantly better than female users using chatbot (T=1.77, p=0.04). On average, the performance of male users is 61.92 ± 12.71 points compared to 52.52 ± 22.06 points for female users using the chatbot. However, male users perform almost the same as the female users without using the chatbot (i.e., 52.106 ± 16.04 point vs 49.24 ± 18.15 points) and there is no statistical significant difference between their performance without using the chatbot (T=0.577, p=0.283). Also, the performance improvement for male users using the chatbot is higher than female users (9.81± 14.45 points compared to 3.27 ± 22.41 points) and this difference is approaching significance (T=1.18, p=0.12, N male=23, N female=28). In total, 78.2% male users have higher performance using the chatbot while 50% female users have higher performance using the chatbot. Our results also show that male users significantly used the chatbot more than the female users (i.e., T(49)=3.21, p<0.001). We can come to the conclusion that our chatbot has a better learning impact on the male users since they are using the chatbot more compared to female users. This result is in line with a previous study that found that male users have greater experience for utilizing the GPT3.5 based conversational agents than females (Wang et al., 2024).

DEPLOYMENT AT CAPE/AFCAA

To complement the lab experiment, we deployed the Daphne Academy tool with 22 cost estimators (8 CAPE users and 14 AFCAA users). The users were provided the four full learning modules and questionnaires described in previous sections and were asked to complete any number of modules with the AI Assistant and fill out the post-module questionnaires described above, in addition to providing feedback about the tool. Of note, they did not fill out the pre-task questionnaires and did not do any modules without the AI tool. Moreover, the number of modules each subject completed varies across subjects. Overall, the environment was much less controlled than in the lab experiment. That said, the research team gathered a lot of feedback and some useful data.

The test scores per module are provided in Table 7. Immediately, we notice that the four modules were not of similar difficulty, which was completely expected. We also observe that test scores in modules 1-3 are higher than in the lab experiment, which is not the case of module 4 (presumably module 4 was a bit too advanced). This could also suggest that by the time some users got to the fourth module, they experienced some fatigue.

Table 7: Number of users and test scores statistics per module

Module	Introduction to Space	Space Environment and Orbits	Spacecraft Subsystems	Remote Sensing Payloads
# of users who took the module	12	9	7	5
Test scores (mean $\pm$ stdev)	87.5 ± 8.036	64.44 ± 14.61	65.98 ± 9.99	51.0 ± 12.806

The distribution of user workload measures across different modules is shown in Table 8 below.

Table 8: Distribution of cognitive load measures for different dimensions as per NASA TLX questionnaires. The highest cognitive load and the lowest cognitive load measures across each dimension are marked as red and green respectively

NASA TLX Questions	Introduction to Space	Space Environment and Orbits	Spacecraft Subsystems	Remote Sensing Payloads	Range
Mental Demand: How mentally demanding was the task?	8.42 ± 3.92	10.0 ± 0.0	11.66 ± 4.71	13.0 ± 4.54	1-20
Temporal Demand: How hurried or rushed was the pace of the task?	5.42±3.69	10.0 ± 0.0	10.0 ± 0.0	10.0 ± 5.715	1-20
Performance: How successful were you in accomplishing what you were asked to do?	15.0 ± 4.34	6.0 ± 0.0	11.0 ± 5.65	9.33 ± 7.58	1-20
Effort: How hard did you have to work to accomplish your level of performance?	8.857 ± 3.719	5.0 ± 0.0	11.0 ± 3.74	7.66 ± 3.85	1-20
Frustration: How insecure, discouraged, irritated, stressed and annoyed were you?	5.85 ± 5.02	10.0 ± 0.0	7.33 ± 4.71	9.66 ± 4.49	1-20
Aggregated NASA TLX: Cumulative scores of all five workload measures	33.57 ± 16.62	49.0 ± 0.0	49.0 ± 10.61	51.0 ± 19.20	5-100

Workload is lowest for module 1 and highest for module 4, which is consistent with test score results.

At the end of each module, we asked this question to the users “How confident do you feel about your understanding of the material you have learned?”, and converted the answers to a 1-5 Likert scale (1: lowest confidence, 5: highest confidence). The distribution of self-confidence measures for each module are shown in Table 9.

Table 9: Distribution of self-confidence measures of users across each module

Module	Introduction to Space	Space Environment and Orbits	Spacecraft Subsystems	Remote Sensing Payloads
Confidence in learning	3.85 ± 0.98	2.0 ± 0.0	2.0 ± 0.816	2.33 ± 1.24

We see that confidence is high for the first module, but low for the other modules.

At the end of each module, we asked this question to the users "To what extent did you use the AI chatbot in this module?". The users were asked to select any of the 5 given options to answer this question including "Never", "Sometimes", "About half the time", "Most of the time", and "Always". We also asked this question to the users "How much did you trust AI chatbot?" The users were asked to select any of the 5 given options to answer this question including "Not at all", "Slightly", "Moderately", "Very", "Extremely". Following that, we converted the answer to these questions on 1-5 Likert scale (1: lowest measure, 5: highest measure). The distribution of user trust measures and usage of chatbot measures for each module are shown below in Table 10.

Table 10: Distribution of frequency of AI usage and trust in chatbot measures of users across each module and the total number of users who used the chatbot in each module. User trust measure is invalid for Remote Sensing Payload module since none of the users used the chatbot for this module

Module	Introduction to Space	Space Environment and Orbits	Spacecraft Subsystems	Remote Sensing Payloads
Frequency of AI usage	2.42 ± 1.17	2.0 ± 0.0	1.33 ± 0.471	1.0 ± 0.0
User trust in chatbot	3.5 ± 1.25	2.0 ± 0.0	5.0 ± 0.0	NA
# of users who used the chatbot	6	1	1	0

We captured the perceived usability of the chatbot via three questions on a 1-5 Likert scale, "To what extent did you use AI chatbot in this module? AI chatbot was useful in effectively answering my questions? AI chatbot enabled me to accomplish tasks (e.g., answer questions) more quickly". The answer to these questions was aggregated to yield a final usability score between 0-15. We also measured the user engagement with the chatbot by the question, "How engaged did you feel when using AI chatbot?", which we measured on a 1-5 Likert scale. The distribution of usability and user engagement distribution across each module are shown in Table 11 below. Please note, since no user used the chatbot in the Remote Sensing Payloads module, the distribution of usability of AI and Engagement with AI values are listed as NA in Table 11.

Table 11: Distribution of usability of AI and engagement with chatbot measures of users across each module. The measurements are invalid for the "Remote Sensing Payloads" module since none of the users used the chatbot for this module

Module	Introduction to Space	Space Environment and Orbits	Spacecraft Subsystems	Remote Sensing Payloads
Usability of AI	10.66 ± 3.49	5.0 ± 0.0	11.0 ± 0.0	NA
Engagement with AI	3.33 ± 1.24	2.0 ± 0.0	4.0 ± 0.0	NA

Results indicate a moderate to high user engagement with the chatbot, as well as moderate to high perceived usability of the AI in the first and third module, but not in the second module. This suggests that the design and content of these modules were effective in capturing and maintaining user interest, and that the AI was able to facilitate the learning process in these areas.

QUALITATIVE FEEDBACK

At the end of each module, we asked this open-ended question to the users "Please share your thoughts regarding the training material, the AI chatbot, and the study overall?". In general, users have a **positive impression about our chatbot "*...Daphne overall really impressed me! I tried them out frequently...*".** One of the users recommended **adding a voice over to the slides** "*...Provide a voice over. I learn better when I am able to read and listen...*".

A few **technical glitches** were also observed. One of the users pointed out that the chatbot referred them to the wrong slide number "*...When asking the virtual assistant (AI chatbot) a question, sometimes the virtual assistant would reference the wrong slide number. The wrong slide number would always be off by one (e.g. slide 21 versus slide 20 with slide 20 being the correct slide)....*". One of the users also noted that the chatbot did not open for them after completing the exam. Our team will work on fixing these issues. In general, the **users appreciated the depth of the content** of the slides as they commented: "*I thought the content of the training was really good, especially for someone who knows nothing about space systems and was actually trying to learn something*"; "*The material was laid out well and at a really good level of detail*". However, they suggested that the chatbot should not only summarize the materials in the slide, rather it should **cover the topics more in-depth**. Users commented that "*Training the chatbot with additional resources may help it explain content more in depth rather than summarizing things that a user could find in the lecture slides themselves.*"; "*I asked Daphne an in-depth question and they weren't able to answer me. Just repeated what is exactly on the slide. It seems to be they can only re-word text to try to present it in a different way (which can sometimes be useful).*" In future work, we plan to fine-tune our model on additional data sources so that it can incorporate additional context into their answers.

SOFTWARE AND DOCUMENTATION

The Daphne Academy software is located on private GitHub repositories:

https://github.com/seakers/academy-interface
https://github.com/seakers/academy-brain

Detailed instructions to install the server and client parts are provided in a separate "Transition Plan" document. This "Transition Plan" also contains detailed instructions for other use cases of the tool, such as adding or modifying users, instructional materials, assigning courses to users, etc.

CONCLUSIONS

The research team has demonstrated proof of concept for a CA that can improve workforce training in the context of cost estimation. Leveraging the recent developments in large language models, we developed a flexible and scalable AI agent accompanied by a simple web-based front end. We also developed four different learning modules around the topic of space systems as a commodity in cost estimation.

We validated the AI tool in two studies: a controlled experiment in the lab with 51 TAMU students as users, and a validation study with 22 CAPE and AFCAA personnel as users. Results of the lab experiment show that the AI agent resulted in an increase in test scores of about 6 points on average, which was statistically significant. This increase could be attributed at least partially to the subjects spending more time interacting with the instructional materials when they have access to the AI assistant. The CAPE study did not compare AI vs no AI conditions due to the less controlled environment. However, some useful qualitative feedback was gathered that can improve future development of this or similar tools. In particular, users noted that the assistant is limited in the answers it gives to the content provided in the slides. This is by design as the slides are the only context that the agent was provided with and to avoid hallucinations, we forced the agent to answer questions based on the context. However, in future work, additional data sources could be provided so that the AI assistant has access to a richer context.

A transition plan document has been delivered to the sponsor together with the software tool. The plan provides detailed instructions to install the tool in a new system and perform a number of routine maintenance tasks, such as adding a new learning module, adding users, assigning users to learning modules, etc.

While the tool developed is still just a prototype and several shortcomings were noted, we believe the results show promise of the technology for improving training outcomes. Ultimately, this work could result in significant improvements in the efficiency and effectiveness of workforce training in the DoD.

Additional recommendations beyond the immediate implications of this project include studying the adoption of AI agents that adapt to individual user needs and provide a QA system for other applications in systems engineering and acquisition.

APPENDIX A. LIST OF PUBLICATIONS RESULTED

No peer-reviewed papers have been published yet as a direct result of this project. However, the following is a list of closely related publications and presentations whose authors were partially sponsored by this project:

G. Apaza and D. Selva. Leveraging Large Language Models for Tradespace Exploration. Under review in *Journal of Spacecraft and Rockets*.

A. Demagall and D. Selva. LLM-Based SysML Virtual Assistant. Presented at the 2023 AI4SE & SE4AI workshop. Washington DC. September 27-28 2023.

REFERENCES

A.C. Graesser, et al. “AutoTutor: An Intelligent Tutoring System with Mixed-Initiative Dialogue.” IEEE Transactions on Education, vol. 48, no. 4, 2005, pp. 612–18, doi:10.1109/TE.2005.856149.

Aliasghari, P., Ghafurian, M., Nehaniv, C. L., & Dautenhahn, K. (2021, August). Effect of domestic trainee robots’ errors on human teachers’ trust. In 2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN) (pp. 81-88). IEEE.

Bawack, R. E., Wamba, S. F., & Carillo, K. D. A. (2021). Exploring the role of personality, trust, and privacy in customer experience performance during voice shopping: Evidence from SEM and fuzzy set qualitative comparative analysis. International Journal of Information Management, 58, 102309.

Cheung, B., et al. “SmartTutor: An Intelligent Tutoring System in Web-Based Adult Education.” Journal of Systems and Software, vol. 68, no. 1, Elsevier, 2003, pp. 11–25.

Chien, S. Y., Sycara, K., Liu, J. S., & Kumru, A. (2016, September). Relation between trust attitudes toward automation, Hofstede’s cultural dimensions, and big five personality traits. In Proceedings of the human factors and ergonomics society annual meeting (Vol. 60, No. 1, pp. 841-845). Sage CA: Los Angeles, CA: SAGE Publications.

Corbett, Albert T., et al. “Intelligent Tutoring Systems.” Handbook of Human-Computer Interaction, Elsevier Science B. V., 1997, pp. 849–74, doi:10.1126/science.228.4698.456.

D’Mello, Sidney K., Scotty D. Craig, Amy Witherspoon, et al. “Automatic Detection of Learner’s Affect from Conversational Cues.” User Modeling and User-Adapted Interaction, vol. 18, no. 1, 2008, pp. 45–80, doi:10.1007/s11257-007-9037-6.

D’Mello, Sidney K., Scotty D. Craig, B. Gholson, et al. “Integrating Affect Sensors in an Intelligent Tutoring System.” Affective Interactions: The Computer in the Affective Loop Workshop, International Conference on Intelligent User Interfaces, ACM Press, 2005, pp. 7–13.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS quarterly, 319-340.

Folsom-Kovarik, Jeremiah T., et al. “Tractable POMDP Representations for Intelligent Tutoring Systems.” ACM Transactions on Intelligent Systems and Technology, vol. 4, no. 2, 2013, pp. 1–22, doi:10.1145/2438653.2438664.

John, O. P. & Srivastava, S. (1999). The Big-Five trait taxonomy: History, measurement,

Kasneci, Enkelejda, et al. “ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education.” Learning and Individual Differences, vol. 103, Elsevier, 2023, p. 102274.

Kim, Byungsoo, et al. “AI-Driven Interface Design for Intelligent Tutoring System Improves Student Engagement.” ArXiv Preprint ArXiv:2009.08976, 2020.

Koedinger, Kenneth R., et al. Intelligent Tutoring Goes To School in the Big City. 1997, pp. 30–43.

Kojima, Takeshi, et al. "Large Language Models Are Zero-Shot Reasoners." Advances in Neural Information Processing Systems, vol. 35, 2022, pp. 22199–213.

Mayo, Michael John. Bayesian Student Modelling and Decision-Theoretic Selection of Tutorial Actions in Intelligent Tutoring Systems. 2001, http://ir.canterbury.ac.nz/bitstream/10092/2565/1/thesis_fulltext.pdf.

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. Information systems research, 13(3),334-359.

Nye, Benjamin D. "Barriers to ITS Adoption: A Systematic Mapping Study." International Conference on Intelligent Tutoring Systems, Springer, 2014, pp. 583–90.

Nye, Benjamin D. "Intelligent Tutoring Systems by and for the Developing World: A Review of Trends and Approaches for Educational Technology in a Global Context." International Journal of Artificial Intelligence in Education, vol. 25, no. 2, 2015, pp. 177–203, doi:10.1007/s40593-014-0028-6.

Oksanen, A., Savela, N., Latikka, R., & Koivula, A. (2020). Trust toward robots and artificial intelligence: An experimental approach to human–technology interactions online. Frontiers in Psychology, 11, 568256.

Ong, James, and Sowmya Ramachandran. "Intelligent Tutoring Systems: Using AI to Improve Training Performance and ROI." Networker Newsletter, vol. 19, no. 6, 2003, pp. 1–6.

Rossi, A., Dautenhahn, K., Koay, K. L., & Walters, M. L. (2018). The impact of peoples' personal dispositions and personalities on their trust of robots in an emergency scenario. Paladyn, Journal of Behavioral Robotics, 9(1), 137-154.

Sandra G. Hart, & Lowell E. Staveland: Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research, Advances in Psychology, Elsevier, Volume 52, 1988, Pages 139-183.

Sarrafzadeh, Abdolhossein, et al. "'How Do You Know That I Don't Understand?' A Look at the Future of Intelligent Tutoring Systems." Computers in Human Behavior, vol. 24, no. 4, 2008, pp. 1342–63, http://ovidsp.ovid.com/ovidweb.cgi?T=-JS&PAGE=reference&D=psyc6&NEWS=N&AN=2008-05428-005.

Sibley, C., Coyne, J., & Baldwin, C. (2011, September). Pupil dilation as an index of learning. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 55, No. 1, pp. 237-241). Sage CA: Los Angeles, CA: SAGE Publications.

Singhal, Karan, et al. "Towards Expert-Level Medical Question Answering with Large Language Models." ArXiv Preprint ArXiv:2305.09617, 2023.

Siyuan Chen, Julien Epps, Fang Chen: A comparison of four methods for cognitive load measurement, OZCHI'11: Proceedings of the 23rd Australian Computer-Human Interaction Conference, November 2011, Pages 76-79.

Virós-i-Martin, Antoni, and Daniel Selva. "Improving Designer Learning in Design Space Exploration by Adapting to the Designer's Learning Goals." *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. Vol. 86236. American Society of Mechanical Engineers, 2022.

Wang, J., Ivrissimtzis, I., Li, Z., & Shi, L. (2024, March). Impact of personalised ai chat assistant on mediated human-human textual conversations: Exploring female-male differences. In Companion Proceedings of the 29th International Conference on Intelligent User Interfaces (pp. 78-83).

Wei, Jason, et al. “Finetuned Language Models Are Zero-Shot Learners.” ArXiv Preprint ArXiv:2109.01652, 2021.

Zhao, Wayne Xin, et al. “A Survey of Large Language Models.” ArXiv Preprint ArXiv:2303.18223, 2023.