ACQUISITION INNOVATION
RESEARCH CENTER

Cognitive Assistant for Training Cost Estimators

EXECUTIVE SUMMARY AND REPORT

SEPTEMBER 2023

PRINCIPAL INVESTIGATOR:
Daniel Selva, *Texas A&M University*

PRINCIPAL INVESTIGATOR:
Theodora Chaspari, *Texas A&M University*

PRINCIPAL INVESTIGATOR:
Alejandro Salado, *The University of Arizona*

SPONSORS:
OFFICE OF COST ASSESSMENT AND PROGRAM EVALUATION (CAPE), OFFICE OF THE SECRETARY OF DEFENSE (OSD)
OFFICE OF THE ASSISTANT SECRETARY OF DEFENSE FOR ACQUISITION (ASD(A)), OFFICE OF THE UNDER SECRETARY OF DEFENSE FOR ACQUISITION AND SUSTAINMENT (OUSD(A&S))

RESEARCH TEAM

NAME	ORG.	LABOR CATEGORY
Daniel Selva	Texas A&M University	Principal Investigator (PI)
Theodora Chaspari	Texas A&M University	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)
David Asatryan	Texas A&M University	Undergraduate Research Assistant

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.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

The goal of this research project is to develop a cognitive assistant to support training of new 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 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.

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 been adopted yet 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. This document reports on the results of the first year of Phase II ($134k, Sep 2022—Sep 2023). In this time, we have 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 have started thorough testing of the effectiveness of the tool at Texas A&M University (TAMU).

Future research plans include delivering a longer 5-module version of the online course and testing it with real users at the Office of Cost Assessment and Program Evaluation (CAPE). In addition, the research team will refine our estimates of the development and maintenance costs.

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 to 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).

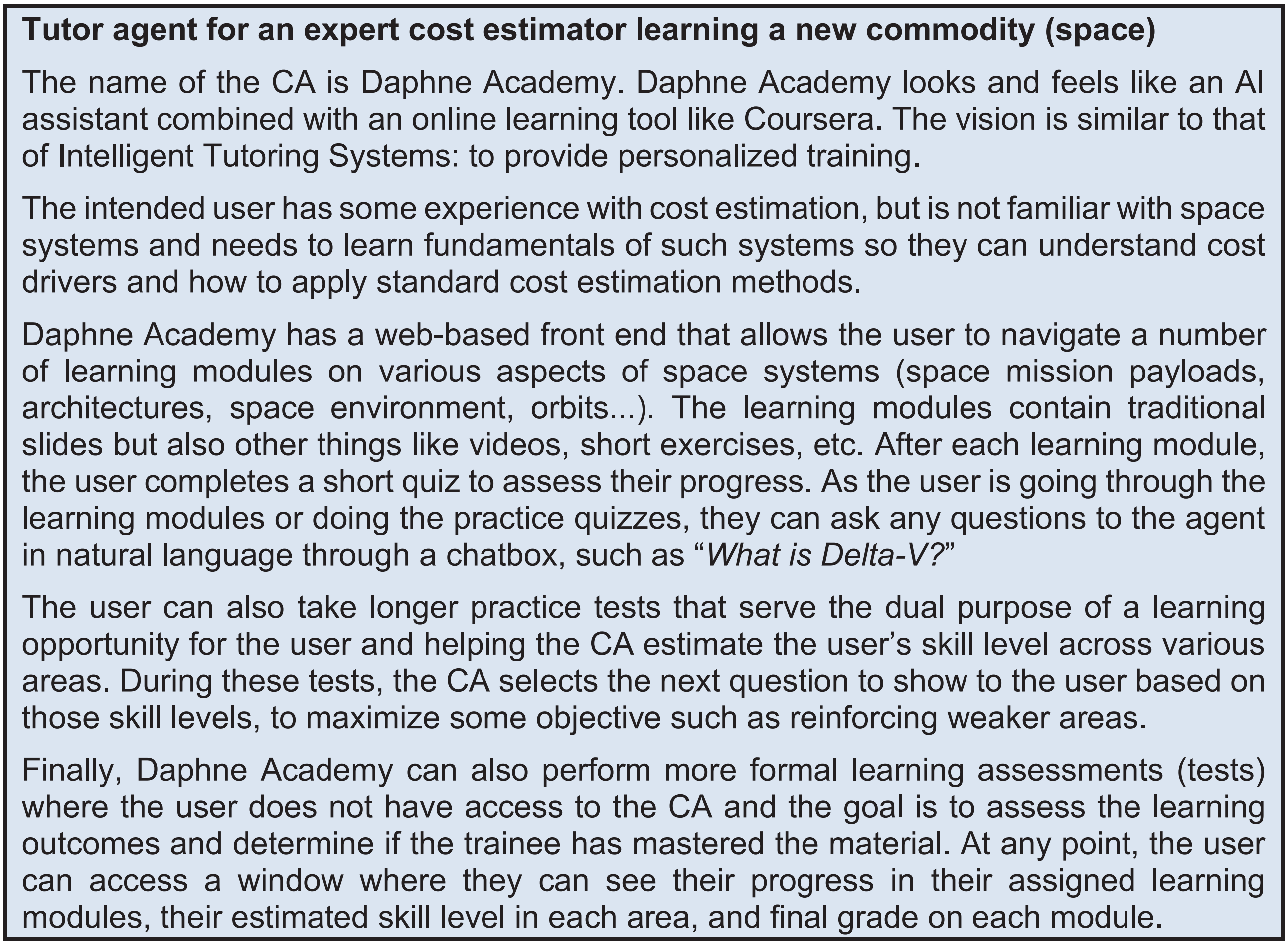
Finally, Large Language Models (LLMs) (Zhao et al.) have recently emerged as a game-changing technology in artificial intelligence (AI) due to their abilities to reason (Kojima et al.) and learn without providing examples (Wei et al.). Beyond the widespread interest this technology generated, it has 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

1.1 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 tradespace 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

1.2 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). During the past year, the repercussions of using AWS were discussed and it was decided that it was acceptable to use. However, the impact of implementing this in an in-house cloud environment will be addressed in the task related to estimating the cost of developing and maintaining such a tool in the DoD (see future research plan in the conclusion).

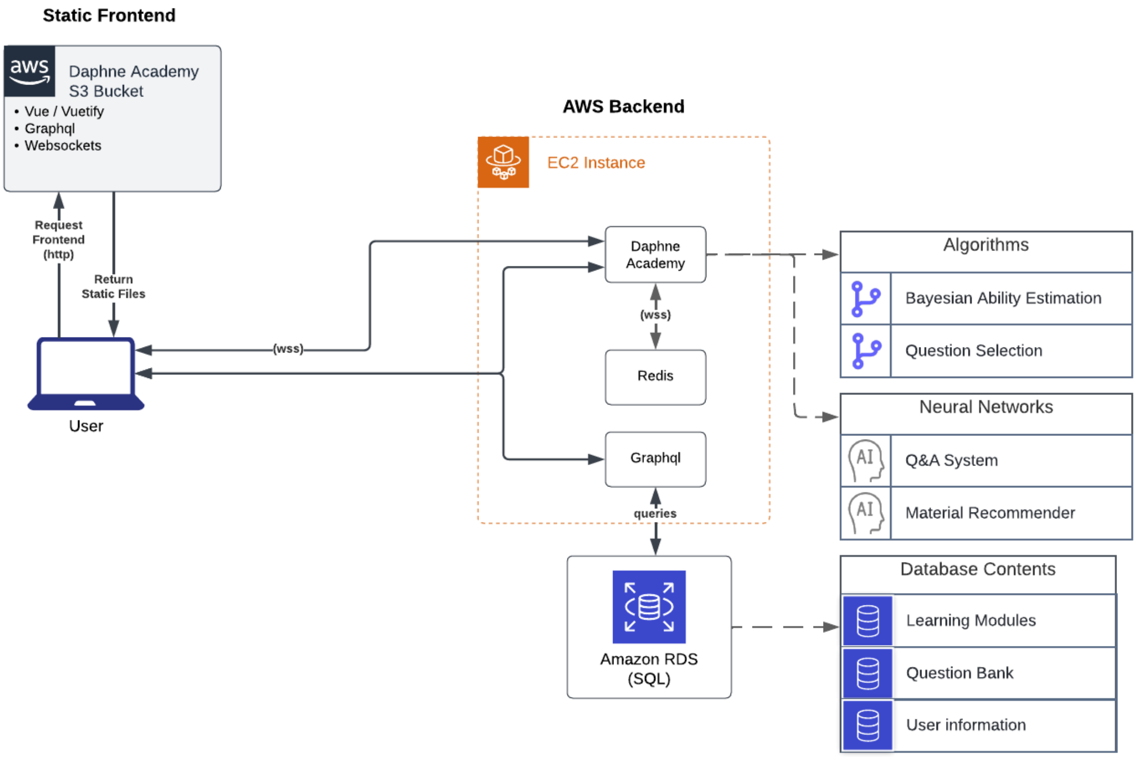


Figure 2 . Software Architecture

1.3 Question Answering System

The Question Answering (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 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 new QA system is shown in Figure 3.

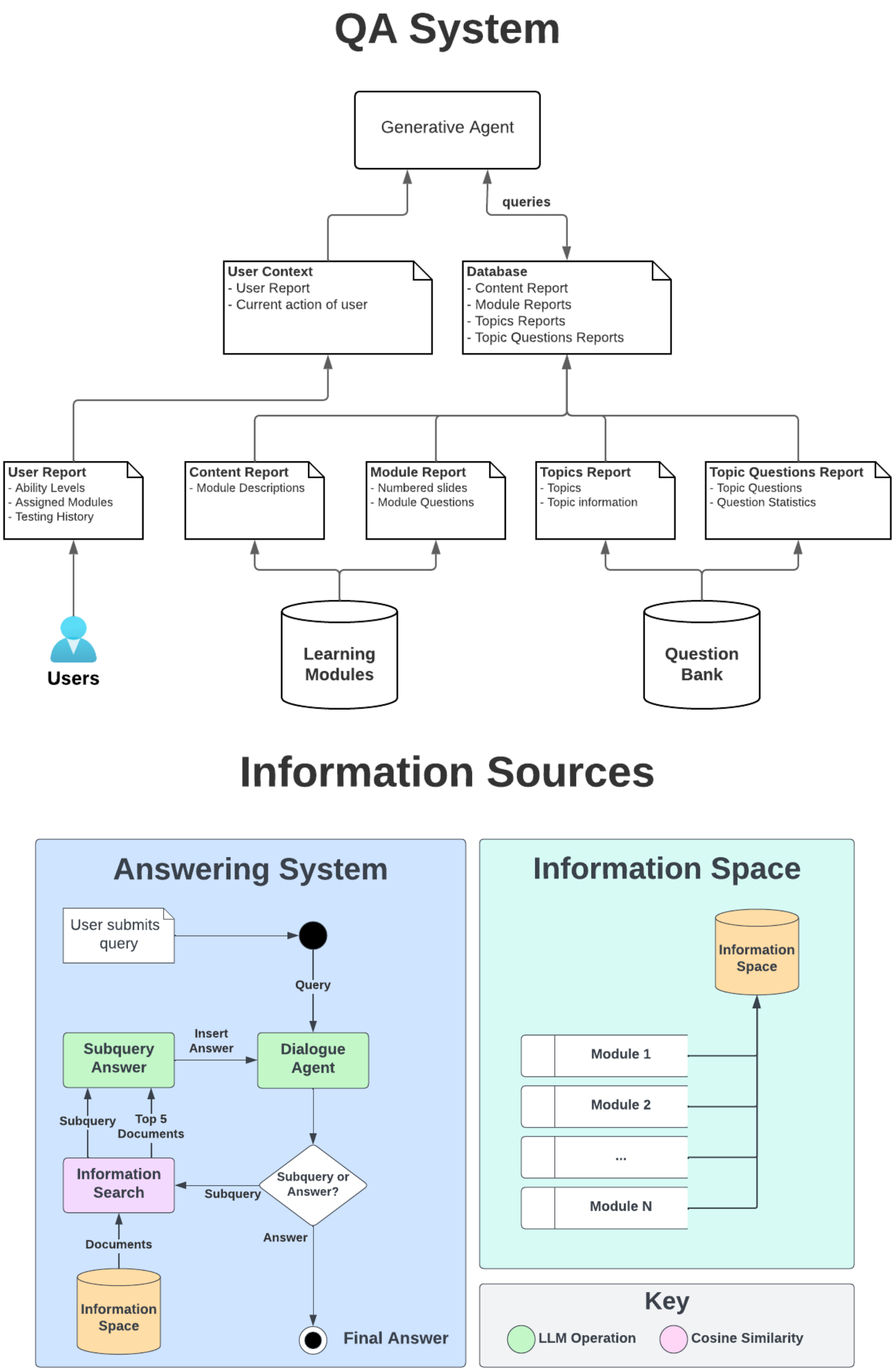


Figure 3 . Architecture of the new Question Answering System leveraging a generative pre-trained transformer model (GPT-4)

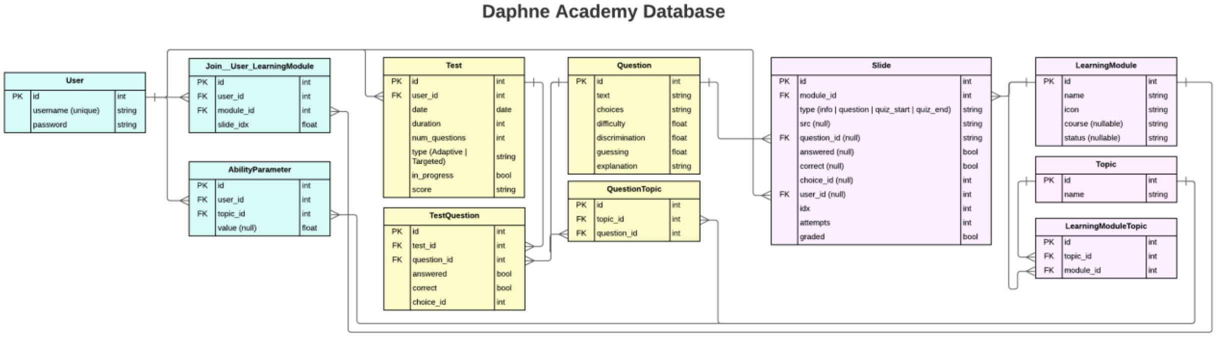
The QA system utilizes GPT4 with the backend database to fuse data sources that encompass information about a given user (e.g., assigned learning modules and current ability estimates) including the material content they have been assigned (e.g., textual content from learning module slides). As this 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 been 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.

1.4 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. While the schema is largely unchanged from Phase 1, 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.



User		
PK	id	int
	username (unique)	string
	password	string

Join__User_LearningModule		
PK	id	int
FK	user_id	int
FK	module_id	int
	slide_idx	float

AbilityParameter		
PK	id	int
FK	user_id	int
FK	topic_id	int
	value (null)	float

Test		
PK	id	int
FK	user_id	int
	date	date
	duration	int
	num_questions	int
	type (Adaptive | Targeted)	string
	in_progress	bool
	score	string

TestQuestion		
PK	id	int
FK	test_id	int
FK	question_id	int
	answered	bool
	correct	bool
	choice_id	int

Question		
PK	id	int
	text	string
	choices	string
	difficulty	float
	discrimination	float
	guessing	float
	explanation	string

QuestionTopic		
PK	id	int
FK	topic_id	int
FK	question_id	int

Slide		
PK	id	int
FK	module_id	int
	type (info | question | quiz_start | quiz_end)	string
	src (null)	string
FK	question_id (null)	string
	answered (null)	bool
	correct (null)	bool
	choice_id (null)	int
FK	user_id (null)	int
	idx	int
	attempts	int
	graded	bool

LearningModule		
PK	id	int
	name	string
	icon	string
	course (nullable)	string
	status (nullable)	string

Topic		
PK	id	int
	name	string

LearningModuleTopic		
PK	id	int
FK	topic_id	int
FK	module_id	int

Figure 4 . Schema of the Daphne Academy Database

1.5 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 a_j, b_j, c_j are 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.

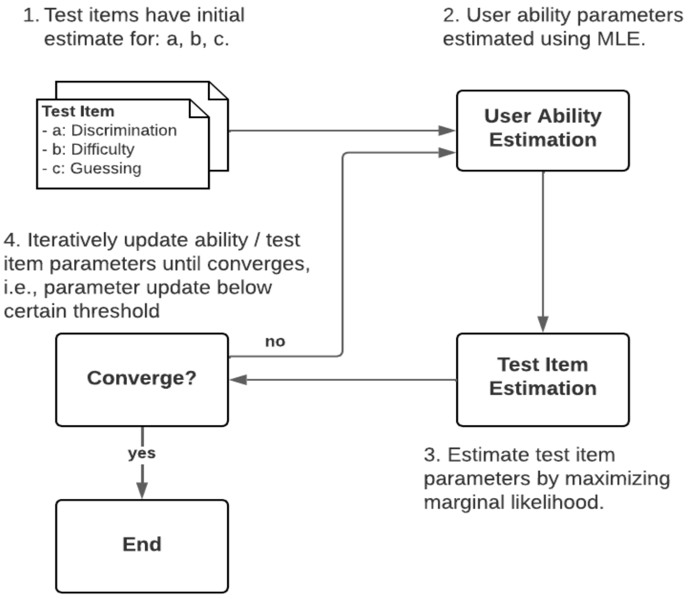

1. Test items have initial estimate for: a, b, c.
2. User ability parameters estimated using MLE.
Test Item
- a: Discrimination
- b: Difficulty
- c: Guessing
User Ability Estimation
4. Iteratively update ability / test item parameters until converges, i.e., parameter update below certain threshold
no
Converge?
Test Item Estimation
yes
3. Estimate test item parameters by maximizing marginal likelihood.
End

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.

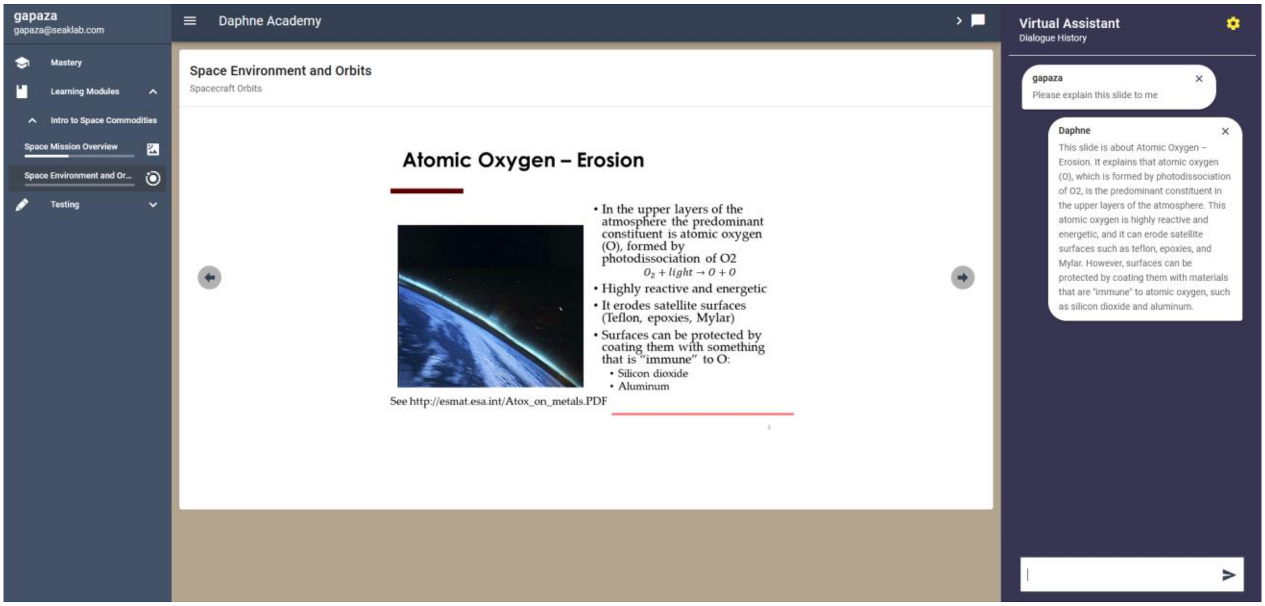
The question selection algorithm is unchanged from Phase 1. It selects the next question in two steps: 1) determine the user's weakest topic area, and 2) select a question from the determined topic area to maximize information gained with respect to the user's ability level. Step one is completed by 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.

1.6 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.



Atomic Oxygen – Erosion

- In the upper layers of the atmosphere the predominant constituent is atomic oxygen (O), formed by photodissociation of O2
  $O_2 + light \rightarrow O + O$
- Highly reactive and energetic
- It erodes satellite surfaces (Teflon, epoxies, Mylar)
- Surfaces can be protected by coating them with something that is "immune" to O:
  - Silicon dioxide
  - Aluminum

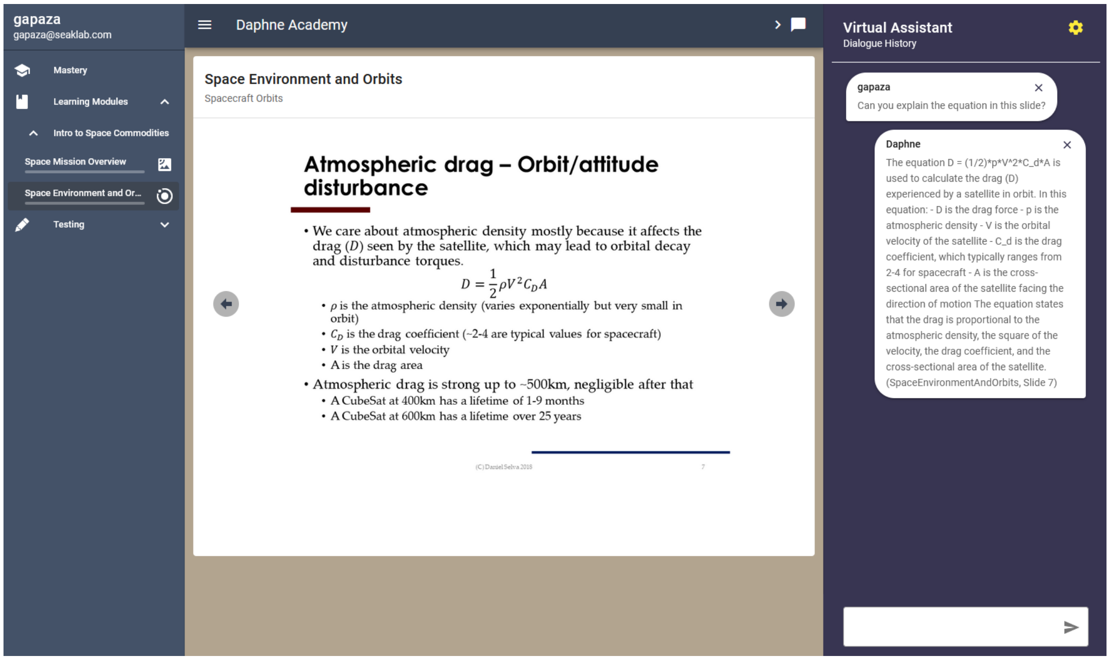
See http://esmat.esa.int/Atox_on_metals.PDF

gapaza: Please explain this slide to me

Daphne: This slide is about Atomic Oxygen – Erosion. It explains that atomic oxygen (O), which is formed by photodissociation of O2, is the predominant constituent in the upper layers of the atmosphere. This atomic oxygen is highly reactive and energetic, and it can erode satellite surfaces such as teflon, epoxies, and Mylar. However, surfaces can be protected by coating them with materials that are "immune" to atomic oxygen, such as silicon dioxide and aluminum.

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.



Atmospheric drag – Orbit/attitude disturbance

- We care about atmospheric density mostly because it affects the drag (D) seen by the satellite, which may lead to orbital decay and disturbance torques.

$$D = \frac{1}{2}\rho V^2 C_D A$$

  - ρ is the atmospheric density (varies exponentially but very small in orbit)
  - C_D is the drag coefficient (~2-4 are typical values for spacecraft)
  - V is the orbital velocity
  - A is the drag area
- Atmospheric drag is strong up to ~500km, negligible after that
  - A CubeSat at 400km has a lifetime of 1-9 months
  - A CubeSat at 600km has a lifetime over 25 years

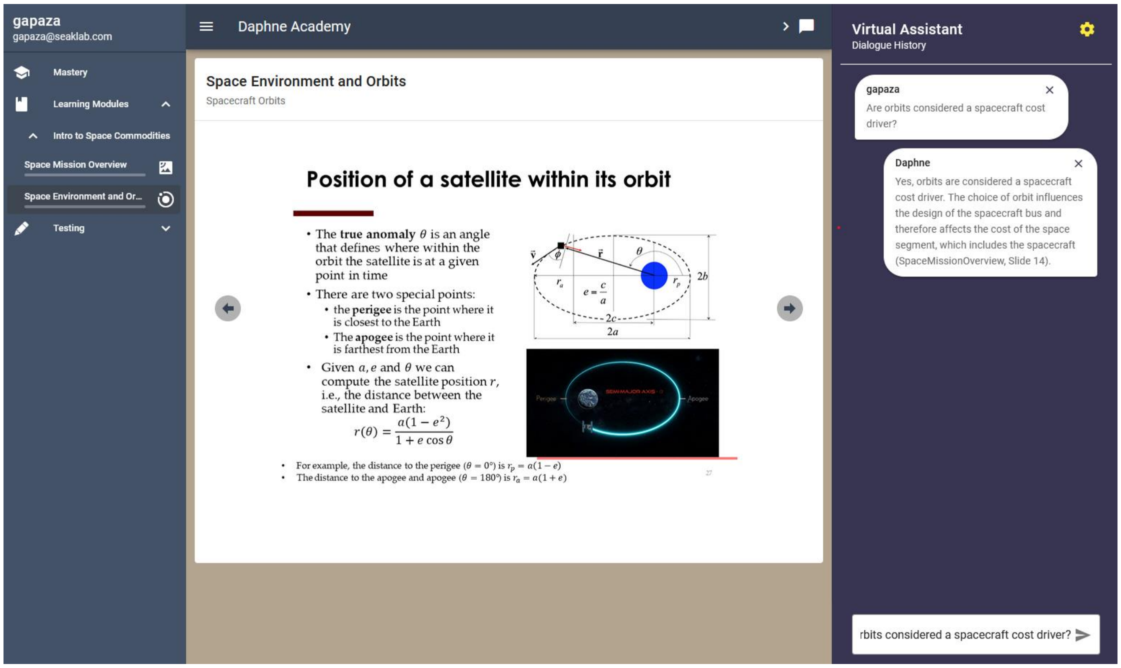
(C) Daniel Selva 2018 7

gapaza: Can you explain the equation in this slide?

Daphne: The equation D = (1/2)*p*V^2*C_d*A is used to calculate the drag (D) experienced by a satellite in orbit. In this equation: - D is the drag force - p is the atmospheric density - V is the orbital velocity of the satellite - C_d is the drag coefficient, which typically ranges from 2-4 for spacecraft - A is the cross-sectional area of the satellite facing the direction of motion The equation states that the drag is proportional to the atmospheric density, the square of the velocity, the drag coefficient, and the cross-sectional area of the satellite. (SpaceEnvironmentAndOrbits, Slide 7)

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.



Space Environment and Orbits

Spacecraft Orbits

Position of a satellite within its orbit

- The **true anomaly** θ is an angle that defines where within the orbit the satellite is at a given point in time
- There are two special points:
  - the **perigee** is the point where it is closest to the Earth
  - The **apogee** is the point where it is farthest from the Earth
- Given a, e and θ we can compute the satellite position r, i.e., the distance between the satellite and Earth:

$$r(\theta) = \frac{a(1-e^2)}{1+e\cos\theta}$$

- For example, the distance to the perigee ($\theta = 0°$) is $r_p = a(1-e)$
- The distance to the apogee and apogee ($\theta = 180°$) is $r_a = a(1+e)$

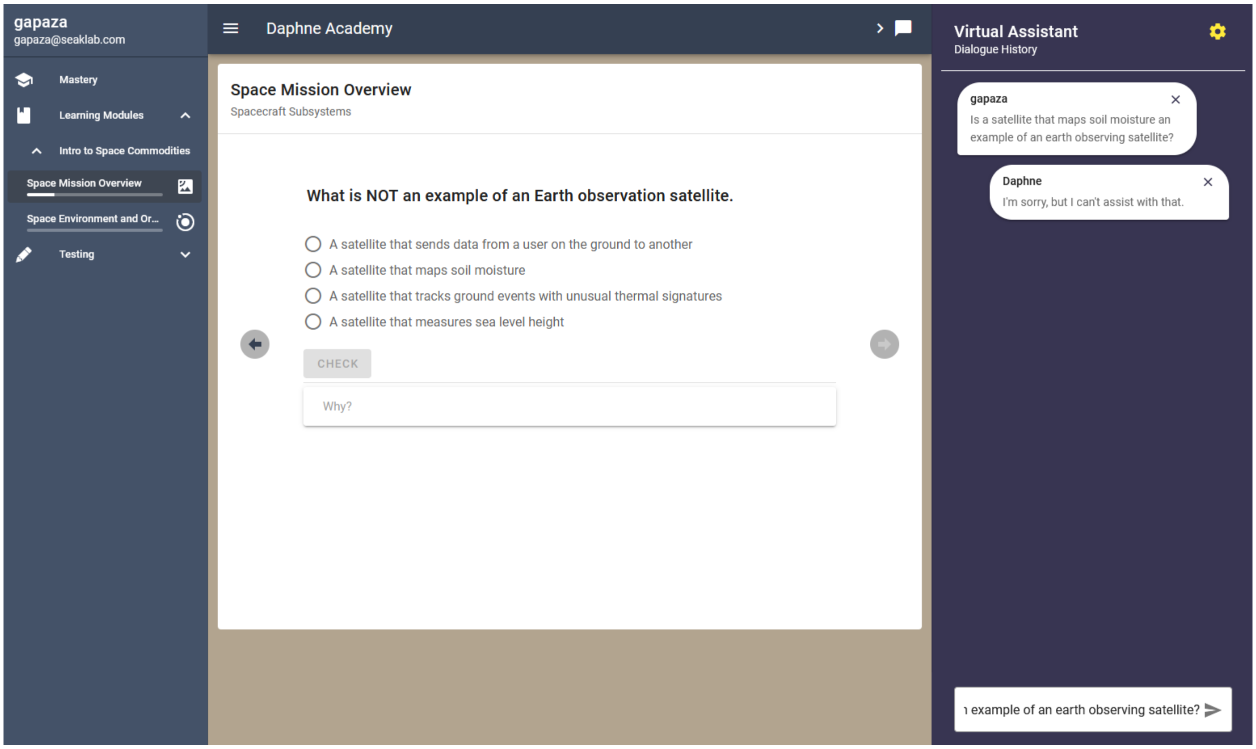
gapaza
Are orbits considered a spacecraft cost driver?

Daphne
Yes, orbits are considered a spacecraft cost driver. The choice of orbit influences the design of the spacecraft bus and therefore affects the cost of the space segment, which includes the spacecraft (SpaceMissionOverview, Slide 14).

rbits considered a spacecraft cost driver?

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.



Space Mission Overview

Spacecraft Subsystems

What is NOT an example of an Earth observation satellite.

○ A satellite that sends data from a user on the ground to another
○ A satellite that maps soil moisture
○ A satellite that tracks ground events with unusual thermal signatures
○ A satellite that measures sea level height

gapaza
Is a satellite that maps soil moisture an example of an earth observing satellite?

Daphne
I'm sorry, but I can't assist with that.

ı example of an earth observing satellite?

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

An important part of the research plan is 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 plan on performing a more systemic validation effort, organized in two thrusts:

> 1) a small number of highly representative users (actual cost estimators from the DoD) using the tool with a realistic learning module over a period of a few days;
> 2) a larger number of less representative users (students from Texas A&M University (TAMU)) using the tool with a simplified learning module over a period of a few hours.

The focus for the first experiment will be to assess whether the tool actually improves learning compared to not having the tool. The goal of the second experiment is to understand more subtle aspects such as how the difficulty of the learning module affects the impact and usefulness of the agent.

In this first half of the Phase II, while we were developing the new version of the agent, we have also carefully designed and planned a controlled experiment with human subjects to be performed at Texas A&M University for the second thrust mentioned above. The protocol for the lab experiment is provided in Table 1. A Tobii eye tracker is used to get objective measures of parameters that are known to correlate with pupil diameter, blinks, etc. and can affect user learning, such as level of engagement, and workload. A questionnaire is used to obtain information about the user's background, personality traits, prior knowledge of space systems, and experience and attitudes towards AI agents. A between-subjects design will assess the effectiveness of the CA in enhancing user learning and will include two conditions: (1) Control condition, where the user will follow the tutorial with the questions without using the CA; and (2) Experimental condition, where the user will follow the same tutorial and questions, while being able to use the CA. Each participant will be randomly assigned to one of the two conditions. The participants assigned to the Experimental condition will go through a training that discusses the basic capabilities of the CA. All users complete two learning modules, one with low difficulty and one with high difficulty. After each learning module, the user takes a test to assess their learning. They also complete another survey to measure user confidence in their learning, engagement, trust in automation, workload and usability of the tool. The order of the difficulty condition is counterbalanced. Currently, we are performing a small pilot study to obtain preliminary data and iron out any issues with the protocol.

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.)	5min	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 (Low Difficulty: Space environment)* NASA TLX Survey Engagement Survey Usability Survey Trust Perceived confidence in material *Task 2 (High Difficulty: Orbits)* 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

Table 1: Protocol for the lab experiment at TAMU

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 will be submitted and must be approved before data collection can start as the source of funds will change between the first and second half of the Phase II. We will also use this amendment to include any minor changes to the protocol identified as a result of the pilot study currently in progress.

CONCLUSIONS

Over the Phase I and Phase II stages of this project, the research team has demonstrated proof of concept for a CA that can improve training for cost estimators. The irruption of generative AI agents last year provided us with an opportunity to revisit the architecture of the agent and make it substantially more scalable and flexible while maintaining accuracy. The current status of the project is that development of this second prototype is complete, including 3 learning modules on space systems, and validation has started at TAMU with a pilot study.

During the next year (Phase II-B), we will focus on validation both at TAMU and CAPE, and on supporting what an in-house development and maintenance of such an agent at the DoD would look like. More specifically, the following tasks will be performance during Phase II-B:

Task 1: Finalize 2 additional learning modules on space payloads.

Task 2: Characterize accuracy of the LLM-based QA system.
Task 3: Validate skill estimation algorithm.
Task 4: Conduct lab experiment at TAMU.
Task 5: Deploy Daphne Academy with potential users at CAPE.
Task 6: Refine cost estimates for deploying and maintaining this system operationally.
Task 7: Prepare software for delivery (e.g., documentation, code cleaning, quality checks).

Recommendations from this phase of the project include: 1) study the adoption of AI tools that adapt to individual user needs and provide a QA system for training cost estimators; 2) study the feasibility of hosting this kind of tool in web-based services such as AWS within the government IT infrastructure; 3) ensure that there are people in the relevant department organization with the AI skills to maintain this kind of tool; and 4) explore the use of hybrid systems leveraging LLMs for other applications.

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.

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

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

ACRONYMS AND ABBREVIATIONS

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
OHRO	Office of Human Research Oversight
PL	Parameter Logistic
QA	Question Answering
TAMU	Texas A&M University

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.

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.

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.

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.

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.

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.

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.

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.

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

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.

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#0480.

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