Bridging Intelligent Tutoring Systems and Chatbots: Development and Evaluation of a Conversational AI Tutor (CAIT)

Ryan Nguyen
Cameron Robbins
Cody Rueda

Supervised by Dr Neil Heffernan

Co-supervised by Eamon Worden

Department of Computer Science

Worcester Polytechnic Institute

March, 2025

A MQP submitted in partial fulfilment of the requirements for the degree of B.S. in Computer Science.

Abstract

ASSISTments, a platform dedicated to enhancing classroom learning through intelligent tutoring systems (ITS), has developed the Conversational AI Tutoring (CAIT) chatbot. CAIT leverages randomized control trials (RCTs) to evaluate various instructional strategies and improve student learning outcomes. In this project, we extended CAIT's functionality with a focus on assessing its effectiveness in real classroom settings. By collecting data from real student interactions while solving middle school math problems, we developed a conversational analyzer to process and interpret these conversations. Although the results were inconclusive due to limited student data, this work lays the foundation for future research, with the goal of optimizing CAIT's impact on learning.

Contents

List of Figures

1

Introduction

Technological advancements in education have profoundly reshaped how students learn, and teachers deliver instruction. Over the past two decades, online platforms and intelligent tutoring systems (ITS) have emerged as powerful tools to enrich classroom learning, provide immediate feedback, and help educators tailor instruction to each student's needs. Despite these innovations, many existing systems rely heavily on structured, pre-programmed feedback, which may not fully capture the complexities of real-time human interaction. At the same time, conversational agents—commonly called chatbots—have evolved from rudimentary text-based programs into sophisticated, AI-driven tools capable of engaging in dynamic, human-like conversations. Leveraging large language models (LLMs) offers a promising avenue for bridging the gap between structured educational platforms and the more fluid, interactive nature of one-on-one tutoring. Modern LLMs can interpret complex student queries, generate tailored hints or explanations, and even adapt the tone or detail of a response to suit a learner's understanding. This capability allows for unprecedented personalization, giving students just-in-time support while freeing teachers to focus on higher-level instructional tasks. This report details the development of CAIT (Conversational AI Tutor), a new framework designed to merge traditional ITS functionalities with the conversational flexibility of advanced chatbots. We explore how CAIT integrates teacher-written content—hints, explanations, and detailed solutions—with novel LLM-driven scaffolding, which breaks problems into smaller steps to guide students toward mastery. Additionally, we discuss how CAIT employs randomized control trials to evaluate which feedback forms are most effective for fostering long-term retention. In the following pages, the Background section provides historical context on ITS and chatbots, introduces key milestones in artificial intelligence, and outlines the capabilities and limitations of LLMs. The Methodology then explains the core technical components of CAIT and how we structured our randomized experiments. Next, we describe the Research and Testing phase, detailing our data collection and evaluation procedures. Finally, we present Results from the initial trials and offer insights on how this approach might shape the future of AI-driven education.

2

Background

2.1 | Introduction

The evolution of Intelligent Tutoring Systems (ITS) and advancements in artificial intelligence have transformed the landscape of personalized education. From early cognitive tutoring tools to modern AI-powered chatbots, these systems have increasingly mimicked human tutoring by providing adaptive feedback and real-time support. This section explores the development of ITS, the impact of platforms like ASSISTments, and the rise of chatbots and large language models (LLMs) in education. By understanding the history and capabilities of these technologies, we can better appreciate their role in enhancing learning experiences.

2.2 | Intelligent Tutoring Systems

Intelligent tutoring systems (ITS) are computer-based instructional platforms designed to emulate the effectiveness of a one-on-one human tutor. By offering personalized feedback and targeted instruction, ITS can adapt to each student's pace and learning style, similar to the benefits of a traditional classroom setting. Research indicates that ITS can achieve learning gains comparable to those of expert human tutors, making them a powerful tool for broadening access to personalized education (VanLEHN). The main advantage lies in real-time data collection and analytics, which enable continuous improvement in both the system's teaching strategies and content delivery (ScienceDirect).

2.3 | ASSISTments

ASSISTments, founded by Neil Heffernan and Cristina Heffernan in 2003, originated from a practical challenge many educators faced: tracking individual student performance while still delivering whole-class instruction. As former middle school math teachers, the Heffernans recognized that teachers needed a more efficient way to collect data on each student's skill level and deliver timely feedback. Before ASSISTments, efforts in using Artificial Intelligence in Education had led to systems like Ms. Lindquist (sometimes referred to as Ms. Linguist), noted as the "first ITS to use the internet to perform experiments on learning" (Heffernan and

Koedinger). Building on these insights, ASSISTments emerged with an important innovation: integrating randomized control trials (RCTs) into the tutoring process (Heffernan abdHeffernan).

2.4 | Overcoming Implementation Barriers

One early obstacle was teacher buy-in. Many educators were hesitant to disrupt their established curriculums with randomized trials. The Heffernans addressed this by aligning ASSISTments with standardized testing preparation for the Massachusetts Comprehensive Assessment System (MCAS). By offering immediate value—helping students practice for high-stakes tests and giving teachers real-time performance data—ASSISTments secured its initial user base.

2.4.1 | Evidence Based Design

Through embedded RCTs, ASSISTments gather high-quality data on which interventions—such as specific hints, scaffolding, or feedback modalities—yield the greatest learning gains.

2.4.2 | Teacher Centric Focus

The platform streamlines the grading process and generates detailed reports, enabling teachers to quickly identify which students need additional assistance on particular skills.

2.4.3 | Scalability

As a not-for-profit, ASSISTments collaborates with researchers and schools nationwide to refine its features and pedagogical methods, ensuring system improvements are grounded in empirical evidence.

By combining the benefits of Intelligent Tutoring Systems with a strong research infrastructure, ASSISTments has demonstrated how adaptive, data-informed educational platforms can effectively support teachers and learners. This approach helps address the perennial challenge of delivering personalized instruction at scale—an essential goal in modern K–12 education.

2.5 | Chatbots

Conversational agents, or chatbots, have significantly evolved since their creation, progressing from simple rule-based programs to sophisticated artificial intelligence systems. Early chatbots like ELIZA relied on pattern matching and keyword substitution to simulate conversation, providing a glimpse into the potential of human-machine interaction despite their limited functionality. Over time, natural language processing (NLP) and machine learning advancements

have enabled chatbots to interpret more complex user inputs and respond with greater contextual understanding. Today, modern chatbots can handle tasks ranging from basic customer service to advanced educational support, reflecting improvements in AI algorithms and the integration of vast training data (Gobiet).

2.5.1 | Early Chatbots: ELIZA/PERRY

The origin of chatbots can be traced back to the 1960s when ELIZA was created. Developed by Joseph Weizenbaum at MIT, ELIZA was defined as a mimic of a Rogerian psychotherapist. Using pattern-matching techniques and substitution methodology, ELIZA reformulated user questions into questions of its own to simulate conversation. Its knowledge and capability were quite limited. However, its simple responses were effective enough to engage users in meaningful conversation, marking one of the first demonstrations of human-machine interaction (Gobiet).

Building upon ELIZA, PARRY developed by psychiatrist Kenneth Colby in 1972, took chatbots a step further. PARRY attempted to simulate patients with schizophrenia by resembling the thinking of an individual with the disease. PARRY works by assigning weights to verbal inputs, which are then changed via a system of assumptions, attributions, and "emotional responses." For the time, PARRY was considered quite robust and useful as when tested with the Turing test and with real psychiatrists, PARRY was indistinguishable from real patients (Gobiet).

2.5.2 | Advancements in NLP: ALICE

Introduced in 1995 and created by Richard Wallace, ALICE improved upon the ELIZA chatbot system by monitoring conversations. ALICE (Artificial Linguistic Internet Computer Entity) made significant progress as a chatbot, and Wallace was determined to improve iteratively upon it. Whenever ALICE was presented with an unknown/unrecognizable phrase, Wallace would add a response to ALICE so that the case could be handled in the future. This schema of iterative improvement made ALICE extremely flexible compared to older chatbots (Gisonna).

2.5.3 | Modern Chabots: SIRI, Alexa and Cortana

Siri, introduced by Apple in 2011, was one of the first voice-activated digital assistants integrated into a smartphone. Siri uses machine learning and NLP to understand user queries and execute tasks such as sending messages, setting reminders, providing navigation assistance, and more. Over the years, Apple has improved Siri's contextual understanding and integration with third-party applications, enhancing its usability (McDonough).

Amazon's Alexa, launched in 2014, became a dominant force in smart home technology. Unlike Siri, which is primarily embedded in mobile devices, Alexa operates through Amazon Echo and other smart devices, allowing users to control home automation systems, play music, check the weather, and more. Alexa's success is largely attributed to its extensibility through

third-party "skills," enabling developers to expand its functionalities beyond basic commands (Volle).

Microsoft's Cortana, released in 2014, was designed to provide productivity-focused assistance, integrating deeply with Microsoft's suite of services, including Outlook and Windows 10. Cortana aimed to enhance workplace efficiency by setting reminders, managing schedules, and providing proactive suggestions. However, due to competitive challenges, Microsoft has gradually scaled back Cortana's presence in consumer products, repositioning it towards enterprise solutions (Corden).

2.5.4 | The Rise of AI: Generative-AI-Based Chatbots

Generative AI-based chatbots, powered by large-scale models like OpenAI's GPT, Google's Bard, and Meta's Llama, have revolutionized conversational AI by enabling dynamic, human-like responses. Their ability to continuously learn and improve over time allows for more personalized and relevant interactions, making them versatile across customer service, education, healthcare, and entertainment domains. However, challenges such as ethical concerns, bias, and misinformation remain crucial research areas as these technologies evolve (Wangsa et al.). Unfortunately, Microsoft has officially discontinued Cortana as a standalone application for Windows 10 and Windows 11. Support for Cortana in these operating systems ended in late 2023 (Corden).

2.6 | Large Language Models

A large language model (LLM) is an advanced artificial intelligence (AI) system designed to interpret, generate, and manipulate human language to closely mimic natural communication. These models are built using deep learning techniques—particularly Transformer architectures—and trained on vast datasets containing text from books, articles, websites, and other sources (OpenAI Mikolov et al.). Given the complexity and variability of human language—including nuances like grammar, syntax, context, and cultural references—LLMs often require billions to trillions of parameters to function effectively. These parameters act as weighted connections in neural networks, enabling the recognition of patterns, inference of meaning, and generation of coherent responses. Unlike traditional rule-based AI models, which follow predefined instructions, or purely statistical NLP models that rely on probability-based predictions, LLMs can adapt dynamically and perform a wide variety of language-based tasks without specific programming for each task (NVIDIA).

When trained on diverse and extensive corpora, LLMs develop a broad understanding of human language, making it possible to answer questions, summarize text, translate languages, generate creative content, and assist in coding or reasoning tasks. This flexibility makes them valuable in numerous domains—from business automation and research to education and entertainment (Good Sapiens and Stanford HAI). The concept of the modern LLM can be traced back to 2001 when Yoshua Bengio first proposed using word embeddings in neural language models (Bengio et al.). However, the idea did not gain widespread popularity

until 2013 with the development of word2vec by Tomáš Mikolov and colleagues, which significantly improved contextual language representation (Mikolov et al. Toward Data Science). This shift from rule-based AI to deep learning-driven NLP set the stage for the revolutionary Transformer architecture introduced by Google in 2017, a design that underpins many contemporary LLMs (Vaswani et al.).

Shortly thereafter, Google unveiled BERT (Bidirectional Encoder Representations from Transformers) in 2018, allowing models to interpret words based on the context of both preceding and succeeding text—an advancement that markedly enhanced tasks like question answering and sentiment analysis (Devlin et al. NVIDIA). In 2020, OpenAI's GPT-3, boasting 175 billion parameters, pushed the boundaries further by introducing few-shot learning, enabling the model to generalize effectively across tasks without extensive fine-tuning (OpenAI). Three years later, GPT-4 refined these capabilities with multimodal processing, allowing it to handle text and images simultaneously while providing enhanced reasoning and factual accuracy (OpenAI). These advancements firmly established LLMs as powerful business, productivity, and research tools. LLMs excel in text generation, language translation, and conversational AI, offering context-aware responses and the ability to handle diverse, domain-specific content (Stanford HAI). They streamline workflows, enhance productivity, and assist in creative and analytical processes. However, these models are computationally expensive to train and run, and they can inadvertently reflect biases present in their training data or produce inaccurate information if prompts are ambiguous. Ethical concerns—such as misinformation, privacy risks, and potential misuse—highlight the need for responsible development and deployment. Despite their sophistication, LLMs do not possess genuine understanding or reasoning; rather, they rely on statistical pattern recognition to generate outputs (OpenAI Stanford HAI).

In summary, LLMs represent a culmination of cutting-edge innovations in deep learning and natural language processing. Their rapid evolution—from early word embedding techniques to massive Transformer-based architectures—has revolutionized how machines process human language. Yet, the field continues to grapple with significant challenges around computational demands, fairness, transparency, and ethical governance. Addressing these concerns will be essential for harnessing LLMs' full potential across education, business, healthcare, and beyond.

3

Methodology

3.1 | Introduction

CAIT was conceived as a next-generation Intelligent Tutoring System that integrates traditional teacher-designed content with the adaptive capabilities of large language models. Whereas earlier ITS solutions often rely on strictly scripted feedback loops, CAIT leverages NLP and AI-driven message classification to provide responses more closely resembling one-on-one tutoring conversations. This hybrid design enables CAIT to maintain the pedagogical rigor of teacher-crafted hints, explanations, and scaffolding while adding the flexibility needed to address diverse student inquiries in real time. At the core of CAIT's approach are four principal components: a Message Classifier, a Message Handler, a Logging System, and a Randomized Control Trial (RCT) Controller. The Message Classifier uses simple pattern-matching algorithms (for profanity filtering and other basic checks) and more advanced LLM-based classification to interpret the intent behind each student query. The Message Handler then delivers the appropriate response—whether a teacher-written hint, a scaffolded breakdown of a complex problem, or a GPT fallback message when the query falls outside predefined categories. Meanwhile, a robust Logging System captures every interaction for subsequent analysis and improvement. Finally, the RCT Controller manages controlled experiments that systematically vary instructional strategies—such as hints versus scaffolding—to measure their impact on learning outcomes. Central to the methodology is an emphasis on continuous refinement. By collecting detailed logs of student interactions and outcomes, the system can be iteratively improved through data-driven insights. Feedback loops with teachers ensure that newly generated scaffolding or GPT responses remain educationally sound, and real-time analytics guide ongoing decisions about which interventions to prioritize. This blend of structured experimentation and dynamic AI functionality sets CAIT apart, allowing it to adapt to individual learners and evolving pedagogical goals.

3.2 | Message Classifier and Handler

CAIT employs a sophisticated message classification system to effectively assist students to determine the best way to provide support. When a student submits a message, it is cate-

gorized into several predefined classifications: hints, explanations, Common Wrong Answer Feedback (CWAF), scaffolding, or GPT fallback. This structured approach ensures that students receive the most relevant guidance through teacher-written hints, step-by-step scaffolding, or AI-generated responses. By leveraging a fine-tuned LLM prompt for fallback cases, CAIT enhances learning while maintaining a structured and engaging tutoring experience.

3.2.1 | Handling Messages Based on Classification

The student's messages can be submitted to the answer box or texted to the CAIT message box. These messages must be classified so CAIT can determine how best to answer them. The classifications are hints, explanations, CWAF, scaffolding, and GPT fallback. These classifications allow CAIT to use preexisting teacher content where applicable while leveraging newfound LLMs. Based on the classification, follow a procedure to help the student in the best way possible.

3.2.2 | Intro Message

When a student first selects a problem CAIT will introduce itself with a welcoming message and greet the student. This message has gone through multiple iterations of testing to maximize emotional neutrality. Past iterations of the intro message were to "friendly" which resulted in students communication with CAIT in a non-educational way.

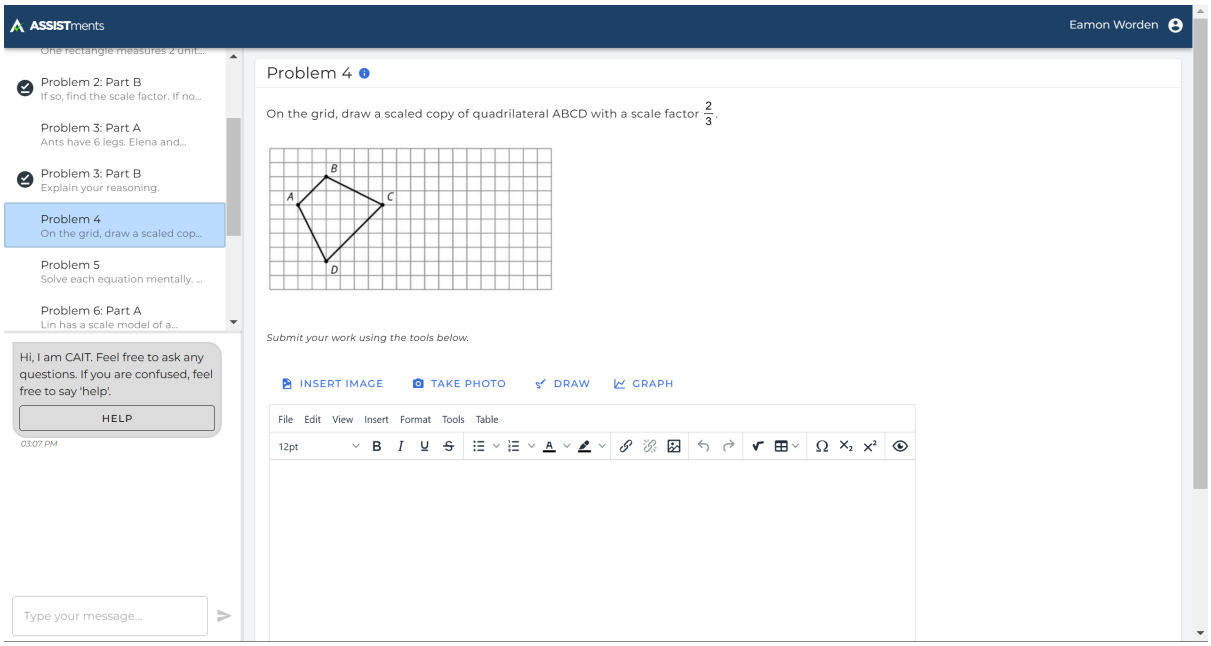


Figure 3.1: Example of a CAIT Intro Message

3.2.3 | Hints

Hints are the most basic form of support a student can get. Questions usually have multiple hints that will help the student get closer to the answer. A hint will give the student a little more information to help them solve the problem. Each hint will reduce a student's score on the problem by one-third of a point out of one. Later on, we will mention our Randomized Control Trials, to determine if hints or scaffolding help students learn better. These hints were written by teachers to try to recreate classroom learning as much as possible.

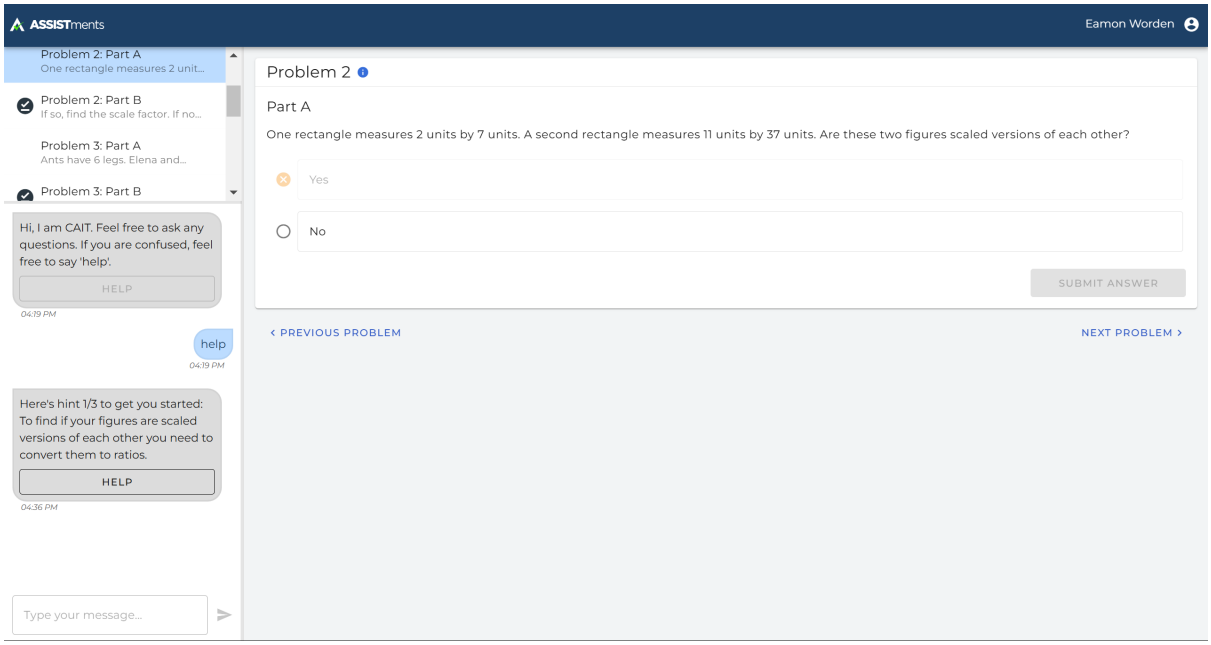


Figure 3.2: Example of a CAIT Hint

3.2.4 | Explanations

Explanations are another kind of support that gives students information about a problem. They are different from hints because they give details about how to solve the whole problem and the final answer. Explanations on the CAIT platform are authored by teachers.

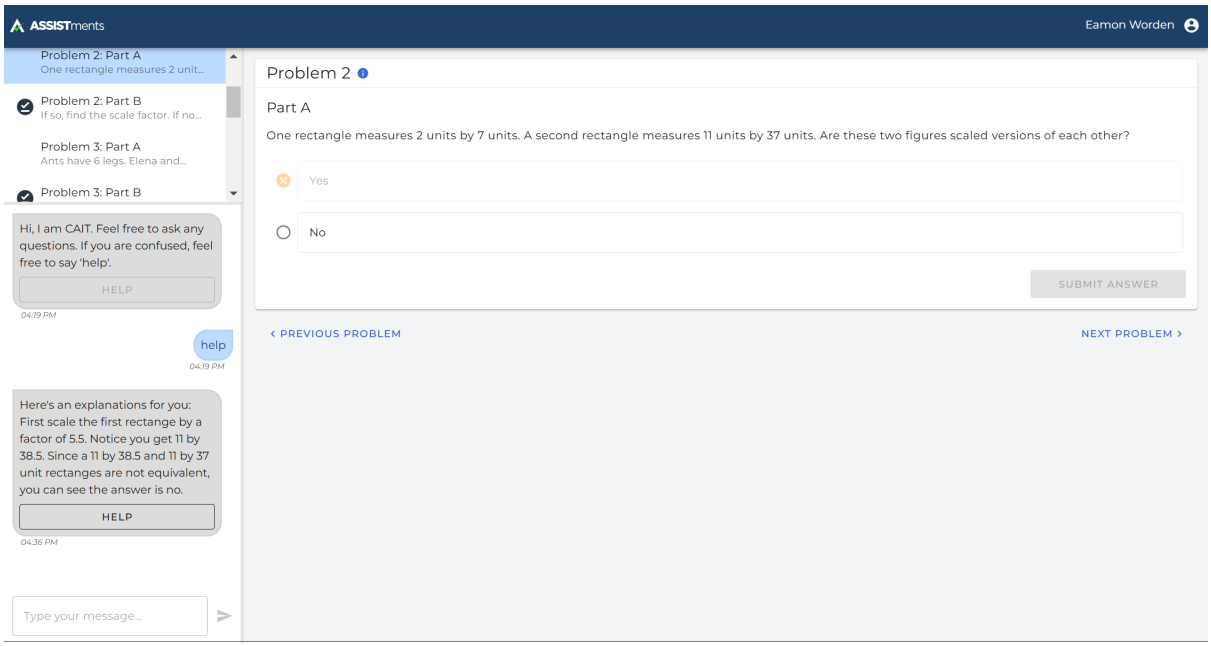


Figure 3.3: Example of a CAIT Explanation

3.2.5 | Common Wrong Answer Feedback

Common Wrong Answer Feedback (CWAF) is a support that aims to help students who enter a common wrong answer. For example, if a question is 1/2 + 1/3 and the student answers 2/7 then there would likely be CWAF for the questions. This support only activates after a student has answered a question and will give a hint to assist the problem where they went wrong so they can correct their mistake. Common wrong answer feedback on that CAIT platform is authored by teachers.

3.2.6 | Scaffolding

Scaffolding is a new support that breaks the problem down into individual steps to help "walk" the student through the problem. This helps a student break a multistep problem into subproblems, that are easier than the overall problem. If the student gets the part right then they can move on, however, if they get it wrong there are hints and explanations for the sub-problems to help the student learn. This way we can teach the student the steps of the problem and the order to take them to solve the original question. The scaffolding is generated by GPT, but teachers have approved it.

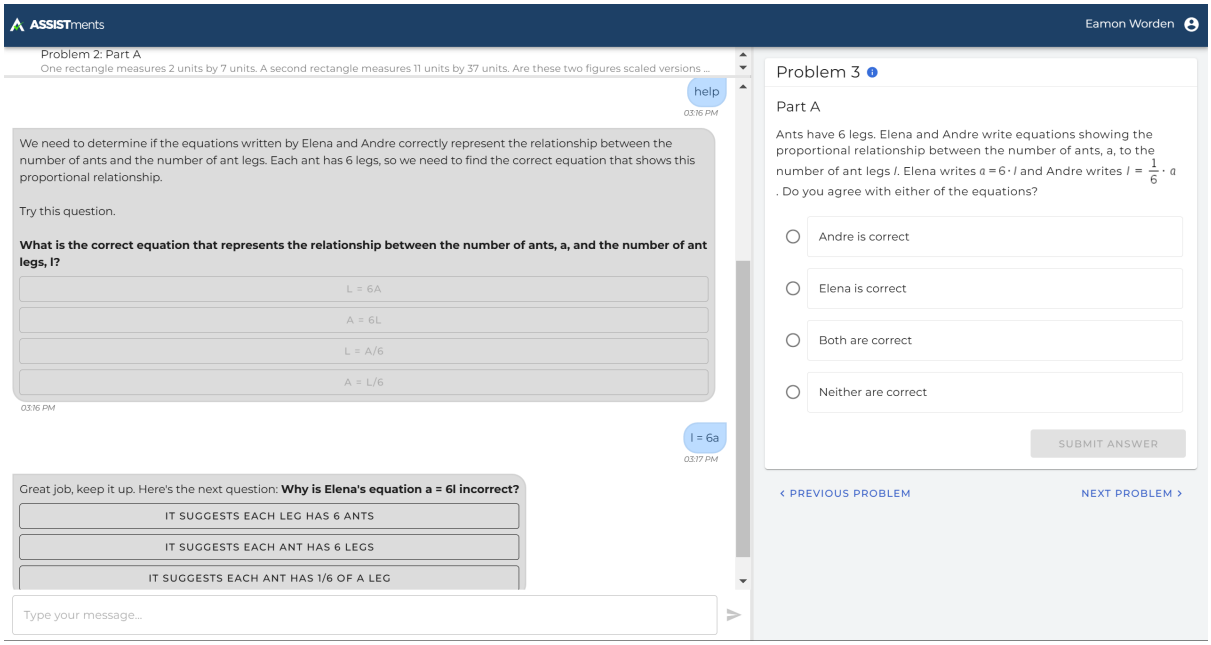


Figure 3.4: Example of a CAIT Scaffolding Support

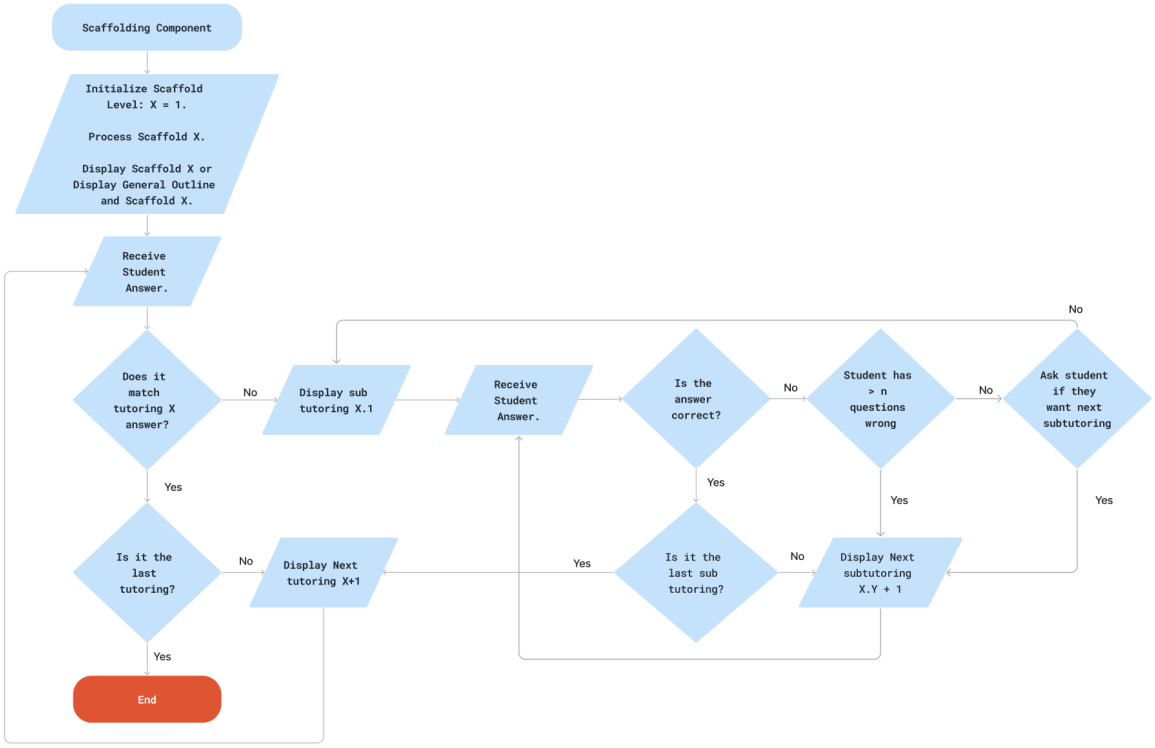


Figure 3.5: Scaffolding Diagram

This is the flowchart of how scaffolding works. It starts with showing either the first scaffolding question or the first question with a general outline. This is a RCT to determine if the general outline helps. Then the action goes to the student, and the student responds to the question. If the student gets the question right and continues to get them right they will never go into sub-tutoring. If a student gets a question wrong the enter the sub-tutoring cycle. CAIT will ask a sub-question and if the student gets the question wrong again CAIT will ask if they want the next sub-tutoring. We learned it is better for a student to ask for help instead of force help upon them, however there is a limit to the amount of times a student can get a sub-tutoring wrong before we end up giving them the solution. We do not want them to be stuck in this loop forever. Scaffolding ends when the student completes all tutoring questions, and the last question's answer will be the answer to the main question they are trying to solve.

3.2.7 | GPT/LLM Fall Back

When the message classifier fails to interpret a message and any of the above categories it is assigned to "GPT/LLM" and is used to give a default handling of the message. This classification aims to improve student learning by sending a tailored prompt to GPT for a unique response that can resemble a human tutor as much as possible.

A very fine tuned prompt is needed to direct the LLM in its task. Below is the current prompt that is used. It starts by setting the frame as telling the LLM its goal and the the current problem as well as the answer. It then gives the LLM ten rules to follow to direct the response. The first five responses have to do with the content of the answer. The first rule is to provide small hints and to guide the student without giving them the answer.

The specific prompt used is:

```
You are CAIT (Conversational AI Tutor), a tutorbot designed to help students learn.
A student is currently working on the following problem:

    Problem: {problem}

    Correct Answer: {answer}

    Please respond to the student appropriately. Your response should:

    - Provide small hints if needed to guide the student without giving away the answer.
    - Clarify terms or concepts if the student seems unsure.
    - Address any misconceptions the student may have.
    - If the student's question or answer is unclear, ask them to clarify their question.
    - Acknowledge if the student is on the right track and provide encouragement.
    - Keep your message concise, aiming for under 20 words.
    - Respond in plain text with no HTML formatting.
    - Redirect the student's attention if it is off topic, such as requesting responses to be
    - NEVER give away the answer

    If you cannot understand the student's question or response, reply with:

    "I'm sorry, but I cannot understand your question."
    This might be because the problem includes an image or the student's input is unclear."
```

This prompt also includes the conversation the student was having before this call was issued.

3.3 | Logging

For each message sent between a user all measurable system and user statistics are recorded. This information is then stored in a SQL database. A survey, sent at the completion of each question if CAIT is used, is also logged.

- **Assignment Xref**: Used to track the current assignment a student is working on
- **Problem Ceri**: Used to track the current problem in an assignment a student is working on
- **Student Xref**: Used to identify the student
- **Student Response**: Used to track student messages inside CAIT
- **CAIT Response**: Used to track CAIT's response to a student response
- **CAIT Veresion**: Used to track the current working version of CAIT
- **Selection Criteria**: Used to identify the message classification technique
- **Time Sent**: Used to identify when a user sent a request
- **Time Received**: Used to identify when CAIT responded
- **Experiment**: Used for randomized control trials (RCT)

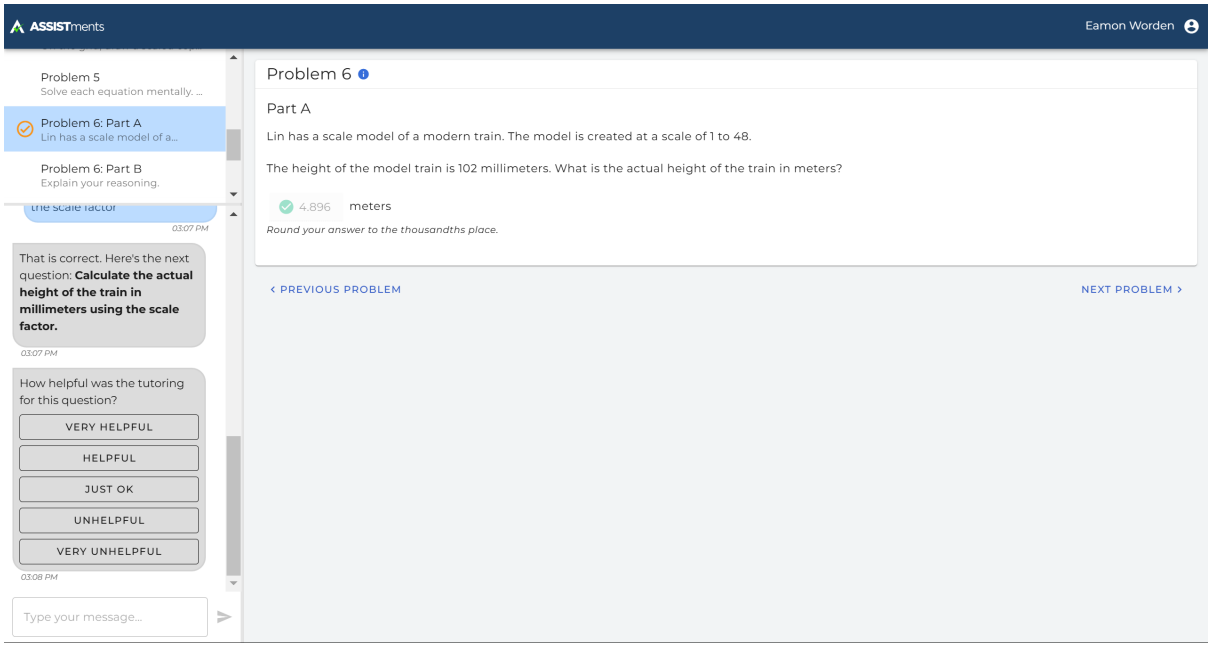


Figure 3.6: Example of a CAIT Survey

3.4 | Randomized Control Trials

ASSISTments integrates Randomized Control Trials (RCTs) into CAIT to systematically evaluate the effectiveness of different instructional supports. In these trials, students can be randomly assigned various types of support within CAIT, such as hints, step-by-step scaffolding, or additional similar but not the same practice problems. This randomization ensures unbiased comparisons between different interventions, allowing for a data-driven assessment of student learning outcomes. Additionally, CAIT introduces an element of randomness in its responses, particularly when students request help. For instance, when a student asks for assistance, CAIT provides scaffolding 7 out of 8 times if it is available, as previous research suggests it is more effective than traditional hints and explanations. Student progress is then tracked to measure the impact of each support on learning outcomes. Additionally, some students may receive surveys to assess their experiences and perceptions of the system. ASSISTments goal to make teaching and learning better through evidence-based claims is supported by leveraging RCTs. By utilizing RCTs, ASSISTments ensures that CAIT continuously improves based on data-driven insights, optimizing the way it enhances student learning

4

Results, Testing, and Future Work

4.1 | Conclusion

To assess the effectiveness and usability of CAIT an evaluation methodology was implemented. This approach aimed to measure CAIT's impact on both teachers and students by examining its integration, usability, and effectiveness in real-world learning environments. The evaluation process was designed to capture both qualitative and quantitative insights through a combination of structured testing, student interactions, teacher comments, and data analysis. By gathering feedback from educators and learners, we aimed to refine CAIT's capabilities and ensure its alignment with classroom needs. The following sections outline the testing procedures, results, and future improvements based on these evaluations.

4.2 | Testing

CAIT was deployed in a controlled environment for evaluation with teachers and students. The testing process included the following stages

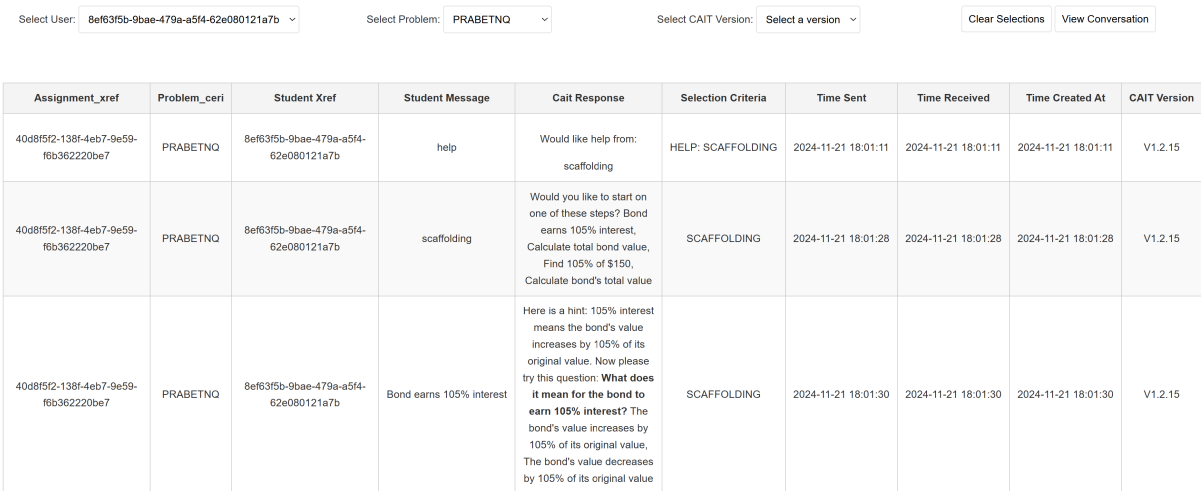
- **Teacher Testing**: Teachers were introduced to CAIT and trained on its features and capabilities to assess its integration within ASSISTments. After training, they used CAIT within ASSISTments lessons, noting its usability, effectiveness, and classroom integration.

  One-on-one interviews were also conducted between the development team and the teachers. During these interviews, teachers were asked to simulate student interactions, evaluate system responses and robustness, and give overall feedback about the system. Their feedback ensured continuous iterative improvements based on real classroom needs.

- **Student Interaction**: Under the supervision of the integrated teacher, the students began using CAIT in real-world learning scenarios. Students were instructed to complete their work normally with the option to use the CAIT system, and were not forced to use it. The system guided them through various questions, offering feedback and support. Then, these interactions were recorded to analyze effectiveness, engagement, and accuracy.

4.3 | Results

A key part of assessing CAIT's effectiveness was a proprietary conversation analyzer that utilized the raw database logs from test runs to provide actionable insights. By examining each message exchanged between students and CAIT–along with information on hints, scaffolding, GPT fallback usage, and correctness–the analyzer provided a snapshot of student progress and engagement. These reports were used to identify how often students asked for help, which hint formats were most effective, and how consistently students reached correct answers.



Assignment_xref	Problem_ceri	Student Xref	Student Message	Cait Response	Selection Criteria	Time Sent	Time Received	Time Created At	CAIT Version
40d8f5f2-138f-4eb7-9e59-f6b362220be7	PRABETNQ	8ef63f5b-9bae-479a-a5f4-62e080121a7b	help	Would like help from: scaffolding	HELP: SCAFFOLDING	2024-11-21 18:01:11	2024-11-21 18:01:11	2024-11-21 18:01:11	V1.2.15
40d8f5f2-138f-4eb7-9e59-f6b362220be7	PRABETNQ	8ef63f5b-9bae-479a-a5f4-62e080121a7b	scaffolding	Would you like to start on one of these steps? Bond earns 105% interest, Calculate total bond value, Find 105% of $150, Calculate bond's total value	SCAFFOLDING	2024-11-21 18:01:28	2024-11-21 18:01:28	2024-11-21 18:01:28	V1.2.15
40d8f5f2-138f-4eb7-9e59-f6b362220be7	PRABETNQ	8ef63f5b-9bae-479a-a5f4-62e080121a7b	Bond earns 105% interest	Here is a hint: 105% interest means the bond's value increases by 105% of its original value. Now please try this question: **What does it mean for the bond to earn 105% interest?** The bond's value increases by 105% of its original value, The bond's value decreases by 105% of its original value	SCAFFOLDING	2024-11-21 18:01:30	2024-11-21 18:01:30	2024-11-21 18:01:30	V1.2.15

Figure 4.1: Conversation analyzer database view

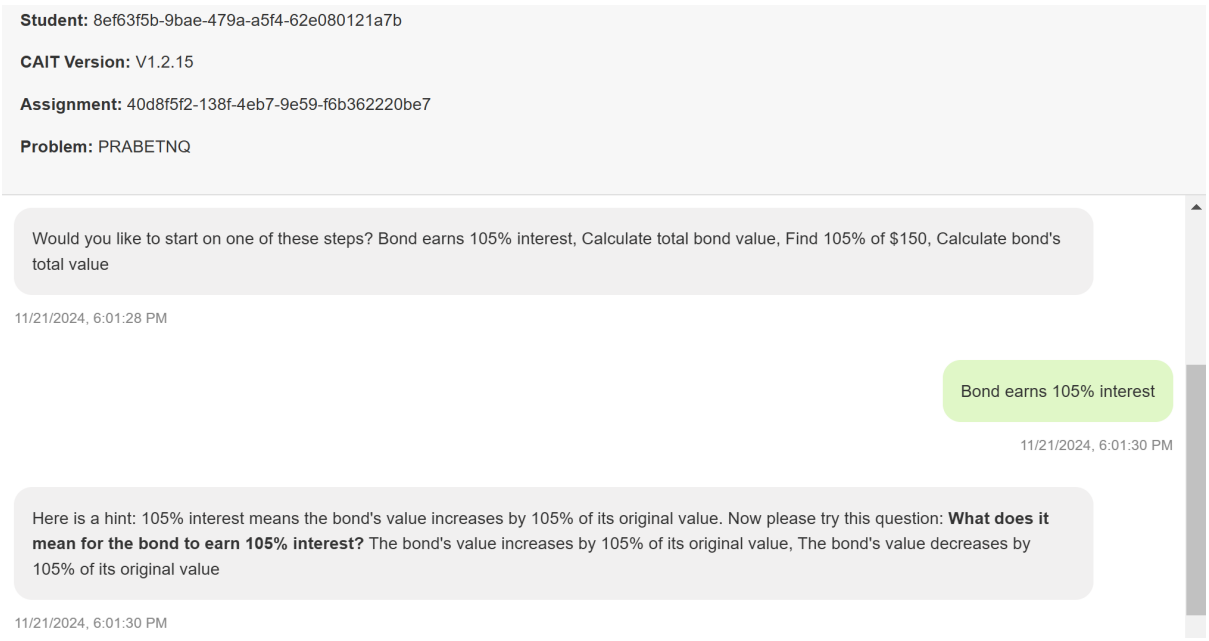


Figure 4.2: Conversation analyzer chat view

4.4 | Future Work

ASSISTments is founded on the idea of continuous research to determine the best ways to help students learn. To continue on this goal it is important to improve on the work we have started. The conversation analyzer is an important tool for judging the effectiveness of CAIT, but currently is inefficient. This is in part due to problems with logging our data. The scores of the students on each problem are not logged properly, which makes it more difficult to judge if a student is learning better with the help of CAIT. By improving logging and collecting more data we can more accurately judge the usefulness of CAIT. The outcome of the RCTs that were implemented are unclear due to the lack of data to analyze. CAIT is still in its early stages and has only been tested a few times in real classrooms. Most of the current conversations are from testers rather than actual students. With more use in the future the results of the currently implemented RCTs will be more clear. They could also lead the way to more RCTs being developed in the future to further research other parts of CAIT.

5

References

1. Gobiet, Marie. "The History Of Chatbots – From ELIZA to ChatGPT". *Onlim* 15 Feb. 2024 https://onlim.com/en/the-history-of-chatbots/

2. Gisonna, Nicholas. "Chatbot". *Britannica* 3 Mar. 2025 https://www.britannica.com/topic/chatbot#ref1325159

3. McDonough, Michael. "Siri". *Britannica* n.d https://www.britannica.com/technology/Siri

4. Volle, Adam. "Amazon Alexa". *Britannica* 27 Feb. 2025 https://www.britannica.com/technology/Amazon-Alexa

5. Corden, Jez. "A brief history of Cortana, Microsoft's trusty digital assistant". *Windows Central* 24 Apr. 2017 https://www.windowscentral.com/history-cortana-microsofts-digital-assistant

6. Wangsa, Ketmanto, Karim, Shakir, Gide, Ergun, Elkhodr, Mahmoud. "A Systematic Review and Comprehensive Analysis of Pioneering AI Chatbot Models from Education to Healthcare: ChatGPT, Bard, Llama, Ernie and Grok". *Windows Central* 22 June 2024 https://www.mdpi.com/1999-5903/16/7/219?utm_source=chatgpt.com

7. Heffernan, Neil T., and Cristina L. Heffernan. "The ASSISTments Ecosystem: Building a Platform That Brings Scientists and Teachers Together for Minimally Invasive Research on Human Learning and Teaching." *International Journal of Artificial Intelligence in Education*, vol. 24, 2014, pp. 470–497. Springer, https://link.springer.com/article/10.1007/s40593-014-0024-x.

8. Heffernan, Neil T., and Kenneth R. Koedinger. "An Intelligent Tutoring System Incorporating a Model of an Experienced Human Tutor." *Computers & Education*, vol. 49, no. 4, 2002, pp. 1037–1063. Elsevier, https://www.sciencedirect.com/topics/computer-science/intelligent-tutoring-system.

9. Koedinger, Kenneth R., and Vincent Aleven. "Exploring the Assistance Dilemma in Experiments with Cognitive Tutors." *Artificial Intelligence in Education: Building Technology-Rich Learning Contexts That Work*, edited by R. Luckin et al., Springer, 2007, pp. 215–222, https://link.springer.com/chapter/10.1007/11774303_7.

10. McDonough, John. *Siri: Apple's Voice-Activated Digital Assistant.* Britannica, 2025, https://www.britannica.c

11. NVIDIA. "BERT: Bidirectional Encoder Representations from Transformers." *NVIDIA Glossary*, https://www.nvidia.com/en-us/glossary/bert/.

12. OpenAI. "Better Language Models and Their Implications." *OpenAI Research*, 2019, https://openai.com/res language-models.

13. OpenAI. "GPT-4: The Latest in Generative AI Technology." *OpenAI Research*, 2023, https://openai.com/rese 4.

14. ScienceDirect. "Intelligent Tutoring System." *ScienceDirect*, n.d., https://www.sciencedirect.com/topics/co science/intelligent-tutoring-system.

15. Stanford HAI. "Ethical Concerns with Large Language Models." *Stanford Institute for Human-Centered AI*, n.d., https://hai.stanford.edu/news/ethical-concerns-large-language-models.

16. VanLehn, Kurt. "The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems." *Educational Psychologist*, vol. 46, no. 4, 2011, pp. 197–221, https://www.tandfonline.com/doi/full/10.1080/00461520.2011.611369.

17. Volle, Erik. *Amazon Alexa: Revolutionizing Smart Home Technology.* Britannica, 2025, https://www.britannica Alexa.

18. Windows Central. "The History of Cortana: Microsoft's Digital Assistant." *Windows Central*, 2017, https://www.windowscentral.com/history-cortana-microsofts-digital-assistant.

19. Mikolov, Tomáš, et al. "Efficient Estimation of Word Representations in Vector Space." *arXiv*, 2013, https://arxiv.org/abs/1301.3781.

20. Good Sapiens. "How LLMs Are Changing the World." *Medium*, 2023, https://goodsapiens.medium.com/h llms-are-changing-the-world-01463f23e3c5.

21. Toward Data Science. "Uncovering the Pioneering Journey of Word2Vec and the State of AI Science: An In-Depth Interview." *Medium*, n.d., https://towardsdatascience.com/uncovering-the-pioneering-journey-of-word2vec-and-the-state-of-ai-science-an-in-depth-interview-fbca93d8f4ff.