

From Passive Watching to Active Learning: Empowering Proactive Participation in Digital Classrooms with AI Video Assistant

Anna Bodonhelyi
Chair of Human-Centered
Technologies for Learning, TUM
Munich, Germany
anna.bodonhelyi@tum.de

Enkeleda Thaqi
Chair of Human-Centered
Technologies for Learning, TUM
Munich, Germany
enkeleda.thaqi@tum.de

Süleyman Özdel
Chair of Human-Centered
Technologies for Learning, TUM
Munich, Germany
ozdelsuleyman@tum.de

Efe Bozkir
Chair of Human-Centered
Technologies for Learning, TUM
Munich, Germany
efe.bozkir@tum.de

Enkelejda Kasneci
Chair of Human-Centered
Technologies for Learning, TUM
Munich, Germany
enkelejda.kasneci@tum.de

Abstract

In online education, innovative tools are crucial for enhancing learning outcomes. SAM (Study with AI Mentor) is an advanced platform that integrates educational videos with a context-aware chat interface powered by large language models. SAM encourages students to ask questions and explore unclear concepts in real time, offering personalized, context-specific assistance, including explanations of formulas, slides, and images. We evaluated SAM in two studies: one with 25 university students and another with 80 crowdsourced participants, using pre- and post-knowledge tests to compare a group using SAM and a control group. The results demonstrated that SAM users achieved greater knowledge gains specifically for younger learners and individuals in flexible working environments, such as students, supported by a 97.6% accuracy rate in the chatbot's responses. Participants also provided positive feedback on SAM's usability and effectiveness. SAM's proactive approach to learning not only enhances learning outcomes but also empowers students to take full ownership of their educational experience, representing a promising future direction for online learning tools.

CCS Concepts

• **Human-centered computing** → **User interface design; User studies; Usability testing; Web-based interaction;** • **Social and professional topics** → **Adult education.**

Keywords

E-Learning, Real-Time Assistant, AI tutor, ChatGPT, User Study

© Owner/Author | ACM 2025. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in CHI Conference on Human Factors in Computing Systems (CHI '25), *CHI '25, April 26-May 1, 2025, Yokohama, Japan*. <https://doi.org/10.1145/3706598.3713513>.

Please use nonacm option or ACM Engage class to enable CC licenses. This work is licensed under a Creative Commons Attribution 4.0 International License. *CHI '25, April 26-May 1, 2025, Yokohama, Japan*
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1394-1/25/04
<https://doi.org/10.1145/3706598.3713513>

ACM Reference Format:

Anna Bodonhelyi, Enkeleda Thaqi, Süleyman Özdel, Efe Bozkir, and Enkelejda Kasneci. 2025. From Passive Watching to Active Learning: Empowering Proactive Participation in Digital Classrooms with AI Video Assistant. In *CHI Conference on Human Factors in Computing Systems (CHI '25)*, April 26-May 1, 2025, Yokohama, Japan. ACM, New York, NY, USA, 21 pages. <https://doi.org/10.1145/3706598.3713513>

1 Introduction

Chatbots have increasingly become valuable tools in educational settings, offering significant enhancements to the learning experience and outcomes [35, 40]. Research consistently shows that these artificial intelligence (AI)-driven tools can increase student engagement, motivation, and knowledge retention by providing personalized support and immediate feedback [8, 23, 40, 49, 72, 78]. Chatbots have been effectively used to deliver course materials, assist with assignments [68], and foster interactive learning environments [35, 40], leading to improved academic performances [49] across various disciplines. However, these tools are not without their limitations. Common drawbacks include the potential for contextually irrelevant or inaccurate responses [25], limited ability to handle complex queries [81], and reliance on pre-programmed content [54], which can restrict the depth and adaptability of interactions.

Building on the growing influence of AI in education [38], our tool, SAM (Study with AI Mentor), significantly enhances the learning experience. SAM is an innovative and context-aware platform designed to assist learners in grasping complex topics through interactive, AI-driven mentoring. SAM is built to function as an intelligent video-watching platform where users can watch educational videos, particularly lectures, and engage with an AI mentor to ask questions related to the content. This real-time assistance aims to deepen students' comprehension by providing immediate, context-relevant feedback, thereby simulating an interactive learning environment. Furthermore, SAM builds on existing literature that demonstrates how asking questions while studying leads to increased engagement and, consequently, improved learning outcomes [14, 62, 80]. By fostering this proactive behavior, SAM contributes to a more effective and engaging learning process.

The core functionality of SAM revolves around its ability to integrate multiple sources of information. Users can provide a link to the lecture video and optionally upload the associated slides, allowing SAM to rely on both the video transcript and the slides to deliver more accurate and informative responses. As the video progresses, users can interact with the AI mentor by asking questions directly related to the content they view. We designed SAM to be responsive, recognizing specific references to slides or concepts and generating detailed explanations. Additionally, SAM employs LaTeX for rendering complex mathematical formulas, ensuring that all visual elements are presented clearly and accurately, which is essential for subjects that rely heavily on precise notation. Upon completing the video, users could download a PDF summarizing their questions and the AI's responses, providing a record of their learning process.

For the purpose of the user study, we tailored SAM to support a crowdsourced user study where participants engaged with a pre-selected lecture video. To this end, the platform featured a simplified interface, allowing participants to start the already uploaded video, ensuring uniform interaction with the content and features across the study. In this study, we focused on measuring the knowledge gain to provide insights into how a dynamic and real-time chatbot assistant like SAM can enhance learning outcomes. By examining the impact of SAM on participants' comprehension and retention, we aimed to better understand the effectiveness of integrating AI-driven, real-time assistance into educational environments.

While chatbots have significantly advanced education, critical gaps remain, particularly in fields requiring complex, context-specific explanations that SAM specifically addresses. SAM represents a new way of learning, where learners take full ownership of their educational journey. By encouraging proactive behavior, SAM not only enhances engagement—a factor already shown to be effective in previous studies [14, 62, 80]—but also fosters a more interactive and responsive learning environment. Through our user study with SAM, we provide valuable insights into future online learning techniques, offering guidance on how to design tools that trigger and sustain interactive learning experiences. The main contributions of this work are as follows:

- Development of SAM, an AI-driven educational tool that enhances learning by integrating real-time question-answering with video content, supporting both text and image-based queries, and accurately rendering complex mathematical formulas.
- Comprehensive evaluation of SAM's impact on knowledge gain and user satisfaction through crowdsourcing, highlighting its effectiveness and positive effect. The version, which was used to this study is available at <https://gitlab.lrz.de/hcti/sam>.
- Identification of key demographic subgroups, such as students and younger users, that demonstrated the highest knowledge gains when using SAM, providing valuable insights for tailoring future educational tools.

2 Related Work

AI has been significantly transforming education by enabling personalized learning experiences, enhancing student engagement,

and optimizing administrative tasks [8, 52]. AI technologies, such as machine learning and natural language processing, allow for tailored educational approaches that cater to individual learning styles and paces, potentially leading to improved student outcomes [33]. Recent studies further highlight the potential of tools like ChatGPT to support education by enabling widespread usage, though careful consideration of student backgrounds and concerns is essential for effective implementation [71]. This section highlights key advancements relevant to SAM, examining online learning tools, chatbots, virtual tutors, their overall impact on learning outcomes, and the relationship between student engagement and question-asking in classrooms.

2.1 Overview of Online Learning Tools

Online learning tools have revolutionized education, providing flexible and accessible options for students and educators worldwide. These digital platforms and software applications facilitate remote learning, content creation, student engagement, and assessment in virtual environments [2, 46]. They also have a crucial role in modern education, especially in the context of engineering disciplines [69], science education [57], language learning [24], and overall digital transformation of education [7]. These tools, ranging from digital collaboration platforms like Zoom [29, 47], Microsoft Teams [55], and Google Classroom [7] to interactive video-based learning and simulation tools [26], have been shown to enhance student engagement, motivation, and learning outcomes in various educational settings [75]. By incorporating features like learner-centered approaches, fast feedback mechanisms, and modern technologies, online learning tools not only improve the quality of education but also make the learning process more efficient for students [51]. Additionally, the use of various approaches and tools like gamification, simulation, and virtual reality have been experimented with to enhance knowledge dissemination and learning experiences [67]. As technology continues to evolve, further research is needed to explore the full potential of these tools in diverse educational contexts and student populations.

One of the most notable trends in online education is the rise of Massive Open Online Courses (MOOCs), which offer high-quality courses for mass delivery through the Internet [19, 20]. Initially coined in 2008 [21, 44], MOOCs have gained popularity globally, attracting learners interested in experiencing top-tier university teaching [44]. These courses have disrupted conventional educational models by offering free, accessible, and diverse learning opportunities to a broad audience, challenging the economic and pedagogical paradigms of traditional on-campus universities [20]. With over 11,000 MOOCs available worldwide as of 2022 [30], these courses have revolutionized education delivery, reaching thousands of students and prompting discussions on their impact on international communication and learning models. Platforms such as Coursera [5], edX [60], and Udacity [4] have expanded access to education by offering courses from top universities and institutions to a global audience [74], often for free or at a low cost. MOOCs cover many subjects and provide learners with flexibility and accessibility, accommodating diverse schedules and learning paces. These platforms also incorporate interactive elements such as peer

assessments, discussion forums, and real-time feedback to enhance the learning experience.

In addition to MOOCs, video-based learning platforms have also gained popularity by offering benefits such as enhanced student engagement and motivation, flexibility and accessibility for learners, and support for teacher professional development [58]. Additionally, using video-based student support platforms, like ORBITS [65] predictive engine, has positively enhanced student learning experiences through AI techniques. Platforms such as Panopto [22] and Kaltura [27] provide comprehensive video hosting and streaming services, which include features like interactive quizzes and analytics to track learner engagement and progress. These tools are designed to cater to various learning styles by incorporating visual and auditory elements, making complex subjects more accessible and engaging for students [66]. While these platforms offer valuable features, SAM goes a step further by providing real-time, personalized support during learning sessions, enabling students to ask questions and receive immediate, context-specific feedback. This level of interactivity helps bridge knowledge gaps more effectively than traditional video-based platforms, enhancing the overall learning experience.

2.2 Chatbots and Virtual Tutors in Learning

Chatbots are increasingly recognized for their transformative educational potential, leveraging AI to enhance learning experiences. They offer personalized and continuous support, which can significantly improve student engagement and motivation, particularly in e-learning environments where traditional interaction may be limited [8, 23, 40, 49, 72]. They can enhance learning by providing course materials [17], assignments and practice questions [68], while also engaging students individually [35] or facilitating group learning activities [40, 73]. Chatbots usually engage learners by utilizing text, speech, visuals, touch, and gestures to support them in different educational activities [40]. Previous research indicates that chatbots can foster higher-order cognitive skills and reduce dropout rates by adapting to individual student needs [49]. A recent study [40] indicates that many previous works have predominantly utilized chatbots in the fields of computer science and language education. In these contexts, chatbots can assume various roles to enhance learning, most commonly serving as teaching [16, 70] or peer agents [18, 43], but also functioning as improving [42] or motivational agents [40, 61, 78]. Teaching agents employ a variety of approaches to enhance learning. These methods include recommending personalized tutorials, initiating discussions based on educational videos or past learning, presenting scenario-based exercises, and providing assessments through multiple-choice questions. On the other hand, peer agent chatbots primarily support students by offering on-demand assistance, including answering specific questions, clarifying terms or concepts, providing guidance on using certain technologies, and facilitating group discussions [40].

Several chatbot implementations highlight these roles, showcasing how chatbots can effectively address diverse educational needs across various subjects. For instance, the FAQ chatbot [32] assists students in MOOC classes by answering administrative questions, leveraging a pre-trained large language model (LLM) based on

information from a FAQ website. Similar to the FAQ chatbot [32], Cipherbot [59] was developed to answer students' questions based on educational materials that were uploaded by educators to minimize hallucinations. Powered by GPT-3.5 Turbo, Cipherbot achieved an 87.0% accuracy rate in its evaluation study. Both tools are designed to respond to course- or material-specific questions; however, while both utilize externally validated sources, their potential for a close integration with a video player remains to be investigated. To bridge this gap, Sara, the lecturer [77], offers scaffolding support during lecture viewing by prompting students with questions and exercises at predefined points within the video. Sara guides learners through these tasks, providing scaffolds until they can solve the posed problems independently. This tool prioritizes engaging learners through inquiry and active questioning rather than directly answering user queries [77]. Unlike these tools, we designed our framework to provide real-time support during lecture viewing by responding to any student question with high-quality, accurate answers related to the lecture content, focusing specifically on seamless integration with the lecture material to enhance understanding.

On the other hand, there are notable challenges associated with using chatbots. Their integration also raises ethical concerns, including issues of privacy, data manipulation, and the potential for academic dishonesty, particularly with tools like ChatGPT [1, 45, 79]. Research indicates that user privacy is a significant issue, with self-disclosure being a primary concern due to the sensitive data shared during interactions with chatbots [28, 37]. Furthermore, the misuse of chatbots can lead to the dissemination of incorrect or harmful safety-related information, emphasizing the need for expert verification and ethical safeguards [45, 82]. Additionally, as LLMs like GPT-4 evolve, challenges in accurately responding to specific user intents emphasize the need for improvements in intent recognition to enhance user satisfaction and ensure more reliable interactions [10]. Overall, while chatbots offer substantial benefits, these concerns highlight the necessity for responsible usage and regulatory frameworks to mitigate risks associated with privacy, ethics, and information accuracy.

While chatbots can facilitate language learning and stimulate curiosity through project-driven tasks, their reliance on extensive knowledge bases poses challenges in development and implementation [23], often leading to difficulties in handling complex messages effectively [81]. This dependency can lead to instances of chatbot fallibility, where the information provided may be inaccurate or misleading [25]. Such reliability issues can undermine the learning experience, necessitating careful oversight and regular updates to ensure that chatbots remain accurate and trustworthy educational tools [56]. Educators must be aware of these limitations and supplement chatbot interactions with human guidance to maintain the quality of instruction. To address these challenges, SAM focuses on increasing answer accuracy by providing context-aware responses. By integrating real-time feedback with relevant contextual information, SAM aims to minimize errors and enhance the reliability of its educational support.

2.3 Engagement Through Questions

Learner engagement, which refers to the investment of students' cognitive and emotional effort in completing a learning task [41],

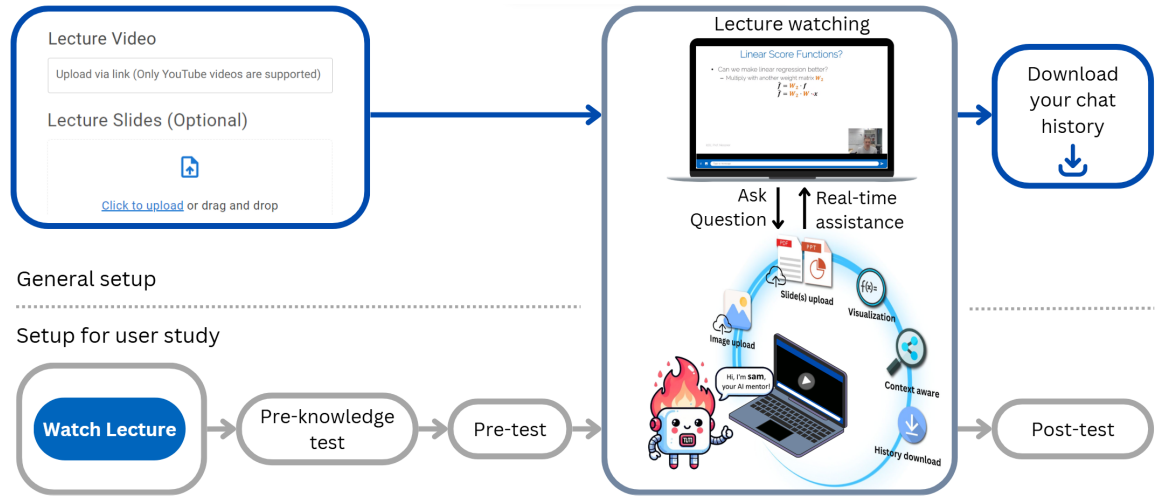


Figure 1: Structure of SAM, with modification for the user study.

has been shown to positively influence key educational outcomes such as academic performance, persistence, satisfaction, and a sense of community [31, 76]. In online settings, engagement often encompasses not only traditional indicators like contributing to discussions and asking questions [14, 80], but also involves consistent participation in virtual activities such as live chats [34], quizzes [53], and collaborative tasks [39]. In online learning, where distractions can be more common [13], engagement becomes crucial to ensure students remain focused, driving better learning outcomes. Effective online engagement also involves leveraging digital tools that promote interaction and personalized learning experiences. AI-driven approaches, such as automated classroom observation protocols, can provide teachers with specific feedback on their practices, streamlining the assessment process while ensuring consistent and valuable insights [36].

Proactive behaviors, such as students actively asking questions, serve as key indicators of engagement in online learning environments [14, 62, 80]. When students take the initiative to ask about course content, it reflects a high level of cognitive engagement, pointing out that they are not passively absorbing information but critically processing it. Questioning is an essential component of active learning, as it demonstrates curiosity and a desire to deepen understanding, driving students to connect new knowledge with prior learning [11, 12].

Taking all into account, SAM aligns with and extends theoretical perspectives on active and interactive learning, particularly as articulated in the ICAP framework [15]. This framework posits that learning outcomes improve progressively from passive to active, constructive, and interactive engagement levels. By integrating real-time and context-aware interactions into video-based learning, SAM elevates the traditional passive consumption of educational content to an interactive learning experience. Unlike previous approaches that rely on pre-scripted prompts or rigidly structured content, SAM’s ability to respond dynamically to user queries fosters higher cognitive engagement and deepens comprehension [14, 62, 80]. This interactive paradigm bridges a critical gap

in the literature by demonstrating how AI-driven pedagogical tools can operationalize theoretical constructs of interactive learning in a scalable, real-world application.

3 SAM Design and Implementation

SAM is an innovative tool designed to enhance educational experiences by integrating AI-driven assistance with video-based learning. SAM operates as an advanced video-watching platform, enabling users to view YouTube lectures while interacting with an AI mentor to pose questions related to the content (Figure 1). The integrated AI mentor is powered by GPT-4o [48], a highly advanced language model that offers sophisticated natural language processing capabilities. The AI mentor not only supports the instant clarification of doubts but also promotes a deeper understanding of complex topics by delivering explanations tailored to learners’ needs. This dynamic and personalized learning experience ultimately helps students grasp the subject matter more effectively, enhancing their overall educational journey.

SAM actively fosters interactive learning, aligning with the principles highlighted in the ICAP framework [15], which identifies interactivity as the most effective level for enhancing learning compared to constructive, active, or passive approaches, thereby ensuring an optimal learning experience. The user interface design of SAM prioritizes simplicity and aesthetic appeal, keeping in mind previous findings that users tend to favor interfaces that are visually attractive and easy to navigate over overly complex designs [63]. Research suggests a strong positive relationship between user satisfaction and key factors such as learnability, visibility, and interactivity [63]. In SAM, the chat window is strategically placed at the bottom of the interface, differing from the conventional right-side placement commonly seen in streaming platforms. This bottom placement allows for a resizable chat window that partially overlaps with the video. However, to enhance user focus, the video automatically pauses when users begin typing, enabling them to engage more effectively with the AI tutor and view detailed responses, including formulas, without distraction.

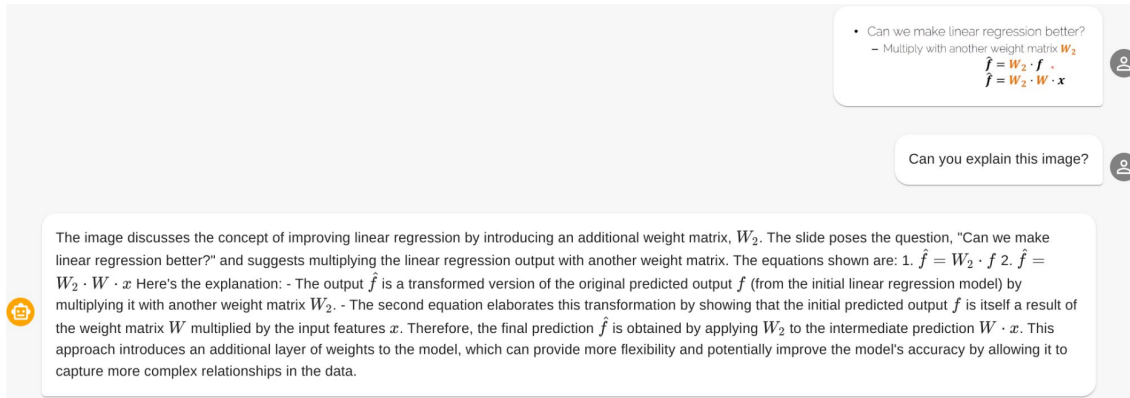


Figure 2: Example of SAM in action: user submission of an image and formula visualization.

The user experience with SAM begins when a lecture link from YouTube is provided. Users may also upload corresponding slides with the possibility to also upload previous slides or relevant related materials, which enriches the AI mentor's ability to deliver accurate and context-aware responses by leveraging both the video transcript and the additional slide information. To achieve highly accurate answers and only answer lecture-related questions, the following prompt was used:

You are roleplaying as an assistant teacher helping students understand their lecture content. Answer their questions based on the video, video transcripts, and slides. If a question is about a lecture unrelated topic, respond with "Please focus on the lecture material." If you don't know the answer, just say that you don't know. Do not attempt to fabricate an answer.

During video playback, users can engage the AI mentor by entering queries in a designated text field and the AI mentor formulates responses to the user inquiries. When a user submits a message, the lecture slides, transcript, the 30 seconds of the transcript before and after the current timestamp, and the chat history are sent to the LLM. The video itself is not forwarded. The timely relevant parts of the transcript are also provided, offering additional context to the embedded language model. Additionally, SAM allows users to upload images (e.g., screenshots from the slides), further aiding the LLM in providing precise and contextually relevant answers (Figure 2). SAM is designed with features to enhance the learning process, including the automatic transmission of the current slide whenever the word "slide" appears in a user's message. For instance, if a user asks, "Explain this slide!" the slide is sent as an image along with the query to the AI mentor, significantly increasing the accuracy and relevance of the response. Prior to the studies, we specifically tested this feature and found that including the slide improved answer accuracy, as the LLM often struggled to pinpoint the exact content of the current slide without this contextual aid. Moreover, SAM supports LaTeX-type formatting for visualizing mathematical formulas, ensuring that complex expressions are displayed clearly, facilitating better comprehension (Figure 2). Upon

completion of the session, users can download a PDF containing the questions posed and the AI mentor's responses, serving as a valuable study resource.

To support the smooth operation of SAM, the platform's architecture (Figure 3) was designed with distinct frontend and backend components, each hosted on separate servers. The frontend, deployed on a dedicated server, is responsible for the user interface and interaction. The backend, hosted on Amazon's Web Services (AWS) server [64], handles the core processing and data management tasks. AWS [64] was chosen for its robust infrastructure, scalability, and reliability, allowing SAM to efficiently manage multiple user sessions simultaneously. Each user's data is stored under a specific but anonymous ID, enabling the platform to securely save their interactions and responses, which are also needed for the resulting downloadable PDF. These records are then compiled into a resulting PDF and JSON files for further analysis. The coding of SAM leverages Node Package Manager for managing dependencies, with the core files written in TSX (TypeScript and JSX), a powerful combination that enhances both the type safety and flexibility of the development process. This technical setup ensures that SAM remains a robust, scalable, and user-friendly tool for enhancing educational experiences.

4 Methodology

To explore how proactive behavior can be encouraged in online learning environments, we designed SAM to provide learners with personalized, real-time support while maintaining their full control over the learning process. This approach aims to investigate the impact of SAM in fostering active participation and deeper engagement during independent learning. We evaluated the effectiveness of SAM in two studies. In the pilot study, we tested SAM among university students with similar educational backgrounds. Our goal was to evaluate the tool within a specific demographic, namely young university students, to gather preliminary insights on its design and usefulness. We then incorporated this feedback to refine our tool before expanding the study to a broader audience. In the main study, we introduced SAM to a diverse audience recruited through an online platform, i.e., Prolific, to assess its effectiveness across a more diverse participant pool with varied learning

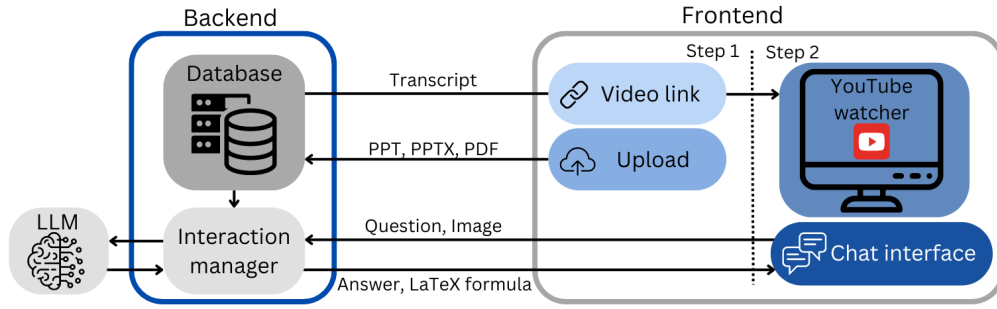


Figure 3: Implemented modules and their connections in the web application of SAM.

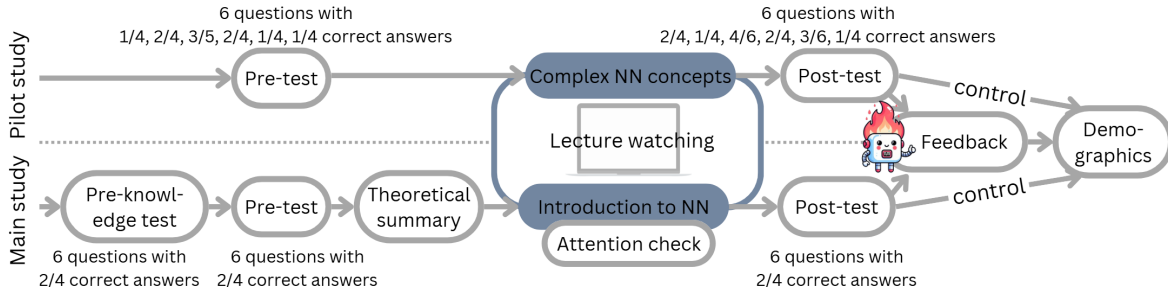


Figure 4: Representation of the structural elements in the pilot and main studies.

backgrounds. Figure 4 represents the structural elements and the differences between the pilot and main studies. We adjusted the study design for the main study to address potential knowledge gaps among participants, as detailed below.

For the purpose of our user studies, we adapted SAM to simplify the user experience. The selected YouTube video and corresponding slides were uploaded into the platform. Participants were only required to click a “Watch Lecture” button to initiate the session, streamlining the process and focusing on the interaction with SAM.

4.1 Study Design

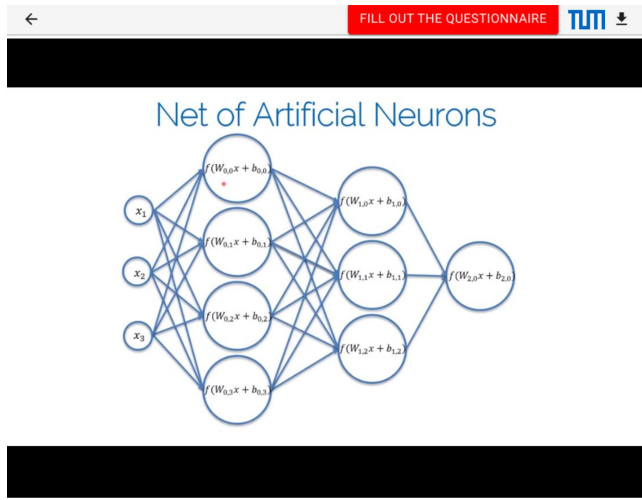
During both studies, we divided our participants into two groups: a test group and a control group. The test group had full access to SAM, meaning they could watch the video and interact with the AI mentor to ask questions anytime during the lecture (Figure 5b). In contrast, the control group also used SAM’s video-watching interface, but the chat functionality was disabled (Figure 5a). This setup (Figure 5) enabled us to compare the learning outcomes between participants who had access to the AI mentor and those who did not, providing insights into the impact of interactive AI support on the learning process. In the pilot study, participants were randomly assigned to groups using a backend function. For the main study, however, due to platform limitations (Prolific and Qualtrics), randomization was achieved by sequentially advertising separate study links, first the test group and then the control group. To ensure that participants could only enroll once, we restricted participation through Prolific. The advertisement included minimal information, primarily detailing exclusion criteria and specifically requesting that individuals with advanced knowledge of machine

learning refrain from participating. After accepting the invitation, participants received additional information about the study.

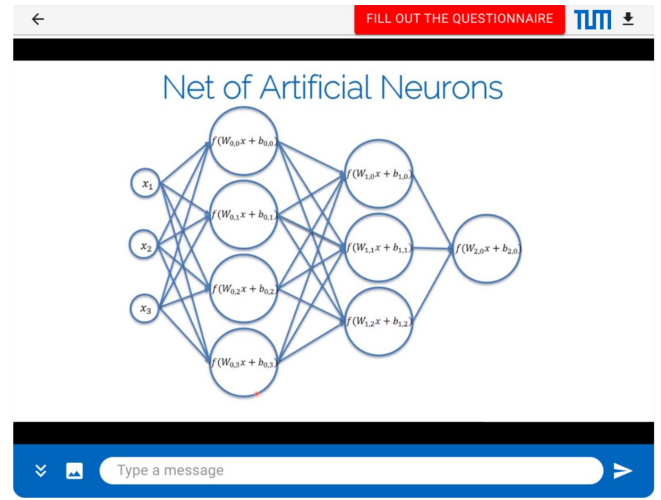
After the pilot study, we made a few modifications to our study design. In the pilot study, all participants were enrolled in the same course and had attended the same previous 11 lectures, ensuring a similar background knowledge level across the group. However, the main study was advertised to a broader pool on Prolific, where participants varied in their prior knowledge. To account for this, we introduced pre-knowledge assessments to establish a baseline for each participant. We also included a brief theoretical overview, specifically an explanation of linear regression (refer to subsection A.4 for the full text), prior to the lecture. Additionally, based on insights from the pilot study, we standardized the test questions by ensuring each question had the same number of answer choices and an equal distribution of correct answers to more effectively measure knowledge gain. In the main study, we also introduced an attention check after 20 minutes of video playback to assess participants’ attentiveness. A window popped up asking, “Would you like to continue to watch the video?” with a button labeled “Continue.” The time taken to click this button was recorded as an indicator of attentiveness.

When designing our user study, we drew inspiration from a study [3] conducted in an elementary school where eleventh graders were divided into two groups: an experimental group with access to ChatGPT and a control group without it, using pre- and post-tests to measure learning outcomes; however, while their study was conducted in a traditional classroom setting over twelve weeks, our research focuses on the context of online learning.

After participants completed the lecture, a red button labeled “Fill out the questionnaire” appeared at the top right corner of the



(a) Exemplary view of the video watcher with disabled chat interface.



(b) Exemplary view of SAM. At the bottom, users can start a conversation with the integrated AI mentor with the possibility of uploading an image.

Figure 5: Exemplary representations from the study.

screen, directing them to the post-test questionnaire. This process differed between the test and control groups. While the control group only completed the post-test and demographics questionnaire, the test group had an additional questionnaire to provide feedback regarding their experience with SAM, offering insights into our tool's effectiveness and usability.

4.2 Procedure

This section outlines the three phases of our user studies, which were consistent across both the pilot and main studies. Each study followed the same sequence of phases to ensure comparability and reliability in the results. The differences between the pilot and main studies are highlighted in each phase and also summarized in Figure 4. The questionnaires used in the studies were designed and administered through the Qualtrics platform.

4.2.1 Phase 1: Pre-tests. In the first phase of the pilot study, participants were asked to complete a pre-test (Appendix subsection A.5.1). The pre-test contained six questions and participants were informed that one or more answers could be correct, which helped measure their initial understanding. The pre-test questions were categorized into three key areas: graph neural networks, generative adversarial networks, and reinforcement learning, with two questions per category.

In the first phase of the main study, participants were asked to complete both a pre-knowledge test (Appendix subsection A.6.1) and a pre-test (Appendix subsection A.6.2). The pre-knowledge test, which consisted of six questions with only one correct answer each, was designed to assess and categorize participants based on their existing knowledge of neural networks (NNs). The pre-knowledge test was introduced to better understand participants' prior knowledge of NNs. This was not required in the pilot study, as it was conducted within a university course where participants had similar levels of familiarity with the topic. The pre-test also

contained six questions with two correct answers each. The instructions remained consistent with the pilot study, indicating that one or more answers could be correct. Similarly, the pre-test questions were categorized into three key areas: the basics of NNs, the structure and training of NNs, and activation functions, with two questions per category. Following these assessments, participants were provided with a brief text on linear regression in the main study to ensure they had the necessary background knowledge before moving on to the video-watching phase of the study (Appendix subsection A.4).

4.2.2 Phase 2: Lecture watching. In the second phase of our studies, participants engaged in a lecture-watching activity. We piloted the platform among participants of a university course named as "Introduction to Deep Learning" at the Technical University of Munich, on the topic of deep learning. This pilot study focused on the 12th lecture of the course with a specific focus on more advanced topics, allowing us to gather preliminary feedback on SAM's usability and measure the participants' knowledge gain. The video had a duration of 1 hour and 56 minutes. As it encompassed the entire lecture content delivered by the instructor, it was not possible to reduce its length. We chose this lecture because NNs and machine learning are increasingly popular topics that resonate with a broad audience, making them ideal for engaging participants. The course at the university had hundreds of students enrolled, providing a substantial pool of potential participants. Participation in this pilot was entirely voluntary.

In the main study, participants watched the lecture titled "Introduction to Neural Networks". For the main study, we edited the original video down to 28 minutes, focusing exclusively on the introduction section. This adjustment was made to accommodate the diverse participant pool by ensuring sustained attention and engagement. Unlike the full-length lecture used in the pilot study, which spanned the entire university course session, the shorter

video was better suited to the objectives of this study and the varied backgrounds of the participants. This approach also allowed us to streamline the content, making it more accessible and relevant for the study’s purposes. We chose this specific lecture because it introduces the fundamentals of NNs, a topic of growing interest in the field of AI. To tailor the content to our study’s needs, we edited the original video to focus exclusively on NNs, ensuring the material was challenging enough to test SAM’s AI-driven mentoring capabilities while still being conceivable by beginners.

4.2.3 Phase 3: Post-tests. In the third phase of our studies, participants were asked to complete a post-test (Appendix subsubsection A.5.2, subsubsection A.6.3) and a feedback questionnaire on SAM’s usability (only for the test group), and share their demographic information. This phase followed the same structure in both studies. The post-test comprised six questions, mirroring the structure of the pre-test, with two questions per topic across the same three categories. Both the control and test groups were required to complete this test in the studies. Participants were informed that one or more answers could be correct, similar to the pre-test questionnaire. This assessment was designed to gauge any improvements in the participants’ understanding of the material after watching the lecture, with a specific focus on whether the AI-driven mentoring of SAM facilitated enhanced learning outcomes.

Key aspects of the feedback questionnaire included the response time of SAM, the quality and reliability of its answers, and whether their comprehension of the three NN topics improved after interacting with SAM. Afterward, we collected demographics from both groups, covering age, gender, and employment status.

4.3 Participant Recruitment

In our main user study evaluating SAM, we recruited participants from the Prolific platform, splitting them into two groups—one utilizing an AI mentor during video watching (i.e., test group) and the other simply watching the video (i.e., control group). We chose Prolific due to its reputation for providing high-quality data [50]. Our sample was gender-balanced, including participants who were 18 or older, located in Germany, and fluent in English. After recruitment, participants were directed to Qualtrics for the study, and upon completion, they returned to Prolific for compensation, which was set at a rate of €12.41/hour. Given the time difference between interacting with the AI mentor and watching the video uninterrupted, the main study duration varied between 42 and 45 minutes. All data was collected anonymously, with digital informed consent obtained from each participant before the study began. Participants could withdraw at any time without further consequences to them. For the main study on Prolific, we recruited $n_{control} = 51$ for the control group and $n_{test} = 59$ for the test group. For the pilot study in the university course, we recruited $n_{control} = 15$ for the control group and $n_{test} = 15$ for the test group.

We excluded participants who completed the study in an unusually short amount of time. Given the time differences between the two groups and after analyzing outliers based on the total study duration, we set minimum time thresholds of 20 minutes for the control group and 35 minutes for the test group. These thresholds were consistent with testing in smaller groups, which identified the optimal time range for each phase. Participants who spent less time

Table 1: Sample sizes and demographic information for pilot and main studies.

Study	Group	n	Age	Men	Women	Diverse
Pilot	control	11	23.9 ± 2.3	9	2	0
	test	14	23.4 ± 1.9	11	2	1
Main	control	38	29.9 ± 5.8	20	18	0
	test	42	28.4 ± 5.3	21	21	0

than this likely did not fully engage with the questions or the lecture, making their responses less reliable. We excluded participants who failed the attention check, which involved responding to a message that appeared during the video. They were excluded from the study if they did not click on this message within eight minutes. Additionally, participants who provided low-quality answers were excluded, as they directly entered the post-test questions into SAM, resulting in responses that did not accurately reflect their comprehension. After filtering out participants with age discrepancies greater than three years or mismatches in gender (e.g., listed as a man in one place and woman in another) between Prolific (if applicable) and Qualtrics, we ended up with the final sample sizes that are presented in Table 1. We also provide more detailed demographics in the Appendix subsection A.1. Ethical considerations were strictly adhered to, with all user data collected anonymously, and an electronic consent obtained for storing, using, and potentially sharing the conversational data for research purposes.

4.4 Analysis

We evaluated the collected results through a comprehensive analysis of multiple aspects, including knowledge gain, user satisfaction, and answer accuracy. By examining these key metrics, we aim to assess the overall effectiveness of SAM in enhancing the learning experience. We assessed participants’ overall performance by assigning points to each question based on the number of correct answers it contained. Participants earned points for correctly selecting valid answers and for refraining from selecting incorrect ones. We initially analyzed the total points achieved by participants in both the pre- and post-test phases. Additionally, we examined the number of incorrect selections made, providing insight into participants’ accuracy and comprehension before and after using SAM. After the pilot study, a pre-knowledge test was included because self-reported assessments of prior knowledge, rated on a 5-point scale from “no knowledge” to “expert knowledge,” can be subjective and inconsistent. To obtain a more accurate measure of participants’ understanding of NNs, we included six objective pre-knowledge questions in the main study. These questions provided a standardized basis for grouping participants according to their actual level of expertise, ensuring a more reliable analysis of the results. Participants who scored 1-2 points on the pre-knowledge test were categorized as having no prior knowledge, while those who scored 3-4 points were considered to have basic knowledge, and those who scored 5-6 points were classified as having intermediate knowledge.

In the pilot study, the maximum available points varied based on the number of correct answers per question across six pre-test

Table 2: Analytical data and performance analysis from the pilot and main studies.

	Pilot Study				Main Study			
	Control		Test		Control		Test	
	Pre	Post	Pre	Post	Pre	Post	Pre	Post
avg. pre-knowledge points	-	-	-	-	3.6 ± 1.2	-	4.0 ± 1.2	-
avg. collected points	17.5 ± 3.1	23.5 ± 1.6	18.4 ± 2.6	23.9 ± 2.7	13.1 ± 2.5	14.7 ± 2.9	14.8 ± 2.9	14.9 ± 2.5
avg. mistakes	2.6 ± 1.3	1.3 ± 0.9	2.6 ± 1.6	1.4 ± 1.8	4.0 ± 2.0	3.2 ± 1.5	3.2 ± 1.5	2.5 ± 1.3

and post-test questions, whereas in the main study, each of the six questions in both the pre-test and post-test featured four answer options with two correct answers each, standardizing the point distribution. In the pilot study, the pre- and post-test questions were validated by an instructor of the course. The difficulty level of the pre- and post-test questions in the main study was evaluated by 10 unrelated machine learning experts, who rated the questions on a scale of 1 to 3, with 1 standing for easy and 3 standing for difficult. The average difficulty level for both the pre- and post-tests was 1.65, indicating that the questions were optimized for participants with little to no prior knowledge of the field, while still aiming to challenge their understanding. In both the pilot and main studies, the pre- and post-test questions varied, but consistently addressed the same topics covered in the lecture.

To account for the small sample size and variance differences, we used the Baumgartner-Weiss-Schindler (BWS) test [6] to compare knowledge gains between the test and the control groups. This statistical analysis allowed us to assess whether the differences in knowledge gain were significant. The corresponding p-values are reported in the results section (section 5), providing insight into the effectiveness of SAM in enhancing learning outcomes. In addition to assessing SAM’s impact on knowledge gain, we are exploring which demographic groups—such as age and employment status—benefit most from its use. This investigation addresses a gap identified in a prior meta-analysis [78], which emphasized the importance of determining which learner groups are best suited for AI chatbot-supported learning. Furthermore, we analyzed the frequently asked questions to understand user focus areas and identify key learning challenges, alongside a detailed examination of all hallucinations generated by SAM. This included evaluating every hallucination individually as well as the accuracy of the responses, their potential causes, and how users reacted to these inaccuracies. Additionally, we collected user feedback on our tool’s usability and satisfaction using a 5-point Likert scale.

5 Results

In this section, we present the findings from our study, focusing on the evaluation of SAM’s effectiveness as an educational tool. By comparing participants’ performance in the test and control groups and in demographic subgroups, we aim to offer insights into how triggered interactions enhance learning outcomes. Additionally, we explore how these insights can inform the design of future educational applications for more effective learning experiences.

On average, participants in the main study interacted with SAM (lecture watching and/or chatting) for 31.71 ± 10.91 minutes and

31.78 ± 17.64 minutes in the control and test groups, respectively, reflecting consistent usage patterns across groups. While the average duration in the test and control groups are similar, the difference in standard deviation indicates that SAM users tended to spend more time actively engaging with the system. In both groups, participants frequently used playback controls such as speed adjustments and rewinds, which accounts for variations in interaction duration. The interaction duration in the pilot study was more diverse due to students utilizing SAM to prepare for the exam and note-taking while stopping the video.

5.1 Quantitative Findings

To provide a clear understanding of the impact SAM had on participants’ learning, we first present the detailed pre- and post-test scores. The exact mean values achieved by individuals, as detailed in Table 2, show the pre- and post-test results for both the test and control groups in the pilot and main studies. The results from the main study indicate that participants in the test group made fewer incorrect selections after watching the video with SAM. We further analyzed these findings by calculating knowledge gains to draw meaningful conclusions, ensuring a comprehensive evaluation of SAM’s impact on learning outcomes.

The statistical analysis of knowledge gains in both the pilot and main studies showed no significant differences between the test and control groups, as indicated by the high variance in the data (see Figure 6a). A comparison of the data distribution between the main and pilot studies revealed a significant difference with $p < .001$, suggesting that university students actively enrolled in a lecture employ different learning strategies compared to the diverse participant pool recruited via Prolific. To better understand the impact of SAM’s features, we conducted a deeper analysis by examining the data across various subgroups, such as age and employment status. This involved comparing the performance of younger participants against older ones, as well as analyzing differences between students and full-time workers. Additionally, we looked at the test group in more detail, focusing on the number and nature of questions asked, which helped us identify how different participants engaged with SAM’s context-aware capabilities.

To assess knowledge gains across different demographic groups, we focused on the main study due to the limited demographic diversity in the pilot study, where most participants were university students under 30 years old. The analysis of age groups revealed a significant difference ($p = .049$) in data distribution for participants under 28 years old between the test group (19 participants) and the control group (11 participants). Younger participants in the test group exhibited greater knowledge gains compared to their

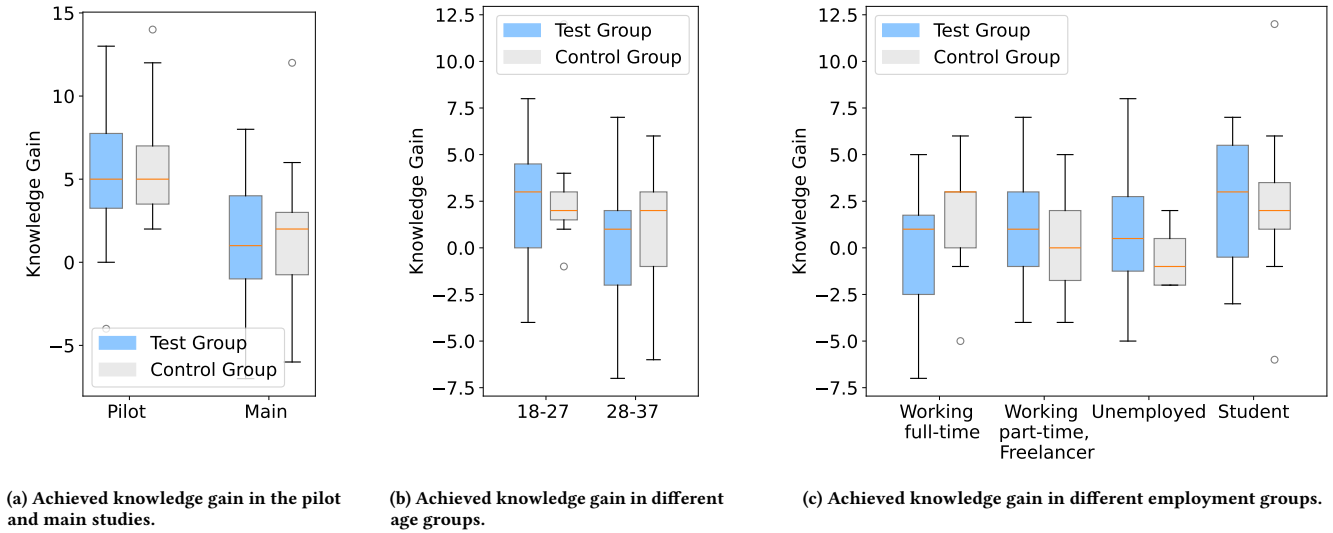


Figure 6: Box plots showing knowledge gains in test and control groups and across demographic subgroups.

counterparts in the control group. Conversely, for participants aged 28–37 (21 in the test group, 23 in the control group), the trend was reversed, with the control group achieving higher knowledge gains ($p = .042$) (Figure 6b). Given the limited number of participants older than 37 (only two and four participants in the test and control groups, respectively), we focused only on the first two age groups to ensure representative data across all four subgroups. This trend suggests that our tool, SAM, which was primarily designed with students in mind, is particularly effective for younger learners, many of whom fall within the student demographic. This could indicate that younger participants, who are likely more accustomed to interactive learning technologies, benefit more from the AI-driven mentoring provided by SAM.

In our analysis of the main study, we examined knowledge gains in relation to employment status. We specifically considered groups with at least four participants to ensure meaningful comparisons. In Figure 6c, we also represent part-time workers and freelancers, and unemployed, but based on their low representation, four and nine for part-time workers and freelancers, and four and four for unemployed participants in the control and test groups, respectively, we do not draw conclusions for these employment groups. The test group showed higher knowledge gain for students with 15 participants (mean of 1.8 ± 3.6) compared to the control group with 11 participants (mean of 1.4 ± 3.9) (Figure 6c), although this difference is not statistically significant ($p > .05$), the observed trend and mean values provide support for our findings. However, our tool's effectiveness was less pronounced among those working full-time, whereas the control group performed better with 19 participants (mean of 1.9 ± 2.7), compared to the test group with 14 participants (mean of 0.1 ± 3.3) with a significance of $p = .022$. This finding suggests that students who have flexible work schedules can more easily engage with and benefit from SAM's features, likely due to having more time to interact with the AI mentor, which is further supported by the observation that students in the test group spent,

on average, more time interacting with SAM (35.83 ± 13.82 minutes) compared to full-time workers (30.58 ± 24.65 minutes). Our results are similar to the results from a previous study, which showed higher flexibility in learning enables learners to persist through challenges and adopt a wider array of cognitive and metacognitive strategies [9]. These results underscore SAM's potential as a valuable learning aid, particularly for students and other groups with adaptable learning environments.

By further analyzing the data of the test group, Figure 7a shows that participants with intermediate knowledge asked more questions (mean of 5.9 ± 5.3), than participants with basic knowledge (mean of 3.4 ± 2.3), although the distributions do not differ significantly ($p > .05$), a clear trend supports our findings. Based on the limited number of participants with no prior knowledge (six participants), we only made assumptions regarding participants with basic knowledge (20 participants) and intermediate knowledge (16 participants). The tendency for participants with prior knowledge to ask more questions further supports the idea that SAM is particularly useful for those looking to deepen their understanding, as these users were more likely to engage with the AI mentor to clarify and build upon the concepts presented in the lecture.

Additionally, we analyzed when these questions occurred during lecture watching. In Figure 7b, we illustrate how often participants asked questions at each minute of the video, using different colors to represent the various topics. Notably, this figure presents the data user-wise, indicating the number of participants who asked at least one question at a given time point. A clear pattern emerged, with the highest number of questions being asked 2–3 minutes after a new topic was introduced. This suggests that participants took some time to process the material before seeking clarification. This observation highlights the importance of providing timely support during critical moments of learning, as participants are more likely to engage actively with the material once they have had time to familiarize themselves with a new concept. Also, a significant peak

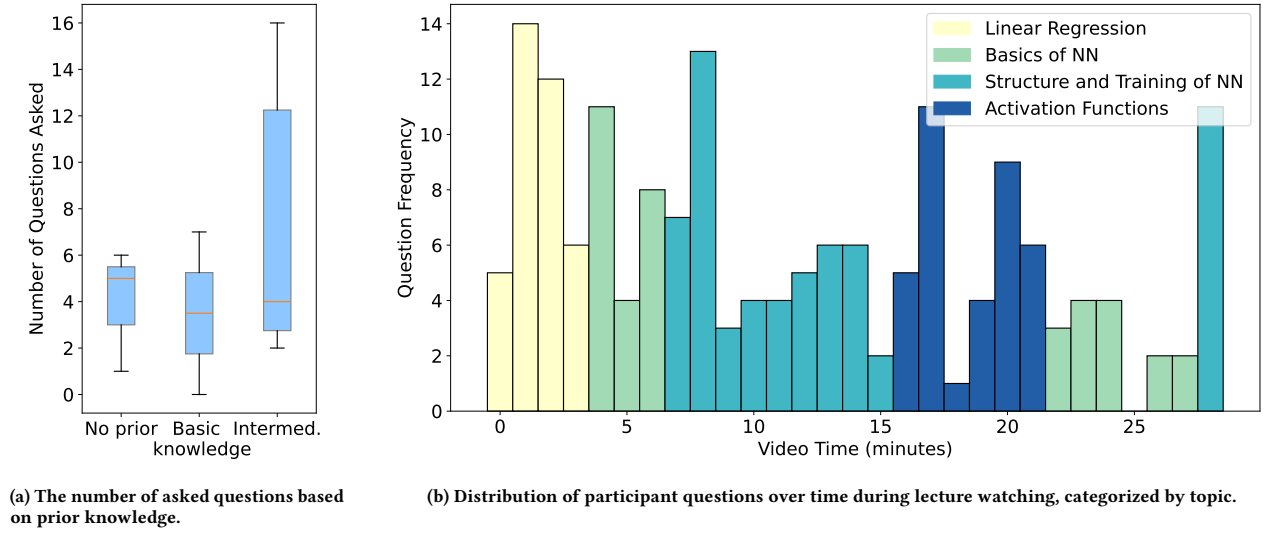


Figure 7: Analysis of the asked questions.

is observed at the last minute of the video, where participants asked for summaries or clarifications to reinforce their understanding before taking the post-test.

We also analyzed the frequently asked questions to better understand the areas where participants sought clarification or additional information while interacting with SAM. The three most common topics identified were weights in NNs, general information about NNs, and activation functions. To offer a thorough overview, we highlight the eight most frequently asked question categories, accompanied by two representative examples from each, detailed in the subsection A.2. While most participants focused their questions on key topics and core concepts, a notable number inquired about basic elements, such as the CIFAR-10 database, indicating limited prior knowledge in machine learning.

5.2 Qualitative Feedback

Overall, users in the pilot and main studies reported positive experiences with SAM, rating it as helpful and engaging on a 5-point Likert scale (see Table 3), while a 3-point Likert scale was specifically used to assess whether participants avoided asking questions due to long response times. This is reflected in the generally high satisfaction scores (3.4 ± 0.7 in the pilot and 4.0 ± 0.5 in the main study) and the willingness of participants to use SAM in the future (3.6 ± 0.7 in the pilot and 4.3 ± 0.8 in the main study). In general, participants rated high response quality (3.6 ± 0.9 in the pilot and 4.0 ± 0.6 in the main study) and trust in the accuracy of answers (3.8 ± 0.7 in the pilot and 4.2 ± 0.7 in the main study), demonstrating SAM’s effectiveness in providing real-time assistance. In both studies, participants found personalized learning support and feedback to be the most valuable feature of SAM. In the pilot study, this was followed by the ability to request summaries, receive formula explanations, and upload images to inquire about figures and formulas.

Table 3: User feedback from the pilot and main studies is presented as mean ratings with standard deviations on a 5-point Likert scale, except for “Avoided asking due to delayed responses”, measured on a 3-point scale.

Feedback category	Pilot	Main
Satisfaction with Response time	3.2 ± 1.0	4.2 ± 0.7
Avoided asking due to delayed responses	1.3 ± 0.5	1.2 ± 0.4
Response quality	3.6 ± 0.9	4.0 ± 0.6
Trust in accuracy of answer	3.8 ± 0.7	4.2 ± 0.7
Enhanced learning experience	3.5 ± 0.5	4.1 ± 0.6
Satisfaction	3.4 ± 0.7	4.0 ± 0.5
Helpfulness	3.5 ± 0.9	4.0 ± 0.9
Use in the future	3.6 ± 0.7	4.3 ± 0.8

In the main study, the order of preference shifted slightly, with formula explanations being the second most valued, followed by the ability to upload images for clarification on figures and formulas, and the option to request summaries.

Lastly, we evaluated the free-text feedback collected from participants, divided into two fields: feature suggestions and general messages to the developers. Suggestions included simplifying responses, improving the interface, enabling customizable interaction levels, adding voice recognition, providing visual aids, and integrating a video transcript with interactive explanations. Participants generally praised SAM as an innovative and helpful tool, especially for topics they had some prior knowledge of. They highlighted its ability to simplify concepts, generate formulas, and customize responses. While some found it less effective for entirely new topics, many appreciated its speed and functionality, with several expressing enthusiasm for its potential in education.

Table 4: Participant responses to the question: “Have you encountered any instances where SAM provided incorrect answers?”.

Study	Never	Occasionally	Frequently	Unnoticed
Pilot	7	2	0	4
Main	23	2	2	15

5.3 AI tutor answer analysis

In both the pilot and main studies, we asked participants to report whether they encountered any incorrect answers provided by SAM. Most participants indicated that they did not find or notice any inaccuracies. This feedback suggests that participants generally perceived SAM’s responses as accurate and reliable. The detailed responses regarding the perceived accuracy of SAM’s answers are summarized in Table 4, further supporting the tool’s effectiveness in delivering correct and contextually relevant information.

Since the lecture material was mostly new to the participants, their responses regarding whether they encountered incorrect answers may not be fully reliable, necessitating a manual evaluation of SAM’s provided answers to ensure accuracy. In the main study, we collected all user-generated questions and had them evaluated by two machine learning experts to assess the correctness of the responses provided by SAM. Two experts carefully reviewed each answer, considering not only its general correctness but also its relevance to the specific context of the lecture material. For instance, SAM occasionally provided accurate but contextually inappropriate answers, such as offering a general explanation of matrix indices when the question was specifically related to the indices within a neural network’s weight matrix, a concept tied closely to the lecture’s content on network structure. Because the majority of the provided answers were correct, resulting in unbalanced labels, Cohen’s κ cannot accurately reflect the inter-rater reliability. The experts achieved 96.0% agreement, and after discussion, they found that the LLM provided a factually correct answer in 97.6% of the cases. This high level of accuracy reflects SAM’s effectiveness in responding to user queries within the context of the lecture material. A more detailed analysis, focusing on the factual correctness in each category, is provided in Table 5.

We further analyzed the hallucinations that the experts marked and how the users reacted to them. We analyzed each and every one of them individually. Most hallucinations did not contain outright false information but were often unrelated to the lecture material, which could still mislead users. In one case, the LLM responded with “Please focus on the lecture material.” instead of answering the question. On one occasion, the mistake was corrected in the following message. Overall, the majority of the responses did not contain factual errors. Interestingly, users never directly addressed or highlighted errors in their messages. For factual inaccuracies, it is likely that users did not realize the information was incorrect. We collected all hallucinations with explanations in the Appendix subsection A.3.

6 Discussion and Limitations

SAM is an AI-driven educational tool designed to enhance learning by integrating real-time, context-aware question-answering with video content. A main user study involving 80 participants evaluated SAM’s effectiveness, focusing on knowledge gain, user satisfaction, and answer accuracy. However, variability in participants’ prior knowledge, engagement levels, or motivation could have influenced the results. In our main study, we used a video lecture from a university course as the main instructional material. This video, being the third in the course series, may not have been sufficient to bring all participants to a similar baseline understanding of the topic at hand. By providing a summary of the concept to the participants before watching the lecture helped address this issue. This gap is evident in the test group, where many participants sought further explanations (e.g., on linear regression) during the study. Although we instructed participants not to use external resources during the tests, we cannot be entirely certain that this was adhered to. We attempted to mitigate this by monitoring the time participants spent on the study.

Moreover, because the pilot study was integrated into the course, we lacked clear insights into the participants’ prior knowledge. Since the question, “Do you have previous knowledge in the following areas: Graph Neural Networks, Generative Models, or Reinforcement Learning?” relies on self-reporting, we were unable to objectively measure how much they had absorbed from previous lectures or related courses in machine learning and deep learning offered by the same university. Feedback from the pilot study was valuable and informed adjustments for the main study, though some areas for improvement may have remained unaddressed. Another limitation identified was SAM’s inconsistent performance in answering questions. In rare cases, SAM responded with “I don’t know,” even when it had previously provided correct answers to differently phrased questions. There were also some instances where SAM failed to understand the specific context of a question, offering general but irrelevant answers to the current lecture content.

In designing SAM, we aligned its functionality with the ICAP framework [15], ensuring it operates at the interactive learning level, where learners actively engage through real-time interactions, such as they could discuss the given justifications, ask questions, and discuss similarities and differences if needed. In contrast, the control group was limited to the active learning level, allowing them only to manipulate the video without additional support. However, not all participants utilized the option to interact with the video, resulting in some engaging solely in passive learning. Future work could explore alternative support for the control group, such as a blog for asynchronous assistance. However, this approach was beyond the scope of this study due to practical constraints with Prolific, as varying interaction times would disrupt the real-time scenario we aimed to simulate. SAM’s primary goal is to provide immediate, real-time support, and offering instant responses through a blog would compromise the integrity of the results. Also, future work could explore how SAM’s design principles inform broader educational theories, particularly in asynchronous or hybrid learning contexts.

Furthermore, while SAM was specifically designed with students in mind and demonstrated effectiveness for this subgroup, it did not

Table 5: Analysis of the answers provided by SAM in the main study with respect to the three main categories.

	Basics of NN	Structure & Training of NN	Activation functions	Other
num. of correct answers	60	95	40	47
num. of incorrect answers	0	4	2	0
Correctness rate	100%	97.0%	96.0%	97.9%

perform as well for full-time employees. The structure of the tool emphasizes flexibility and interactive engagement, and it aligns well with the learning habits and needs of students who are accustomed to using digital platforms for their studies. However, full-time employees, who may have different learning preferences or time constraints, did not benefit from SAM as much. These findings suggest that while SAM is a powerful tool for enhancing student learning, its design might need to be adjusted to accommodate different user groups better, such as full-time professionals who require more tailored educational resources.

The evaluation in our study focused on a single topic, neural networks, which is highly structured and well-suited for AI-driven mentoring. The tool’s applicability to disciplines requiring more abstract or discussion-oriented learning remains unexplored. Testing SAM in these diverse areas could provide insights into its versatility and effectiveness across different disciplines. In general, SAM’s flexible design allows for adjustments to accommodate various subjects and lecture materials, making it adaptable to diverse educational contexts. Another limitation of SAM is the need for a thorough investigation into how lecture video quality impacts learning outcomes both with and without SAM’s support. The present study relied on high-quality instructional videos with clear audio, well-organized content, and accompanying slides. However, the impact of varying video quality on SAM’s effectiveness has not yet been systematically examined. Factors such as audio clarity, visual fidelity, and the coherence of lecture materials may influence both user engagement and the AI’s ability to provide accurate, contextually relevant responses.

Additionally, although SAM is flexible in its use, it is currently optimized for videos available on YouTube. This is a limitation given that many educational institutions utilize other platforms, such as Panopto [22], for hosting and streaming lecture content. Future iterations of SAM should consider integrating with a broader range of video-watching tools to enhance its applicability and reach within academic environments. Expanding SAM’s compatibility with widely used educational platforms would not only increase its utility but also support its adoption across various learning contexts, ensuring that more students and educators can benefit from its capabilities.

While this study primarily measured short-term knowledge gains and user satisfaction, future research should assess SAM’s impact on long-term knowledge retention and higher-order cognitive skills, such as critical thinking and problem-solving. Metrics such as retention rates, frequency of subsequent interactions with SAM, and the transferability of learned concepts to novel scenarios would provide a more comprehensive evaluation of the tool’s educational value. Additionally, longitudinal studies that track learners’ progress over time could elucidate the sustained impact of SAM on

academic achievement and self-directed learning behaviors. These investigations would not only validate SAM’s effectiveness but also contribute to the design of adaptive learning systems that promote lifelong learning.

Although SAM currently utilizes GPT-4o [48] as its underlying LLM, this model is easily exchangeable, allowing for flexibility in integrating other commercial or open-source LLMs. Our primary focus has been on SAM’s long-term impact in fostering personalized, real-time learning support rather than being tied to a specific model. Investigating the most reasonable LLM for SAM’s continued development is a future direction for this project.

Overall, SAM introduces a novel approach to real-time support by delivering highly accurate answers combined with personalized learning experiences tailored to individual needs. This innovation showcases the potential of adaptive AI systems to enhance engagement and effectiveness in educational tools.

7 Conclusion

We introduced SAM, a context-aware AI chatbot that allows students to ask questions and explore unclear parts of the lecture. User studies with a total of 105 participants (80 in the main and 25 in the pilot study) revealed its effectiveness as an educational tool focusing on proactive teaching. SAM enhances learning outcomes by encouraging students to take full ownership of their learning experience and fosters a more interactive and engaged learning environment. The real-time, personalized assistance provided by SAM allows learners to delve deeper into complex topics, reinforcing their understanding of the lecture content. The studies highlighted SAM’s potential as an effective learning tool, and the detailed analysis of demographic groups revealed that SAM has a great potential, especially for young learners and individuals in flexible working environments, such as students, to increase knowledge gains. We also found that participants who had intermediate-level prior knowledge asked more questions during lecture watching. According to expert raters, SAM maintained a high answer correctness rate of 97.6%, and overall, participants provided positive feedback regarding user satisfaction, response quality, and the tool’s helpfulness, with many expressing a willingness to use SAM in the future. These findings underscore SAM’s potential to transform education by promoting interactivity and learner independence. Future versions could expand its use beyond STEM subjects to areas that focus on discussion, interpretation, and critical thinking, while also incorporating features like summarizing argumentative texts to support learning in non-STEM disciplines.

Acknowledgments

We extend our gratitude to Prof. Dr. Matthias Nießner and his group for providing the lecture videos and allowing us to pilot our study

in his course. We also thank students Rahmiye Büşra Büyükgebiz and Ozan Aydın for their invaluable contributions in developing the first prototype of SAM. Finally, we deeply appreciate Márton Szép for his insightful ideas, feedback, and assistance with writing.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
- [2] Dimah Al-Fraihat, Mike Joy, Jane Sinclair, et al. 2020. Evaluating E-learning systems success: An empirical study. *Computers in human behavior* 102 (2020), 67–86. <https://doi.org/10.1016/j.chb.2019.08.004>
- [3] Saif Alneyadi and Yousef Wardat. 2023. ChatGPT: Revolutionizing student achievement in the electronic magnetism unit for eleventh-grade students in Emirates schools. *Contemporary Educational Technology* 15, 4 (2023), ep448. <https://doi.org/10.30935/cedtech/13417>
- [4] FN Anyatasia, HB Santos, and K Junus. 2020. An evaluation of the Udacity MOOC based on instructional and interface design principles. In *Journal of Physics: Conference Series*, Vol. 1566. IOP Publishing, 012053. <https://doi.org/10.1088/1742-6596/1566/1/012053>
- [5] Arshia Ayoub, Raashida Amin, and Zahid Ashraf Wani. 2020. Contribution of developed countries towards MOOCs: an exploration and assessment from a representative platform Coursera. *Asian Association of Open Universities Journal* 15, 2 (2020), 251–262. <https://doi.org/10.1108/AAOUJ-03-2020-0016>
- [6] W Baumgartner, P Weiß, and H Schindler. 1998. A nonparametric test for the general two-sample problem. *Biometrics* (1998), 1129–1135.
- [7] Kizzy Beaumont. 2018. Google Classroom: An online learning environment to support blended learning. *Compass: Journal of Learning and Teaching* 11, 2 (2018), 1–6. <https://pdfs.semanticscholar.org/1537/78fb6948d9a4f5fc798a359d754f09c09781.pdf>
- [8] Luciana Benotti, Mara Cecilia Martinez, and Fernando Schapachnik. 2017. A tool for introducing computer science with automatic formative assessment. *IEEE transactions on learning technologies* 11, 2 (2017), 179–192. <https://doi.org/10.1109/TLT.2017.2682084>
- [9] Per Bernard Bergamin, Simone Ziska, Egon Werlen, and Eva Siegenthaler. 2012. The relationship between flexible and self-regulated learning in open and distance universities. *International Review of Research in Open and Distributed Learning* 13, 2 (2012), 101–123. <https://doi.org/10.19173/irrodl.v13i2.1124>
- [10] Anna Bodonheli, Efe Bozkir, Shuo Yang, Enkelejda Kasneci, and Gjergji Kasneci. 2024. User intent recognition and satisfaction with large language models: A user study with chatgpt. *arXiv preprint arXiv:2402.02136* (2024).
- [11] Ricardo Böheim, Maximilian Knogler, Christian Kosel, and Tina Seidel. 2020. Exploring student hand-raising across two school subjects using mixed methods: An investigation of an everyday classroom behavior from a motivational perspective. *Learning and Instruction* 65 (2020), 101250. <https://doi.org/10.1016/j.learninstruc.2019.101250>
- [12] Ricardo Böheim, Tim Urdan, Maximilian Knogler, and Tina Seidel. 2020. Student hand-raising as an indicator of behavioral engagement and its role in classroom learning. *Contemporary Educational Psychology* 62 (2020), 101894. <https://doi.org/10.1016/j.cedpsych.2020.101894>
- [13] Babette Bühler, Efe Bozkir, Hannah Deininger, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. 2024. On Task and in Sync: Examining the Relationship between Gaze Synchrony and Self-Reported Attention During Video Lecture Learning. *Proceedings of the ACM on Human-Computer Interaction* 8, ETRA (2024), 1–18. <https://doi.org/10.1145/3655604>
- [14] Babette Bühler, Ruikun Hou, Efe Bozkir, Patricia Goldberg, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. 2023. Automated hand-raising detection in classroom videos: A view-invariant and occlusion-robust machine learning approach. In *International Conference on Artificial Intelligence in Education*. Springer, 102–113. https://doi.org/10.1007/978-3-031-36272-9_9
- [15] Michelene TH Chi and Ruth Wylie. 2014. The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist* 49, 4 (2014), 219–243. <https://doi.org/10.1080/00461520.2014.965823>
- [16] Miguel Coronado, Carlos A Iglesias, Álvaro Carrera, and Alberto Mardomingo. 2018. A cognitive assistant for learning java featuring social dialogue. *International Journal of Human-Computer Studies* 117 (2018), 55–67. <https://doi.org/10.1016/j.ijhcs.2018.02.004>
- [17] Samuel Cunningham-Nelson, Mahsa Baktashmotlagh, and Wageeh Boles. 2019. Visualizing student opinion through text analysis. *IEEE Transactions on Education* 62, 4 (2019), 305–311. <https://doi.org/10.1109/TE.2019.2924385>
- [18] Jeferson da Silva Oliveira, Danubia Bueno Espindola, Regina Barwaldt, Luciano Maciel Ribeiro, and Marcelo Pias. 2019. Ibm watson application as faq assistant about moodle. In *2019 IEEE Frontiers in education conference (FIE)*. IEEE, 1–8. <https://doi.org/10.1109/FIE43999.2019.9028667>
- [19] Nada Dabbagh, Angela D Benson, André Denham, Roberto Joseph, Maha Al-Freih, Ghania Zgheib, Helen Fake, Zhetao Guo, Nada Dabbagh, Angela D Benson, et al. 2016. Massive open online courses. *Learning Technologies and Globalization: Pedagogical Frameworks and Applications* (2016), 9–13. https://doi.org/10.1007/978-3-319-22963-8_2
- [20] José Manuel Martins Ferreira. 2016. Massive Open Online Courses (MOOCs). *Cross-Border Higher Education and Quality Assurance: Commerce, the Services Directive and Governing Higher Education* (2016), 203–219. https://doi.org/10.1057/978-1-137-59472-3_12
- [21] Antonio Fini. 2009. The technological dimension of a massive open online course: The case of the CCK08 course tools. *International Review of Research in Open and Distributed Learning* 10, 5 (2009). <https://doi.org/10.19173/irrodl.v10i5.643>
- [22] Seyum Getenet, Niharika Singh, Eseta Tualaulelei, Yosheen Pillay, and Sue Worsley. 2022. Students' behavioural engagement with recorded lecture videos: Panopto video analytics. (2022). <https://www.learnlib.org/p/221675/>
- [23] Masood Ghayoomi. 2023. Applications of Chatbots in Education. In *Trends, Applications, and Challenges of Chatbot Technology*. IGI Global, 80–118. <https://doi.org/10.4018/978-1-6684-6234-8.ch004>
- [24] Marta González-Lloret. 2020. Collaborative tasks for online language teaching. *Foreign Language Annals* 53, 2 (2020), 260–269. <https://doi.org/10.1111/flan.12466>
- [25] Rachel S Goodman, J Randall Patrinely, Cosby A Stone, Eli Zimmerman, Rebecca R Donald, Sam S Chang, Sean T Berkowitz, Avni P Finn, Eiman Jahangir, Elizabeth A Scoville, et al. 2023. Accuracy and reliability of chatbot responses to physician questions. *JAMA network open* 6, 10 (2023), e2336483–e2336483. <https://doi.org/10.1001/jamanetworkopen.2023.36483>
- [26] Aldo Gordillo, Daniel López-Fernández, and Edmundo Tovar. 2022. Comparing the effectiveness of video-based learning and game-based learning using teacher-authored video games for online software engineering education. *IEEE Transactions on Education* 65, 4 (2022), 524–532. <https://doi.org/10.1109/TE.2022.3142688>
- [27] Reginald Gerald Govender and Simon Bheki Khoza. 2022. First-hand user experience: Can Kaltura video come to the rescue of Moodle during/post COVID-19? *African Identities* (2022), 1–18. <https://doi.org/10.1080/14725843.2022.2082376>
- [28] Ece Gumusel. 2024. A literature review of user privacy concerns in conversational chatbots: A social informatics approach: An Annual Review of Information Science and Technology (ARIST) paper. *Journal of the Association for Information Science and Technology* (2024). <https://doi.org/10.1002/asi.24898>
- [29] Guntur Gunawan, Muhammad Kristiawan, Eko Risdianto, and Ririn Eka Monicha. 2021. Application of the zoom meeting application in online learning during the pandemic. *Education Quarterly Reviews* 4, 2 (2021). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3819839
- [30] Shikha Gupta and Samarth Gupta. 2022. Massive Open Online Courses: Promoting Intercultural Communication. In *Handbook of Research on Teaching Strategies for Culturally and Linguistically Diverse International Students*. IGI Global, 289–305. <https://doi.org/10.4018/978-1-7998-8921-2.ch015>
- [31] Lisa R Halverson and Charles R Graham. 2019. Learner engagement in blended learning environments: a conceptual framework. *Online learning* 23, 2 (2019), 145–178. <https://eric.ed.gov/?id=EJ1218398>
- [32] Songhee Han and Min Kyung Lee. 2022. FAQ chatbot and inclusive learning in massive open online courses. *Computers & Education* 179 (2022), 104395. <https://doi.org/10.1016/j.compedu.2021.104395>
- [33] Alexandara Harry and Sayudin Sayudin. 2023. Role of AI in Education. *Interdisciplinary Journal and Humanity (INJURY)* 2, 3 (2023), 260–268. <https://doi.org/10.58631/injury.v2i3.52>
- [34] Khe Foon Hew, Weijiao Huang, Jiahui Du, and Chengyuan Jia. 2023. Using chatbots to support student goal setting and social presence in fully online activities: learner engagement and perceptions. *Journal of Computing in Higher Education* 35, 1 (2023), 40–68. <https://doi.org/10.1007/s12528-022-09338-x>
- [35] Sebastian Hobert and Raphael Meyer von Wolff. 2019. Say hello to your new automated tutor—a structured literature review on pedagogical conversational agents. (2019). <https://aiselaisnet.org/wi2019/track04/papers/2/>
- [36] Ruikun Hou, Tim Fütterer, Babette Bühler, Efe Bozkir, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. 2024. Automated Assessment of Encouragement and Warmth in Classrooms Leveraging Multimodal Emotional Features and ChatGPT. In *International Conference on Artificial Intelligence in Education*. Springer, 60–74. https://doi.org/10.1007/978-3-031-64302-6_5
- [37] Carolin Ischen, Theo Araujo, Hilde Voorveld, Guda van Noort, and Edith Smit. 2020. Privacy concerns in chatbot interactions. In *Chatbot Research and Design: Third International Workshop, CONVERSATIONS 2019, Amsterdam, The Netherlands, November 19–20, 2019, Revised Selected Papers* 3. Springer, 34–48. https://doi.org/10.1007/978-3-030-39540-7_3
- [38] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences* 103 (2023), 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- [39] Yunhwan Kim, Michael Glassman, and Michael Steven Williams. 2015. Connecting agents: Engagement and motivation in online collaboration. *Computers in Human Behavior* 49 (2015), 333–342. <https://doi.org/10.1016/j.chb.2015.03.015>

- [40] Mohammad Amin Kuhail, Nazik Alturki, Salwa Alramlawi, and Kholood Alhejori. 2023. Interacting with educational chatbots: A systematic review. *Education and Information Technologies* 28, 1 (2023), 973–1018. <https://doi.org/10.1007/s10639-022-11177-3>
- [41] Shui-fong Lam, Bernard PH Wong, Hongfei Yang, and Yi Liu. 2012. Understanding student engagement with a contextual model. In *Handbook of research on student engagement*. Springer, 403–419. https://doi.org/10.1007/978-1-4614-2018-7_19
- [42] Edith Law, Parastoo Baghaei Ravari, Nalin Chhibber, Dana Kulic, Stephanie Lin, Kevin D Pantoasdo, Jessy Ceha, Sangho Suh, and Nicole Dillen. 2020. Curiosity notebook: A platform for learning by teaching conversational agents. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*. 1–9. <https://doi.org/10.1145/3334480.3382783>
- [43] Lap-Kei Lee, Yin-Chun Fung, Yau-Wai Pun, Ka-Kin Wong, Maverick Tai-Yin Yu, and Nga-In Wu. 2020. Using a multiplatform chatbot as an online tutor in a university course. In *2020 international symposium on educational technology (ISET)*. IEEE, 53–56. <https://doi.org/10.1109/ISET49818.2020.00021>
- [44] Tharindu Rekha Liyanagunawardena. 2015. Massive open online courses. *Humanities* 4, 1 (2015), 35–41. <https://doi.org/10.3390/h4010035>
- [45] Chung Kwan Lo, Khe Foon Hew, and Morris Siu-yung Jong. 2024. The influence of ChatGPT on student engagement: A systematic review and future research agenda. *Computers & Education* (2024), 105100. <https://doi.org/10.1016/j.compedu.2024.105100>
- [46] Florence Martin and Doris U Bolliger. 2018. Engagement matters: Student perceptions on the importance of engagement strategies in the online learning environment. *Online learning* 22, 1 (2018), 205–222. <https://doi.org/10.24059/olj.v22i1.1092>
- [47] Shahid Minhas, Tasaddaq Hussain, Abdul Ghani, Kiran Sajid, and L Pakistan. 2021. Exploring students online learning: A study of zoom application. *Gazi University Journal of Science* 34, 2 (2021), 171–178. <https://doi.org/10.35378/gujs.691705>
- [48] OpenAI. 2024. *Hello GPT-4o*. <https://openai.com/index/hello-gpt-4o/> Accessed: 2024-08-30.
- [49] S Pappagallo. 2024. Chatbots in Education: A Dual Perspective on Innovation and Ethics. *Journal of Digital Pedagogy* 3, 1 (2024), 3–10. <https://doi.org/10.61071/JDP.2420>
- [50] Eyal Peer, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. 2022. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods* 54, 4 (2022), 1643–1662. <https://doi.org/10.3758/s13428-021-01694-3>
- [51] Pavlo Pokotylo. 2023. Some of basic key point of online learning educational tools. *Problems of Education* 1 (98) (2023), 219–230. <https://doi.org/10.52256/2710-3986.1-98.2023.14>
- [52] Muh Putra Pratama, Rigel Sampelolo, and Hans Lura. 2023. Revolutionizing education: harnessing the power of artificial intelligence for personalized learning. *Klasikal: Journal of education, language teaching and science* 5, 2 (2023), 350–357. <https://doi.org/10.52208/klasikal.v5i2.877>
- [53] Annelies Raes, Pieter Vanneste, Marieke Pieters, Ine Windey, Wim Van Den Noortgate, and Fien Depaepe. 2020. Learning and instruction in the hybrid virtual classroom: An investigation of students' engagement and the effect of quizzes. *Computers & Education* 143 (2020), 103682. <https://doi.org/10.1016/j.compedu.2019.103682>
- [54] Bhavika R Ranoliya, Nidhi Raghuwanshi, and Sanjay Singh. 2017. Chatbot for university related FAQs. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 1525–1530. <https://doi.org/10.1109/ICACCI.2017.8126057>
- [55] Ahmad Ridho Rojabi. 2020. Exploring EFL Students' Perception of Online Learning via Microsoft Teams: University Level in Indonesia. *English Language Teaching Educational Journal* 3, 2 (2020), 163–173. <https://eric.ed.gov/?id=EJ1268365>
- [56] Estefania Roldan-Vasquez, Samir Mitri, Shreya Bhasin, Tina Bharani, Kathryn Capasso, Michelle Haslinger, Ranjna Sharma, and Ted A James. 2024. Reliability of artificial intelligence chatbot responses to frequently asked questions in breast surgical oncology. *Journal of Surgical Oncology* (2024). <https://doi.org/10.1002/jso.27715>
- [57] Nico Rutten, Wouter R Van Joolingen, and Jan T Van Der Veen. 2012. The learning effects of computer simulations in science education. *Computers & education* 58, 1 (2012), 136–153. <https://doi.org/10.1016/j.compedu.2011.07.017>
- [58] Marija Sablić, Ana Miroslavljević, and Alma Škugor. 2021. Video-based learning (VBL)—past, present and future: An overview of the research published from 2008 to 2019. *Technology, Knowledge and Learning* 26, 4 (2021), 1061–1077. <https://doi.org/10.1007/s10758-020-09455-5>
- [59] Joni Salminen, Soon-gyo Jung, Johanne Medina, Kholoud Aldous, Jinan Azem, Waleed Akhtar, and Bernard J Jansen. 2024. Using Ciperbot: An Exploratory Analysis of Student Interaction with an LLM-Based Educational Chatbot. In *Proceedings of the Eleventh ACM Conference on Learning@Scale*. 279–283. <https://doi.org/10.1145/3657604.3664690>
- [60] Sandra Sanchez-Gordon, Sergio Luján-Mora, et al. 2016. How could MOOCs become accessible? The case of edX and the future of inclusive online learning. (2016). <https://rua.ua.es/dspace/handle/10045/54410>
- [61] Dylan GM Schouten, Fleur Venneker, Tibor Bosse, Mark A Neerincx, and Anita HM Cremers. 2017. A digital coach that provides affective and social learning support to low-literate learners. *IEEE Transactions on Learning Technologies* 11, 1 (2017), 67–80. <https://doi.org/10.1109/TLT.2017.2698471>
- [62] Klara Sedova, Martin Sedlacek, Roman Svaricek, Martin Majcik, Jana Navratilova, Anna Drexlerova, Jakub Kychler, and Zuzana Salamounova. 2019. Do those who talk more learn more? The relationship between student classroom talk and student achievement. *Learning and instruction* 63 (2019), 101217. <https://doi.org/10.1016/j.learninstruc.2019.101217>
- [63] Gayathri Senevirathne and Kalpani Manathunga. 2021. Impact of E-Learning system user interface design on user satisfaction. In *2021 IEEE 9th Region 10 Humanitarian Technology Conference (R10-HTC)*. IEEE, 01–06. <https://doi.org/10.1109/R10-HTC53172.2021.9641570>
- [64] Amazon Web Services. [n.d.]. *Amazon Web Services (AWS) - Cloud Computing Services*. <https://aws.amazon.com/> Accessed: 2024-08-30.
- [65] Shady Shehata, David Santandreu Calonge, Philip Purnell, and Mark Thompson. 2023. Enhancing video-based learning using knowledge tracing: Personalizing students' learning experience with ORBITS. In *Proceedings of the 18th workshop on innovative use of NLP for Building Educational Applications (BEA 2023)*. 100–107. <https://doi.org/10.18653/v1/2023-bea-1.8>
- [66] B. Sheridan and L. Campbell. 2024. Using Panopto to Encourage Deeper Learning and Reflection in a Writing Course. In *Revitalising Higher Education*, T. Bowell et al. (Eds.). Cardiff University Press, Cardiff. <https://doi.org/10.18573/conf2.o>
- [67] Anamica Sinha. 2024. Tools Involved in Online Education. In *Reconfiguring Pedagogy and Curriculum Practice in Light of Online Teaching: A Machine-Generated Literature Overview*. Springer, 133–160. https://doi.org/10.1007/978-3-031-37974-1_5
- [68] Sharob Sinha, Shyanka Basak, Yajushi Dey, and Anupam Mondal. 2020. An educational chatbot for answering queries. In *Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018*. Springer, 55–60. https://doi.org/10.1007/978-981-13-7403-6_7
- [69] Subarna Sivapalan, Michael J Clifford, and Sarah Speight. 2016. Engineering education for sustainable development: using online learning to support the new paradigms. *Australasian Journal of Engineering Education* 21, 2 (2016), 61–73. <https://doi.org/10.1080/22054952.2017.1307592>
- [70] Donggil Song, Eun Young Oh, and Marilyn Rice. 2017. Interacting with a conversational agent system for educational purposes in online courses. In *2017 10th international conference on human system interactions (HSI)*. IEEE, 78–82. <https://doi.org/10.1109/HSI.2017.8005002>
- [71] Christian Stöhr, Amy Wanyu Ou, and Hans Malmström. 2024. Perceptions and usage of AI chatbots among students in higher education across genders, academic levels and fields of study. *Computers and Education: Artificial Intelligence* 7 (2024), 100259. <https://doi.org/10.1016/j.caeai.2024.100259>
- [72] Olga Tapalova and Nadezhda Zhiyenbayeva. 2022. Artificial intelligence in education: AIED for personalised learning pathways. *Electronic Journal of e-Learning* 20, 5 (2022), 639–653. <https://eric.ed.gov/?id=EJ1373006>
- [73] Stergios Tegos, Stavros Demetriadis, and Thrasylvoulos Tsiatsos. 2014. A configurable conversational agent to trigger students' productive dialogue: a pilot study in the CALL domain. *International Journal of Artificial Intelligence in Education* 24 (2014), 62–91. <https://doi.org/10.1007/s40593-013-0007-3>
- [74] Alexandros Tsironis, Christos Katsanos, and Michail Xenos. 2016. Comparative usability evaluation of three popular MOOC platforms. In *2016 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, 608–612. <https://doi.org/10.1109/EDUCON.2016.7474613>
- [75] Emma J Vermeulen and Monique LL Volman. 2024. Promoting student engagement in online education: online learning experiences of Dutch university students. *Technology, Knowledge and Learning* (2024), 1–21. <https://doi.org/10.1007/s10758-023-09704-3>
- [76] Ming-Te Wang and Jessica Degol. 2014. Staying engaged: Knowledge and research needs in student engagement. *Child development perspectives* 8, 3 (2014), 137–143. <https://doi.org/10.1111/cdep.12073>
- [77] Rainer Winkler, Sebastian Hobert, Antti Salovaara, Matthias Söllner, and Jan Marco Leimeister. 2020. Sara, the lecturer: Improving learning in online education with a scaffolding-based conversational agent. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14. <https://doi.org/10.1145/3313831.3376781>
- [78] Rong Wu and Zhonggen Yu. 2024. Do AI chatbots improve students learning outcomes? Evidence from a meta-analysis. *British Journal of Educational Technology* 55, 1 (2024), 10–33. <https://doi.org/10.1111/bjet.13334>
- [79] Song Yang, Ying Dong, and Zhong Gen Yu. 2024. ChatGPT in Education: Ethical Considerations and Sentiment Analysis. *International Journal of Information and Communication Technology Education (IJICTE)* 20, 1 (2024), 1–19. <https://doi.org/10.4018/IJICTE.346826>
- [80] Yvonne Zeegers and Katrina Elliott. 2019. Who's asking the questions in classrooms? Exploring teacher practice and student engagement in generating engaging and intellectually challenging questions. *Pedagogies: An International Journal* 14, 1 (2019), 17–32. <https://doi.org/10.1080/1554480X.2018.1537186>

- [81] Ruby Wenjiao Zhang, Xiaoning Liang, and Szu-Hsin Wu. 2024. When chatbots fail: exploring user coping following a chatbots-induced service failure. *Information Technology & People* 37, 8 (2024), 175–195. <https://doi.org/10.1108/ITP-08-2023-0745>
- [82] Jianlong Zhou, Heimo Müller, Andreas Holzinger, and Fang Chen. 2023. Ethical ChatGPT: Concerns, challenges, and commandments. *arXiv preprint arXiv:2305.10646* (2023).

A Appendix

A.1 Participant details

Table 6 represents the detailed demographic distribution based on the questionnaires.

A.2 Frequently asked questions

Here we present examples of frequently asked questions that highlight the most common topics of interest among participants. These examples provide insight into the areas where users sought additional clarification or deeper understanding while interacting with SAM.

- Linear regression:
 - “What is linear regression?”
 - “Define “linear regression” in really easy terms”
- Neural networks:
 - “what are neural networks in simple terms”
 - “Define a neural network”
- Weights:
 - “how to calculate the weight?”
 - “what is a weight matrix”
- Hidden layer:
 - “what is a hidden layer”
 - “what is the hidden layer?”
- Activation function:
 - “what do activation functions do?”
 - “define activation function”
- Tanh:
 - “what is a tanh function”
 - “what is tanh on the screen”
- CIFAR-10 : (image database with 10 classes)
 - “what is cifar?”
 - “What is CIFAR-10?”
- Summary:
 - “explain the whole video”
 - “please summarize the whole video shortly”

A.3 Hallucinations

We conducted a detailed analysis of all hallucinations that occurred during user interactions with SAM. We documented the specific inaccuracies in the LLM’s responses, providing commentary on why the given answer was incorrect. Additionally, we examined user reactions to these hallucinations, offering insights into how such errors impacted their experience and engagement with the system.

- Question: is it possible to see a visual example of an output (Context: “output” refers to the output of a last layer in the neural network.)
 Answer: Please focus on the lecture material.
 Incorrect: This was part of the lecture material.

User reaction: ah nice, in this slide he’s explaining what I’ve asked you before about the activation functions right?
 Comment: the participant did not ask further question on this topic, moved on with the lecture. No inaccuracies were stated by SAM.

- Question: what are indices (Context: matrix indices for W and b)

Answer: In the context of your lecture, indices are used to refer to specific elements within arrays, matrices, or tensors. Each index corresponds to a dimension of the array:

- (1) **Single Index:** If you have a vector (a one-dimensional array), a single index will refer to a specific element in that vector. For example, if \mathbf{v} is a vector, v_i would refer to the i -th element of \mathbf{v} .
- (2) **Multiple Indices:** When dealing with matrices (two-dimensional arrays) or higher-dimensional tensors, multiple indices are used. For example:
 - For a matrix \mathbf{M} , M_{ij} refers to the element in the i -th row and j -th column.
 - In neural networks, as mentioned in your transcript, a bias term b might have two indices. The first index b_l could indicate the layer in the neural network, and the second index $b_{l,n}$ could represent the specific neuron or node within that layer.

By using indices, you can refer to and manipulate specific elements within more complex data structures, which is essential for performing operations like matrix multiplication and applying transformations in neural networks.

Incorrect: The answer itself is correct, however it is misleading related to the lecture. In NN, the matrix indices do not necessarily have two, sometimes only the layer is mentioned.
 User reaction: define activation function (Comment: the participant did not ask further question on this topic, moved on with the lecture.)

- Question: can you explain what max represents in this video (Context: The maximum function is commonly used in various activation functions.)

Answer: In simple terms, an activation function in a neural network is a rule that decides whether a particular neuron should be activated or not. Think of it as a decision-making step that determines if the input information is important enough to pass on to the next layer of the network. The activation function introduces non-linearity to the model, enabling the neural network to learn and model complex data patterns. Here’s a very basic analogy: Imagine you are sorting emails, and your job is to decide whether each email should go to the “Important” folder or the “Trash” folder. The rules you use to make this decision (like checking if the email is from your boss or if the subject line contains certain keywords) are similar to what an activation function does. It evaluates the input and decides whether to pass the information along based on certain criteria. Would you like more details about the types of activation functions or their specific roles?

Incorrect: This answer aligns with the lecture, as it correctly highlights that while activation functions are generally non-linear to enable learning complex patterns. However, the

Table 6: Detailed Demographics data.

		Pilot Study		Main Study	
		Control	Test	Control	Test
Age		23.9 ± 2.3	23.4 ± 1.9	29.4 ± 5.8	28.4 ± 5.3
Gender	Man	9	11	20	21
	Woman	2	2	18	21
	Diverse	0	1	0	0
Employment	Working full-time	0	0	19	14
	Working part-time	2	1	1	7
	Unemployed	0	0	4	4
	Freelancer	0	0	3	2
	Student	9	13	11	15
Prior ML	No experience	6	4	17	14
	User	5	8	19	27
	Programmer	0	2	2	1

answer does not address the role of the maximum function in activation functions or provide specific examples.

User reaction: The conversation stopped.

- Question: how do we estimate the value of the hidden layer?
Answer: Estimating the value (or size) of the hidden layer in a neural network often involves a combination of theoretical understanding and empirical testing. Here are some common approaches:

(1) **Theoretical Guidelines:**

- **Input and Output Dimensions:** Consider the dimensions of your input and output. For example, if your input is high-dimensional, you might need a larger hidden layer to capture the complexity of the data.
- **Complexity of the Problem:** More complex problems might require larger or multiple hidden layers to learn the intricate patterns in the data.

(2) **Empirical Methods:**

- **Cross-Validation:** Use techniques like cross-validation to test different sizes for the hidden layer and determine which size provides the best performance on validation data.
- **Grid Search:** Perform a grid search over a range of values for the hidden layer size to systematically evaluate performance.

(3) **Common Practices:**

- A common starting point is to use hidden layer sizes that are between the input size and the output size. For instance, if your input layer has 16384 units and your output layer has 10 units, you might start with hidden layers having sizes anywhere from a few hundred to a few thousand units.
- Use powers of 2: Sizes such as 128, 256, 512, 1024, etc., are frequently used.

(4) **Regularization and Dropout:**

- When using larger hidden layers, it is essential to incorporate regularization techniques like dropout, L2 regularization, etc., to prevent overfitting.

In the context of the lecture, a hidden layer was given an arbitrary dimension, such as 1000, which is a reasonable choice for experimentation. The exact size would ideally be determined through empirical testing and cross-validation to find the optimal configuration for your specific task.

Incorrect: It seems likely that the participant was referring to the calculation of the dimensions of the hidden layer rather than the process of choosing its size. While the provided answer was technically correct, it addressed how to determine the size of a hidden layer, which may have been slightly misleading in the context of the question. A more precise explanation could help clarify the distinction between calculating dimensions and selecting the size based on specific criteria.

User reaction: how to find the neuron to weight connection

Comment: The participant continued the conversation, slightly changing the topic.

- Question: what value does the scalar add to this equation? (Context: The participant is referring to the bias term.)

Answer: The scalar in the context of the lecture adds a bias term to the equation. Biases are critical in neural networks because they allow the model to better fit and approximate complex functions. Let's break it down with an example: Consider the output of a neuron in a neural network layer, which can be calculated as: $y = W \cdot x + b$

- W is the weight matrix,
- x is the input vector,
- b is the bias term (scalar),
- y is the output.

The bias term b is added after the dot product of W and x . This addition shifts the activation function to better fit the data. Without a bias, the neuron would always output zero when all inputs are zero. The bias allows the activation function to be shifted left or right, which enables the neural network to fit the data more accurately. In the context of the video transcript:

“important when you’re talking about these biases B has two indices so 0 0 means there’s two indices and this is a scala that is added on top after you did this, this dot product here so this is a scaler and you have two indices here first one tells you the layer second one tells you the norm right this is a scalar this one is a vector and this one is also a vector right and doing a whole layer together is matrix vector multiplication plus a vector of biases” The scalar bias component ensures that each neuron can output a non-zero value even when the input is zero, thereby enhancing the model’s ability to learn complex data patterns.

Incorrect: The activation function is missing from the formula.

User reaction: sam, can you prepare an image for me of this structure with proper labelling of all componenets? *Comment: Asking for further explanation, which contains the activation function.*

- Question: where else could someone use the MaxOut function? (*Context: This is an introduced activation function in the lecture.*)

Answer: The MaxOut function can be used in various scenarios beyond traditional neural networks. Here are a few examples:

- (1) **Complex Neural Network Architectures:** MaxOut can be used in convolutional neural networks (CNNs) for image recognition tasks, recurrent neural networks (RNNs) for sequence prediction tasks, and other deep learning models to enhance model performance and representational capacity.
- (2) **Ensemble Methods:** MaxOut can also be useful in ensemble methods where multiple models’ outputs are combined. By selecting the maximum output among several model predictions, MaxOut can help in making more robust predictions.
- (3) **Generative Models:** In generative models like Generative Adversarial Networks (GANs), MaxOut can be employed in both generator and discriminator networks to improve the quality of generated samples.
- (4) **Automated Machine Learning:** In AutoML systems, MaxOut can serve as one of the activation function options that the system can explore to find the best model configuration automatically.
- (5) **Regularization Techniques:** MaxOut can act as a form of implicit regularization, providing robustness against overfitting by its inherent property of choosing the maximum activation.

These are just a few examples, and the use of MaxOut can extend to any domain where neural networks and machine learning models are applied.

Incorrect: The answer only collects NN related applications, without mentioning a broader field. The provided information was correct.

User reaction: The conversation stopped.

A.4 Linear Regression Summary from the Pre-Test Questionnaire

The following text appeared for each participant before watching the video:

“Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. The simplest form, simple linear regression, involves a single independent variable and models the relationship with the equation $y = m \cdot x + b$, where y is the predicted value, x is the independent variable, m is the slope of the line (indicating the relationship strength and direction between x and y), and b is the y-intercept (the value of y when x is zero). In multiple linear regression, the model includes multiple independent variables and extends to

$$y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + \dots + b_n \cdot x_n. \quad (1)$$

The coefficients b_i are estimated using methods such as Least Squares, which minimizes the sum of the squared differences between the observed values and the values predicted by the model. The goal is to find the best-fitting line that explains the variation in the dependent variable based on the independent variables, enabling predictions of y for given values of x .

Linear regression connects to classification tasks by providing a foundation for understanding how inputs (such as images of cats and dogs) can be used to predict outputs (labels - "cat" or "dog"). While linear regression predicts continuous values, classification tasks involve predicting discrete class labels. Techniques similar to linear regression, such as logistic regression, extend this concept by applying a transformation to the linear equation to handle binary or multiclass classification. This approach models the probability that a given input belongs to a particular class, enabling the assignment of class labels based on the highest predicted probability.”

A.5 Questionnaires - Pilot Study

A.5.1 Pre-Test. Topic 1: Graph Neural Networks

- (1) In a graph neural network, what do nodes and edges represent?
 - ☐ Nodes represent features, and edges represent labels.
 - ☐ Nodes represent computational units, and edges represent the flow of information or dependencies.
 - ☐ Nodes represent latent variables, and edges represent probabilistic dependencies.
 - ☐ Nodes represent entities with attributes, and edges represent the connections or correlations among them. ✓
- (2) What are significant challenges when working with graph neural networks?
 - ☐ Handling fixed-size input data.
 - ☐ Ensuring invariance to node permutations. ✓
 - ☐ Avoiding the use of activation functions.
 - ☐ Capturing the complex relationships in graph-structured data. ✓

Topic 2: Generative Adversarial Networks

- (3) Which of the following statements correctly describe the loss functions used in training Generative Adversarial Networks (GANs)?

- ☐ The generator's loss aims to minimize the Kullback-Leibler (KL) divergence between the generated data distribution and the real data distribution.
 - ☐ *The discriminator's loss aims to maximize the likelihood of correctly distinguishing between real and generated data. ✓*
 - ☐ *The generator's loss typically includes a term that maximizes the probability of the discriminator being incorrect. ✓*
 - ☐ The discriminator's loss is computed using the mean squared error between the real and generated data labels.
 - ☐ *Both generator and discriminator use cross-entropy loss functions in their original formulation. ✓*
- (4) Which of the following are key characteristics of generative models?
- ☐ They require labeled data for training.
 - ☐ *They generate new samples from the same distribution as the training data. ✓*
 - ☐ *They involve optimization techniques. ✓*
 - ☐ They are only used for regression tasks.

Topic 3: Reinforcement Learning

- (5) Which of the following best describes the difference between supervised learning, unsupervised learning, and reinforcement learning?
- ☐ *Supervised learning uses labeled data, unsupervised learning uses unlabeled data, and reinforcement learning uses reward signals to learn. ✓*
 - ☐ Supervised learning uses reward signals, unsupervised learning uses labeled data, and reinforcement learning uses unlabeled data.
 - ☐ Supervised learning uses unlabeled data, unsupervised learning uses labeled data, and reinforcement learning uses both labeled and unlabeled data.
 - ☐ Supervised learning uses labeled data, unsupervised learning uses reward signals, and reinforcement learning uses unlabeled data.
- (6) In reinforcement learning, what is a policy?
- ☐ A function that maps actions to states.
 - ☐ *A function that maps states to actions. ✓*
 - ☐ A reward signal over time.
 - ☐ A state transition model.

A.5.2 Post-Test. Topic 1: Graph Neural Networks

- (1) Which of the following are applications of graph neural networks?
- ☐ Autonomous vehicle navigation
 - ☐ *Protein structure prediction ✓*
 - ☐ Stock market prediction
 - ☐ *Social network analysis ✓*
- (2) In graph neural networks, what does the process of message passing involve?
- ☐ Passing fixed-size vectors through convolutional layers.
 - ☐ *Aggregating information from neighboring nodes to update node representations. ✓*
 - ☐ Using recurrent units to maintain state information.
 - ☐ Applying dropout for regularization.

Topic 2: Generative Adversarial Networks

- (3) Which of the following are generative models?

- ☐ *Variational Autoencoder ✓*
- ☐ Support Vector Machine
- ☐ *Generative Adversarial Network ✓*
- ☐ *Fully Visible Belief Nets ✓*
- ☐ Convolutional Neural Network
- ☐ *Boltzmann Machine ✓*

- (4) What are the primary differences between Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs)?

- ☐ VAEs use a discriminator to distinguish real from fake data, while GANs do not.
- ☐ *VAEs optimize a likelihood function, while GANs use a minimax game between generator and discriminator. ✓*
- ☐ VAEs require labeled data for training, while GANs do not.
- ☐ *VAEs explicitly model the data distribution, while GANs learn to generate data through adversarial training. ✓*

Topic 3: Reinforcement Learning

- (5) Which of the following concepts are fundamental to reinforcement learning?
- ☐ *Reward ✓*
 - ☐ Supervised learning
 - ☐ *Policy ✓*
 - ☐ Loss function
 - ☐ *Environment ✓*
 - ☐ Gradient Descent
- (6) Which of the following best describes the Markovian assumption in reinforcement learning?
- ☐ *The future state depends only on the current state and action, not on the sequence of events that preceded it. ✓*
 - ☐ The future state depends on the entire history of states and actions.
 - ☐ The future reward depends on the initial state of the environment.
 - ☐ The future state and reward are independent of the current action.

A.6 Questionnaires - Main Study

A.6.1 Pre-Knowledge Test.

- (1) What does 'overfitting' mean in the context of machine learning?
- ☐ When a model performs poorly on the training data
 - ☐ When a model performs poorly on the testing data
 - ☐ When a model performs equally well on both training and testing data
 - ☐ *When a model performs well on the training data but poorly on new, unseen data ✓*
- (2) Which algorithm is commonly used for training neural networks?
- ☐ *Gradient Descent ✓*
 - ☐ K-means Clustering
 - ☐ Principal Component Analysis
 - ☐ Apriori Algorithm
- (3) What is training in machine learning?
- ☐ Teaching people to use computers
 - ☐ *Improving a model with data ✓*
 - ☐ Writing a program to solve specific tasks

- ☐ Eliminating mistakes from a computer program
- (4) In neural networks, what is the purpose of an activation function?
 - ☐ To initialize the weights
 - ☐ To transform the input signal into an output signal ✓
 - ☐ To measure the error of the model
 - ☐ To update the weights during backpropagation
- (5) What is a dataset?
 - ☐ A set of computer programs
 - ☐ A type of database system
 - ☐ A collection of data for training ✓
 - ☐ Information available on Google
- (6) What is the primary goal of supervised learning?
 - ☐ To find patterns in unlabeled data
 - ☐ To generate new data
 - ☐ To perform dimensionality reduction
 - ☐ To find patterns in labeled data ✓

A.6.2 Pre-Test. Topic 1: Basics of Neural Networks

- (1) What are the main goals of a machine learning model?
 - ☐ Fit a function to the training data ✓
 - ☐ To approximate the data distribution
 - ☐ Optimize for the highest accuracy on the training set
 - ☐ Generalize from the training set ✓
- (2) What roles do non-linearities play in a neural network?
 - ☐ They enhance computational efficiency
 - ☐ They increase model complexity ✓
 - ☐ They ensure a more straightforward optimization process
 - ☐ They allow the network to approximate any continuous function ✓

Topic 2: Structure and Training of Neural Networks

- (3) Why is it ineffective to merely combine multiple linear models in a sequential or stacked fashion?
 - ☐ Multiplying matrices keeps the model linear ✓
 - ☐ Sequential stacking of linear models creates a non-linear model by default.
 - ☐ Optimization does not affect linearity
 - ☐ Stacking linear models cannot capture non-linear interactions or complexities. ✓
- (4) What is the role of weights in a neural network?
 - ☐ They adjust the learning rate
 - ☐ They are parameters that help the network learn ✓
 - ☐ They determine the strength of connections between neurons ✓
 - ☐ They define the network's architecture and layer structure

Topic 3: Activation Functions

- (5) Which of the following are common activation functions in neural networks?
 - ☐ Linear Function
 - ☐ Sigmoid Function ✓
 - ☐ Inverse Tangent Function
 - ☐ Hyperbolic Tangent Function ✓
- (6) How does the ReLU activation function operate?
 - ☐ Clamps negative values to zero ✓
 - ☐ Normalizes all values to a range between -1 and 1
 - ☐ Passes positive values unchanged ✓

- ☐ Transforms negative values to positive values while leaving positive values unchanged

A.6.3 Post-Test. Topic 1: Basics of Neural Networks

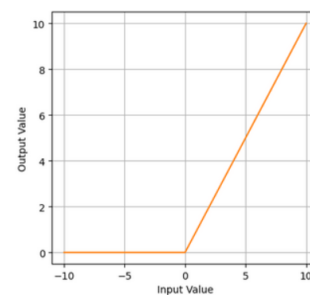
- (1) What is the primary function of a neural network?
 - ☐ To organize data linearly
 - ☐ To separate nonlinearities in data ✓
 - ☐ To approximate complex functions ✓
 - ☐ To execute matrix multiplication for feature extraction
- (2) Why are nonlinearities important in neural networks?
 - ☐ They enhance the network's ability to generalize ✓
 - ☐ They reduce the computation time
 - ☐ They normalize input data by scaling features
 - ☐ They enable approximation of arbitrary non-linear functions ✓

Topic 2: Structure and Training of Neural Networks

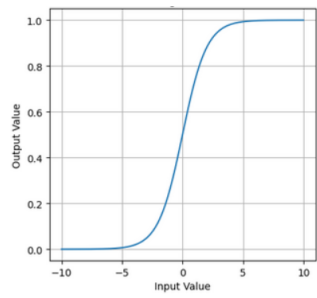
- (3) How is the dimensionality of weights (W) determined in a neural network?
 - ☐ By the number of layers in the network
 - ☐ By the number of neurons in the input and output layers ✓
 - ☐ By the total count of biases across all layers
 - ☐ By the product of neurons in adjacent layers ✓
- (4) How does a neuron in a neural network compute its output?
 - ☐ By averaging the weighted inputs and using this average to adjust the learning rate
 - ☐ By clustering inputs into groups and selecting outputs based on clusters
 - ☐ By performing a dot product of its inputs and weights ✓
 - ☐ By applying a nonlinearity to the input ✓

Topic 3: Activation Functions

- (5) Why are activation functions critical in neural networks?
 - ☐ They enable the network to model intricate data distributions ✓
 - ☐ They speed up training by accelerating convergence
 - ☐ They introduce non-linearities, allowing the network to learn complex patterns ✓
 - ☐ They normalize the data inputs for better performance
- (6) Which of the following is showing/describing the Sigmoid activation function?
 - ☐ a)



- ☐ b) $f(x) = \max(0.1x, x)$
- ☐ c) ✓



□ d) $f(x) = \frac{1}{1 + e^{-x}}$ ✓