Can an AI Partner Empower Learners to Ask Critical Questions?

Pratyusha Maiti*
pratyusha.maiti@gmail.com
Georgia Institute of Technology
Atlanta, Georgia, USA

Ashok K. Goel
ashok.goel@cc.gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Abstract

Jill Watson is an LLM-powered conversational AI partner integrated with instructor-provided courseware, offering learners contextually relevant and immediately applicable support. This study examines learner-generated questions as part of organic interactions with Jill embedded within classroom Learning Management System and investigates whether Jill empowers learners to ask higher-order questions. Leveraging Bloom's Taxonomy to assess question complexity, we collected over 5500 student questions from classroom deployments across three academic semesters and two educational settings. Student questions were classified using a fine-tuned BERT model and regression models were used to analyze the trends of complexity of the questions over time. Our results reveal a significant proportion of higher-order questions being asked in our classrooms, exceeding typical educational distributions. We also found a statistically significant increase in higher-order questioning with sustained interaction with Jill. These findings demonstrate that Jill empowers learners to engage in critical questioning, thereby enhancing their educational experience by promoting depth, relevance, and application of course concepts. Further research is recommended with larger and more diverse samples to generalize these findings.

CCS Concepts

• **Applied computing** → **Interactive learning environments**; • **Human-centered computing** → *Empirical studies in HCI*.

Keywords

Conversational AI Agents, Question Answering, Virtual Teaching Assistant

ACM Reference Format:

Pratyusha Maiti and Ashok K. Goel. 2025. Can an AI Partner Empower Learners to Ask Critical Questions? . In *30th International Conference on Intelligent User Interfaces (IUI '25), March 24–27, 2025, Cagliari, Italy*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3708359.3712134

*Currently at NVIDIA

IUI '25, March 24–27, 2025, Cagliari, Italy

ACM ISBN 979-8-4007-1306-4/25/03
https://doi.org/10.1145/3708359.3712134

1 Introduction

Conversational agents powered by Large Language Models (LLMs) are increasingly being integrated into educational settings to support student learning, shifting the focus toward more interactive and personalized experiences. An extensive part of current research focuses on analyzing the quality and effectiveness of LLM-generated responses, examining how well these tools provide accurate, contextually relevant, and pedagogically sound answers to specific learning questions and objectives [30]. However, there is a paucity of studies that examine the actual nature of the questions that students pose to these tools in a classroom setting. Studies have shown that student-initiated question-asking is a critical part of classroom learning, as it promotes active engagement, cognitive development, and deeper understanding of material through higher-order thinking skills. In fact, classrooms that promote higher-order thinking enable students to develop transferable skills, which are necessary for success in complex, real-world situations [42]. As AI technologies become increasingly prevalent in classrooms, understanding the dynamics of real interactions between students and an AI tool becomes crucial. This exploratory study aims to bridge this gap by investigating the complexity of questions initiated by students in an AI-assisted learning environment without any guided instruction or specially designed learning activities to promote the use of the tool. Specifically, we focus on whether the simple integration of an AI instructional partner called Jill Watson can foster an environment where learners feel empowered to ask more critical and higher-order questions, thus enhancing their engagement and promoting a deeper understanding of course concepts.

Jill Watson is an AI conversational instructional partner that answers students questions and engages them in extended conversations based on teacher-prescribed courseware using retrieval augmented generation for prompting off-the-shelf LLMs in the backend. Jill has been equipped with OpenAI's GPT3.5 Turbo model, accessed via the OpenAI API[1], and coupled with several other technologies to facilitate more nuanced, context-aware, and safe interactions with students [37]. Jill has been deployed in both online and hybrid classrooms across different educational institutes and courses [27]. We collect on over 5500 student interactions from these deployments to investigate the higher-order questions asked by students to Jill Watson and whether increased engagement with Jill influences the complexity of these questions asked.

Specifically, we focus on the following two research questions to explore the nature and trends of students' questions to Jill under organic classroom settings without any guided instructions:

- RQ1: Do students ask a higher proportion of higher-order questions to Jill Watson as compared with traditional classroom settings?

[1]https://openai.com/blog/ChatGPT/

- RQ2: Does the complexity of questions posed by learners to Jill Watson increase over time?

In subsequent sections, we provide theoretical motivation for our evaluation and describe the datasets, tools and analysis employed to study the complexity of student questions. Section 2 explores previous work on exploring student questions in classrooms and discusses our work in this context. Section 3 details the various sources of data and tools for analyzing student interaction with the tool as well as the evaluation metrics focusing on the cognitive dimensions of student queries. In section 4, we present our findings in the context of our two main research questions based on observations from different educational settings. We conclude with a summary of key takeaways, limitations and future work on studying student interactions with AI tools embedded in classrooms.

2 Related Work

2.1 Student Interaction with AI Tools

Previous research has explored the integration of AI tools into classroom environments, focusing on their potential to serve as virtual teaching assistants or personalized tutors. These tools have been designed to support learning in various ways, including answering student questions, facilitating discussions, and providing feedback, offering scalable, real-time assistance to educators and learners alike. In 2018, Goel and Polepeddi introduced a virtual teaching assistant, Jill Watson[22] built on top of IBM's Watson platform[17, 21]. Jill answered students' questions on course logistics in online discussion forums and was trained on historic human TA-student interactions. Since then Jill Watson has evolved from a virtual teaching assistant into an AI instructional partner, specifically created to deliver real-time, context-sensitive responses based on the course materials.

Research in this area has highlighted both the benefits and challenges of integrating AI tools into education. For instance, Gilson et al. [20] observed that responses generated by ChatGPT-based tool are structured in a manner that could lead to more in-depth questioning, stimulating students to leverage their knowledge and reasoning abilities. Conversely, Rudolph et al. [36] warned against over-reliance on these tools, cautioning that it should not be used as a substitute for critical thinking and originality. While AI tools can provide quick and reliable answers to factual questions, educators must carefully manage how these tools are used to support and foster, rather than diminish or replace, higher-order thinking.

In most of the work exploring role of AI tools in classrooms, the focus has been on evaluating AI responses to predefined question types and very little research has examined the nature of questions students ask AI tools in organic classroom settings. Understanding the types of questions students ask without guided interventions can provide insights into how they engage with AI, whether they leverage these tools for higher-order thinking or primarily for recall-based tasks. Our goal is to study, using data collected across multiple semesters, how students interact with an AI tool under no specific guidance and whether learners ask higher-order questions under such conditions.

Finally, existing literature predominantly relies on surveys and self-reports to gauge student usage of AI tools, which, while useful, may fail to capture the intricacies of real-time interactions between students and AI assistants [11]. In contrast, this study focuses on analyzing organic student-AI interactions to explore how students naturally engage with the AI tool in the classroom setting. By studying these interactions, we aim to provide a clearer understanding of how AI tools are used in live educational environments and their potential to foster critical thinking and deeper learning.

2.2 Classroom Questioning and Impact on Learning

A significant amount of literature on classroom questioning focuses on teacher-initiated questions and their impact on student learning[35][38][6][7][31][13]. In comparison, there has been relatively little research on an equally important component of classroom learning: student-initiated questions in classrooms. This could be because 'investigators can scarcely find any student questions' (Dillon (1988)[16]), potentially due to systemic conditioning or students finding alternate channels for question-asking, like among peers, or simply in their mind. However, student-generated questions hold a significant value in the learning process as it encourages independent thinking[8]. The early works by King (1994)[28], Rosenshine (1996)[34], Zoller(1987)[43] and Pizzini(1991)[32] show that students who engage in higher-order questioning demonstrate significantly improved problem-solving abilities. Later studies expanded this, finding that the quality and depth of student inquiries are strong predictors of cognitive growth[12]. Zohar and Dori (2003) [42] further demonstrated that encouraging higher-order thinking through questioning strategies leads to long-term improvements in critical thinking and science literacy. Several other works have rigorously studied the effects of higher-order questioning and concluded that it promotes deeper engagement and understanding of complex concepts [18][33][10][9].

In our study, we aim to investigate how students interact with an AI teaching assistant, Jill Watson by initiating questions. Specifically, we are interested in whether students pose higher-order questions to Jill Watson, which could indicate deeper engagement and critical thinking. This approach allows us to explore the role of Jill Watson in facilitating organic classroom interactions, particularly in promoting enhanced learning outcomes through student-driven inquiry. By examining these interactions, we aim to understand whether Jill Watson encourages students to engage with complex concepts, thereby extending the benefits of higher-order questioning previously highlighted by studies in human-led classrooms.

2.3 Frameworks to Examine Question Complexity

Previous research explores several different frameworks for assessing the type and needs of student questions, but most of these frameworks are heavily context-dependent and assess student questions for specific educational settings and outcomes. Pizzini and Shepardson (1991)[32] proposed a schema for classifying student questions into input (recall), processing (forming relationships), and output (evaluate, hypothesize, create) types based on cognitive levels, but this framework lacks detailed descriptions of cognitive processes and does not capture the full range of cognitive complexity. Other frameworks, like Pedrosa de Jesus et al. (2003)[14], analyze questions based on inquiry phases (e.g., confirmation or exploration), and Watts, Gould, & Alsop (1997)[39] focus on conceptual change in student questions by classifying them into consolidation, exploration and elaboration phases. In our study, we will use Bloom's Taxonomy[5] to define the complexity of student questions.

The application of Bloom's Taxonomy[5] in educational settings has been extensively explored, mainly for designing educational objectives and assessment tasks. The original taxonomy consists of six levels — Knowledge, Comprehension, Application, Analysis, Evaluation, and Synthesis. Each level represents a step in the cognitive process, from basic recall of facts to higher-order thinking skills such as critical analysis and creative problem-solving. Krathwohl (2002)[29] provided a foundational revision of Bloom's original cognitive taxonomy by emphasizing the dynamic nature of learning objectives and categorized them into cognitive processes that are crucial for structuring educational content and assessments. Since different types of student-initiated questions engage and challenge the mind to varying degrees, they can be categorized based on the level of cognitive effort required to formulate an answer. Previous studies[12] on exploring student questions also support the intuition behind using Bloom's taxonomy for identifying types of student questions. It provides an established and descriptive framework for analyzing cognitive complexity of questions without us having to define coding schemes across the different educational settings we have deployed Jill Watson in. Questions that demand higher-order thinking skills (i.e, fall under Application, Analysis, Evaluation, and Synthesis levels) reveal a deeper level of understanding and engagement with the educational content.

3 Methodology

To study whether Jill Watson empowers learners to ask critical questions, we deployed Jill across multiple classrooms and collected comprehensive data on organic student interactions with the tool. We analyze these interactions based on the framework provided by Bloom's Taxonomy and report the trends in cognitive complexity of questions asked.

3.1 Agent Design

Jill Watson, described by Taneja et al. [37], features a modular, Retrieval Augmented Generation (RAG)-based pipeline for question-answering. Figure 1 shows the question-answering pipeline employed by Jill. A new Jill Watson partner is created per course: each partner is pre-configured with a knowledge base by processing verified and relevant courseware from instructors. When a student asks Jill a question, Jill retrieves conversation history for that student from memory, and relevant information from the knowledge base, and constructs a prompt for OpenAI API to generate a response strictly restricted to the given context. The response is validated before being made available to the student: if a student asks a question that is not relevant to the course or contains toxic comments, Jill refuses to answer the question, encouraging students to stay on-topic. Our previous study has shown that by adding these guardrails and strictly restricting Jill's responses to instructor-vetted courseware, Jill outperforms ChatGPT in accuracy, relevance and safety[37]. Further design modifications are discussed in Kakar et al.[27] to tailor Jill for classroom integration. These adjustments helped scale Jill's

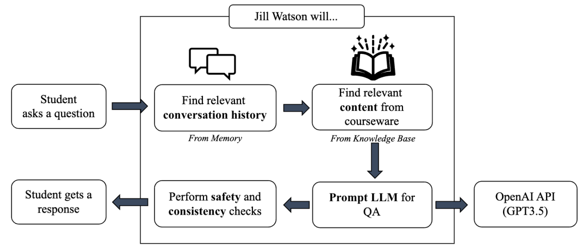

Figure 1: Architecture of Jill Watson. Adapted from Taneja et al.[37]

deployment across diverse educational environments, providing rich insights into student interactions.

3.1.1 *Deployments.*

Table 1 shows the courses and institutes where Jill Watson was deployed during Fall 2023, Spring 2024 and Summer 2024. The table also lists the courseware used to generate the knowledge base for a Jill Watson partner for each course on which all responses are grounded. The Knowledge-Based Artificial Intelligence (KBAI), Introduction to Cognitive Science (CogSci) and Human-Computer Interaction (HCI) courses are part of Georgia Institute of Technology (Georgia Tech) Online Master of Science in Computer Science program, designed primarily for part-time graduate students who typically maintain full-time employment. These courses aim to provide advanced knowledge and skills in Computer Science through a flexible online format. In contrast, Wiregrass Georgia Technical College (Wiregrass Tech) and Southern Regional Technical College (SRTC) are two-year community colleges within the Technical College System of Georgia (TCSG) recognized for their workforce development programs, and have integrated Jill into their undergraduate English: Composition and Rhetoric (English) courses. This course teaches various modes of writing and includes a review of standard grammatical and stylistic usage in proofreading and editing.

In each of these courses, Jill is deployed as a private chat interface within the Learning Management System (LMS) using Learning Tools Interoperability (LTI)[2]. Jill can be accessed through the course's Canvas[3] or Blackboard[4] page. This integration ensures that students receive assistance within their familiar course platform, promoting a seamless and supportive learning experience. Once a question is asked on the platform, a student typically receives a structured response within a few seconds. Instructors introduce Jill to students during the first week of classes and encourage them to interact freely with the tool anytime and anywhere. To set expectations for tool usage, students are provided with sample questions they could ask Jill, such as factual inquiries, comparison and contrast questions, or requests for relevant examples of concepts. No additional prompts or guided activities are provided. This deployment strategy aimed to observe natural, spontaneous interactions between students and the AI tool, minimizing external influences on question-asking behavior.

[2]https://www.1edtech.org/standards/lti
[3]https://www.instructure.com/canvas/
[4]https://www.blackboard.com/

Semester	Course/ Institute	Enrolled Students	Course Structure	Supported Courseware
Fall 23	KBAI/Georgia Tech	198	Coding assignments, journal writing, exams	Syllabus, Textbook
	Eng/Wiregrass Tech	100	Writing assignments	Syllabus, Textbook, MLA Guide, Instructor Handouts
Spring 24	KBAI/Georgia Tech	201	Coding assignments, journal writing, exams	Syllabus, Textbook
	CogSci/Georgia Tech	58	Weekly quizzes, writing assignments, final project	Syllabus, Textbook, Video Transcripts
	Eng/Wiregrass Tech	60	Writing assignments	Syllabus, Textbook, MLA Guide, Instructor Handouts
Summer 24	HCI/Georgia Tech	256	Coding assignments, quizzes, exams	Syllabus, Textbook, Weekly readings, Research papers
	Eng/Wiregrass Tech	25	Writing assignments	Syllabus, Textbook, MLA Guide, Instructor Handouts
	Eng/SRTC	30	Writing assignments, topic research	Syllabus, Schedule, Textbook, MLA Guide, Instructor Handouts

Table 1: Jill Watson deployments across two major settings: Georgia Tech and Community Colleges (Wiregrass Tech and SRTC) under the Technical College System of Georgia (TCSG)

3.1.2 Participants. This study included students from two distinct educational institutions: Georgia Tech (Georgia Tech) and Community Colleges (Wiregrass Tech and SRTC), offering a diverse cross-section of student demographics. To show a representative sample, Tables 2 and 3 show the demography of students enrolled in the KBAI course at Georgia Tech and the English course at Wiregrass Tech. At Georgia Tech, the majority of participants were Asian (51.25%) and White (33.54%), with a predominantly male (71.46%) cohort. Most students fell within the 25-34 age group (64.79%), indicating a mature student base balancing education with professional responsibilities. In contrast, Wiregrass Tech featured a larger proportion of female participants (78.00%) and a more ethnically diverse group, with Black/African American (40.00%) and White (42.00%) students forming the majority. The 17-24 age group (67.50%) dominated, highlighting a younger, more traditional college demographic.

3.2 Data Collection

We collected and stored exhaustive student interaction data for each deployment of Jill Watson in a persistent memory on the hosting server. This dataset includes de-anonymized student IDs, timestamps, questions, responses, along with inputs and outputs for each component of the question-and-answer (QA) pipeline. The tool also provides student with the option of marking a response as helpful or not helpful, and we collect this student feedback as well. Data collection adhered to strict privacy protocols, ensuring that all student information was anonymized and securely stored. Ethical approval was obtained from the institutional review board (IRB), and informed consent was secured from all participants prior to the study. This structured data collection allows for a robust analysis of usage patterns.

3.3 Fine-tuning BERT for Question Complexity Classification

To address our first research question on understanding the type and complexity of questions asked in classrooms, we use cognitive levels defined by Bloom's Revised Taxonomy[29]. We compiled a dataset of 3,600 labeled questions derived from various publicly accessible sources[19, 40]. This dataset contains questions across various disciplines and labels them into one of the six cognitive levels defined in the taxonomy- Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation, which range from basic recall of facts to generating new ideas. First, we investigated zero-shot classification techniques using GPT-3.5, GPT-4, and BERT[15] for categorizing student queries. Given the nuanced nature of this classification, zero-shot methods proved ineffective, with accuracies lower than 50%, prompting us to explore fine-tuning strategies on a curated dataset to improve performance.

For fine-tuning, we divided the dataset with 3600 questions into training and testing subsets with a 75:25 split. Standard preprocessing methods were applied to each question to ensure consistency across the data. These methods included tokenization, stopword removal, and text normalization, which were essential for minimizing variability and enhancing the efficiency of model training. We opted for a fine-tuning strategy on the *bert-base-uncased* BERT model to classify questions based on cognitive complexity. The model underwent fine-tuning over five epochs, using a batch size of 32 and a learning rate of 2e-5, parameters selected from initial tests that indicated optimal results. We settled on using the fine-tuned model that achieved an accuracy of 91.9%. We employ this model to classify all student questions collected from Jill Watson deployments across Fall 2023 to Summer 2024, and will use this classified result for our subsequent analysis.

Institute	Race/Ethnicity (%)							Gender (%)	
	American Indian /Alaskan Native	*Asian*	*Black/ African American*	*Hispanic /Latino*	*Two or more*	*Unk*	*White*	*Female*	*Male*
Georgia Tech	0	51.25	3.54	6.45	2.71	2.50	33.54	28.54	71.46
Wiregrass Tech	1.00	1.00	40.00	11.50	4.5	0	42.00	78.00	22.00

Table 2: Student ethnicity and gender across educational settings: Georgia Tech and Wiregrass Tech

Institute	Age Groups (%)				
	17-24	*25-34*	*35-44*	*45-54*	*55-64*
Georgia Tech	12.50	64.79	17.08	4.16	1.45
Wiregrass Tech	67.50	23.50	6.00	2.00	1.00

Table 3: Student age across educational settings: Georgia Tech and Wiregrass Tech

3.4 Analysis of Increasing Question Complexity

To answer our second research question, we want to investigate whether the complexity of question increases over the course of learners' interaction with Jill Watson throughout the semester. First, to simplify regression analysis, we categorize student questions into lower-order (comprising 'Knowledge' and 'Comprehension') and higher-order (comprising 'Application', 'Analysis', 'Evaluation' and 'Synthesis') types. Rather than analyzing the overall class output—where students might randomly ask a single recall-type question at any point during the semester—it is more meaningful to examine the usage patterns among students who have interacted with Jill Watson multiple times. When determining a threshold to classify users as "frequent," selecting an appropriate statistical measure—mean or median—is crucial. In our case, some students have very high usage while others have very low, resulting in a skewed distribution. Therefore, the median is a preferable threshold, as it is robust against skewness and provides a better measure of central tendency.

To assess the statistical significance of any increase in question complexity, we employ two statistical models:

- Fixed-Effects Logistic Regression (LR)[1]: This model assesses whether each additional interaction increases the likelihood of a student asking a higher-order question, controlling for individual-specific characteristics. By focusing on each student individually, the model accounts for characteristics that are unique to each person and don't change over time (like their baseline knowledge or learning style). This means we're comparing each student's progress against themselves, rather than against other students. Since we have a small number of frequent users, this model is suitable because it doesn't rely on large sample sizes to provide meaningful results.
- Generalized Estimating Equations (GEE)[24]: This model evaluates the population-averaged effect of additional interactions on question complexity. It looks at the overall trend across all students to see if, on average, additional interactions with Jill lead to asking more complex, higher-order questions.

3.4.1 Justification of Model Use. Since LR method only compares students against themselves, it eliminates the potential biases introduced by between-student differences, making it a powerful tool for identifying individual trends in question complexity as students engage more with the Jill. However, it assumes that individual characteristics remain static over time, which might introduce bias if student learning strategies evolve during the semester. Despite this limitation, LR is particularly valuable in our context where the sample size is small or highly unbalanced, as it does not rely on a large number of frequent users to generate meaningful insights.

GEE has been widely used for longitudinal data analysis [26] [41] [2] [23]. In educational data, where student engagement can be inconsistent, GEE is especially useful because it handles missing or unbalanced data effectively. This robustness is crucial in our setting where some students engaged with Jill more frequently than others. However, the efficiency of the GEE model depends on correctly specifying the correlation structure of the data, which can impact the precision of the estimates. Thus, we conclude that GEE is particularly well-suited for this type of repeated-measures data, where interactions by students in the same classroom are not independent but are correlated across time.

3.4.2 Effect Size. The effect size for both LR and GEE models is typically reported as odds ratios (ORs) as they directly quantify the likelihood of a particular outcome (e.g., asking a higher-order question) occurring with each additional interaction. Unlike other standardized effect sizes, ORs align with the logit link function used under the hood in these models, providing a consistent measure of association across contexts. ORs are calculated by exponentiating the model coefficients for each course.

3.4.3 Correcting for Multiple Tests. To address the heightened risk of Type I errors associated with multiple testing, we employed the Bonferroni correction[4]. This method adjusts the original significance level by dividing it by the number of hypotheses tested, thereby establishing a more stringent threshold for statistical significance. This approach is consistent with established literature that underscores the necessity of controlling for false positives in multiple comparisons[3].

By combining the aforementioned methods, we are able to examine both individual trajectories and overall classroom trends in question complexity. We report p-values for both models and along with the corrected p-values from Bonferroni test for each course to determine whether there is a statistically significant increase in question complexity over time. This dual approach allows us to rigorously assess whether the AI tool encourages deeper cognitive engagement and promotes higher-order thinking skills, which are critical to achieving improved learning outcomes.

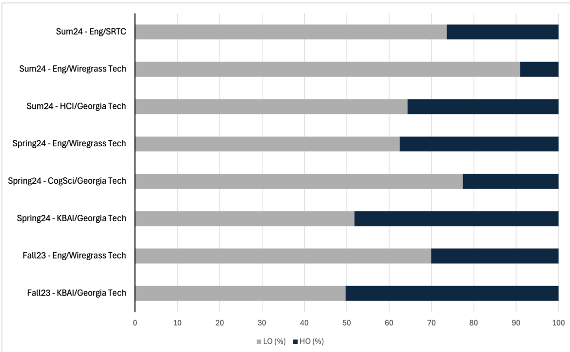

Figure 2: Proportion of Question by Complexity Per Semester Per Course

4 Results

We classify all relevant student questions across three semesters into complexity levels using the fine-tuned BERT. Next, we use the statistical models discussed to study the trends in complexity of questions over extended student interactions with Jill Watson.

4.1 What proportion of learner questions are of a higher order?

Table 5 shows the total number of enrolled students, total questions asked and the distribution of questions by complexity for each deployment under study. To address our first research question, our hypothesis is that if the proportion of higher-order questions (Application, Analysis, Evaluation, and Synthesis) posed to Jill Watson is significantly higher than what would be expected in a typical learning environment without such an AI assistant, it suggests that the presence of Jill could empower learners to ask more critical questions.

Previous studies [12][42][25] have established that in traditional classroom settings, the majority of student questions are of lower order, primarily seeking clarifications or recalling factual information. Only approximately 10–20% of questions in these environments are categorized as higher-order. In contrast, as seen in Figure 2 our study consistently observed a higher percentage of higher-order questions across various classes. Note that in the English course at Wiregrass Tech, the total number of questions was too limited to draw significant conclusions. However, in all other classes examined, higher-order questions constituted between 30–40% of the total inquiries posed to Jill.

Table 4 lists some of the real student questions observed as part of this study to show examples for questions at each level of cognitive complexity. These findings indicate that the presence of the Jill may create an environment where students feel more comfortable and empowered to ask more sophisticated questions. It is important to acknowledge that while we report numbers higher than related studies have shown, the contexts of previous studies differ, as they analyzed student questions in traditional classroom settings without AI tools and the definition of "higher-order" varies across studies, with varied coding systems being used based on the specific learning outcomes of each educational setting. Although a direct comparison could be made with a traditional setting at the same university, such a comparison would not be practical or equitable. In conventional environments, questions are usually posed in discussion forums, during office hours, or in class, which does not provide a valid basis for comparison with the continuous 1:1 interactions facilitated by a 24/7 non-human agent.

In this regard, our study contributes a baseline measure of question complexity within an organic classroom environment augmented by an AI tool, without any guided interventions. This baseline serves as a reference point for understanding how the introduction of AI agents in a classroom can influence the nature of student inquiries in real-world educational contexts.

4.2 Does the complexity of questions increase over sustained interaction?

To investigate whether the complexity of questions posed by learners to Jill Watson increases over time—reflecting growing confidence and initiative in self-directed learning—we conducted Fixed-Effects Logistic Regression and Generalized Estimating Equations (GEE) analyses for multiple courses across three semesters. In case of English courses at Wiregrass Tech, there were signs of quasi-separation, meaning that some variables could perfectly predict outcomes. Quasi-separation typically occurs when there is a very small number of observations, as seen in case of English at Wiregrass Tech, which may lead to inflated coefficients and inaccurate inferences. Hence, we have dropped that course from our analysis. Table 6 summarizes the coefficient estimates, standard errors, z-values, and p-values for the interaction predictor from both models for the rest of the courses. To account for possible over-inflation in significance reporting due to multiple testing, we performed a Bonferroni correction and reported the corrected p-values in Table 7.

4.2.1 Knowledge-Based Artificial Intelligence (KBAI) at Georgia Tech. In Fall 2023, the KBAI course exhibited a significant positive relationship between subsequent interactions and question complexity. The logistic regression yielded a coefficient of 0.0099 (std. err. = 0.002), with a z-value of 3.970 and a highly significant p-value ($P>|z| = 0.000$). Similarly, the GEE model confirmed this finding with a coefficient of 0.0111 (std. err. = 0.004), z-value of 3.131, and $P>|z| = 0.002$. The p-values adjusted for multiple testing also indicate a significant correlation. This suggests that as students interact with Jill more frequently, they ask increasingly complex questions, reflecting greater confidence and initiative in learning. In Spring 2024, the KBAI course again showed a significant positive effect in the logistic regression model, with a coefficient of 0.0103 (std. err. = 0.003), z-value of 3.226, and $P>|z| = 0.001$. However, the GEE model did not find this relationship to be significant (coef. = 0.0020, std. err. = 0.004, z = 0.523, $P>|z| = 0.601$), suggesting that while some individual students show a pattern of increasing question complexity, this trend does not hold across the overall population.

4.2.2 Introduction to Cognitive Science (CogSci) at Georgia Tech. For the CogSci course in Spring 24, the logistic regression indicated a significant positive association between session number and question complexity (coef. = 0.0222, std. err. = 0.010, z = 2.326, $P>|z| = 0.020$, corrected $P>|z| = 0.0399$ after Bonferroni correction). The

Dimension	Real Student Question
Knowledge	Can you help me define heuristics and give me an example?
Comprehension	Can you summarize the second chapter for me?
Application	Would rotating an image or finding symmetry in an image be considered conceptual or imagistic represention?
Analysis	What is the difference between control knowledge and heuristics?
Evaluation	Fear can be interpreted as a shifting of the allocation of computational resources to expand CRUM. Computational resources, such as memory, attention, etc., are limited and fear can allocate those precious resources to the urgent situation that is relevant to surviving goal. Is there anything wrong in my proposition?
Synthesis	Consider an extended version of the scientific method that incorporates conceptual and simulation models. How might scientists pairing conceptual models with simulation models of the same concepts be used as part of the scientific method?

Table 4: Examples of real student questions per cognitive dimension, showing the evolving complexity of questions asked by students.

Semester	Course/ Institute	Total Qs	Know (%)	Comp (%)	App (%)	Anlys (%)	Eval (%)	Synth (%)	LO (%)	HO (%)
Fall23	KBAI/Georgia Tech	2311	35.91	13.80	14.76	24.02	6.66	4.85	**49.71**	**50.29**
	Eng/Wiregrass Tech	163	64.42	5.52	3.07	3.68	16.56	6.75	**69.94**	**30.06**
Spring24	KBAI/Georgia Tech	1920	35.89	15.94	7.24	24.84	7.86	8.23	**51.83**	**48.17**
	CogSci/Georgia Tech	332	61.45	15.96	5.12	10.54	3.31	3.61	**77.41**	**22.59**
	Eng/Wiregrass Tech	32	50.00	12.50	3.12	3.12	18.75	12.50	**62.50**	**37.50**
Summer24	HCI/Georgia Tech	561	49.38	14.97	4.81	9.81	13.90	7.13	**64.35**	**35.65**
	Eng/Wiregrass Tech	11	72.73	18.18	0.00	0.00	9.09	0.00	**90.91**	**9.09**
	Eng/SRTC	235	55.74	17.87	4.68	5.53	8.51	7.66	**73.62**	**26.38**

Table 5: Distribution of type of student questions across courses and semesters grouped by question complexity. Last two columns show questions grouped into lower order ('Knowledge' and 'Comprehension') and higher order ('Application', 'Analysis', 'Evaluation' and 'Synthesis').

Semester	Course / Institute	Median	Logistic Regression					GEE Regression				
			coeff	std err	z	P>|z|	OR	coeff	std err	z	P>|z|	OR
Fall 23	KBAI/Georgia Tech	10.50	0.0099	0.002	3.970	**0.000**	1.0099	0.0111	0.004	3.131	**0.002**	1.0111
Spring 24	KBAI/Georgia Tech	2.50	0.0103	0.003	3.226	**0.001**	1.0103	0.0020	0.004	0.523	**0.006**	1.0020
	CogSci/Georgia Tech	4.00	0.0222	0.010	2.326	**0.020**	1.0224	0.0182	0.010	1.740	0.082	1.0183
Summer 24	HCI/Georgia Tech	2.00	0.0060	0.007	0.795	0.427	1.0060	0.0057	0.008	0.727	0.467	1.0057
	Eng/SRTC	4.00	0.0077	0.010	0.754	0.451	1.0077	0.0100	0.006	1.564	0.118	1.0100

Table 6: Fixed-Effects Logistic Regression and Generalized Estimating Equations Analyses: This table presents the coefficient estimates, standard errors, z-values, p-values and odds ratio (OR) for the session number predictor from both models.

GEE model showed a positive coefficient (coef. = 0.0182), but the effect was not statistically significant (std. err. = 0.010, z = 1.740, P>|z| = 0.082). This result indicates that individual students, on average, tend to ask more complex questions over time, but the population-wide effect is not as strong.

4.2.3 Human Computer Interaction (HCI) at Georgia Tech. In Summer 24, the HCI course did not exhibit a significant relationship between session number and question complexity in either model. The logistic regression yielded a coefficient of 0.0060 (std. err. = 0.007, z = 0.795, P>|z| = 0.427), and the GEE model produced similar non-significant results (coef. = 0.0057, std. err. = 0.008, z = 0.727, P>|z| = 0.467). This suggests that in this course, there is no clear evidence of increasing question complexity as students interact more frequently with Jill.

4.2.4 English: Composition and Rhetoric (English) at SRTC. The English course at SRTC during Summer 24 also showed no significant increase in question complexity over time. The logistic regression coefficient was 0.0077 (std. err. = 0.010, z = 0.754, P>|z| = 0.451), and the GEE model coefficient was 0.0100 (std. err. = 0.006, z = 1.564, P>|z| = 0.118). Both models suggest that session number did not significantly predict the complexity of questions in this setting.

4.2.5 Interpretation of Findings. The findings partially support the hypothesis that the complexity of questions posed by learners to Jill Watson increases over time, reflecting growing confidence and

Semester/ Course/ Institution	Logistic Regression		GEE Regression	
	p-value	corrected p-value	p-value	corrected p-value
Fall 2023/ KBAI/ Georgia Tech	**0.000**	**0.0028**	**0.002**	**0.0035**
Spring 2024/ KBAI/ Georgia Tech	**0.001**	**0.0025**	**0.006**	**0.0130**
Spring 2024/ Cogsci/ Georgia Tech	**0.020**	**0.0399**	0.082	0.1636
Summer24/ HCI/ Georgia Tech	0.427	15.7887	0.467	0.9346
Summer24/ English/ SRTC	0.451	4.5095	0.118	0.2358

Table 7: Corrected p-value from Bonferroni correction on LR and GEE models.

initiative in self-directed learning. A small positive odds ratio supports positive association in each case. The significant positive associations observed in the KBAI course suggest that sustained interactions with the AI assistant can enhance learners' propensity to engage in higher-order questioning. However, the lack of consistent significant results across all courses indicates that this effect may be context-dependent. Factors such as course content, instructional design, and initial student engagement levels might influence how learners interact with Jill Watson over time.

5 Discussion

5.1 Interaction Between Students and Jill Watson

Tables 8 and 9 show real conversations logged between a student in CogSci class at Georgia Tech and Jill Watson.

In the first example, in a 4-turn conversation, a student asks for the basic components of CRUM, and Jill provides a detailed explanation of representational structures and computational procedures. The questions get progressively complex, as the student explores interpretations of a human emotion within CRUM and comparing with other emotions. Finally, Jill evaluates the student's interpretation, affirming its correctness while offering additional nuances. Each turn demonstrates progressively higher levels of cognitive engagement, from recalling basic facts to analyzing complex ideas and evaluating a hypothesis. Similarly, in the second example, the student starts with a simple question asking for a definition, and gradually moves towards more complex, synthesis-level queries.

6 Limitations

While we have identified a trend towards higher-order question-asking in certain course structures, the study does not explore the underlying factors driving these behaviors. More granular investigations into how specific pedagogical approaches, course designs, and external factors like digital literacy influence student interaction with AI tools are necessary to draw more precise conclusions. Second, while our analysis covers a range of question types and cognitive levels, it does not fully address the quality of the responses provided by Jill. Future studies should include a more detailed assessment of response accuracy and pedagogical soundness to ensure that the tool not only engages students but also supports effective learning. Third, to study the impact of sustained conversations on learning outcomes, we need a larger sample size for more robust findings. While the findings are promising, the study's exploratory design and limited sample size may limit the generalizability of the results. Future research with larger and more diverse populations is necessary to validate these initial observations. Future studies could investigate the impact of structured prompts on the types of questions asked, or explore how different AI tool features influence student engagement and cognitive processes. Finally, we are unable to share the the classroom data used in our analysis due to strict confidentiality protocols.

7 Ethical Considerations

Before each deployment, we take into consideration an instructor's need for the tool. We engage with instructors to discuss the format and content of the relevant course material that can be used to develop Jill Watson, ensuring that responses are grounded in instructor-vetted material. We also provide instructors with a detailed demonstration of the tool that allows them to understand how it works and provide relevant guidance to students. We ensure complete clarity and transparency regarding the technology and capabilities of the tool when communicating with instructors. Regarding data handling, we strictly adhere to ethical standards and IRB protocols. All student data is collected with informed consent, and access is restricted to authorized personnel. To protect privacy, the data is anonymized prior to analysis, ensuring that individual student identities remain confidential throughout the research process.

8 Conclusion

Through this exploratory study, we examined the cognitive complexity of student questions posed to an AI instructional partner (Jill Watson) embedded in the classroom LMS. Without guided instruction or specific learning activities, over 30-50% of classroom questions were higher-order, significantly surpassing traditional classroom benchmarks of around 20%. This indicates that conversational AI partners like Jill have the potential to foster critical thinking by encouraging students to engage with more complex concepts through active, self-driven knowledge construction. We also found that students who interacted with the AI tool more frequently exhibited a trend of increasing higher-order questions over time. This trend was more pronounced among individual students than at the overall classroom level. While we did not directly measure learning outcomes, prior work has consistently shown that classrooms that foster higher-order questioning see improvements in student problem-solving abilities and critical thinking [12]. This suggests that conversational AI partners like Jill could be valuable assets in supporting not only student engagement but also

#	Student Question and Jill Watson Response	Level
1	Student: Basic components of CRUM? Jill Watson: The basic components of the Computational-Representational Understanding of Mind (CRUM) include: 1. Representational Structures: CRUM posits that the mind uses representational structures to encode and manipulate information. These structures can take various forms, such as symbolic representations or neural networks, and they capture the content and organization of mental states. 2. Computational Procedures: CRUM suggests that the mind operates through computational procedures that manipulate the representational structures. These procedures involve the transformation, combination, and manipulation of mental representations to produce cognitive behaviors. [...] It is important to note that CRUM is a broad framework that encompasses various theories and models within cognitive science. The specific details and interpretations of these components may vary among different researchers and theories within the CRUM framework.	Know
2	Student: Can we interpret Fear as a shifting of the allocation of computational resources to expand CRUM? Jill Watson: Yes, fear can be interpreted as a shifting of the allocation of computational resources within the framework of CRUM. When experiencing fear, our attention and cognitive resources become focused on the fear-inducing stimulus, such as a threatening situation or object. This allocation of resources is a response to the perceived danger and is aimed at enhancing our ability to respond effectively. [...] By interpreting fear as a shifting of computational resources, CRUM can provide insights into how emotions influence cognitive processes. It highlights the dynamic nature of cognitive functioning and how emotions can modulate the allocation of resources to adaptively respond to different situations.It is worth noting that this interpretation is one way to incorporate emotions, such as fear, into the CRUM framework. Different researchers and theories within CRUM may have their own specific interpretations and models of how emotions interact with computational processes.	Comp
3	Student: Common patterns between fear and anger in CRUM is that they're both related to computational resources? Jill Watson: Yes, within the CRUM framework, fear and anger share a common pattern in that they both involve the allocation and redirection of computational resources. In the case of fear, the allocation of computational resources is focused on the fear-inducing stimulus. This allocation allows for heightened attention, memory retrieval, and problem-solving related to the perceived threat. The goal is to effectively respond to the threat and ensure survival or safety. Similarly, anger involves the reallocation of computational resources towards addressing the interference or obstruction that is preventing the achievement of a goal. This reallocation directs attention, memory, and problem-solving towards resolving the conflict and removing the obstacle. The goal is to restore goal-directed behavior and overcome the hindrance. Both fear and anger can be seen as adaptive responses within the CRUM framework, as they involve the mobilization of cognitive resources to deal with specific challenges or threats. They highlight the dynamic nature of cognitive processes and the role of emotions in modulating the allocation of computational resources to achieve desired outcomes.	Anlys
4	Student: Fear can be interpreted as a shifting of the allocation of computational resources to expand CRUM. Computational resources, such as memory, attention, etc., are limited and fear can allocate those precious resources to the urgent situation that is relevant to surviving goal. Is there anything wrong in my proposition? Jill Watson: Fear can indeed be interpreted as a shifting of the allocation of computational resources within the CRUM framework, but it is important to clarify that this interpretation does not necessarily expand CRUM itself. Instead, it highlights how fear influences the allocation of computational resources in cognitive processes. By interpreting fear as a shifting of computational resources, CRUM can provide insights into how emotions impact cognitive processes. It emphasizes the dynamic nature of cognitive functioning and how emotions can modulate the allocation of resources to adaptively respond to different situations. However, it is important to note that this interpretation of fear as a shifting of computational resources is just one perspective within the CRUM framework. Different researchers and theories within CRUM may have their own specific interpretations and models of how fear interacts with computational processes.	Eval

Table 8: Conversation between a student and Jill Watson in CogSci in Georgia Tech in Spring 2024.

#	Student Question and Jill Watson Response	Level
1	Student: What does model mean in cognitive science? Jill Watson: In cognitive science, a model refers to an abstract representation or simulation of a cognitive process, system, or phenomenon. It is a simplified and formalized version of the cognitive processes being studied, designed to capture the essential features and mechanisms involved. A cognitive model can take various forms, depending on the specific research question and the theoretical framework being used. It can be a computational model,which uses algorithms and data structures to simulate cognitive processes. Alternatively, it can be a conceptual model, which describes the underlying principles and relationships between different components of the cognitive system.[...]	Know
2	Student: A simulation model uses simplifying assumptions to generate a prediction of the dynamics and outcome of a modeled scenario. Based on your current knowledge, how might simulation models be incorporated into the scientific method? Jill Watson: Simulation models can be incorporated into the scientific method in several ways: 1. Hypothesis generation: By specifying the assumptions and parameters of the model, researchers can simulate different scenarios and observe the predicted dynamics. These predictions can then be used to formulate hypotheses about the underlying mechanisms or relationships in the system. 2. Experiment design: Researchers can use simulation models to explore different experimental conditions, identify critical factors, and optimize the experimental design to maximize the information gained. [...] Overall, simulation models can be a valuable tool within the scientific method, aiding in hypothesis generation, experiment design, data generation, parameter estimation, hypothesis testing, and theory development. They provide a means to explore complex systems, make predictions, and test hypotheses in a controlled and iterative manner.	App
3	Student: Consider an extended version of the scientific method that incorporates conceptual and simulation models. How might scientists pairing conceptual models with simulation models of the same concepts be used as part of the scientific method? Jill Watson: Pairing conceptual models with simulation models of the same concepts can be a powerful approach within an extended version of the scientific method. Here are some ways in which scientists can use this pairing as part of the scientific method: 1. Conceptual model development: [...] 2. Simulation model construction: Based on the conceptual model, scientists can then construct a simulation model that translates the abstract concepts and relationships into a computational framework. The simulation model incorporates simplifying assumptions and mathematical or computational algorithms to simulate the dynamics and behavior of the system. 3. Hypothesis formulation: The conceptual model can be used to generate hypotheses about the behavior or outcomes of the system. Scientists can formulate specific hypotheses based on the conceptual model, which can then be tested using the simulation model. The simulation model allows for the generation of predictions and the exploration of different scenarios to evaluate the validity of the hypotheses. [...] By pairing conceptual models with simulation models, scientists can integrate theoretical understanding with computation simulations, allowing for a more comprehensive and iterative approach to scientific inquiry. This pairing enables scientists to generate hypotheses, design experiments, generate data, estimate parameters, test hypotheses, and refine theories in a systematic and rigorous manner.	Synth

Table 9: Conversation between a student and Jill Watson in CogSci in Georgia Tech in Spring 2024.

advanced cognitive development. However, the pattern of increasing question complexity varied across different contexts, indicating that the effectiveness of such tools may depend on specific course designs or student populations. Additionally, we acknowledge a limitation of our study: we did not evaluate the quality of the AI partners's responses to student questions. Understanding how the quality of AI-generated responses impacts student learning is essential and should be explored in future research. These findings demonstrate the potential of AI partners like Jill to promote advanced learning outcomes by offering students opportunities for complex, on-demand interactions. Future research should explore the long-term impacts of integrating AI instructional partners in education, consider studies that address different educational contexts for more generalizable outcomes, and optimize AI functionalities to support a wider range of educational objectives.

Acknowledgments

This research has been supported by NSF Grants #2112532 and #2247790 to the National AI Institute for Adult Learning and Online Education headquartered at Georgia Tech, and a Bill & Melinda Gates Foundation Grant to Georgia Tech. This research was done

when both authors were affiliated with Georgia Tech's Design Intelligence Laboratory.

References

[1] Paul D. Allison. 2009. *Fixed effects regression models.* SAGE Publications.

[2] Gary A. Ballinger. 2004. Using generalized estimating equations for longitudinal data analysis. *Organizational Research Methods* 7, 2 (2004), 127–150.

[3] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57, 1 (1995), 289–300. http://www.jstor.org/stable/2346101

[4] J M. Bland and D. G Altman. 1995. Statistics notes: Multiple significance tests: the Bonferroni method. *BMJ* 310, 6973 (Jan. 1995), 170–170. https://doi.org/10.1136/bmj.310.6973.170

[5] Benjamin S. Bloom and David R. Krathwohl. 1956. *Taxonomy of Educational Objectives: The Classification of Educational Goals.* Vol. Handbook I: Cognitive Domain. Longmans, Green, New York, NY. By a committee of college and university examiners.

[6] Patricia E. Blosser. 1995. *How to Ask the Right Questions.* National Science Teachers Association, Arlington, VA.

[7] Jo Boaler and Karin Brodie. 2004. The importance, nature and impact of teacher questions. In *Proceedings of the twenty-sixth annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*, Vol. 2. Citeseer, 774–782.

[8] M. H. Bowker. 2010. Teaching Students to Ask Questions Instead of Answering Them. *The NEA Higher Education Journal* (2010).

[9] John D. Bransford, Ann L. Brown, and Rodney R. Cocking. 2000. *How people learn: Brain, mind, experience, and school.* National Academy Press.

[10] Susan M. Brookhart. 2010. *How to assess higher-order thinking skills in your classroom.* ASCD.

[11] Cecilia Ka Yuk Chan and Wenjie Hu. 2023. Students' voices on generative AI: perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education* 20, 1 (July 2023). https://doi.org/10.1186/s41239-023-00411-8

[12] Christine Chin and Jonathan Osborne. 2008. Students' questions: a potential resource for teaching and learning science. *Studies in Science Education* 44, 1 (March 2008), 1–39. https://doi.org/10.1080/03057260701828101

[13] Marilyn Cochran-Smith and Susan L Lytle. 1998. Teacher research: The question that persists. *International Journal of Leadership in Education Theory and Practice* 1, 1 (1998), 19–36.

[14] Helena Pedrosa De Jesus, José J. C. Teixeira-Dias, and Mike Watts. 2003. Questions of chemistry. *International Journal of Science Education* 25, 8 (Aug. 2003), 1015–1034. https://doi.org/10.1080/09500690305022

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/ARXIV.1810.04805

[16] J. T. Dillon. 1988. The Remedial Status of Student Questioning. *Journal of Curriculum Studies* 20, 3 (1988), 197–210.

[17] Bobbie Eicher, Lalith Polepeddi, and Ashok Goel. 2018. Jill Watson Doesn't Care if You're Pregnant: Grounding AI Ethics in Empirical Studies. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New Orleans, LA, USA) *(AIES '18)*. Association for Computing Machinery, New York, NY, USA, 88–94. https://doi.org/10.1145/3278721.3278760

[18] Robert Fisher. 1998. *Teaching thinking: Philosophical enquiry in the classroom.* Bloomsbury Publishing.

[19] Mohammed Osman Gani and Anbuselvan Sangodiah. 2023. Exam Question Datasets. https://figshare.com/articles/dataset/Exam_Question_Datasets/22597957. https://doi.org/10.6084/m9.figshare.22597957.v3

[20] Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, and David Chartash. 2023. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Medical Education* 9 (Feb. 2023), e45312. https://doi.org/10.2196/45312

[21] Ashok Goel. 2020. AI-Powered Learning: Making Education Accessible, Affordable, and Achievable. https://doi.org/10.48550/ARXIV.2006.01908

[22] Ashok Goel and Lalith Polepeddi. 2018. Jill Watson: A Virtual Teaching Assistant for Online Education. In *Education at Scale: Engineering Online Teaching and Learning*, C. Dede, J. Richards, and B. Saxberg (Eds.). Routledge, New York.

[23] James A. Hanley and Abdissa Negassa. 2003. Statistical analysis of correlated data using generalized estimating equations: an orientation. *American Journal of Epidemiology* 157, 4 (2003), 364–375.

[24] James W. Hardin and Joseph M. Hilbe. 2012. *Generalized Estimating Equations.* Chapman and Hall/CRC. https://doi.org/10.1201/b13880

[25] William Hare and John D. Pulliam. 1980. Effective questions in classroom teaching. *Educational Leadership* 38, 3 (1980), 190–194.

[26] Alan E. Hubbard, Jennifer Ahern, Nancy L. Fleischer, Mark Van Der Laan, Sheri A. Lippman, Nicholas Jewell, Tim Bruckner, and William A. Satariano. 2010. To GEE or not to GEE: comparing population averaged and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology* 21, 4 (2010), 467–474.

[27] Sandeep Kakar, Pratyusha Maiti, Karan Taneja, Alekhya Nandula, Gina Nguyen, Aiden Zhao, Vrinda Nandan, and Ashok Goel. 2024. Jill Watson: Scaling and Deploying an AI Conversational Agent in Online Classrooms. In *Proceedings of the 20th International Conference on Intelligent Tutoring Systems (ITS 2024)*. Springer, Thessaloniki, Greece.

[28] Alison King. 1994. Autonomy and question asking: The role of personal control in guided student-generated questioning. *Learning and Individual Differences* 6, 2 (Jan. 1994), 163–185. https://doi.org/10.1016/1041-6080(94)90008-6

[29] David Krathwohl. 2002. A Revision of Bloom's Taxonomy: An Overview. *Theory Into Practice - THEORY PRACT* 41 (Nov. 2002), 212–218. https://doi.org/10.1207/s15430421tip4104_2

[30] Lasha Labadze, Maya Grigolia, and Lela Machaidze. 2023. Role of AI chatbots in education: systematic literature review. *International Journal of Educational Technology in Higher Education* 20, 1 (Oct. 2023). https://doi.org/10.1186/s41239-023-00426-1

[31] Youngju Lee and Mable B Kinzie. 2012. Teacher question and student response with regard to cognition and language use. *Instructional science* 40 (2012), 857–874.

[32] Ernest L. Pizzini and Daniel P. Shepardson. 1991. Student Questioning in the Presence of the Teacher During Problem Solving in Science. *School Science and Mathematics* 91 (1991), 348–352.

[33] Lauren B. Resnick. 1987. *Education and learning to think.* National Academies Press.

[34] Barak Rosenshine, Carla Meister, and Saul Chapman. 1996. Teaching Students to Generate Questions: A Review of the Intervention Studies. *Review of Educational Research* 66, 2 (1996), 181. https://doi.org/10.2307/1170607

[35] Mary Budd Rowe. 1987. Using Wait Time to Stimulate Inquiry. In *Questions, Questioning Techniques, and Effective Teaching*, William W. Wilen (Ed.). National Education Association, Washington, DC, 95–106.

[36] J. Rudolph, T. Samson, and T. Shannon. 2023. ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning & Teaching* 6, 1 (Jan. 2023). https://doi.org/10.37074/jalt.2023.6.1.9

[37] Karan Taneja, Pratyusha Maiti, Sandeep Kakar, Pranav Guruprasad, Sanjeev Rao, and Ashok Goel. 2024. Jill Watson: A Virtual Teaching Assistant powered by ChatGPT. In *Proceedings of AIED 2024*.

[38] Kenneth Tobin. 1987. The Role of Wait Time in Higher Cognitive Level Learning. *Review of Educational Research* 57 (1987), 69–95.

[39] Mike Watts, Gillian Gould, and Steve Alsop. 1997. Questions of Understanding: Categorising Pupils' Questions in Science. *The School science review* 79 (1997), 57–63. https://api.semanticscholar.org/CorpusID:141492359

[40] Anwar Ali Yahya. 2011. Bloom's Taxonomy Cognitive Levels Data Set. http://dx.doi.org/10.13140/RG.2.1.4932.3123. Accessed on May 05, 2024.

[41] Scott L. Zeger and Kung-Yee Liang. 1986. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 42, 1 (1986), 121–130.

[42] Anat Zohar and Yehudit J. Dori. 2003. Higher Order Thinking Skills and Low-Achieving Students: Are They Mutually Exclusive? *Journal of the Learning Sciences* 12, 2 (April 2003), 145–181. https://doi.org/10.1207/s15327809jls1202_1

[43] U. Zoller. 1987. The Fostering of Question-Asking Capability: A Meaningful Aspect of Problem-Solving in Chemistry. *Journal of Chemical Education* 64 (1987), 510–512.