April 22, 2024

The Return of Jill Watson

Ashok Goel
goel@cc.gatech.edu

Summary: Georgia Tech's Design Intelligence Laboratory and NSF's National AI Institute for Adult Learning and Online Education have developed a new version of the virtual teaching assistant named Jill Watson that uses OpenAI's ChatGPT, performs better that OpenAI's Assistant service, enhances teaching and social presence, and correlates with improvement in student grades. Insofar as we know, this is the first time a Chatbot has been shown to improve teaching presence in online education for adult learners.

History: The initial version of Jill Watson in 2016 used IBM's Watson platform and answered students' questions on online discussions forums (such as Piazza) based on class syllabi (Goel & Polepeddi 2018). The initial Jill answered new questions based on a digital library of previously asked questions and answers given by human teaching assistants. Chronicle of Higher Education called it world's first virtual teaching assistant. Wall Street Journal, New York Times, Washington Post, and Wired published stories on it.

However, Jill of that time was biased because of the data on which it was trained was skewed depending on student demographics in previous classes (Eicher, Polepeddi & Goel 2018). An intermediate version of Jill Watson in 2019 switched to Google's BERT as the platform because it was available as open-source software and thus could be tuned for Jill. This version of Jill answered questions based natural language documents. It operated on online discussion forums (such as EdStem) as well as directly on learning management systems (such as Canvas). It was also embedded in interactive learning environments where it answered questions based on user guides (Goel et al. 2024). Given the more recent rise of Generative AI, and especially OpenAI's release of ChatGPT, a new version of Jill Watson uses ChatGPT as the backend and answers students' questions and engages in extended conversations based on the courseware such as class websites, class syllabi and schedule, textbooks and user guides, video transcripts and presentation slides (Kakar et al. 2024; Taneja et al. 2024).

Theory: The *Community of Inquiry* (CoI) is a conceptual framework for understanding adult learning and online education (Garrison, Anderson, and Archer 1999; 2001). It identifies three distinct but related elements of a successful educational experience. *Teaching presence* is the design and facilitation of educational experiences by the instructor, which includes planning learning activities, communicating requirements, and guiding discussions. *Social presence* is the ability of learners to establish and maintain a sense of connectedness with one another and the instructor; and cognitive presence is the ability of learners to construct meaning through reflection, discourse, and the application of concepts to real-world situations. *Cognitive presence* is a key component of critical thinking, often taken to be the goal of higher education,

Commented [AG1]: Ashok Goel & Lalith Polepeddi. (2018) Jill Watson: A virtual teaching assisant for online education. In Dede, C., Richards, J., & Saxberg, B., (Editors) Education at scale: Engineering online teaching and learning. NY: Routledge.

Commented [AG2]: Bobbie Eicher, Lalith Polepeddi, and Ashok Goel. Jill Watson Doesn't Care if you are Pregnant: Grounding AI Ethics in Experiments. In Procs. First AAA - ACM Conference on AI, Ethics and Society. New Orleans, February 2018.

Commented [AG3]: Ashok Goel, Vrinda Nandan, Eric Gregori, Sungeun An, & Spencer Rugaber. (2024). Explanation as Question Answering Based on User Guides. In S. Tulli & D. W. Aha (Eds.), Explainable Agency in Artificial Intelligence: Research and Practice. CRC Press.

Commented [AG4]: Sandeep Kakar, Pratyusha Maiti, Alekhya Nandula, Gina Nguyen, Karan Taneja, Aiden Zhong, Vrinda Nandan, and Ashok Goel. (2024) Jill Watson – A Conversational Virtual Teaching Assistant. Accepted for publication in Procs. 20th International Conference on Intelligent Tutoring Systems, Greece, June 2024.

Commented [AG5]: Karan Taneja, Sandeep Kakar, Pratyusha Maiti, Pranav Guruprasad, and Sanjeev Rao, and Ashok Goel. (2024) Jill Watson: A Virtual Teaching Assistant powered by ChatGPT. Accepted for publication for in Procs. 25th International Conference on AI in Education, Brazil, July 2024.

Commented [AG6]: D. Randy Garrison, Terry Anderson, and Walter Archer. 1999. Critical Inquiry in a Text-Based Environment: Computer Conferencing in Higher Education. The Internet and Higher Education 2, 2 (March 1999), 87–105.

D. Randy Garrison, Terry Anderson, and Walter Archer. 2001. Critical thinking, cognitive presence, and computer conferencing in distance education. American Journal of Distance Education 15, 1 (Jan. 2001), 7–23.

and therefore the most fundamental of the three presences. Social presence supports cognitive presence in the sense that a student's engagement with their community of inquiry promotes critical thinking. Teaching presence supports both social and cognitive presence via the design and facilitation of a successful learning experience. The goal of the Jill Watson project is to (virtually) enhance teaching presence. CoI distinguishes between two kinds of teaching presence: Direct Instruction, and Facilitation and Design & Organization.

Technology: Figure 1 illustrates the run-time architecture of Jill Watson. There are three main components: knowledge base, agent memory, and question-answering pipeline. Knowledge Base The knowledge base (KB) is generated by pre-processing the verified courseware (class website, syllabus and schedule, textbooks and user guides, video transcripts and presentation slides) in PDF format (not shown in Figure 1). The agent memory is a MongoDB database that stores information about all the interactions between the agent and users, serving as a source for question history to ensure conversationality.
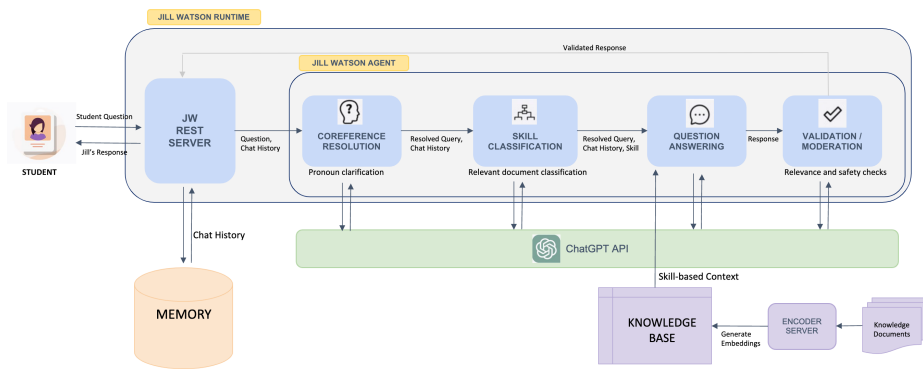

Figure 1: The run-time architecture of Jill Watson consists of coreference resolution for continuity in dialog, skill classification for relevant document selection, question-answering pipeline for generating an appropriate prompt for ChatGPT to respond to student query, and moderation to prevent harmful content generation. This assumes pre-processing of courseware. Note the role of Generative AI in the form of ChatGPT. (Figure 1 is adapted from Kakar et al. 2024).

Given a student's query, Jill uses co-reference resolution to enrich ambiguous queries with entities such as nouns or noun phrases sourced from the conversation history. It then uses skill classification to select the relevant document from the KB that can be used to answer the query. Queries that are based on the configured KB are forwarded to the Contextual Answering Skill, out-of-domain questions are forwarded to the Irrelevant Skill, and salutations are answered by the Greetings Skill. Each skill is associated with a different response generation technique. Jill

uses the context-aware response generation technique for questions identified as in-domain. To ensure conversationality, Jill extract's the conversation history for the current user from the MongoDB memory. To ensure that the responses are generated solely based on information sourced from verified courseware, Jill passes specific contexts to ChatGPT (and declines to respond if the information is insufficient). Jill confirms the consistency of the generated response with the context through textual entailment. The validated response is then returned to the student.

Experiments: In Fall 2023, we deployed Jill Watson in both a Georgia Tech Online Master of Science on Computer Science (OMSCS) class in AI consisting of more than 600 hundred students and a basic English class at Wiregrass College that is part of the Technical College System of Georgia (TCSG). Further, in the OMSCS class on AI we conducted an A/B experiment, randomly dividing the class into sections one of which had access to Jill Watson while the other did not (though it too has access to ChatGPT). We are in the middle of conducting similar experiments in five classes in Spring 2024, including a Georgia Tech class on Introduction to Programming.

For Fall 2023, we collected data from Jill Watson about all student conversations with the AI agent. We also conducted several student surveys throughout the term, including the well-established Community of Inquiry survey for measuring teaching, social, and cognitive presence. In addition, we obtained data about student performance from the learning management system and student demographics from the registrar (all in accordance with an established IRB protocol). We used mixed methods to analyze all this data.

Table 1: Accuracy of Jill Watson's answers to multiple synthetic datasets prior to deployment.
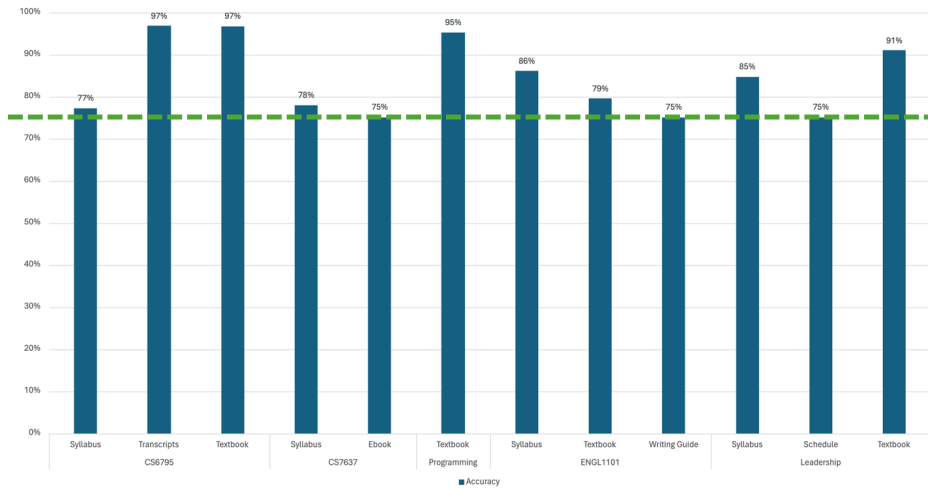
We also found that Jill Watson performs much better than OpenAI's Assistant service for building AI assistants as indicated in Table 2 (Taneja et al. 2024). Jill Watson's answers "passed"

Results: First, as Table 1 indicates, the accuracy of Jill Watson's answers on multiple synthetic datasets we use for testing prior to deployment ranged from about 75% to about 97% for both the OMSCS class in AI and the TCSG class in English (Kakar et al 2024). The variation in accuracy seems to depend on the source of the answers: for sources such as syllabi, Jill gets about 75-80% questions right, while for textbooks it typically gets more than 90% accurate. Note that for class on Introduction on Programming (shown in the center of the figure), Jill gets about 95% of the test questions based on the textbook right.

Second, as Table 2 illustrates Jill Watson performs better than Open-AI's Assistant on a synthetic data: While Open-AI's Assistant answered only about 30% of the questions correctly, Jill Watson more than 75% of the questions correctly as judged by a human evaluator. Questions on which Jill's answers failed to pass the judgement of human evaluators, OpenAI Assistant's performance was substantially worse.

Table 2: Comparison of Jill Watson using ChatGPT and Open-AI Assistant.

Method	Pass	Failures: Harmful	Failures: Confusing	Failures: Retrieval
JW-GPT	78.7%	2.7%	54.0%	43.2%
OpenAI-Assistant	30.7%	14.4%	69.2%	68.3%

Table 3: Differences between cognitive presence (CP), social presence (SP), teaching presence – facilitation and direct instruction (TP-FDI), and teaching presence – design and organization (TP-DO) between the treatment group with access to Jill Watson and the control group without.

Grouping	Treatment					Control				
	Mean	SD	Skewness	Kurtosis	α	Mean	SD	Skewness	Kurtosis	α
CP	3.8481	0.6140	-0.6283	1.4212	0.9072	3.7817	0.7217	-0.8956	1.8058	0.9342
SP	3.8061	0.5724	0.0570	-0.1748	0.8481	3.6749	0.7886	-1.0111	1.9807	0.9353
TP-FDI	3.6056	0.8029	-0.4584	0.5245	0.9243	3.5974	0.7990	-0.6411	0.8994	0.9261
TP-DO	4.1510	0.6938	-1.3075	3.9434	0.8356	4.0080	0.7684	-1.3987	3.2578	0.8983

Third, from the A/B experiment in the OMSCS class on AI we found that Jill Watson helps enhance teaching presence in the form of design and organization of the instruction and learning in the class as well as social presence (Lindgren et al. 2024). This illustrated in Tables 3 and 4. However, we did not find statistically meaningful impact of Jill Watson on teaching presence in the form of direct instruction or facilitation, or on cognitive presence. (These are topics for our plans for the near future).

Table 4: The p-values for the four parameters cognitive presence (CP), social presence (SP), teaching presence – facilitation and direct instruction (TP-FDI), and teaching presence – design and organization (TP-DO).

Factor	Statistic	p	Cohen's d
CP	0.8927	0.1863	0.0809
SP	1.7002	*0.0450	0.1576
TP-FDI	0.0936	0.4628	0.0092
TP-DO	1.7657	*0.0392	0.1668

Finally, from the A/B experiment we found a correlation between access to Jill and student performance in the class (Kakar et al. 2024). We found that the students in the section with access to Jill has more A grades than the section without access to it (about 66% to 62%) and fewer C grades (about 3% to 7%). Tables 5a and 5b illustrate this. The differences in the student grades are large enough to be intriguing but small enough to be cautious.

Table 5: Differences between student grades in the section that had access to Jill Watson (blue) and that did not (red). (The A grades start from 0% of the class but are shown here as starting from 60% to highlight the difference in numbers of A's between the two sections.)
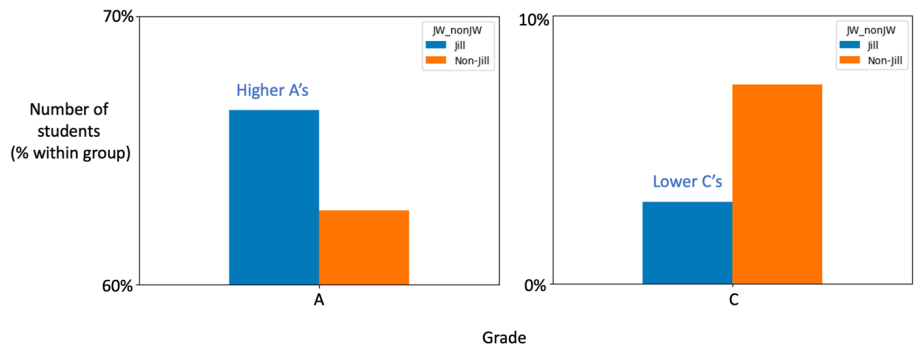
Plans: First, we want to replicate the above results in several classes, especially in introductory computing classes in colleges (including advancement placement in high schools, and technical and community colleges). Second, we want to investigate if Jill Watson can help enhance cognitive presence, especially critical thinking and deeper understanding of the subject matter. Above we noted that we did not observe statistically meaningful impact of Jill on cognitive presence. However, we did notice that some students in the OMSCS class thought that Jill helped them with critical thinking and deeper understanding. Here is a direct quote from a student: "The Jill Watson upgrade is a leap forward. With persistent prompting I managed to coax it from explicit knowledge to tacit knowledge. That's a different league right there, moving beyond merely gossip (saying what it has been told) to giving a thought-through answer after analysis. I didn't take it through a comprehensive battery of tests to probe the limits of its capability, but it's definitely promising. Kudos to the team." This and similar preliminary, qualitative, anecdotal evidence encourages us to explore improvements to Jill so that it might help enhance students' cognitive presence. Third, we want to examine the impact of Jill Watson on teaching. We expect that Jill not only offers a conversational courseware to students anytime anyplace, but it also helps the teaching team by offloading routine question answering to Jill.

Core members of the Jill Watson team include Sandeep Kakar, Pratyusha Maiti, Robert Lindgren, and Karan Taneja.